

ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE  
**CERN** EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

# **Beam Loss and Accelerator Protection**

## **2014 Joint International Accelerator School**

Newport Beach, United States  
5–14 November 2014

**Proceedings**

Editor: R. Schmidt

ISBN 978-92-9083-426-7 (paperback)


ISBN 978-92-9083-427-4 (PDF)

ISSN 0007-8328

DOI <http://dx.doi.org/10.5170/CERN-2016-002>

Available online at <https://e-publishing.cern.ch/> and <https://cds.cern.ch/>

Copyright © CERN, 2016

 Creative Commons Attribution 4.0

Knowledge transfer is an integral part of CERN's mission.

This CERN Yellow Report is published in Open Access under the Creative Commons Attribution 4.0 license (<http://creativecommons.org/licenses/by/4.0/>) in order to permit its wide dissemination and use.

The submission of a contribution to a CERN Yellow Report shall be deemed to constitute the contributor's agreement to this copyright and license statement. Contributors are requested to obtain any clearances that may be necessary for this purpose.

This report is indexed in: CERN Document Server (CDS), INSPIRE, Scopus.

This report should be cited as:

Proceedings of the Joint International Accelerator School: Beam Loss and Accelerator Protection, Newport Beach, United States, 5–14 November 2014, edited by R. Schmidt, CERN-2016-002 (CERN, Geneva, 2016), <http://dx.doi.org/10.5170/CERN-2016-002>

A contribution in this report should be cited as:

[Author name(s)], in Proceedings of the Joint International Accelerator School: Beam Loss and Accelerator Protection, Newport Beach, United States, 5–14 November 2014, edited by R. Schmidt, CERN-2016-002 (CERN, Geneva, 2016), pp. [first page] – [lastpage], <http://dx.doi.org/10.5170/CERN-2016-002>. [first page]

Group Photo courtesy of Alfonse N. Pham

## **Abstract**

Many particle accelerators operate with very high beam power and very high energy stored in particle beams as well as in magnet systems. In the future, the beam power in high intensity accelerators will further increase. The protection of the accelerator equipment from the consequences of uncontrolled release of the energy is essential. This was the motivation for organizing a first school on beam losses and accelerator protection (in general referred to as machine protection). During the school the methods and technologies to identify, mitigate, monitor and manage the technical risks associated with the operation of accelerators with high-power beams or subsystems with large stored energy were presented. At the completion of the school the participants should have been able to understand the physical phenomena that can damage machine subsystems or interrupt operations and to analyze an accelerator facility to produce a register of technical risks and the corresponding risk mitigation and management strategies.



## Preface

The U.S. Particle Accelerator School (USPAS), the CERN Accelerator School (CAS), the Asian Committee on Future Accelerators and the Budker Institute for Nuclear Physics in Russia are partners in organizing the Joint Accelerator School. This partnership has resulted in a series of specialized courses in accelerator physics and technology, the first of which was in 1985. "Beam Loss and Machine Protection" is the thirteenth course in this series, and was held in Newport Beach, California in November 2014.

The success of this particular course was made possible through the fruitful collaboration between the USPAS and the CAS, with the full support of their respective managements. USPAS took care of all local arrangements in California while the CAS took responsibility for the production of these proceedings. The definition of the program and the speakers, as well as the running of the school was a joint effort between the two schools.

Special thanks must go to the lecturers for the preparation and presentation of the lectures, even more so to those who have written a manuscript for these proceedings.

The enthusiasm of the 58 participants, from institutes in 12 countries, provides convincing proof of the usefulness and success of the course.

For the production of the proceedings we are indebted to the efforts of Rudiger Schmidt and to the CERN E-Publishing Service, especially Valeria Brancolini for her very positive and efficient collaboration. These proceedings have been published in paper (black and white with some figures in colour) and electronic form. The electronic version, with full colour figures, can be found at <https://publishing.cern.ch/index.php/CYR/issue/view/12>.

Roger Bailey, CERN Accelerator School  
Bill Barletta, US Particle Accelerator School

**Joint International Accelerator School on Beam Loss and Accelerator Protection**

**November 5-14, 2014**

| Time  | Wednesday<br>Nov. 5           | Thursday<br>Nov. 6   | Friday<br>Nov. 7   | Saturday<br>Nov. 8  | Sunday<br>Nov. 9 | Monday<br>Nov. 10   | Tuesday<br>Nov. 11  | Wednesday<br>Nov. 12   | Thursday<br>Nov. 13                         | Friday<br>Nov. 14        |
|-------|-------------------------------|--|--|---|------------------|---|---|--|---|--------------------------|
| 8:30  |                               | Introduction to Accelerator Protection Course<br>Rudiger Schmidt | Beam Material Interaction, Heating & Activation (Part I)<br>Nikolai Mokhov (2 hrs)                     | Beam Transfer and Machine Protection<br>Verena Kain                                     |                  | Detection of Equipment Failures Before Beam Loss<br>John Galambos | Machine Protection and Interlock Systems for LHC<br>Rudiger Schmidt                 | Machine Protection and Operation for LHC<br>Jorg Wenninger                         | Personnel Protection Systems<br>Sayed Rokni |                          |
| 10:00 |                               |  |  |   | <b>F R E E</b>   |   |   |  |   |                          |
| 10:30 |                               | Beam Dynamics and Beam Losses - Circular Machines<br>Verena Kain | <b>COFFEE</b><br>Beam Material Interaction, Heating & Activation (Part II)<br>Francesco Cerutti (1 hr) | Beam Induced Damage Mechanisms and Their Calculation (Part I)<br>Alessandro Bertarelli  |                  | Controls and Machine Protection<br>Enzo Carrone                   | Machine Protection and Interlock Systems - Linear Machines<br>Marc Ross             | <b>COFFEE</b><br>Machine Protection and Operation for Linear Machines<br>Marc Ross | Medical Facilities<br>Anthony Mascia        |                          |
| 12:00 |                               |  |  |   | <b>D A Y</b>     |   |   |  |   |                          |
| 13:30 |                               | Beam Dynamics and Beam Losses - Linear Machines<br>Mike Plum     | <b>LUNCH</b><br>Reliability and Availability<br>Ferdinand Willeke                                      | Beam Induced Damage Mechanisms and Their Calculation (Part II)<br>Alessandro Bertarelli |                  | Instrumentation for Machine Protection<br>Tom Shea (2 hrs)        | Protection of Hardware: Powering Systems (PC, NC and SC Magnets)<br>Howard Pfeiffer | Beam Cleaning and Collimation Systems<br>Stefano Redaelli (2 hrs)                  |   | <b>D E P A R T U R E</b> |
| 15:00 |                               |  |  |   | <b>F R E E</b>   |   |   |  | Present Case Studies                        |                          |
| 17:00 |                               | High Intensity Synchrotron Radiation Effects<br>Yusuke Suetsugu  | <b>STUDY</b><br>Intro to Risk Management of Complex Systems<br>John Thomas                             | Protection Related to High Power Targets<br>Mike Plum                                   |                  | Beam Loss Monitors at LHC<br>Bernd Dehning (1 hr)                 | Protection of Hardware: RF Systems<br>Sang Ho Kim                                   | <b>STUDY</b><br>Advanced Collimators for Future Colliders<br>Tom Markiewicz (1 hr) |   | <b>D A Y</b>             |
| 18:30 |                               |  |  |   | <b>D A Y</b>     |   |   |  |   |                          |
| 20:00 | Dinner, Registration and Talk |  |  |   |                  |   |   |  |   |                          |
| 21:30 |                               |  | Case Studies<br>(Background material for the TT40 Groups)  |   |                  |   | Case Studies  |  | Final Exam                                  |                          |

# Contents

|  |     |
|--|-----|
| Preface  |     |
| <i>R. Bailey</i> .....   | v   |
| Introduction to Machine Protection   |     |
| <i>R. Schmidt</i> .....  | 1   |
| Beam Dynamics and Beam Losses–Circular Machines  |     |
| <i>V. Kain</i> .....   | 21  |
| Beam Loss in Linacs  |     |
| <i>M.A. Plum</i> .....   | 39  |
| High-Intensity Synchrotron Radiation Effects   |     |
| <i>Y. Suetsugu</i> .....   | 63  |
| Beam–Material Interactions   |     |
| <i>N. Mokhov and F. Cerutti</i> .....  | 83  |
| Reliability Considerations for the Operation of Large Accelerator User Facilities                          |     |
| <i>F. Willeke</i> .....  | 111 |
| Beam Transfer and Machine Protection   |     |
| <i>V. Kain</i> .....   | 137 |
| Beam-Induced Damage Mechanisms and Their Calculation   |     |
| <i>A. Bertarelli</i> .....   | 159 |
| Protection Related to High-power Targets   |     |
| <i>M.A. Plum</i> .....   | 229 |
| Detection of Equipment Faults Before Beam Loss   |     |
| <i>J. Galambos</i> .....   | 253 |
| Controls and Machine Protection Systems  |     |
| <i>E. Carrone</i> .....  | 271 |
| Beam Loss Monitors at LHC  |     |
| <i>B. Dehning</i> .....  | 303 |
| Machine Protection and Interlock Systems for Circular Machines–Example for LHC                             |     |
| <i>R. Schmidt</i> .....  | 319 |
| Protection of Hardware: Powering Systems (Power Converter, Normal Conducting, and Superconducting Magnets) |     |
| <i>H. Pfeffer et al.</i> .....   | 343 |
| Protection of Accelerator Hardware: RF systems   |     |
| <i>S. Ho Kim</i> .....   | 361 |
| Machine Protection and Operation for LHC   |     |
| <i>J. Wenninger</i> .....  | 377 |
| Beam Cleaning and Collimation Systems  |     |
| <i>S. Redaelli</i> .....   | 403 |
| List of Participants .....   | 439 |





## Introduction to Machine Protection

*R. Schmidt*

CERN, Geneva, Switzerland

### Abstract

Protection of accelerator equipment is as old as accelerator technology and was for many years related to high-power equipment. Examples are the protection of powering equipment from overheating (magnets, power converters, high-current cables), of superconducting magnets from damage after a quench and of klystrons. The protection of equipment from beam accidents is more recent, although there was one paper that discussed beam-induced damage for the SLAC linac (Stanford Linear Accelerator Center) as early as in 1967. It is related to the increasing beam power of high-power proton accelerators, to the emission of synchrotron light by electron–positron accelerators and to the increase of energy stored in the beam. Designing a machine protection system requires an excellent understanding of accelerator physics and operation to anticipate possible failures that could lead to damage. Machine protection includes beam and equipment monitoring, a system to safely stop beam operation (e.g. dumping the beam or stopping the beam at low energy) and an interlock system providing the glue between these systems. The most recent accelerator, LHC, will operate with about  $3 \times 10^{14}$  protons per beam, corresponding to an energy stored in each beam of 360 MJ. This energy can cause massive damage to accelerator equipment in case of uncontrolled beam loss, and a single accident damaging vital parts of the accelerator could interrupt operation for years. This lecture will provide an overview of the requirements for protection of accelerator equipment and introduces various protection systems. Examples are mainly from LHC and ESS.

### Keywords

Machine protection; interlock system; high-power accelerator; beam loss; accident.

## 1 Introduction to the school

This is the first school on beam losses and accelerator protection (in general referred to as machine protection). The school is intended for physicists and engineers who are or may be engaged in the design, construction and/or operation of accelerators with high-power particle or photon beams and/or accelerator subsystems with large stored energy.

We will present the methods and technologies to identify, mitigate, monitor and manage the technical risks associated with the operation of accelerators with high-power beams or subsystems with large stored energy, if failures can result in damage to accelerator systems or interruption of operations. At the completion of the school the participants should be able to understand the physical phenomena that can damage machine subsystems or interrupt operations and to analyse an accelerator facility to produce a register of technical risks and the corresponding risk mitigation and management strategies.

Excellent textbook material and a large number of proceedings from CAS schools are readily available for many topics in accelerator physics and technology (e.g. beam dynamics, synchrotron radiation, superconductivity for accelerators). This is not the case for beam losses and protection of accelerators; there are no books, only a limited amount of lectures in accelerator schools and a few invited contributions to accelerator conferences [1–4].

This school is focused on the protection of equipment. Similar strategies are used for the design of systems to protect people and the environment. One lecture during this school will present the challenges for protection of people [5].

An efficient way to learn about machine protection is to consider past accidents and near misses. However, it is not common to present what went wrong during accelerator operation, and it is acknowledged that several examples of past accidents are presented during this school.

The programme of this school follows some questions.

- What can go wrong when operating an accelerator?
- What are the consequences when something goes wrong?
- What mitigation methods can be applied?
- What aspects of controls and operation are relevant for machine protection?

## 2 Introduction to this lecture

In the first part of this lecture we discuss observations from more than 30 years from various accelerators relevant to the design of machine protection systems. The first time that beam-induced damage was discussed in a paper for the SLAC linac as early as in 1967 [6]. Some basic principles for machine protection at today's accelerators are derived. Challenges for machine protection are briefly illustrated with two examples, the CERN Large Hadron Collider (LHC) and the European Spallation Source (ESS). In the second part the performance of accelerators is discussed in relation to hazards and machine protection. A short introduction to machine protection follows. In the last part some accidents at different accelerators are presented.

Accelerators, as all other technical systems, must respect some general principles with respect to safety and protection:

- Protection of people from different threats (radiation, electrical, oxygen deficiency, etc) has always the highest priority. The main strategy to protect people during accelerator operation is to keep everyone out of defined boundaries around an accelerator when beam is running, ensured by a personnel access system. Usually, protection of people is regulated by the governing bodies;
- Protection of the environment (e.g. following legal requirements);
- Protection of accelerator equipment and experiments (the investment).

In general, risks come from energy stored in a system (measured in joules) as well as from power when operating the system (measured in watts). Particle accelerators are examples of such systems, since many accelerators operate with a large amount of electrical power (from a few to many MW). The energy and power flow need to be controlled. An uncontrolled release of the energy or an uncontrolled power flow can lead to unwanted consequences such as damage or activation of equipment and loss of time for operation.

Several questions are addressed in this lecture.

- What accelerators need protection?
- What needs to be protected?
- What are the hazards/risks?
- What can be the consequences of an accident?
- What can be done to prevent accidents?

Accelerators can be divided into two classes: accelerators operating with a large amount of stored energy in the particle beam such as synchrotrons and storage rings, and accelerators operating with large beam power such as high-power proton accelerators, free-electron lasers (FELs) and linear colliders.

The energy stored in a particle beam is given by  $E_{\text{beam}} = N \cdot E_{\text{particle}}$ , with  $N$  the number of particles stored in the accelerator and  $E_{\text{particle}}$  the kinetic energy of a particle. The beam power is given by the energy per unit of time,  $P_{\text{beam}} = N \cdot E_{\text{particle}}/t$ .

For synchrotrons and storage rings, the energy stored in the beam increased over the years (at CERN from the Intersecting Storage Ring ISR to LHC, with an energy of 362 MJ stored in one beam with nominal LHC parameters).

For linear accelerators and fast-cycling machines, the beam power increased over the years. Not only energy or power are relevant but also the beam size. The smaller the beam, the higher is the energy density in case the beam is deflected into equipment. The damage potential increases with decreasing beam size. For accelerators such as future linear colliders, the emittance is expected to become much smaller (down to a beam size of a nanometre) resulting in very high power density ( $\text{W mm}^{-2}$ ) [7–9].

In order to get an idea of what this amount of energy means, some examples are given. The energy of a pistol bullet is about 500 J; the energy of 1 kg of TNT is about 4 MJ. The energy of 1 l of fuel is about 36 MJ; to melt 1 kg of steel about 800 kJ is required (the energy to melt 1 kg of copper is similar). An accidental release of an energy above 1 MJ can cause significant damage. Even an accidental release of a small amount of energy (order of some hundreds of joules) can lead to some (limited) damage if the energy is released in sensitive equipment such as radio-frequency (RF) cavities.

Protection must be considered during all phases of operation, with the accelerator operating with or without beam. Not only damage to accelerator components needs to be considered, but also to the experiments [10].

Several systems operating with high power or a large amount of stored energy, such as the RF system, power converters, the magnet system and the cryogenic system, are usually commissioned long before beam operation starts.

### 3 Synchrotrons and storage rings

In a synchrotron the beams are injected at low energy and the energy is increased while ramping the magnetic field. The particles are accelerated by RF fields in RF cavities.

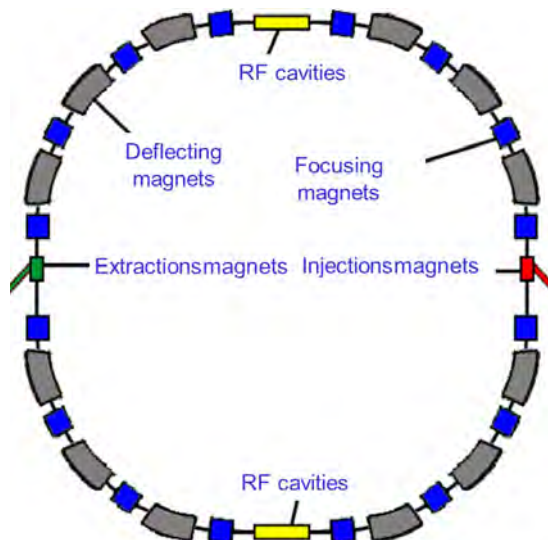
The main components of a synchrotron are deflecting magnets, magnets to focus the beam and correction magnets (Fig. 1). These magnets are operating with slowly changing fields, on a time-scale of seconds. Pulsed magnets are required for injection and extraction. Power supplies provide the magnet current. Radio-frequency cavities accelerate the beam; the power is provided by components in the RF system (e.g. modulators and klystrons). Other systems are beam instrumentation and the control system. The vacuum system ensures very low pressure for the beam circulating in the vacuum chamber.

Beams are injected at low energy and accelerated to higher energy. Depending on the accelerator, the energy is increased by a factor of 10 to 40. At top energy, the beam is extracted during a plateau to fixed-target experiments or to another accelerator, if the synchrotron is used as pre-accelerator.

For colliders operating with counter-rotating beams, the beams are brought into collision and the length of the plateau is extended to many hours while the beams are colliding (depending on the accelerator, from a few to several tens of hours). For particle physicists, the particle energy and total number of events that are collected are the most important performance parameters. The number of events per unit time is proportional to the luminosity:

$$\frac{N}{\Delta T} = L [\text{cm}^{-2} \text{s}^{-1}] \cdot \sigma [\text{cm}^2], \quad (1)$$

with the luminosity  $L [\text{cm}^{-2} \text{s}^{-1}]$  and the cross-section  $\sigma [\text{cm}^2]$ . For head-on collisions the luminosity is given by the number of particles per bunch  $N$ , the number of bunches per beam  $n_b$ , the revolution frequency  $f_{\text{rev}}$  and the rms beam sizes at the interaction point  $\sigma_x$  and  $\sigma_y$ :



**Fig. 1:** Illustration of a circular accelerator with typical components

**Table 1:** Peak luminosity at different colliders [11, 12]

| Accelerator                                   | Peak luminosity                                   |
|---|---|
| PETRA $e + e^-$ (DESY)                        | $2 \times 10^{31} [\text{cm}^{-2} \text{s}^{-1}]$ |
| LEP $e + e^-$ (CERN)                          | $3 \times 10^{31} [\text{cm}^{-2} \text{s}^{-1}]$ |
| Tevatron $p + p^-$ (FERMILAB)                 | Some $10^{32} [\text{cm}^{-2} \text{s}^{-1}]$     |
| SuperKEK-B $e + e^-$ (in construction at KEK) | $10^{36} [\text{cm}^{-2} \text{s}^{-1}]$          |
| FCC $e + e^-$ and $pp$ (study at CERN)        | $5 \times 10^{34} [\text{cm}^{-2} \text{s}^{-1}]$ |
| LHC (CERN)                                    | $1 \times 10^{34} [\text{cm}^{-2} \text{s}^{-1}]$ |
| HL-LHC (CERN)                                 | $2 \times 10^{35} [\text{cm}^{-2} \text{s}^{-1}]$ |

$$L = \frac{N^2 \cdot f_{\text{rev}} \cdot n_b}{4 \cdot \pi \cdot \sigma_x \cdot \sigma_y}. \quad (2)$$

The total number of events, which is what matters for particle physicists, is given by the integrated luminosity, that is, the integral of the luminosity over the time when beams are colliding and experiments take data:

$$N = \sigma \cdot \int L(t) \cdot dt. \quad (3)$$

If the maximum luminosity is limited, either due to accelerator parameters or due to limitations of data taking in the experiments, the efficiency of machine operation plays an essential role for collecting events. The peak luminosity that has been achieved, or that is planned, is given in Table 1.

When a fill is ended, since the luminosity decreased to too low a value, the magnets are ramped down to injection energy and the next cycle starts. The entire process from end collisions to next collisions takes some time (between, say, 30 min and a few hours). Depending on the energy stored in the beams, it might be acceptable that the beams are lost in an uncontrolled way during the down-ramp, or the beams must be extracted to safely deposit the energy in a beam dump block. If the beam is lost during a fill due to a failure a new cycle starts, but the efficiency for data taking is reduced.

### 3.1 Luminosity and consequences for machine protection

The performance of an accelerator is determined by beam parameters and parameters of the hardware systems. For a circular accelerator, energy and luminosity are the most important parameters. The energy stored in the beam as a function of luminosity and accelerator parameters can be approximated by rewriting the luminosity equation:

$$E_{\text{beam}} = \frac{\sigma EC}{c} \cdot \sqrt{\frac{4\pi L}{\delta T}}, \quad (4)$$

$$E_{\text{beam}} = \sigma E \cdot \sqrt{\frac{4\pi L n_b}{f_{\text{rev}}}}, \quad (5)$$

with  $\sigma$  the beam size at the interaction point assuming round beams,  $E$  the energy,  $C$  the circumference,  $c$  the speed of light,  $L$  the luminosity,  $f_{\text{rev}}$  the revolution frequency and  $\delta T$  the time between two bunches. This is an approximation since it assumes that the machine is filled without any large gaps between bunches. Such gaps, e.g. with a length of some 100 ns or a few  $\mu\text{s}$ , are required for injection and for extraction. As an example, we assume the nominal parameters for LHC: luminosity,  $L = 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ ; number of bunches per beam,  $n_b = 2808$ ; revolution frequency,  $f_{\text{rev}} = 11 \text{ kHz}$ ; energy,  $E = 7 \text{ TeV}$ ; and beam size at the interaction point,  $\sigma = 16 \mu\text{m}$ . With the approximate equation the energy stored in the beam is  $E_{\text{beam}} = 407 \text{ MJ}$ ; the exact calculation yields 362 MJ.

In order to reach high energy, the particles are deflected with superconducting magnets. The Lorentz force on a charged particle is proportional to charge, electric field and the vector product of velocity and magnetic field. An approximation for the energy stored in superconducting magnets is given by the energy of the magnetic field in the vacuum chamber (this is a lower limit; the equation is for a superconducting magnet with two vacuum chambers as in LHC):

$$E_{\text{magnets}} = \frac{2 \cdot \text{Length} \cdot R^2 \cdot \pi \cdot B^2}{\mu_0}. \quad (6)$$

For the LHC, the length of the dipole magnet systems is about 20 km, the radius of the vacuum chamber of  $R = 28 \text{ mm}$  and the magnetic field of  $B = 8.3 \text{ T}$  for an energy of  $E = 7 \text{ TeV}$ . This approximation gives an energy stored in the LHC dipole magnets of 4.8 GJ. The exact calculation gives 8.82 GJ ( $\mu_0$  is the permeability).

The energy stored in the beam for different accelerators as well as the energy stored in the LHC magnet system is shown in Fig. 2.

## 4 High-power hadron accelerators

There is a large interest in the exploitation of high-power hadron accelerators. In spallation sources high-intensity proton beams are accelerated and directed to a target. The protons interact with the target material and spallation neutrons are produced. Other accelerators are using high-intensity proton beams for neutrino production. Rare-isotope beams are produced by accelerating ions (e.g. the Facility for Rare Isotope Beams, FRIB, at Michigan University is a folded linac to accelerate ions). Accelerator-driven systems (ADS) are being developed with several projects around the world. A very energetic particle beam is used to stimulate a reaction in a subcritical reactor, which in turn releases enough energy to power the particle accelerator and leaves an energy profit for power generation. Figure 3 shows beam current, beam power and particle momentum for different high-power proton accelerators.

There is a difference between accelerators operating with high-power beams and those with large stored energy. For hadron colliders, the energy stored in the beams can be very high, as has been shown for the LHC. In case of a failure, the energy stored in beam and magnets needs to be safely deposited.

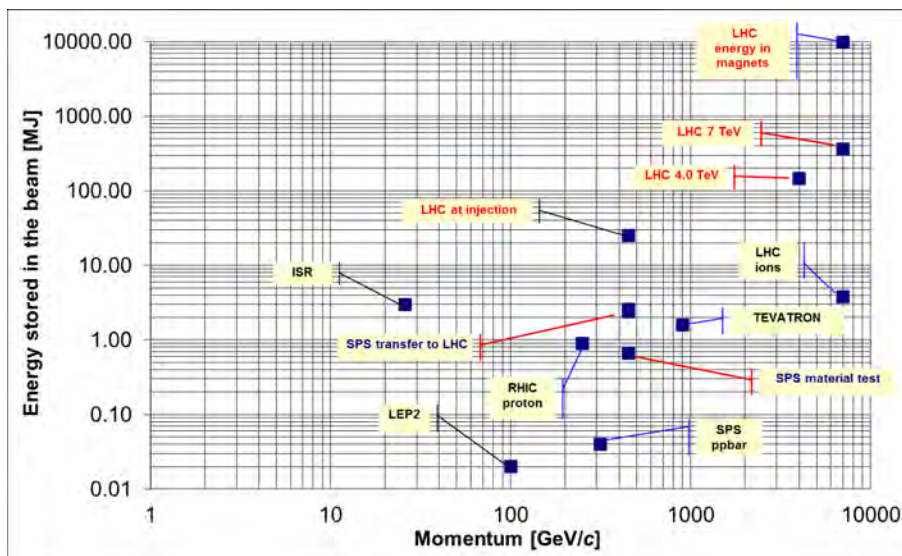


Fig. 2: Energy stored in the beams for different accelerators and the energy stored in the LHC magnet system

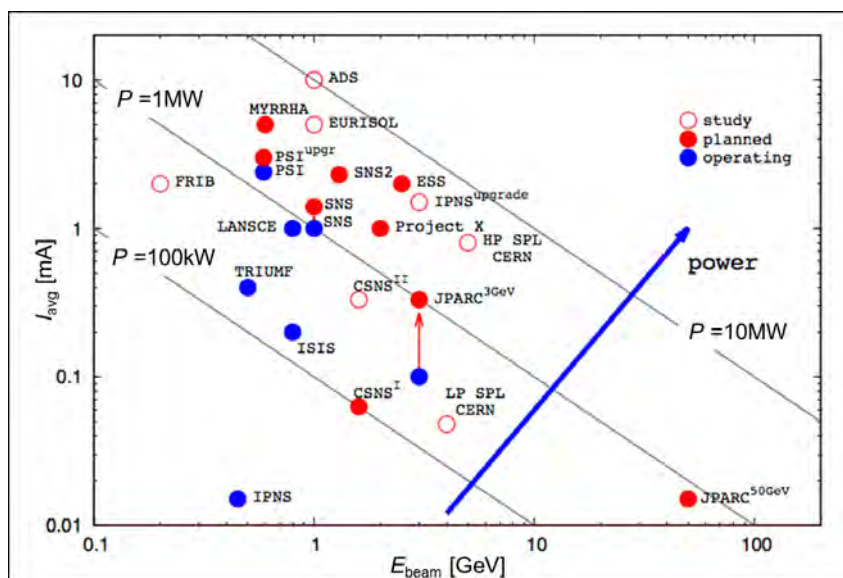


Fig. 3: Current versus particle momentum for high-power proton accelerators around the world

For high-power accelerators, the beam power increases along the accelerating structure proportionally to the particle momentum. The energy stored in the particles present in the accelerator at one moment in time is small. In a case of failure, e.g. causing beam losses in one linac section, the beam must be stopped. It is straightforward to stop the production of particles at the source. After stopping the particle production in the source there are still particles between the source and the location of the beam loss to be considered.

### 5 Lessons learned at some accelerators

In this section, observations related to machine protection for various circular accelerators are presented. These considerations allow us to draw some conclusions that were fundamental for the design of the LHC machine protection systems, as well as for the design of the protection systems for high-power linear accelerators such as ESS.

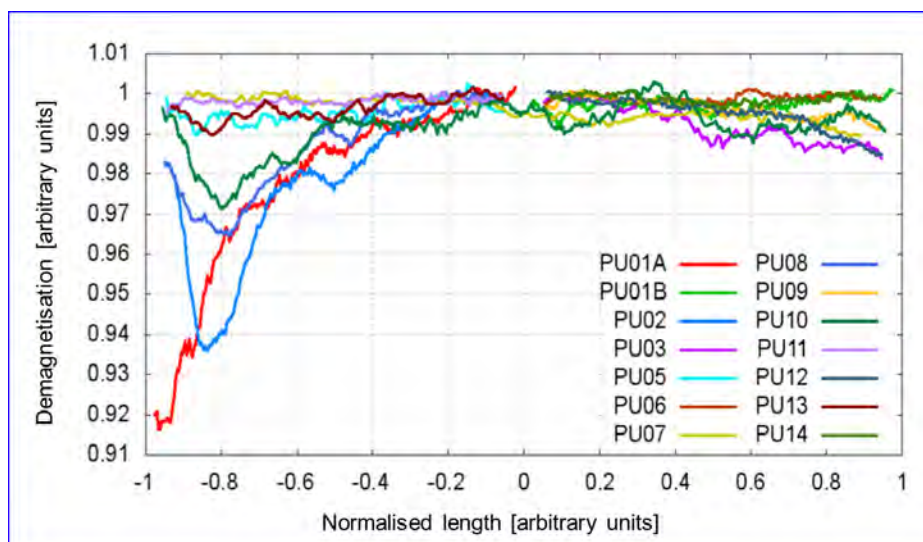


Fig. 4: Demagnetization of undulators at the PETRA storage ring from [13]

- DESY PETRA I,  $e^+ e^-$  collider (1978 to 1986) and PETRA III (in operation).
- CERN SPS, proton–antiproton collider (1982 to 1990).
- CERN SPS, proton synchrotron (starting from 1978, still operating).
- CERN LEP, large electron–positron collider (1989 to 2000).

### 5.1 PETRA, an electron–positron collider at DESY

PETRA was built as an electron–positron collider (PETRA I). Later PETRA II was used as injector for HERA (Hadron Accelerator Ring Anlage) and now it operates as an advanced synchrotron light source. PETRA has a length of 2304 m and the particles are deflected with normal-conducting magnets. PETRA I operation started in 1978 and the particle momentum was up to 21 GeV/c. PETRA I was operating with four bunches per beam, each beam with a current of about 6.5 mA. Frequently the beam was lost, sometimes for unknown reasons, in particular during the energy ramp, which is always very critical since parameters such as betatron tune and chromaticity need to be accurately controlled. When the beams at PETRA I were lost there was no risk of damage, since the stored beam energy was very small.

One way of stopping the beams in an electron–positron accelerator is to switch off the RF field. The particles are lost gradually due to the energy loss by the emission of synchrotron radiation. The particles are distributed around the circumference and the energy deposition in accelerator equipment is very low.

Equipment protection at PETRA I was important. An example of an incident: for measuring beam polarization by analysing the Compton-scattered photon distribution, a high-power laser beam was sent into the vacuum chamber through a glass window. The glass cracked when the laser passed the window, the vacuum pressure increased and an intervention was required. The window could be sealed and operation continued without major impact.

PETRA III has been operating since 2008 as a world-class synchrotron light source. Injection is at operating energy and no energy ramp is required. A degradation of the performance of undulator magnets was observed, a de-magnetization that is likely due to beam losses (Fig. 4, taken from [13]).

## 5.2 CERN-SPS synchrotron and proton–antiproton collider

The SPS was built as a normal-conducting proton synchrotron and operation started in 1978. A proton beam was accelerated to 450 GeV and directed onto a target for fixed-target experiments. Since the beams are circulating for only a few tens of seconds, no ultra-high vacuum was required. The SPS accelerator was transformed into a proton–antiproton collider during 1980 to 1982, and continued operating in this mode until 1990. Antiprotons are very rare and it took a long time to produce them in sufficient quantity in the pre-accelerators to fill the SPS with antiproton beam. If a fill was lost in the SPS, it took many hours to accumulate enough antiprotons for refilling the machine. Therefore, protecting the beam (avoiding an accidental loss of the fill) was of high priority. The energy stored in the beam was not very high, in particular at injection energy; however, on one occasion, injected beam was deflected for a period of about 10 min into the UA2 experiment and led to a degradation of sensitive parts of the experiment (see below).

From 1990 the SPS continued to operate as a synchrotron for fixed-target physics, neutrino production and as injector for LHC. The parameters for the operation as a synchrotron are very different from collider parameters with much higher beam intensity. The beam current in the SPS was constantly being increased over the years. The cycle time was of the order of some seconds to some tens of seconds. If the beam is accidentally lost during one cycle, the efficiency is hardly affected since the next cycle follows a few seconds later. However, beam losses risk to damage components and activate accelerator equipment. Protecting the equipment from beam-induced damage is required and the beams must be extracted into a beam dump in case of a failure. The SPS is operating in different modes with different extraction lines. Beams are extracted with different energies and safe operation requires a complex protection and interlock system. During the operation of the SPS as a synchrotron, magnets and other equipment were damaged several times due to accidental beam losses. In general, it was possible to repair the damage within a short time (< one day), e.g. replacing a magnet since the accelerator is normal conducting. Such repair would be very different for a superconducting machine; the consequence of damaging a superconducting element requires several months of repair. The SPS protection systems have been upgraded following the development of the LHC machine protection; since then there have been no accidents.

## 5.3 CERN-LEP electron–positron collider

LEP (Large Electron Positron collider), operating from 1989 until 2000, was installed in the tunnel that is now used for LHC. The particle energy was up to 104 GeV; the beams were injected at an energy of 20 GeV. Initially, the maximum energy was limited to about 50 GeV and it was sufficient to dump the beams after the end of a fill by switching off the RF field.

Improvement of LEP performance (higher energy and higher intensity) made it necessary to dump LEP beams in a fully controlled way. Fast kicker magnets, close to defocusing QL8 quadrupoles, vertically deflected the bunches; other quadrupoles gave an additional vertical deflection to send the beams into absorbers [14].

In June 1996, operation was just about to begin with the upgraded machine when an unexpected problem appeared. Operators were injecting beam, but it was not getting around the whole accelerator. After careful investigation, the cause was found: a pair of Heineken beer bottles wedged into the beam pipe. Beam pipes at LEP were easily accessible and repairs could be made fairly quickly.

An issue for  $e^+ + e^-$  beams is the emission of synchrotron radiation by charged particles with increasing energy. The power emitted by one particle circulating in a dipole field with the bending radius  $\rho$ , the mass  $m$ , the charge  $e_0$  and the energy  $E$  is given by

$$P_s = \frac{e_0^2 \cdot c}{6\pi\epsilon_0 \cdot (mc)^4} \cdot \frac{E^4}{\rho^2}, \quad (7)$$

where  $c$  is the speed of light and  $\epsilon_0$  the vacuum permeability.



The radiation power increases with the energy as the power of four. This became an issue when the energy of LEP was increased from 50 GeV to more than 100 GeV [15]. Assuming a beam current of 5 mA per beam, the power emitted by both beams at 45 GeV is 770 kW and at 90 GeV it is 12 MW. When the beam passes through a wiggler magnet (a series of magnets designed to periodically laterally deflect the beam), the power density is further increased. Wiggler magnets were used at LEP and are commonly used in synchrotron light sources and free-electron lasers such as the European XFEL at DESY.

When the energy of LEP was increased above 80 GeV, lead stoppers located in front of the aluminium windows of the polarimeter were installed to protect the device when it is not in use. After 30 days running at 80.5 GeV per beam, several of these blocks were found melted and needed to be replaced by tungsten. Other equipment damage was observed: beam instrumentation, electrostatic separators, vacuum equipment and, in particular, wiggler magnets.

Emission of synchrotron radiation is a feature of normal operation. This radiation can damage equipment and needs to be taken into account in the design and during upgrades. This is not limited to LEP, but a potential issue for all synchrotron light sources and XFELs.

At LEP, the beam was frequently lost, e.g. during the energy ramp, without understanding the mechanisms for the loss. No adequate diagnostic system was available to record beam and hardware parameters. A system recording the equipment and beam parameters, including when the beam was lost, would have been very useful (post-mortem system).

#### 5.4 Some lessons for the design of new accelerators

From the experience at these accelerators, as well as from many other machines not mentioned here, some lessons can be derived.

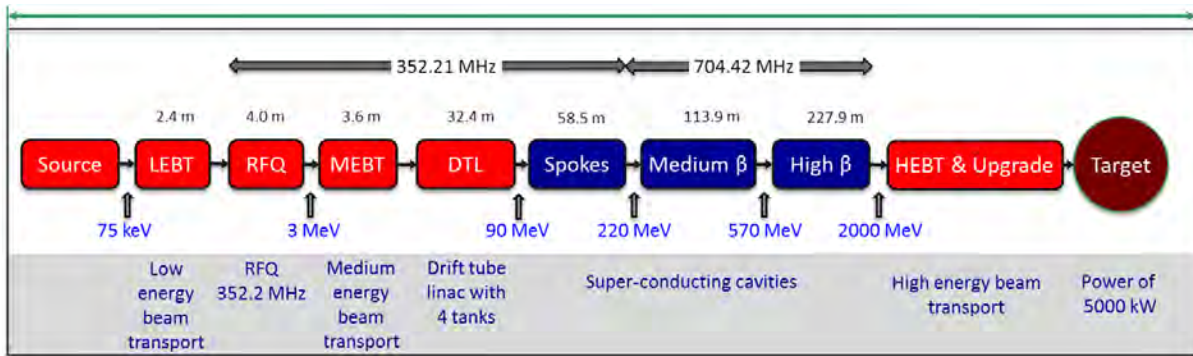
- Protection of equipment is required when there is a significant amount of energy stored in an accelerator system (e.g. superconducting magnets), or if the accelerator operates with high power (e.g. RF systems).
- Accelerator equipment as well as experiments require protection from uncontrolled beam losses in case of a significant amount of energy stored in the beam, or for high-power beams.
- For lepton accelerators, the power of synchrotron radiation needs to be considered in order to avoid possible damage.
- There is the risk of performance degradation for undulators using permanent magnetic material.
- It is important to understand what happens, e.g. when the beam is lost.

## 6 Accelerators with machine protection challenges

### 6.1 The CERN-LHC large hadron collider

The LHC is designed to operate at a momentum of 7 TeV/c with 2808 bunches, each bunch with a nominal intensity of  $1.15 \times 10^{11}$  protons. Machine protection is required during all phases of operation, since the LHC is the first accelerator with injected beam far above the threshold for damage. The energy stored in the nominal LHC beam of 362 MJ corresponds to the energy of a 200 m long fast train at 155 km per hour and to the energy stored in 90 kg of TNT. Surprisingly, this is the same as the energy stored in 15 kg of chocolate; it matters most how easily and fast the energy is released. The energy in an accelerator beam can be released in some 10  $\mu$ s.

At 7 TeV/c, fast beam loss with an intensity of about 5% of a single ‘nominal bunch ( $10^{11}$  protons)’ could damage equipment (e.g. superconducting coils). The only component that can stand a loss of the full beam is the beam dump block. All other components would be severely damaged. The LHC beams must *always* be extracted into the beam dump blocks at the end of a fill as well as in case of a failure.



**Fig. 5:** Layout of the ESS accelerator: the source, low energy beam transport and RFQ are followed by the medium energy beam transport. The protons are accelerated by a normal-conducting linac, followed by three sections of superconducting cavities. In the high energy beam transport line the protons are transported to the target.

The LHC is a two-ring collider (two different vacuum chambers with opposite magnetic fields for deflecting the particle trajectories) using superconducting magnets. Superconducting magnets store a large amount of energy and need to be protected in case of a quench. Protection of superconducting magnets is part of the design of the magnets, taking into account that most magnets are powered in series [16]. For the LHC, quench heaters and energy extraction systems are used to prevent magnet damage after a quench. Quench heaters are stainless steel strips installed in the magnet that are heated by discharging the energy stored in capacitors into the strips when a quench is detected. Energy extraction is provided by switching a resistor in series with the magnet chain to extract the energy from the magnets.

Protecting equipment from uncontrolled release of beam after a failure requires very complex systems. Detection of failures is being done by many different monitors providing interlock signals. There are potentially many thousands of such interlock signals from various systems around the accelerator. As an example, in case of a failure in the magnet powering system, e.g. after a quench, the beams must be extracted. Managing of the different interlocks with the objective to build a coherent system is required.

The development of the machine protection system requires expertise in several domains: physicists defining failure scenarios using computer tools, e.g. for particle tracking and particle-material interactions, systems experts working on machine equipment, and physicists or engineers responsible for the operation of such a complex machine.

The machine protection systems traverse the organization. Therefore, at CERN the co-ordination of all activities related to LHC machine protection was assigned to a team in 2000, the Machine Protection Working Group that co-ordinated design, implementation, commissioning and operation of the machine protection system.

## 6.2 The European Spallation Source

The European Spallation Source (ESS) being built at Lund, Sweden is designed to accelerate a proton beam with an average power of 5 MW and to direct the protons onto a target. Operation of the ESS will be at a frequency of 14 Hz, with a pulse length of 2.86 ms and a peak power of 125 MW. The layout of the ESS accelerator is shown in Fig. 5.

In case of an uncontrolled beam loss during, say, 1 ms at ESS the deposited energy is up to 130 kJ, for 1 s up to 5 MJ. It is required to inhibit the beam after detecting uncontrolled beam loss as fast as possible. There is some delay between detection of a failure (e.g. detection of beam losses by a beam loss monitor) and ‘beam off’. Figure 6 shows the time to melt copper and steel in the case where the proton beam hits a metal surface between 3 and 80 MeV/c [17]. For example, after the Drift Tube normal-conducting Linac (DTL), the proton energy is 78 MeV/c. In case of a beam size of 2 mm radius,

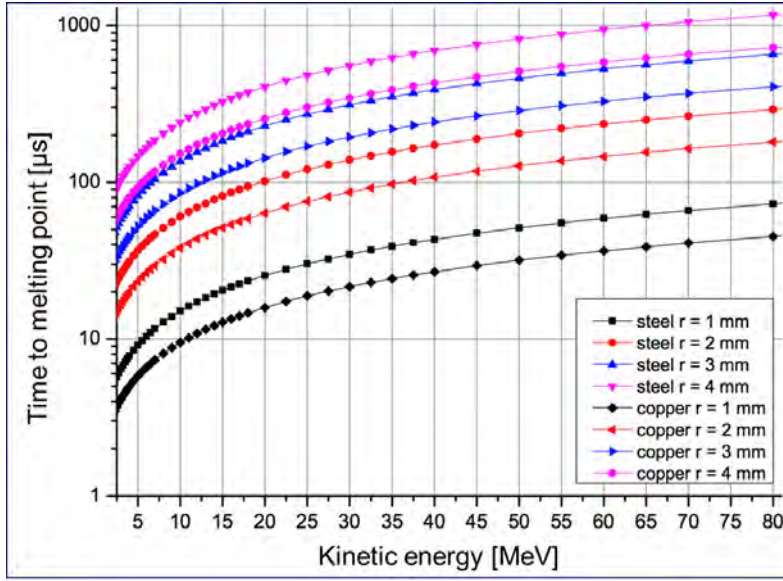


Fig. 6: Time to melt copper and steel, as a function of proton momentum for different beam sizes [17]

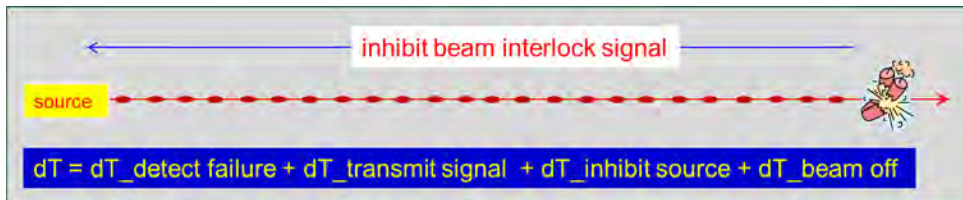


Fig. 7: Time from detecting a failure until no particles are present in the accelerator: time to detect a failure:  $dT_{\text{detect failure}}$ , time to transmit the signal to the source:  $dT_{\text{transmit signal}}$ , time to inhibit proton production:  $dT_{\text{inhibit source}}$ , time until no protons are in the accelerator:  $dT_{\text{beam off}}$ .

melting would start after a beam impact of about  $200 \mu\text{s}$ . Inhibiting of the beam after a failure is detected should be in about 10% of this time (Fig. 7, see [18]).

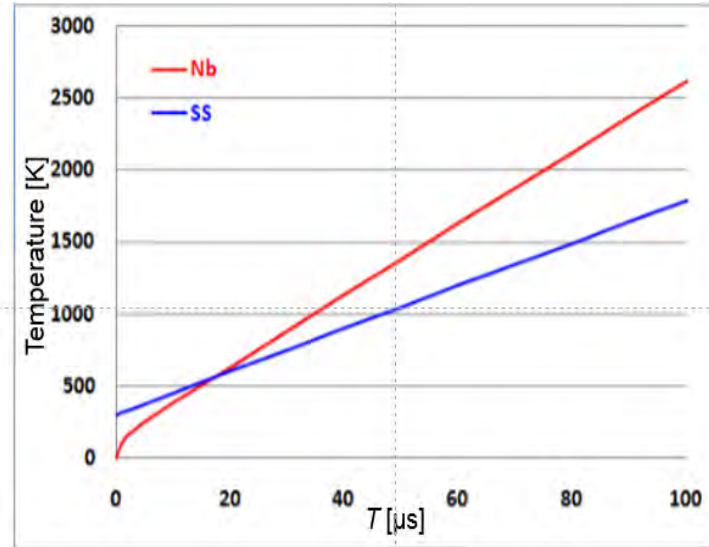
As a comparison, a prediction of the damage limit from FRIB beams is shown in Fig. 8.

## 7 The performance of an accelerator and its availability

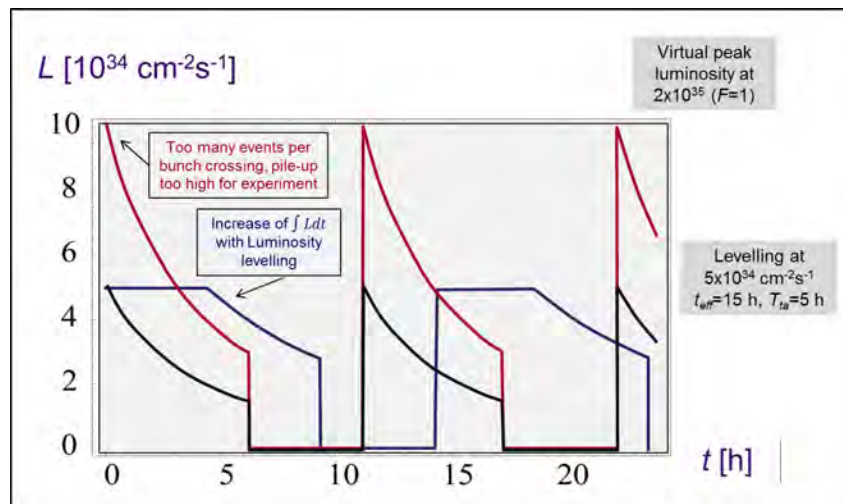
For particle colliders, the total number of events is proportional to the integrated luminosity. For many years the most relevant parameter for the operation of a collider was the peak luminosity. The most significant improvements of the integrated luminosity was achieved by increasing the peak luminosity, e.g. at LEP, at the SPS proton-antiproton collider and at the Tevatron [12]. The accelerators were not as complex as today and availability was not a major concern. With more complex accelerators and limitation of the maximum luminosity this is changing.

HL-LHC (high-luminosity LHC) is a project aiming at an increase of the integrated luminosity of the LHC per year by a factor of 10, from  $20\text{--}30 \text{ fb}^{-1}$  to more than  $300 \text{ fb}^{-1}$  per year. In principle, with the HL-LHC beam parameters a luminosity of  $2 \times 10^{35} [\text{cm}^{-2} \text{ s}^{-1}]$  could be achieved. With a luminosity of  $5 \times 10^{34} [\text{cm}^{-2} \text{ s}^{-1}]$  the number of events per bunch crossing is of the order of 140 (today it is about 30). For the LHC experiments a further increase of the pile-up beyond, say, 140, is not acceptable. In order to achieve the target integrated luminosity, it is planned to level the luminosity during the fill (Fig. 9).

The integrated luminosity depends therefore on the time that the beams are colliding and the experiments take data. The time is directly related to the availability of the accelerator.



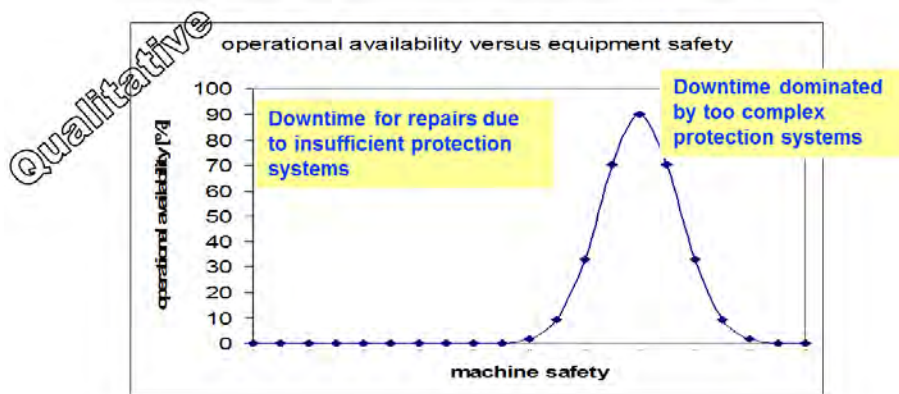
**Fig. 8:** Temperature versus time of stainless steel (SS) from 300 K and niobium (Nb) from 2 K, after being hit by a uranium beam, 100 MeV/u, 200 kW and beam rms radius 1 mm [19].



**Fig. 9:** Luminosity levelling at HL-LHC

For neutron and neutrino sources an important figure of merit is the integrated number of protons on target. Similar quantities can be defined for other accelerators (e.g. synchrotron light sources). For accelerators with many small experiments that take data only for a few days, the availability is important for another reason: if the accelerator is down and not providing beam, users can lose their entire data-taking period.

Damage due to accidental release of energy stored in the beam or in equipment has a major impact on the availability. Machine protection systems prevent such damage. However, machine protection is not an objective in itself; it is to maximize operational availability by minimizing downtime (quench, repairs, waiting for cool-down to access equipment) as well as to avoid expensive repair of equipment and irreparable damage. Since all technical systems cause some downtime, machine protection systems will also contribute to downtime. For an accelerator with limited risk, the presence of protection systems might reduce the overall availability. If the risk is large, protection systems are needed. Side effects from the machine protection systems compromising operational efficiency must be minimized.



**Fig. 10:** Availability versus protection (qualitative graph). For too little protection, failures will lead to damage and reduce availability. For too much protection, the accelerator risks not being able to switch on and the availability is zero. There is some optimum in between.

Figure 10 illustrates the availability for operation as a function of machine protection, for an accelerator with high beam power/large stored beam energy and considerable risk. If there is no protection system, failures will lead to damage and reduce availability. If there are too many and too complex protection systems, the accelerator risks not being able to switch on due to too many interlocks.

### 8 Hazards and risks

A hazard is a situation that poses a level of threat to the accelerator. Hazards are dormant or potential, with only a theoretical risk of damage. Once a hazard becomes ‘active’ it becomes an incident or accident. An accident is defined as an unfortunate incident that happens unexpectedly and unintentionally, typically resulting in damage or injury.

Risk is a quantity that allows us to measure the threat of a hazard, by multiplying consequences and probability for a hazard becoming active:

$$\text{Risk} = \text{Consequences} \times \text{Probability}. \tag{8}$$

Related to accelerators, consequences and probability of an uncontrolled beam loss need to be estimated to evaluate the risk. Machine protection systems prevent damage to equipment and reduce the risk, either by preventing a failure from occurring, or by mitigating the consequences of a failure. The higher the risk, the more effort is needed to prove that the protection system is sufficiently robust. Machine protection needs to be considered during design, construction and operation of the accelerator.

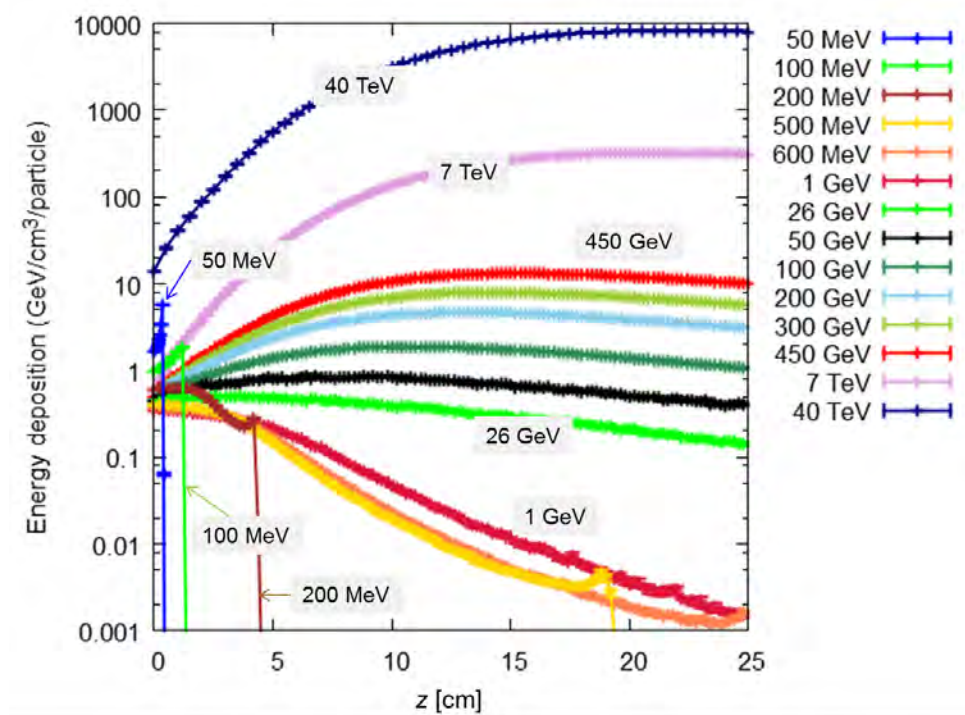
In the following sections, we discuss different types of hazards:

- Particle beams and their effects;
- Electromagnetic energy stored in magnets and RF systems;
- Other sources of energy.

#### 8.1 Hazards related to accelerator systems

Even without operating with beam there can be hazards to be considered.

- A large amount of energy is stored in superconducting magnets. A complex magnet protection system is required [16].



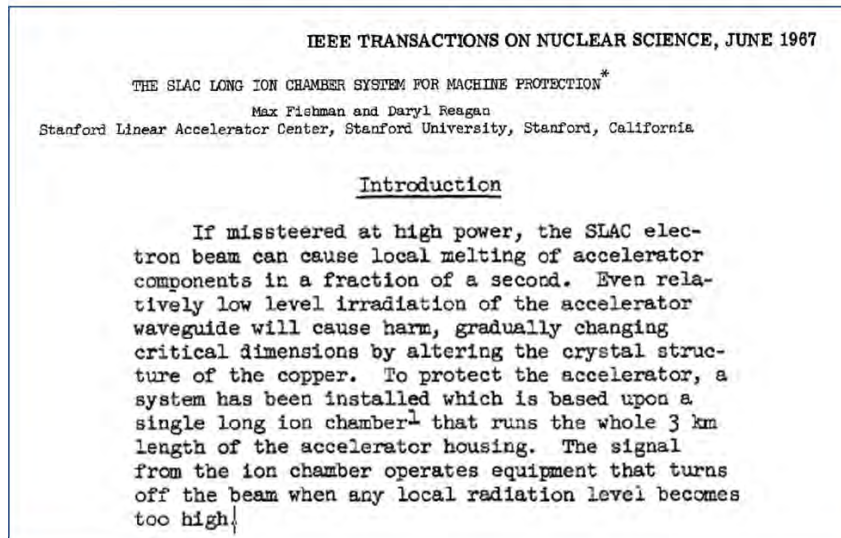
**Fig. 11:** Proton energy deposition for different energies for the impact of protons on copper [21]

- Powering normal-conducting magnets can overheat the magnets if water cooling fails. Monitoring of the temperature is required and interlocks to verify if the cooling works correctly. In case of a cooling problem the power converter is switched off [16].
- Power in RF systems (modulators, klystrons, wave-guides, cavities): the presence of high-voltage arcs can damage the structure. Protection requires complex and fast interlock systems. For high-intensity beams, the RF system has to cope with fast transitions from beam on to beam off [20].
- For high-voltage systems (e.g. kicker magnets) there is always the risk of arcing.
- Powering systems (power converters, power distribution, electrical network [16]).

## 8.2 Hazards related to particle beams

When high-energy particles traverse matter, the particles deposit energy. The energy deposition depends on material, particle type and momentum. In Fig. 11, the energy deposition of one proton as a function of depth is shown for a wide range of energies [21]. The energy deposition leads to a temperature increase. Material can change its material properties, deform, melt, or vaporize. Equipment can become activated. Superconducting magnets can quench (a quench is the transition from superconducting state to normal-conducting state).

- Regular beam losses during operation need be considered, since beam losses lead to activation of equipment and possibly quenches of superconducting magnets. Beam losses can also cause radiation-induced effects in electronics (single-event effects).
- Accidental beam losses due to failures: understand hazards, e.g. mechanisms for accidental beam losses. Hazards become accidents due to a failure; machine protection systems mitigate the consequences.
- Understand effects from synchrotron radiation that potentially lead to damage of equipment.



**Fig. 12:** Title and abstract of a very early paper on accelerator protection [23]

- It is required to understand the interaction of particle beams with the environment, due to the impedance of the vacuum chamber and other accelerator equipment around the beam (kicker magnets, cavities, collimators, etc). This can lead to heating of equipment close to the beam and possibly to damage.

Even if the beam power or stored energy is not large, small beam sizes can cause localized damage. If the beam is lost, say, in a superconducting cavity, the consequences can be serious, as SNS (Spallation Neutron Source at Oak Ridge Laboratory) demonstrated [22].

Particle energy and particle type play an important role. For hadron accelerators beam losses lead to activation of the accelerator equipment. The emitted power of synchrotron radiation is very small and does not lead to a hazard. For electrons and positrons, activation is small, but the emitted power can be very high.

Not only the intensity, but also the beam/bunch structure (number of bunches, repetition frequency and bunch length) determines the interaction with the surrounding equipment due to the interplay of the electromagnetic fields with these structures.

## 9 Strategy for machine protection

Machine protection has been on the agenda for a long time, but only in recent years has it become a significant topic at particle accelerator conferences. A very early paper was published in 1967: The SLAC long ion chamber system for machine protection [23], already illustrating the risks when operating with high-power beams (Fig. 12).

There are some principles for machine protection that need to be considered (sometimes this is referred to the  $P^3$  rule for machine protection):

- Protect the machine;
- Protect the beam;
- Provide the evidence.

**Protect the machine:** The highest priority is to avoid damage to accelerator equipment.

**Protect the beam:** The objective is to maximize beam time, but complex protection systems reduce the availability of the machine. The number of ‘false’ interlocks stopping operation must be minimized. This is a trade-off between protection and operation. A ‘false’ interlock or ‘false’ beam dump is defined as an interlock that stops operation even though there is no risk (example: a temperature sensor reading a wrong value, therefore switching off the power converter of a magnet and stopping beam operation).

**Provide the evidence:** If the protection systems stop operation (e.g. dump the beam or inhibit injection), clear diagnostics should be provided [24]. If something goes wrong (leading to damage, but also a near miss), it should be possible to understand the event. This needs synchronized transient recording of all the important parameters in all relevant systems, as well as long-term logging of parameters with reduced frequency (such as 1 Hz). Examples are the currents in all magnets, beam position, beam losses and beam intensity. The frequency of transient recording depends on the system and can be between Hz and MHz.

For beam operation, a list of all possible failures that could lead to beam loss into equipment should be considered. This is not obvious, since there is a nearly infinite number of mechanisms for losing the beam. However, the most likely failure modes and in particular the worst-case failures and their probabilities must be taken into account for the design of the protection system.

For a specific failure, the consequences of the failure need to be estimated, in terms of damage to equipment (repair requiring investment, e.g. in money), in downtime of the accelerator (e.g. in days) and in radiation dose to personnel accessing equipment (e.g. in mSv). In the estimation of downtime of the accelerator for repairs, the availability of spare parts needs to be considered. If the accelerator was operating for a long time with high-intensity beams, radioactive activation of material must be taken into account. It may be necessary to wait for cool-down of irradiated components to reduce the dose before accessing the equipment.

The second factor entering into the risk is the probability of such a failure happening (e.g. measured in number of failures per year).

## 9.1 Active and passive protection

In case of a failure, it takes a certain time until the beam is affected. The time constant is essential for the design of machine protection systems. We distinguish between two types of failures: failures where there is enough time to detect and mitigate (active protection), and failures where the time is too short for any mitigation (passive protection).

**Active protection** requires the detection of the failure by a sensor in an equipment system as early as possible, or by beam instrumentation detecting when the beam starts to be affected by the failure (for example, increased beam losses or a different trajectory). When a failure is detected, beam operation must be stopped. For synchrotrons and storage rings the beam is extracted by a fast kicker magnet into a beam dump block. The target must be designed to accept the beam pulse without being damaged. Injection must be stopped. For linacs, the beam is stopped in the low-energy part of the accelerator by switching off the source, by deflecting the low-energy beam by electrostatic plates (‘choppers’), or by switching off the Radio Frequency Quadrupole (RFQ) for proton linacs. In the case of an accelerator complex with a chain of several accelerators, injection of beam into the next stage of the accelerator complex should be prevented.

Experience from LHC shows that for most types of failures a careful and fast monitoring of hardware parameters allows stopping beam operation before the beam is affected. Parameters monitored include state signals, other parameters, etc. As an example, a trip of a magnet power converter should be detected as early as possible.

It is not always possible to detect failures at the hardware level. The second method is to detect the initial consequences of a failure with beam instrumentation and to stop the beam before equipment is damaged. This requires reliable beam instrumentation. An electronic system (beam interlock system) links the different protection systems. It ensures that the beam is extracted from a synchrotron, injection



is stopped, or RF acceleration might be stopped (for linacs). The interlock system might include complex logic that depends on the operational state.

**Passive protection:** There are failures (e.g. ultra-fast losses) when active protection is not possible. One example is the protection against mis-firing of an injection or extraction kicker magnet. A beam absorber or collimator is required to stop the mis-kicked beam in order to avoid damage. All possible beam trajectories for such failures must be considered and the absorbers must be designed to absorb the beam energy without being damaged. Another example is a fast extraction of a high-intensity beam from a circular accelerator into a transfer line. When the extraction takes place, the parameters of the transfer line, e.g. the current of the magnets, must be correctly set since for a wrong magnet current the beam would be deflected and possibly damage the vacuum chamber and other components.

There are a certain number of failures that can be completely eliminated. As an example, fast diagnostic kicker magnets that could deflect the beams into the vacuum chamber wall should only be installed in high-intensity machines if they are indispensable.

## 10 Accidents at accelerators: looking into the past

In this section and other lectures in this school, some examples of accidents are given:

- Damage to silicon detector in UA2 at the SPS proton–antiproton collider [25];
- Vacuum chamber in SPS extraction line incident (see presentation in this school [26]);
- Tevatron proton–antiproton collider (see presentation in this school [27]);
- LHC magnet powering (see presentation in this school [28]);
- CERN-LINAC 4 during commissioning at 3 MeV LINAC 4 (2013) at very low energy: beam hit a bellow and a vacuum leak developed;
- J-PARC radioactive material leak accident (see presentation in this school [5]).

In Fig. 13, the deposited beam or magnetic energy is shown for different events. It ranges from about 1 J sufficient to quench superconducting magnets to several 100 MJ that caused major damage to the LHC for the accident during powering tests in 2008.

### 10.1 Damage to silicon detector in UA2 at the SPS proton–antiproton collider

The silicon detector in the UA2 experiment at the SPS proton–antiproton collider was damaged during the injection process. Because of beam–beam effects, the beams needed to be separated at injection energy of 26 GeV and during the energy ramp to 315 GeV. This was done with electrostatic separators; their strengths were also ramped with the energy.

One day, during injection at 26 GeV, the separators were left accidentally at the settings for 315 GeV corresponding to a much too large angle and the beam was injected but not circulating, see Fig. 14. The separators directed the beam directly into UA2 using part of the detector as a beam dump during some minutes until the problem was understood [25].

### 10.2 CERN-LINAC4 during commissioning at 3 MeV

On 12 December 2013 a vacuum leak on a bellow developed in the MEBT (medium energy beam transfer) line (see Fig. 15). The analysis showed that the beam has been hitting the bellow during a special measurement with very small beams in the vertical plane but large in the horizontal plane. About 16% of the beam was lost for about 14 min and damaged the bellow. The consequences were minor since LINAC4 is still being commissioned and not used in the chain of LHC injectors. The event demonstrates that even beams with very low power can cause damage.

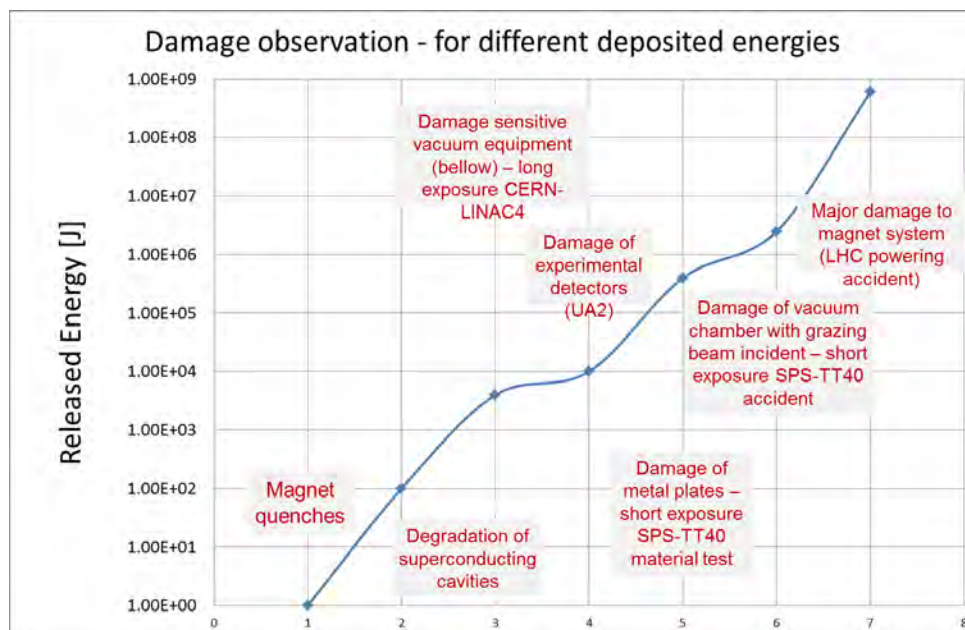


Fig. 13: Deposited energy during quenches of magnets and damage of components for different events

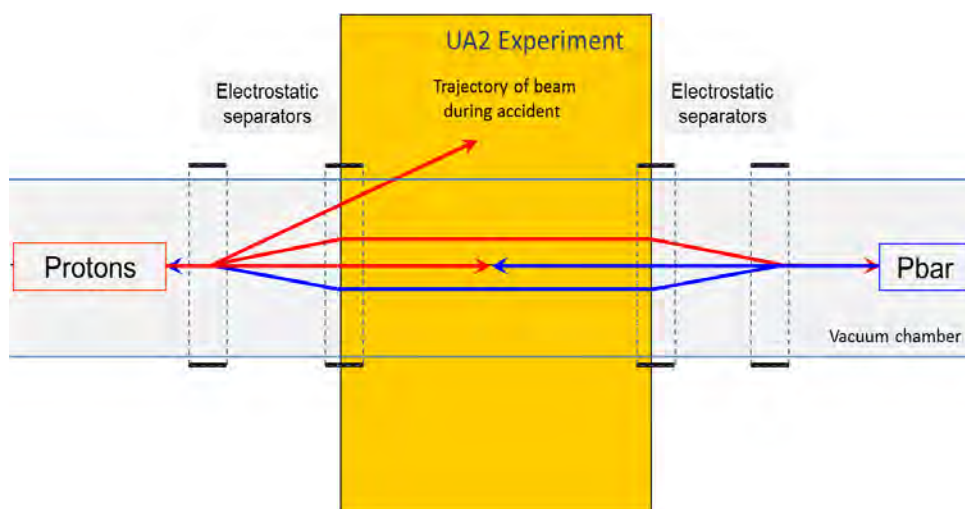
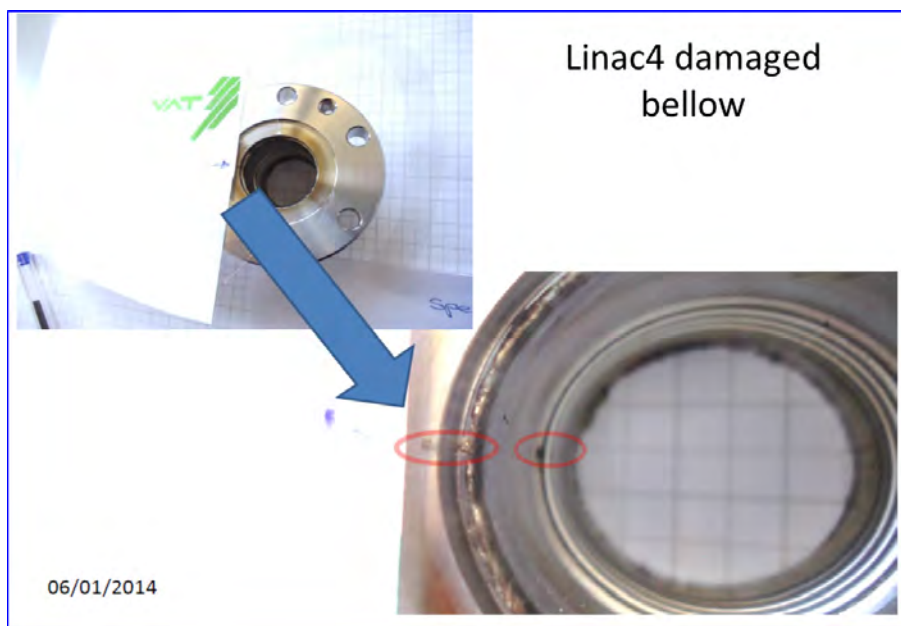


Fig. 14: Illustration of trajectories of proton and antiproton beams for collisions, during injection and during the accident.

### 10.3 LHC magnet powering accident in 2008

During the first phase of operation between 2009 and 2013 the magnetic field in the dipole magnets was limited, and therefore LHC was operating with a momentum of up to 4 TeV/c and the maximum stored beam energy was up to about 140 MJ. This was the consequence of the 2008 LHC accident that happened during test runs without beam. A magnet interconnect was defective and the circuit opened. An electrical arc provoked a helium pressure wave damaging about 600 m of the LHC and polluting the beam vacuum over more than 2 km. An overpressure from the expansion of liquid helium damaged the structure. A total of 53 magnets had to be repaired. A detailed description of the accident is given in [28].



**Fig. 15:** Damaged bellow showing signs of damage. After the damage a vacuum leak was observed and repair was required.

### Acknowledgements

I wish to thank many colleagues from CERN, ESS and the authors of the listed papers for their help and for providing material for this paper.

### References

- [1] C. Sibley, Machine protection strategies for high power accelerators, Particle Accelerator Conf., Portland, OR, USA, 2003, p. 607. <http://dx.doi.org/10.1109/pac.2003.1288989>
- [2] J. Wenninger, State-of-the-art and future challenges for machine protection systems, 5th Int. Particle Accelerator Conf., IPAC 2014, Dresden, Germany, 15–20 June 2014.
- [3] R. Schmidt, Machine protection, in Proceedings of the CAS – CERN Accelerator School: Course on Beam Diagnostics, Dourdan, France, 28 May–6 June 2008, edited by D. Brandt, CERN-2009-005 (CERN, Geneva, 2009), pp. 581-596. DOI: <http://dx.doi.org/10.5170/CERN-2009-005.581>
- [4] R. Schmidt, Machine protection, in Proceedings of the CAS – CERN Accelerator School: Advanced Accelerator Physics, Trondheim, Norway, 18–29 August 2013, edited by W. Herr, CERN-2014-009 (CERN, Geneva, 2014), pp. 221-243. DOI: <http://dx.doi.org/10.5170/CERN-2014-009.221>
- [5] S. Rokni, Personnel protection systems, JAS – Joint International Accelerator School on Beam Loss and Accelerator Protection, Newport Beach, CA, USA, 2014.
- [6] R. Koontz, Multiple beam pulse capability of the SLAC injector, 2nd Particle Accelerator Conf., Washington, DC, USA [*IEEE Trans. Nucl. Sci.* **14** (1967) 104. <http://dx.doi.org/10.1109/tns.1967.4324532>]
- [7] M. Ross, Single pulse damage in copper, LINAC 2000, Monterey, CA, USA, August 2000.
- [8] M. Ross, The next linear collider machine protection system, 1999 Particle Accelerator Conf., New York, NY, USA, 1999.
- [9] M. Jonker, H. Schmickler, R. Schmidt, D. Schulte and M. Ross, Machine protection issues and solutions for linear accelerator complexes, Linac 2012, Tel Aviv, Israel, 9–14 September 2012.

- [10] R. Appleby, B. Goddard, A. Gomez-Alonso, V. Kain, T. Kramer, D. Macina, R. Schmidt and J. Wenninger, *Phys. Rev. Spec. Top. Accel. Beams* **13** (2010) 061002. <http://dx.doi.org/10.1103/PhysRevSTAB.13.061002>
- [11] A. Piwinski, Dependence of the luminosity on various machine parameters and their optimization at PETRA, Proc. Particle Accelerator Conf., PAC 1983, Santa Fe, NM, USA, 21–23 March 1983. <http://dx.doi.org/10.1109/tns.1983.4332822>
- [12] V. Shiltsev, Achievements and lessons from TEVATRON, 2nd Int. Particle Accelerator Conf., IPAC 2011, San Sebastian, Spain, 4–9 September 2011.
- [13] P. Vagin, O. Bilani, A. Schöps, M. Tischer, S. Tripathi and T. Vielitz, Radiation damage of undulators at PETRA III, 5th Int. Particle Accelerator Conf., IPAC 2014, Dresden, Germany, 15–20 June 2014.
- [14] E. Carlier *et al.*, The LEP beam dumping system, 4th European Particle Accelerator Conf., London, England, 27 June–1 July 1994.
- [15] R. Bailey *et al.*, Synchrotron radiation effects at LEP, 6th European Particle Accelerator Conf., EPAC-98, Stockholm, Sweden, 22–26 June 1998.
- [16] H. Pfeffer, Protection of hardware: powering systems, these proceedings.
- [17] L. Tchelidze, In how long the ESS beam pulse would start melting steel/copper accelerating components, ESS AD Technical Note ESS/AD/0031 (2012).
- [18] A. Nordt, A. Apollonio and R. Schmidt, Overview on the design of the machine protection system for ESS, 5th Int. Particle Accelerator Conf., IPAC 2014, Dresden, Germany, 15–20 June 2014.
- [19] Y. Zhang, D. Stout and J. Wei, Analysis of beam damage to FRIB driver linac, Proc. SRF 2011, Chicago, IL, USA, 2011.
- [20] S.H. Kim, Protection of hardware: RF systems, these proceedings.
- [21] F. Burkart, V. Chetvertkova, R. Schmidt, R. Rossel and D. Wollmann, Damage potential for proton beams in the momentum range from 50 MeV/c to 40 TeV/c, 6th Int. Particle Accelerator Conf., IPAC 2015, Richmond, VA, USA, 3–8 May 2015.
- [22] M. Plum, Beam dynamics and beam losses – linear machines, these proceedings.
- [23] M. Fishman *et al.*, The SLAC long ion chamber system for machine protection, US Particle Accelerator Conf., Washington, DC, USA, 1–3 March 1967, CERN-ATS-2011-064 (1967).
- [24] M. Zerlauth *et al.*, The LHC post mortem analysis framework, ICALEPCS 2009, Kobe, Japan, 2009.
- [25] E. Beuville *et al.*, Measurement of degradation of silicon detectors and electronics in various radiation environments, Technical Report, CERN-LHC-PROJECT-Report-1168, CERN-EF 89/4 (1989).
- [26] V. Kain, Beam transfer and machine protection, these proceedings.
- [27] N. Mokhov, Beam material interaction, heating and activation, these proceedings.
- [28] J. Wenninger, Machine protection and operation for LHC, these proceedings.

## Beam Dynamics and Beam Losses – Circular Machines

*V. Kain*

CERN, Geneva, Switzerland

### Abstract

A basic introduction to transverse and longitudinal beam dynamics as well as the most relevant beam loss mechanisms in circular machines will be presented in this lecture. This lecture is intended for physicists and engineers with little or no knowledge of this subject.

### Keywords

Synchrotron; transverse and longitudinal beam dynamics; beam loss mechanisms; equipment failure consequences.

## 1 Introduction

A vast variety of mechanisms can lead to beam losses in accelerators. Examples are collisions in colliders, beam gas interactions, intra-beam scattering, the Touscheck effect, RF noise, collective effects, transition crossing, equipment failures and many more. Losses have an impact on performance, such as the luminosity in a collider or the brightness of the beam. Losses lead to radio-activation, which can have an impact on machine availability and maintainability. Hands-on maintainability requires radiation of less than 1 mSv/h. High losses can cause downtime of the accelerator, due to quenches in superconducting machines, or even damage to components.

Particles are lost in the vacuum chamber if their transverse trajectory amplitudes are larger than the dimension of the vacuum chamber. It is important to understand the mechanisms that can create large amplitudes, to give input for the design of machine protection reaction times, collimators, absorbers, instrumentation, etc. The important characteristic in this respect is the number of particles lost per unit time  $\Delta N/\Delta t$ . The so-called beam lifetime  $\tau$  is defined with

$$N(t) = N_0 \cdot e^{-\frac{t}{\tau}}, \quad (1)$$

where  $N_0$  is the initial intensity. At the accelerator design stage, such questions as, “What is the minimum possible beam lifetime?” and, “What is the tolerable beam loss rate for accelerator components?” have to be answered. This lecture will discuss some of the typical beam loss mechanisms in circular machines. To make the discussion accessible to non-accelerator physicists, a good part of the lecture will be spent on introducing the most basic concepts of accelerator physics.

## 2 Principles of transverse and longitudinal beam dynamics – synchrotrons

This part of the lecture will only introduce the concepts required later on for the discussion of typical beam loss mechanisms. It is based on the lectures by B. Holzer, F. Tecker and O. Bruning at the CERN Accelerator School [1]. A complete introduction to the subject of accelerator physics can be found in Ref. [2].

A typical layout of a synchrotron is shown in Fig. 1. Bending magnets are used to keep the particles on the synchrotron orbit. Strong focusing from an alternating gradient lattice ensures trajectory stability over many turns. Radio-frequency accelerating structures increase the particle momentum turn by turn. Other insertions are arranged to inject or extract the beam with dedicated equipment.

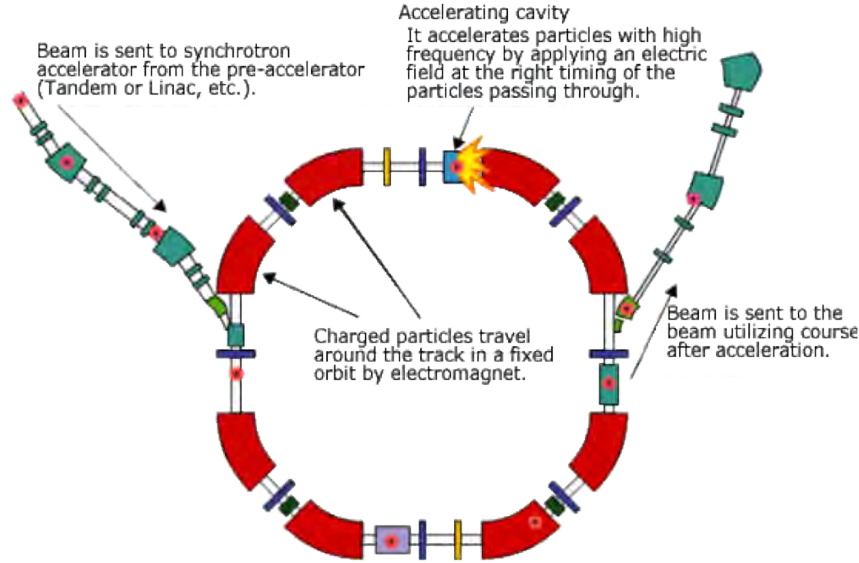


Fig. 1: Typical layout of synchrotron

## 2.1 The transverse plane

The trajectories of charged particles can be manipulated with electromagnetic fields via the Lorentz force:

$$\vec{F} = q \cdot (\vec{E} + \vec{v} \times \vec{B}) . \quad (2)$$

For relativistic particles, the effect of magnetic fields is much enhanced via the product with the velocity, and dipole fields are mainly used as guide fields. For the particles to stay on a circular orbit in a circular machine, the Lorentz force  $F_L$  from the magnetic field has to compensate the centrifugal force  $F_{\text{centr}}$ .

$$\begin{aligned} F_L &= qvB , \\ F_{\text{centr}} &= \frac{mv^2}{\rho} . \end{aligned} \quad (3)$$

From

$$\frac{mv^2}{\rho} = qvB , \quad (4)$$

the well-known relation for the product  $B\rho$ , the beam rigidity, follows:

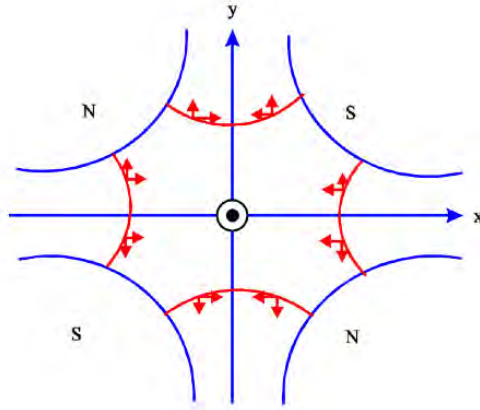
$$\frac{p}{q} = B\rho . \quad (5)$$

A useful formula for ‘back-of-the-envelope’ estimates is

$$\frac{1}{\rho [\text{m}]} \approx 0.3 \frac{B [\text{T}]}{p [\text{GeV}/c]} . \quad (6)$$

Vertical dipole magnets define the design trajectory in the horizontal plane. In a beam of many particles, the trajectories of the particles will deviate from the design trajectory. Without a restoring force, the trajectories will deviate more and more until the particles are eventually lost. Quadrupole magnets provide the required restoring force. They produce a dipole field in the horizontal and vertical plane that increases as a function of the distance from the design trajectory. A schematic cross-section of a quadrupole with its field lines is given in Fig. 2. For example, the vertical field in a quadrupole will be a function of the horizontal position:

$$F(x) = q \cdot v \cdot B(x) . \quad (7)$$



**Fig. 2:** A schematic cross-section of a quadrupole magnet. The red arrows indicate the direction of the force on the particle. A quadrupole that is focusing in the horizontal plane is defocusing in the vertical plane.

It depends linearly on the deviation from the design trajectory:

$$B_y = g \cdot x . \quad (8)$$

The horizontal field is

$$B_x = g \cdot y . \quad (9)$$

A focusing quadrupole in the horizontal plane will be defocusing in the vertical one and vice versa. The characteristic parameter of a quadrupole magnet is its gradient,

$$g = \frac{2\mu_0 n I}{r^2} \left[ \frac{\text{T}}{\text{m}} \right] , \quad (10)$$

where  $r$  is the distance between the quadrupole centre and the pole surface. The normalized gradient is often used to define the strength of the quadrupole:

$$k = \frac{g}{p/e} [\text{m}^{-2}] . \quad (11)$$

### 2.1.1 Equation of motion

To describe the particle trajectories in the synchrotron, the equation of motion has to be solved:

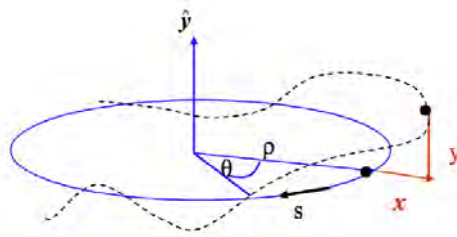
$$F_r = m a_r = e B_y v . \quad (12)$$

To simplify the task, we use the Frenet–Serret coordinate system, see Fig. 3, and the magnetic field is expanded in a Taylor series,

$$B_y(x) = B_{y0} + \frac{\partial B_y}{\partial x} x + \frac{1}{2} \frac{\partial^2 B_y}{\partial x^2} x^2 + \frac{1}{3!} \frac{\partial^3 B_y}{\partial x^3} x^3 + \dots , \quad (13)$$

and normalized with  $p/e$ ,

$$\frac{B_y(x)}{p/e} = \frac{1}{\rho} + k x + \frac{1}{2} m x^2 + \frac{1}{3!} n x^3 + \dots \quad (14)$$



**Fig. 3:** The trajectory coordinates are given with respect to the Frenet–Serret frame, which rotates with the ideal particle around the accelerator. The ideal particle has design momentum  $p_0 = m_0\gamma v$ . It has coordinates  $x = 0$ ,  $y = 0$  for a certain longitudinal location  $s$  lying on the design orbit.

Only the terms linear in  $x$  are kept:

$$\frac{B_y(x)}{p/e} \approx \frac{1}{\rho} + kx. \quad (15)$$

Using the magnetic field as defined in Eq. (15), the equation of motion in the horizontal plane in the Frenet–Serret frame turns out to be

$$x'' + x \left( \frac{1}{\rho^2} - k \right) = x'' + xK = 0, \quad (16)$$

where  $x' = dx/ds$  and  $K$  combines the focusing properties of dipoles and quadrupoles. Assuming that there are no vertical bending magnets, the equation of motion in the vertical plane becomes

$$y'' + ky = 0. \quad (17)$$

Around the accelerator,  $K$  will not be constant, but will depend on  $s$ . However,  $K(s)$  will be periodic with  $L$ ,  $K(s+L) = K(s)$ , where  $L$  is the lattice period. For instance,  $L$  can be the circumference of the accelerator. We then have

$$x''(s) + K(s)x(s) = 0. \quad (18)$$

This type of equation of motion with these characteristics of non-constant but periodic restoring force is called the Hill equation, after George Hill, an astronomer of the 19th century. The general solution of the Hill equation is a quasi-harmonic oscillation:

$$x(s) = \sqrt{\epsilon} \sqrt{\beta(s)} \cos(\psi(s) + \phi). \quad (19)$$

In the case of accelerators, this quasi-harmonic oscillation is called *betatron oscillation*. The amplitude and phase of the oscillation depend on the position in the ring.  $\epsilon$  and  $\phi$  are integration constants and depend on the initial conditions. The so-called *beta function*,  $\beta(s)$ , is a periodic function,  $\beta(s+L) = \beta(s)$ , and is determined by the focusing properties of the lattice, i.e. quadrupole strengths. The *phase advance* of the oscillation between point  $s = 0$  and point  $s$  in the lattice is

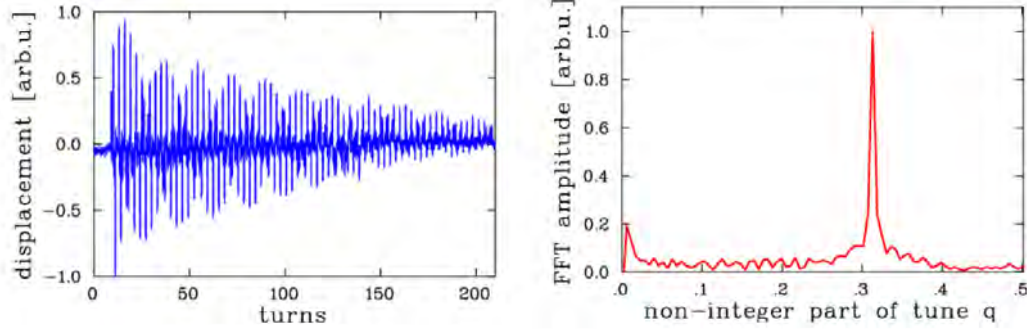
$$\psi(s) = \int_0^s \frac{ds}{\beta(s)}. \quad (20)$$

Two other functions are commonly used:  $\alpha(s)$  and  $\gamma(s)$ . They are defined as

$$\alpha(s) = -\frac{1}{2}\beta'(s), \quad (21)$$

$$\gamma(s) = \frac{1 + \alpha(s)^2}{\beta(s)}. \quad (22)$$





**Fig. 4:** Turn-by-turn oscillation recorded at a beam position monitor after a one-turn excitation with a kicker magnet on the left side. The fast Fourier transform spectrum of this oscillation on the right side. The amplitude peak in the spectrum indicates the tune of the oscillation.

### 2.1.2 The transport matrix

The integration constants  $\phi$  and  $\epsilon$  can be defined from an initial position  $x_0$  and angle  $x'_0$  at location  $s(0) = s_0$  and  $\psi(0) = 0$ , and can be replaced in the equations of position  $x$  and angle  $x'$ ,

$$\begin{aligned} x(s) &= \sqrt{\epsilon} \sqrt{\beta(s)} \cos(\psi(s) + \phi), \\ x'(s) &= -\frac{\sqrt{\epsilon}}{\sqrt{\beta(s)}} \alpha(s) \cos(\psi(s) + \phi) + \sin(\psi(s) + \phi), \end{aligned} \quad (23)$$

such that they become a function of  $x_0$  and  $x'_0$ , as

$$\begin{pmatrix} x \\ x' \end{pmatrix}_{s_1} = M \begin{pmatrix} x \\ x' \end{pmatrix}_{s_0}, \quad (24)$$

where  $M$  is the *transport matrix*:

$$M = \begin{pmatrix} \sqrt{\frac{\beta}{\beta_0}} (\cos \psi + \alpha_0 \sin \psi) & \sqrt{\beta \beta_0} \sin \psi \\ \frac{(\alpha_0 - \alpha) \cos \psi - (1 + \alpha \alpha_0) \sin \psi}{\sqrt{\beta \beta_0}} & \sqrt{\frac{\beta_0}{\beta}} (\cos \psi - \alpha \sin \psi) \end{pmatrix}. \quad (25)$$

Equation (24) is a very useful relation. The trajectory in terms of position and angle can be calculated at any point of the ring as long as the coordinates at a position  $s_0$  and the so-called *Twiss functions*,  $\alpha$  and  $\beta$ , at both longitudinal positions are known.

### 2.1.3 The tune

Another important parameter in a circular accelerator is the so-called *tune*, the number of betatron oscillations per turn

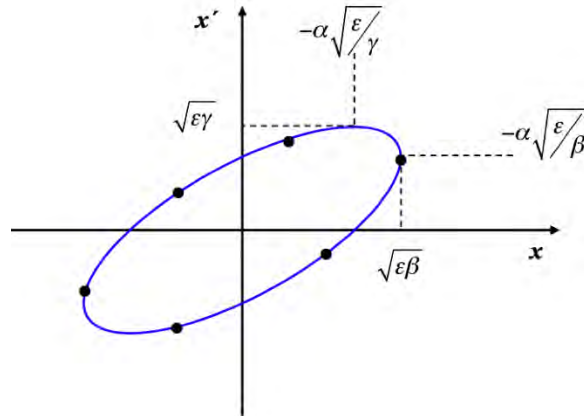
$$Q = \frac{\psi(L_{\text{turn}})}{2\pi} = \frac{1}{2\pi} \oint \frac{ds}{\beta(s)}. \quad (26)$$

As we will see later, an exact knowledge of the tune in both transverse planes and the ability to correct the tune is of great importance for beam stability. The machine tune can be calculated from the turn-by-turn beam position data at a beam position monitor. The tune can then be obtained from the fast Fourier transform of the turn-by-turn data, as indicated in Fig. 4.

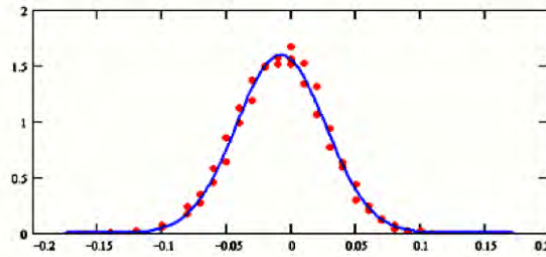
### 2.1.4 Phase-space ellipse and emittance

With  $x$  and  $x'$  in Eq. (23), one can solve for  $\epsilon$ :

$$\epsilon = \gamma(s)x(s)^2 + 2\alpha(s)x(s)x'(s) + \beta(s)x'(s)^2. \quad (27)$$



**Fig. 5:** The trajectory of a particle in phase space  $x, x'$ , turn after turn, is an ellipse. The orientation and shape is defined by the focusing properties of the lattice, whereas the area of the ellipse is an intrinsic property of the beam.



**Fig. 6:** Transverse profile measurements of a particle beam (red dots), using a wire scanner, and Gaussian fit (blue line).

The result in Eq. (27) is the parametric representation of an ellipse in  $x, x'$ -space, see Fig. 5. The shape and orientation of the ellipse are given by the Twiss parameters,  $\beta$ ,  $\alpha$  and  $\gamma$ . The area of the ellipse is  $A = \pi \cdot \epsilon$ , which is a constant of motion according to *Liouville's theorem*. Therefore,  $\epsilon$  is also constant, and is called the *Courant–Snyder invariant*. The area of the ellipse is an intrinsic property of the beam and cannot be changed by the focusing properties of the machine.

Typically, the particles in an accelerator have a Gaussian particle distribution in position and angle. The distribution in position in the horizontal plane, for example, follows the well-known relation

$$\rho(x) = \frac{N}{\sqrt{2\pi}\sigma_x} \cdot e^{-\frac{x^2}{2\sigma_x^2}}. \quad (28)$$

An example of a transverse profile measurement with a wire scanner is shown in Fig. 6. The emittance  $\epsilon$  of a beam of many particles corresponds to the ellipse in phase space that contains 68.3% of the particles, such that the standard deviation of a Gaussian distribution corresponds to

$$\sigma_x = \sqrt{\epsilon\beta_x}. \quad (29)$$

The beam emittance is an invariant, since the ellipse areas in phase-space are invariant. It shrinks, however, during acceleration, as will be shown next.

*Liouville's* theorem from Hamiltonian mechanics states that the volume in phase space is conserved for the canonical variables  $p$  and  $q$ , where  $q$  is typically a position coordinate and  $p$  the momentum:

$$\int p \, dq = \text{const} . \quad (30)$$

For our discussion,  $q$  can be written as  $x$  and  $p = \gamma m v = m c \gamma \beta_x$ , where  $\beta_x = v_x/c$ . The angle  $x'$  can be transformed into

$$x' = \frac{dx}{ds} = \frac{dx}{dt} \frac{dt}{ds} = \frac{\beta_x}{\beta} , \quad (31)$$

such that *Liouville's* theorem from Eq. (30) can be rewritten as

$$\int p \, dq = m c \int \gamma \beta_x dx = m c \gamma \beta \int x' dx = \text{const} . \quad (32)$$

During acceleration,  $\gamma$  and  $\beta$  increase. For the left-hand side of Eq. (32) to remain constant, the area in phase space and therefore also the emittance has to decrease proportionally with  $1/(\beta\gamma)$ . The beam size therefore shrinks during acceleration:

$$\varepsilon = \int x' dx \propto \frac{1}{\beta\gamma} . \quad (33)$$

## 2.2 Longitudinal plane

Circular accelerators allow for multiple application of the same RF accelerating voltage to increase the particle energy. The energy gain per turn for a sinusoidal RF voltage is

$$\Delta E = eV \sin \phi = eV \sin \omega_{\text{RF}} t . \quad (34)$$

The synchrotron has a fixed orbit and bending radius and a magnetic field that increases synchronously with the beam energy. A synchronous RF phase of the RF field exists, for which the energy gain of the particles fits the increase of the magnetic field. A particle that arrives turn after turn at the same phase,  $\phi = \phi_s = \text{const}$ , with respect to the RF field is called a *synchronous particle*. For the acceleration to work, the RF frequency must be locked to the beam revolution frequency,

$$\omega_{\text{RF}} = h \omega_{\text{rev}} , \quad (35)$$

where  $h$  is an integer and is called the harmonic number. The energy gain per turn for the synchronous particle is  $\Delta E = eV \sin \phi_s$ . With  $E^2 = E_0^2 + p^2 c^2 \rightarrow \Delta E = v \Delta p$  and  $v = 2\pi R/T_{\text{turn}}$ , the energy gain can be written as

$$2\pi R \frac{dp}{dt} = q \cdot V \cdot \sin \phi_s . \quad (36)$$

Thus, the stable phase and voltage are changed during acceleration. In the LHC, for example, the energy ramp takes more than 15 min, and the stable phase is close to  $180^\circ$ . The total energy gain per turn is only about 500 keV.

### 2.2.1 The principle of phase stability

As not all particles will go through the accelerating gap at exactly the same time, not all particles will receive the same energy gain; therefore, not all particles will have the same energy. The particles of a beam whose energy is distributed around a mean energy are all accelerated as long as the synchronous phase is chosen adequately (and the energy differences of the different particles are not too large). This is due to the *principle of phase stability*.

Let us assume a group of non-relativistic particles, where the energy increase is still transferred into velocity increase. The particles  $P_1$  and  $P_2$  in Fig. 7 are two different synchronous particles; they

will see the same energy gain turn after turn. The particles  $N_1$ ,  $M_1$ ,  $N_2$  and  $M_2$  represent the particle distribution around the two synchronous particles. The energy gain as a function of time for the different particles coming from the sinusoidal RF field is also shown. The particle  $N_1$  had a larger energy than  $P_1$  in the previous turn. It arrived earlier this turn and will get less energy than  $P_1$ . The particle  $M_1$  on the other hand had less energy than  $P_1$  and arrived later than the synchronous particle. It will get more energy this time and will move closer to the synchronous particle. It will therefore stay synchronous with acceleration.

The situation is different for the particles around synchronous particle  $P_2$ . Particle  $M_2$  arrived earlier, as it had more energy than  $P_2$  and will get even more energy during this passage through the accelerating gap. It will move away from  $P_2$ . Particle  $N_2$  was too slow and will become even slower this time round. These particles will not stay synchronous with the acceleration and the changing magnetic field and will be lost in the vacuum chamber. The two synchronous phases for  $P_1$  and  $P_2$  are not equivalent. The synchronous phase has to be chosen adequately.

If a particle is shifted in momentum, it will run on a different orbit with a different length. The parameter *momentum compaction*,  $\alpha$ , gives the relative orbit length change for a given relative momentum change:

$$\alpha = \frac{dL/L}{dp/p}. \quad (37)$$

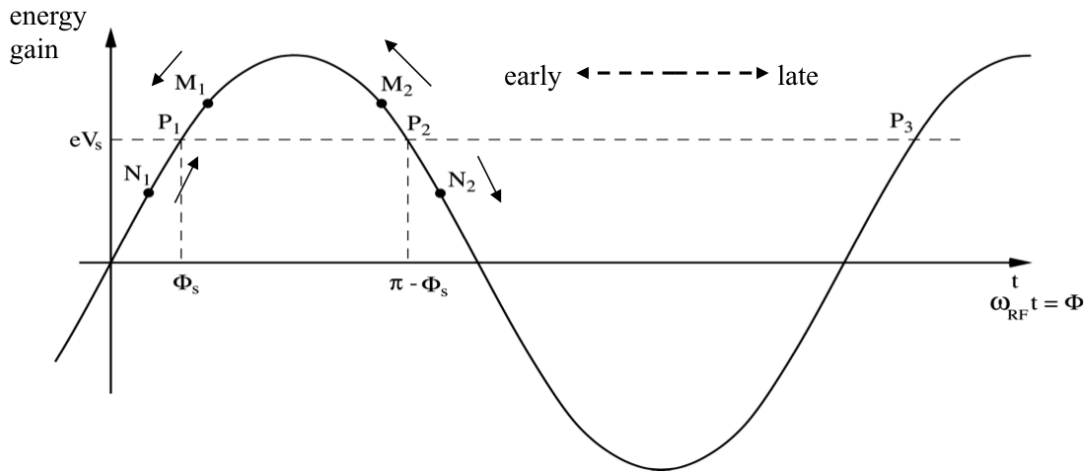
The particle will also have a different velocity and hence a different revolution frequency. The *slippage factor* parameter,  $\eta$ , gives the relative revolution frequency change for a given momentum change.  $\eta$  depends on the *momentum compaction* as

$$\eta = \frac{df_{\text{rev}}/f_{\text{rev}}}{dp/p} = \frac{1}{\gamma^2} - \alpha, \quad (38)$$

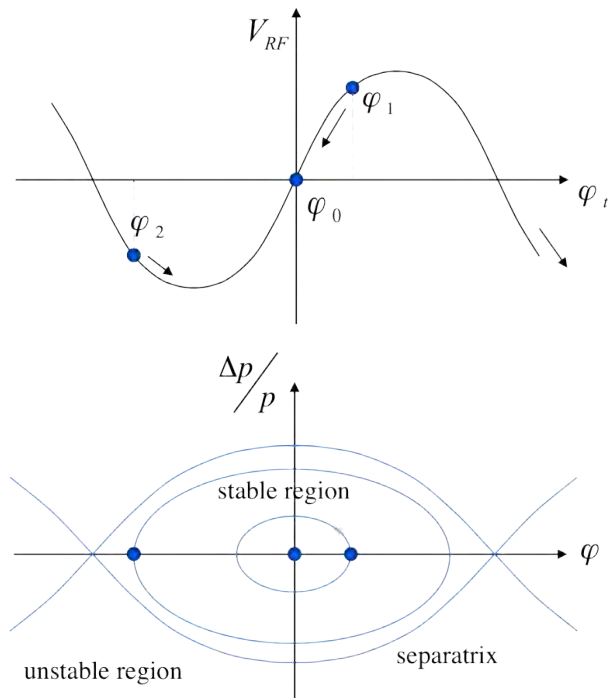
where  $\gamma$  is the relativistic gamma. The energy corresponding to  $\gamma = \gamma_t = 1/\sqrt{\alpha}$  divides the longitudinal motion into two regimes. The energy  $\gamma_t$  is called the *transition energy*. Below transition energy ( $\gamma < \gamma_t$ ,  $\eta > 0$ ), higher momentum corresponds to a higher revolution frequency. Above transition energy ( $\gamma > \gamma_t$ ,  $\eta < 0$ ), higher momentum leads to a lower revolution frequency. Below transition energy, an energy increase still leads to a velocity increase. Above the transition energy, where  $v \approx c$ , the velocity stays roughly constant and the increase in momentum just leads to an increase in path length. In addition, transition crossing during acceleration makes the previously stable synchronous phase unstable. In Fig. 7, the synchronous particle  $P_1$  has a correct stable phase below transition energy; above the transition energy, the synchronous phase of  $P_2$  has the stable phase. The moment of transition during acceleration is delicate. The RF system needs to make a rapid phase change, a *phase jump*, when crossing the transition.

### 2.2.2 The RF bucket and RF acceptance

As in the transverse plane, the particles are oscillating in the longitudinal plane. The particles keep oscillating around the stable synchronous particle varying phase and  $dp/p$ , see Fig. 8. The separatrix defines the region of stable motion, the so-called *bucket*. The entire particle distribution needs to fit into the bucket, to avoid particle losses. The bucket area, called the *RF acceptance*, is measured in electronvolts in  $\Delta E-\Delta t$  space, which is equivalent to  $\Delta p/p-\Delta\phi$  space. The number of buckets around the ring corresponds to the harmonic number  $h$ . The bucket area is largest when the synchronous phase is  $0^\circ$ , or  $180^\circ$ , where the beam is not accelerated. For acceleration, the synchronous phase has to move towards  $90^\circ$  and the buckets become smaller, see Fig. 9. The RF acceptance increases with RF voltage, however. *RF acceptance* plays an important role for losses created by RF capture and stored beam lifetime. During bucket-to-bucket transfer from one machine to another, the bunches might arrive with small momentum and phase errors. If the RF acceptance is too small, part of the injected bunch ends up



**Fig. 7:** The principle of phase stability (Courtesy of F. Tecker)



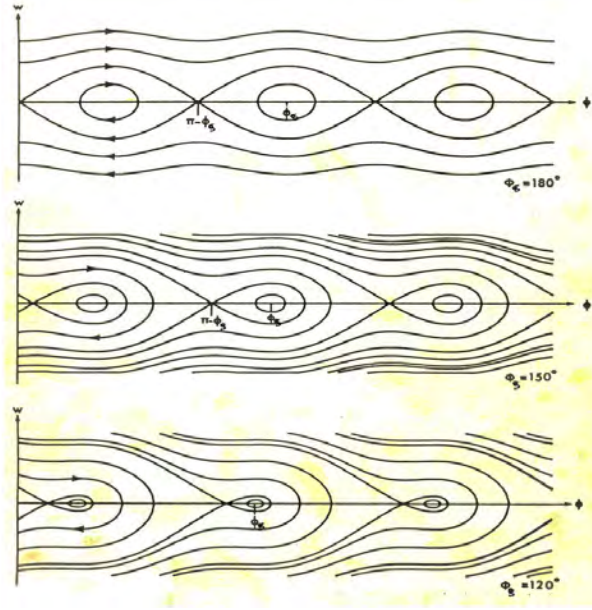
**Fig. 8:** The longitudinal motion in the upper plot follows the trajectory in phase-space in the lower plot. The separatrix defines the limit of stable motion (Courtesy of F. Tecker.)

outside the RF acceptance. This part of the beam will not be accelerated with the rest of the beam when the momentum increases, and will be lost on the vacuum chamber.

### 2.2.3 Synchrotron oscillations

To describe the motion of the particles in the longitudinal plane, their coordinates are expressed with respect to the coordinates of the synchronous particle. Let us call the synchronous particle  $P_s$ ; a particle  $P$  has a phase difference with respect to the synchronous particle of

$$\Delta\phi = \phi - \phi_s, \tag{39}$$



**Fig. 9:** The buckets around the ring shrink during acceleration when the synchronous phase is moved towards  $90^\circ$  (Courtesy of F. Tecker.)

and  $P$  will also have a different revolution frequency:

$$\begin{aligned} \frac{d\Delta\phi}{dt} &= -2\pi h \Delta f_{\text{rev}} , \\ \frac{d^2\Delta\phi}{dt^2} &= -2\pi h \frac{d\Delta f_{\text{rev}}}{dt} . \end{aligned} \quad (40)$$

When the particles cross the RF cavity, the momentum increase of the two particles will be different:

$$\begin{aligned} 2\pi R \frac{dp_s}{dt} &= q \cdot V \cdot \sin \phi_s , \\ 2\pi R \frac{dp}{dt} &= q \cdot V \cdot \sin \phi , \\ 2\pi R \frac{d\Delta p}{dt} &= q \cdot V \cdot \sin \phi - q \cdot V \cdot \sin \phi_s . \end{aligned} \quad (41)$$

Using the definition of the slippage factor,

$$\eta = \frac{df_{\text{rev}}/f_{\text{rev}}}{dp/p} = \frac{\Delta f_{\text{rev}}/f_{\text{rev}}}{\Delta p/p_s} ,$$

Eq. (40) can be transformed to

$$\frac{d^2\Delta\phi}{dt^2} = -2\pi h \frac{d\Delta f_{\text{rev}}}{dt} = -\frac{2\pi\eta h f_{\text{rev}}}{p_s} \frac{d\Delta p}{dt} . \quad (42)$$

Using Eq. (41) for  $d\Delta p/dt$ , the second-order non-linear differential equation describing the synchrotron motion is obtained:

$$\frac{d^2\Delta\phi}{dt^2} + \frac{\eta \cdot f_{\text{RF}}}{R \cdot p_s} q \cdot V (\sin \phi - \sin \phi_s) = 0 . \quad (43)$$

For small amplitude oscillations, with small phase deviations from the synchronous particle, the term  $(\sin \phi - \sin \phi_s)$  can be written as

$$\sin \phi - \sin \phi_s = \sin(\phi_s + \Delta\phi) - \sin \phi_s \cong \cos \phi_s \Delta\phi , \quad (44)$$

and Eq. (43) can be linearized to an equation of an undamped resonator with resonant frequency  $\Omega_s$ ,

$$\frac{d^2\Delta\phi}{dt^2} + \left[ \frac{\eta f_{\text{RF}} \cos \phi_s}{R p_s} qV \right] \Delta\phi = \frac{d^2\Delta\phi}{dt^2} + \Omega_s^2 \Delta\phi = 0, \quad (45)$$

$$\Omega_s = \sqrt{\frac{\eta f_{\text{RF}} \cos \phi_s}{R p_s} qV}. \quad (46)$$

This resonant frequency,  $\Omega_s$ , is called the *synchrotron frequency*. The periodic motion is stable if the expression under the square root of the definition of the synchrotron frequency is larger than zero. This is the case if  $\eta \cdot \cos \phi_s > 0$ . The necessary conditions for the synchronous phase follow from this requirement. Below transition, the condition for the stable synchronous phase is

$$\gamma \leq \gamma_{\text{tr}} \Rightarrow \eta \geq 0 \Rightarrow \cos \phi_s \geq 0 \Rightarrow \phi_s \in [0, \pi/2]. \quad (47)$$

Above transition, the condition for the stable synchronous phase becomes

$$\gamma \geq \gamma_{\text{tr}} \Rightarrow \eta \leq 0 \Rightarrow \cos \phi_s \leq 0 \Rightarrow \phi_s \in [\pi/2, \pi]. \quad (48)$$

### 2.2.4 Dispersion

From this discussion on the behaviour of the particles in the longitudinal plane, it is now clear that not all particles have the same momentum. In fact, a bunch contains a distribution of  $\Delta p/p$ . The typical momentum spread is of the order of  $d p/p \approx 10^{-3}$ . This has been neglected so far in the discussion of transverse motion. Including this fact turns the homogeneous equations of motion in Eq. (16) into the inhomogeneous equation.

$$x'' + x \left( \frac{1}{\rho^2} - k \right) = \frac{\Delta p}{p} \frac{1}{\rho}. \quad (49)$$

The general solution to this equation is the sum of solution to the homogeneous equation  $x_h(s)$  and a solution that fulfils the inhomogeneous equation  $x_i(s)$  with  $x(s) = x_h(s) + x_i(s)$ . The *dispersion* is then defined as

$$D(s) = \frac{x_i(s)}{\Delta p/p}. \quad (50)$$

The *dispersion* is the trajectory an ideal particle would have with  $\Delta p/p = 1$ . The trajectory of any particle is the sum of  $x_\beta(s)$  plus *dispersion*  $\times$  momentum offset.  $D(s)$  is just another trajectory and will therefore be subject to the focusing properties of the lattice. For a particle with momentum offset, the equation for calculating the coordinates  $x, x'$  at any location of the ring becomes

$$\begin{pmatrix} x \\ x' \end{pmatrix}_{s_1} = M \begin{pmatrix} x \\ x' \end{pmatrix}_{s_0} + \frac{\Delta p}{p} \begin{pmatrix} D \\ D' \end{pmatrix}_{s_1}. \quad (51)$$

*Dispersion* also has an effect on the size of the beam. At a given place in the ring, the beam size depends on  $\beta(s)$  and  $D(s)$ , together with the momentum spread of the beam,  $\Delta p/p$ :

$$\sigma = \sqrt{\beta \varepsilon + D^2 \left( \frac{\Delta p}{p} \right)^2}. \quad (52)$$

## 3 Beam loss mechanisms

The tune, the number of betatron oscillations per turn, was introduced in Section 2.1.3. The choice of tune, and hence the focusing properties of the lattice, have important implications for the stability of motion in the presence of linear magnetic field errors and non-linear fields.

### 3.1 Dipole field errors

If the magnetic centre of a quadrupole magnet is not perfectly aligned transversely with the design orbit, or if dipole field errors are present, orbit perturbations around the ring are generated. The orbit at a location  $s$  is the average trajectory over many turns at that location. With a field error  $\Delta x'$  at location  $s_0$  the orbit at location  $s$  changes according to

$$x(s) = \frac{\Delta x'}{2} \cdot \sqrt{\beta(s_0)\beta(s)} \frac{\cos(\pi Q - \psi_{s_0 \rightarrow s})}{\sin(\pi Q)}. \quad (53)$$

The effect of the error,  $\Delta x'$ , will be large at locations with large  $\beta(s)$  and depends on the phase advance between the error location and the orbit location of interest. Also, the larger  $\beta(s_0)$  is at the error location, the larger will be the effect of the error around the ring. The other important lesson to be drawn from Eq. (53) is the dependence on the tune. With the term  $\sin(\pi Q)$  in the denominator, the orbit response around the ring for any dipole error diverges for  $Q = N$ , where  $N$  is an integer.

### 3.2 Gradient errors

Gradient errors will lead to a change in the tune and the beta functions around the ring. The tune change can be calculated by evaluating the distorted one-turn matrix with a small field error  $\Delta k$  over a distance  $l$ , which might be the length of a magnet. The one-turn matrix is the transport matrix from location  $s \rightarrow s + L = s$ , where  $L$  is the circumference of the ring:

$$\begin{aligned} M_{\text{turn}_{\text{dist}}} &= \begin{pmatrix} \cos 2\pi Q + \alpha \sin 2\pi Q & \beta \sin 2\pi Q \\ -\gamma \sin 2\pi Q & \cos 2\pi Q - \alpha \sin 2\pi Q \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ -\Delta k l & 1 \end{pmatrix} \cdot \begin{pmatrix} \cos 2\pi Q_0 + \alpha \sin 2\pi Q_0 & \beta \sin 2\pi Q_0 \\ -\gamma \sin 2\pi Q_0 & \cos 2\pi Q_0 - \alpha \sin 2\pi Q_0 \end{pmatrix}, \end{aligned} \quad (54)$$

where  $Q = Q_0 + \Delta Q$ . With  $\text{Tr}(M_{\text{turn}_{\text{dist}}}) = \text{Tr}(M_{\text{error}} \cdot M_{\text{turn}})$ , the tune change evaluates to

$$\Delta Q = \frac{1}{4\pi} \beta \Delta k \cdot l. \quad (55)$$

The larger the beta function at the location of the gradient error, the larger the tune change. The relative beta function change, the so-called *beta-beat*, due to the gradient error  $\Delta k$  is

$$\frac{\Delta\beta(s)}{\beta(s)} = -\frac{1}{2 \sin(2\pi Q)} \beta(s_0) \cos[2(\psi(s_0) - \psi(s)) - 2\pi Q] \cdot \Delta k \cdot l. \quad (56)$$

As discussed earlier, the beta function is related to the size of the beam with  $\sigma = \sqrt{\beta\epsilon}$ . The beta functions, and hence the beam sizes, diverge with any gradient error if the tune  $Q = N, N/2$ , where  $N$  is an integer, owing to the term  $\sin 2\pi Q$  in the denominator.

### 3.3 Non-linear imperfections

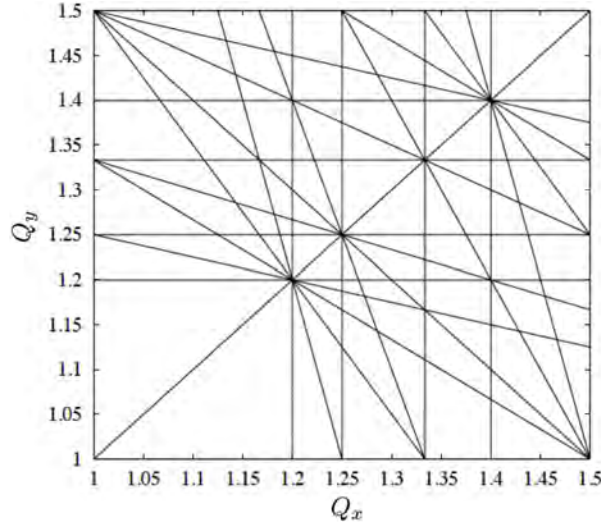
In the Taylor series expansion of the magnetic field for the derivation of the equation motion in the transverse plane, the higher-order components in  $x$  or  $y$  were neglected. Higher-order fields might, however, be present, owing to non-perfect dipole and quadrupole magnets, or they might be introduced on purpose, to stabilize the beam. The magnetic field of multiple of order  $n$  is

$$B_y(x, y) + i \cdot B_x(x, y) = (B_n(s) + iA_n(s)) \cdot (x + iy)^n, \quad (57)$$

where  $B_n(s)$  are the normal coefficients and  $A_n(s)$  are the skew coefficients from the Taylor series expansion,

$$\begin{aligned} B_n(s) &= \frac{1}{(n)!} \frac{\partial^n B_y}{\partial x^n}, \\ A_n(s) &= \frac{1}{(n)!} \frac{\partial^n B_x}{\partial x^n}. \end{aligned} \quad (58)$$





**Fig. 10:** The tune diagram indicates forbidden resonance lines. The chosen machine tunes in the horizontal and vertical plane have to be far from these lines.

For example,  $\partial^2 B_y / \partial x^2$  and  $\partial^3 B_y / \partial x^3$  are the sextupole and octupole component, respectively. In the presence of non-linear fields, the equation of motion becomes a non-linear differential equation. An example for the horizontal plane would be

$$\frac{d^2 x}{ds^2} + K(s) \cdot x = \frac{F_x}{v \cdot p}, \quad (59)$$

where  $F_x$  is the Lorentz force from the non-linear magnetic field.

It was mentioned earlier that the motion becomes unstable for  $Q = N$ ,  $N/2$  in the case of quadrupole field errors. It can be shown for sextupole perturbations that amplitudes increase for  $Q = N$ ,  $N/3$ ; and for octupole perturbations that amplitudes increase for  $Q = N$ ,  $N/2$ ,  $N/4$ . In general, the machine tune has to be chosen such that it does not fulfil the condition

$$nQ_x + mQ_y = N, \quad (60)$$

where  $n$ ,  $m$  and  $N$  are small integers. The forbidden tunes are often summarized in the so-called tune diagram as *resonance lines*. The resonance lines with the lowest order are the most dangerous ones. An example of a tune diagram with low-order resonance lines is shown in Fig. 10.

### 3.3.1 Sextupole fields

The example of the sextupole fields and their effect on particle motion will be briefly discussed to further explain the effect of non-linear fields. The resulting fields of a sextupole coil configuration are

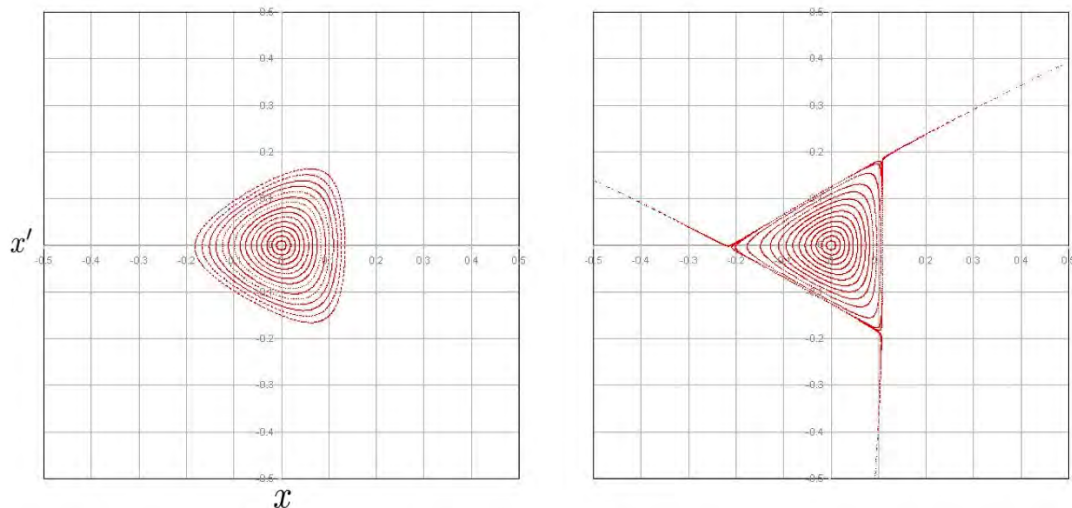
$$\begin{aligned} B_x &= \tilde{g}xy, \\ B_y &= \frac{1}{2}\tilde{g}(x^2 - y^2). \end{aligned}$$

The sextupole fields generate a gradient in both planes, rising linearly with the offset in  $x$  according to

$$\frac{\partial B_x}{\partial y} = \frac{\partial B_y}{\partial x} = \tilde{g}x.$$

The equations of motion in the presence of the sextupole field become

$$x'' + K_x(s) = -\frac{1}{2}m_{\text{sext}}(s)(x^2 - y^2), \quad (61)$$



**Fig. 11:** The trajectories in phase space  $x, x'$  with sextuple fields in the ring. The size of the stable area within the triangle is a function of the distance of the tune from the third-order resonance.

$$y'' + K_y(s) = m_{\text{sext}}(s)xy, \quad (62)$$

where  $m_{\text{sext}}$  is the normalized sextuple strength with  $m_{\text{sext}} = \tilde{g}/(p/e)$ . The effect of a single sextuple around the ring on the particle trajectories can be simulated by adding a sextuple *kick* at the location of the sextuple,

$$\begin{aligned} \Delta x' &= -\frac{1}{2}m_{\text{sext}}l(x^2 - y^2), \\ \Delta y' &= m_{\text{sext}}lxy, \end{aligned}$$

where  $l$  is the length of the sextuple. As a result of these kicks, the phase-space trajectories become more and more distorted for larger amplitudes and are no longer elliptical, see Fig. 11. The motion becomes unstable, meaning that the amplitudes of the particles become larger and larger after each turn. The size of the stable area within the triangle in phase space is proportional to  $(Q - \frac{p}{3})/(m_{\text{sext}})$ , where  $p$  is an integer.

### 3.3.2 Chromaticity

The normalized quadrupole gradient was defined as  $k = g/(p/e)$ . The different particles in a beam have a distribution of momenta around the ideal momentum  $p_0$ . For a given particle,  $p = p_0 + \Delta p$ , and the normalized gradient for this particle is

$$k = \frac{eg}{p_0 + \Delta p} \approx \frac{e}{p_0} \left(1 - \frac{\Delta p}{p_0}\right) g = k_0 + \Delta k,$$

where the gradient error is

$$\Delta k = -\frac{\Delta p}{p_0} k_0.$$

As discussed already, gradient errors result in tune changes. Thus, particles with a different momentum  $p$  distributed around  $p_0$  will all have different tunes. Using Eq. (55), the tune change for a particle of a given momentum offset  $\Delta p/p_0$  is

$$\Delta Q = \frac{1}{4\pi} \beta \Delta k \cdot l = -\frac{1}{4\pi} \frac{\Delta p}{p_0} k_0 \beta l. \quad (63)$$

The parameter *chromaticity* is defined as the ratio of the tune change for a given relative momentum change:

$$\Delta Q = Q' \frac{\Delta p}{p}, \quad (64)$$

and, from Eq. (63), the *chromaticity* of a synchrotron equates to

$$Q' = -\frac{1}{4\pi} \oint k(s)\beta(s)ds. \quad (65)$$

*Chromaticity* is created in the vertical and horizontal planes by the quadrupole fields. Together with the beam momentum spread, it indicates the size of the tune spot in the tune diagram, which will no longer be a single point. The natural chromaticity in the LHC, for example, is 250 units. With a typical momentum spread of  $\Delta p/p = 0.2 \times 10^{-3}$  and fractional injection tune  $Q_x = 0.28$ , the particles would have tunes between  $Q_x = 0.26$  and  $Q_x = 0.33$ . The particle distribution would cross several dangerous resonance lines, leading to beam loss. Chromaticity, therefore, has to be corrected. The sorting of the particle amplitudes in the horizontal plane due to dispersion  $x_D(s) = D(s)\frac{\Delta p}{p}$  is used for this purpose. Sextupole magnets are placed at locations with large dispersion  $D_x$ . The resulting gradient at the sextupole location depends on the particle amplitude in the horizontal plane, and the sextupole strength is chosen such that the chromaticity in both planes is adjusted to a suitable value.  $Q' = 0$  is, however, not necessarily desirable, owing to so-called *collective effects*.

### 3.4 Collective effects

Collective effects can cause beam instabilities, emittance blow-up and beam loss. A typical example of the turn-by-turn trajectory during a beam instability is shown in Fig. 12. There are three main categories of *collective effect*:

**Beam–self:** The beam interacts with itself through *space-charge*, causing a tune spread proportional to  $1/(\beta^2\gamma^3)$ . This effect is one of the main brightness limitations in low-energy machines.

**Beam–beam:** Beams interact with each other at or close to the collision point in colliders or they interact with ambient electron clouds. Colliding beams produce large tune spreads and tune shifts, owing to head-on and long-range collisions.

**Beam–environment:** These are impedance-related instabilities. The beam induces electromagnetic fields in the accelerator environment, such as in the vacuum chambers. These so-called *wake fields* can act back on the trailing beam. The Fourier transform of the *wake field* is called the impedance. Energy is lost, owing to the *wake field*. If the energy remains trapped, it can lead to component heating. If the energy is transferred to the trailing beam, it can cause beam instabilities.

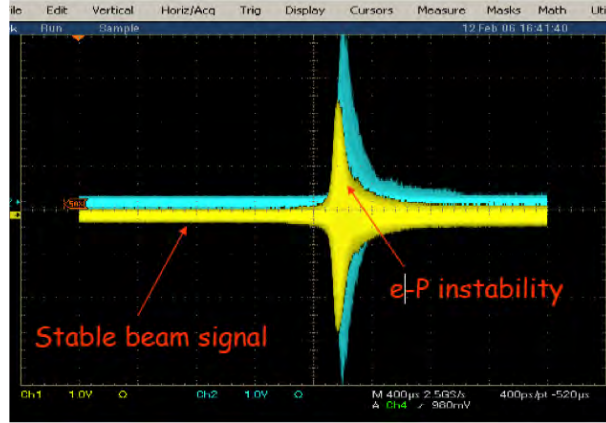
It is possible to stabilize the beam under the influence of collective effects, using such mitigative means as transverse feedback with sufficient bandwidth or the introduction of tune spread through non-zero chromaticity, octupole fields, or the head-on beam–beam effect in colliders for *Landau damping*. A coherent oscillation at a frequency within the beam frequency spread is generally damped.

#### 3.4.1 Space-charge effect

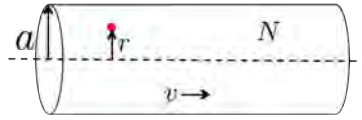
The space-charge effect is the simplest and most fundamental of all collective effects. It will be treated in a simple approximation as an example for collective effects associated with *direct space-charge*.

Let us assume that the beam is a long uniformly charged cylinder of current  $I$ , as indicated in Fig. 13. The force exerted on a particle by the surrounding beam at a distance  $r$  from the beam centre is:

$$F_r = F_E + F_B = \frac{eI}{2\pi c\beta\epsilon_0\gamma^2 a^2} r. \quad (66)$$



**Fig. 12:** Beam position measurement at a beam position monitor turn-by-turn. During the instability, the trajectory amplitude grows exponentially until the beam loses intensity and stabilizes itself.



**Fig. 13:** For a cylindrical beam, the space-charge force on a particle at distance  $r$  from the beam centre can easily be calculated.

For simplicity, it can be assumed that the particle has only a horizontal offset from the beam centre; in the Frenet–Serret coordinate system, Eq. (66) becomes

$$F_x = \frac{eI}{2\pi c\beta\epsilon_0\gamma^2 a^2} x. \quad (67)$$

The influence of the space-charge force on transverse motion is derived by treating it as perturbation of the linear equation of motion, as discussed with the non-linear fields:

$$x''(s) + K(s)x = \frac{F_{SC}}{m\gamma\beta^2 c^2},$$

$$x''(s) + \left( K(s) - \frac{2r_0 I}{ea^2\beta^3\gamma^3 c} \right) x = 0,$$

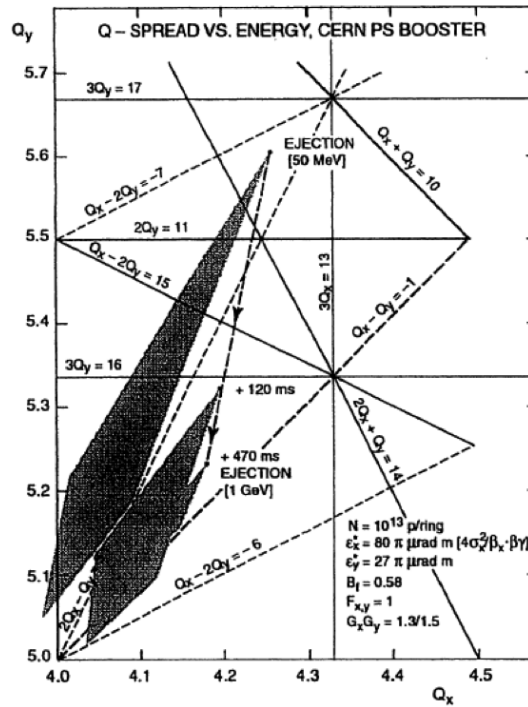
where  $r_0 = e^2/(4\pi\epsilon_0 m_0 c^2)$  is the classical particle radius. As the space-charge force is linear in  $x$ , it introduces a gradient error, and gradient errors lead to tune shift. The tune shift from the space-charge around the whole ring with radius  $R$  is

$$\Delta Q_x = \frac{1}{4\pi} \int_0^{2\pi R} \beta_x(s) \Delta K_{SC}(s) ds = -\frac{r_0 R I}{e\beta^3\gamma^3 c} \left\langle \frac{\beta_x(s)}{a^2(s)} \right\rangle. \quad (68)$$

As  $a$  is related to the transverse size of the cylindrical beam,  $a^2/\beta_x$  is related to the invariant of motion,  $\epsilon_x = a^2/\beta_x$ . The space-charge tune shift can therefore be written as

$$\Delta Q_x = -\frac{nr_0}{2\pi\epsilon_x\beta^2\gamma^3}, \quad (69)$$

with  $I = (ne\beta c)/(2\pi R)$ . The tune shift is larger if the beam has higher brightness  $n/(\epsilon_{x,y})$  and lower energy.



**Fig. 14:** Tune spread due to the space-charge effect for high-intensity beams in the CERN PS Booster. Injection energy, 50 MeV; maximum extraction energy towards the CERN PS, 1.4 GeV.

In realistic beams, the particle density will not be uniform and the different particles will see different space-charge forces, depending on where they are in the beam. This will introduce a tune spread instead of a tune shift. An example of the tune spread due to space-charge in the CERN PS Booster is shown in Fig. 14.

#### 4 Accidental beam loss

Unforeseen aperture limitations can cause beam losses. The aperture limitation may be the result of misalignment of equipment or movable equipment, such as screens, girders or collimators, that is only partially retracted. Aperture limitations may also be introduced by orbit changes from quadrupole misalignment or orbit bumps. Orbit bumps might be needed for extraction systems or to introduce crossing angles in colliders. Orbit bumps might also result from corrections made as a result of false beam position monitor readings.

Fast transverse kickers, such as aperture kicker magnets, tune kicker magnets, crab cavities or transverse feedback systems, can excite significant fractions of the beam to high amplitudes within a single or a few turns. Power supply limitations or special run configurations are put in place to avoid large deflections from the aperture kicker at high intensity in the LHC. Injection kickers or extraction kickers firing asynchronously can be very dangerous for high-intensity machines. This topic is discussed further in Ref. [3].

Beam losses can also originate from uncaptured beams. Uncaptured beam is generated at injection if part of the bunch is injected outside the RF acceptance. Many mechanisms can drive the beam out of the bucket after injection, for example, intra-beam scattering, beam-beam interactions or malfunctioning equipment. Malfunctioning equipment comprises noise in the phase loop, badly adjusted longitudinal blow-up or the switch-off of an RF cavity, resulting in a reduction in the total available voltage and

hence reduction of the RF acceptance. Once uncaptured, the time taken for the particles to get lost on the vacuum chamber depends on the energy loss per turn due to synchrotron radiation, electron cloud or impedance and the so-called *momentum aperture* in  $\Delta p/p$  for the given mechanical aperture, dispersion and beta function.

#### 4.1 Powering failures

The power supplies of the circuits powering the machine elements can fail and the magnetic field seen by the beam of the concerned elements will then decay. Failing dipole magnets will generate an orbit distortion that changes with time. If the failure is sufficiently slow (e.g. in the LHC, over more than 10 turns), the closed orbit formula in Eq. (53) for a time-varying field error can be used to calculate what happens to the beam,

$$\Delta x_{\text{CO}}(s) = \frac{\sqrt{\beta(s)}}{2 \sin(\pi Q)} \left( \frac{I(t)}{I_0} - 1 \right) \sum_i^N \theta_i \sqrt{\beta(s_i)} \cos(\psi(s) - \psi(s_i) + \pi Q), \quad (70)$$

where  $N$  is the number of dipole magnets connected in series to the failing power supply. A power supply has many possible failure cases. The function  $I(t)$  needs to be established. In most powering failure cases, the voltage is set to zero and the current decays as

$$I(t) = I_0 e^{-\frac{t}{\tau}}, \quad (71)$$

where the time constant  $\tau$  is defined by the resistance and inductance of the circuit as  $\tau = L/R$ . In the case of a main dipole magnet quench in the LHC, the current in the quenching magnet would decay as

$$I(t) = I_0 e^{-\frac{t^2}{2\sigma^2}}, \quad (72)$$

where  $\sigma$  has been found to be  $\sigma = 200$  ms. A failing quadrupole circuit will move the tune in the tune diagram with the risk of crossing resonance lines. The beam size will also change as a result of the change of the beta functions, as discussed for gradient errors. If the orbit is not zero in the quadrupole, the failing quadrupole will also change the orbit. This discussion can be extended to cover higher-order circuits. The higher the order, the less critical the circuits tend to be.

#### References

- [1] CERN Accelerator School, Intermediate Level,  
<http://cas.web.cern.ch/cas/Norway-2013/Lectures/Norwaylectures.html>.
- [2] S.Y. Lee, *Accelerator Physics*, 3rd ed. (World Scientific, River Edge, 2011),  
<http://dx.doi.org/10.1142/8335>
- [3] V. Kain, Beam Transfer and Machine Protection, these proceedings.

## Beam Loss in Linacs

*M.A. Plum*

Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

### Abstract

Beam loss is a critical issue in high-intensity accelerators, and much effort is expended during both the design and operation phases to minimize the loss and to keep it to manageable levels. As new accelerators become ever more powerful, beam loss becomes even more critical. Linacs for  $H^-$  ion beams, such as the one at the Oak Ridge Spallation Neutron Source, have many more loss mechanisms compared to  $H^+$  (proton) linacs, such as the one being designed for the European Spallation Neutron Source. Interesting  $H^-$  beam loss mechanisms include residual gas stripping,  $H^+$  capture and acceleration, field stripping, black-body radiation and the recently discovered intra-beam stripping mechanism. Beam halo formation, and ion source or RF turn on/off transients, are examples of beam loss mechanisms that are common for both  $H^+$  and  $H^-$  accelerators. Machine protection systems play an important role in limiting the beam loss.

### Keywords

Linac; beam loss; H-minus proton.

## 1 Introduction

Beam loss is a critical issue in high-intensity linacs, and much work is done during both the design and operation phases to keep the loss down to manageable levels. A generally accepted rule of thumb is to keep the loss to less than approximately 1 W/m to allow for hands-on maintenance. For example, the linac output beam power of the Oak Ridge Spallation Neutron Source (SNS) linac is about 1 MW today, and we plan to increase the power to its design value of 1.4 MW over the next few years and then later to about 3 MW. The fractional loss per metre should then be less than about  $3 \times 10^{-7}$  in the high-energy portion of the linac. The allowable fraction of beam loss will be even lower for the next-generation accelerators, such as the European Spallation Neutron Source.

In general, beam loss in  $H^-$  linacs is more difficult to manage than in  $H^+$  linacs due to the greater number of loss mechanisms, including residual gas stripping,  $H^+$  capture and acceleration, field stripping, black-body radiation and the recently discovered intra-beam stripping (IBSt) mechanism [1]. Mechanisms such as beam halo formation, and ion source or RF turn on/off transients, can cause loss in both  $H^+$  and  $H^-$  linacs.

In this paper we will review beam loss mechanisms in  $H^+$  and  $H^-$  accelerators, drawing mainly from recent work at the SNS, but also including examples from other high-intensity accelerator facilities including the Los Alamos Neutron Scattering Center (LANSCE) facility at Los Alamos National Laboratory, the ISIS facility at Rutherford Appleton Laboratory, the High Intensity Proton Accelerator (HIPA) facility at the Paul Scherrer Institute and the Japanese Proton Accelerator Research Complex (J-PARC).

## 2 Why accelerate $H^-$ beams?

If  $H^-$  beams have so many more loss mechanisms, why should we bother with them? The reason is low-loss injection into storage rings and synchrotrons. Certain applications require multiple beam pulses to be injected into the same RF bucket over multiple turns of injection to obtain a large beam charge per pulse. Example applications include spallation neutron sources and neutrino production facilities. The only way to inject multiple beams with low beam loss is to use charge exchange injection, where

**Table 1:** Beam loss mechanisms for  $H^-$  and  $H^+$  beams

| Beam loss mechanism   | Possible for $H^-$ beam | Possible for $H^+$ beam |
|---|-------------------------|-------------------------|
| Residual gas stripping  | Yes                     | No                      |
| $H^+$ capture and acceleration                                  | Yes                     | No                      |
| Intra-beam stripping  | Yes                     | No                      |
| Field stripping   | Yes                     | No                      |
| Black-body radiation  | Yes                     | No                      |
| Beam halo/tails (resonances, collective effects, mismatch, etc) | Yes                     | Yes                     |
| RF and/or ion source turn on/off transients                     | Yes                     | Yes                     |
| Dark current from ion source                                    | Yes                     | Yes                     |

electrons are stripped off the incoming  $H^-$  ion beam to merge it with the previously injected beam. Charge exchange injection is also required if you want the output beam emittance to be less than the sum of the input emittances (Liouville theorem).

The typical beam loss without charge exchange injection is several percent. This can be acceptable at low-power accelerators, but, for example, at SNS, where the design linac beam power is 1.4 MW, if we lose 2% that would correspond to 28 kW, which is clearly unacceptable. With charge exchange injection the fractional loss is just  $(1-2) \times 10^{-4}$ , so the beam power lost is only 140–240 W.

### 3 Beam loss mechanisms

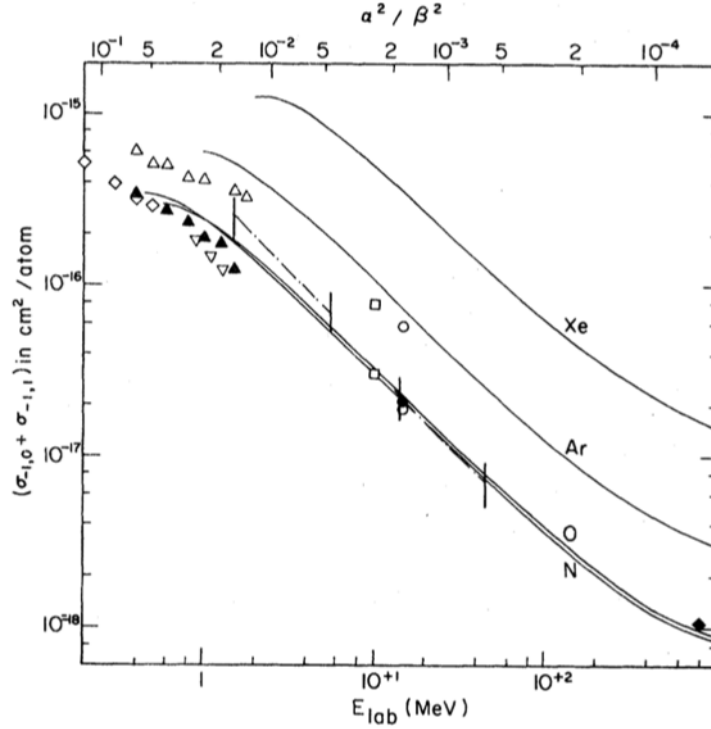
The beam loss mechanisms that we will be discussing are listed in Table 1.

#### 3.1 Residual gas stripping

Residual gas stripping is a significant beam loss mechanism for high-power  $H^-$  ion accelerators but not for proton accelerators. In this mechanism electrons are stripped off the  $H^-$  particles by the residual gas in the accelerator, most likely leaving neutral  $H^0$  particles. The cross-sections are greatest at low beam energies and for gases with high atomic numbers, as shown in Figs. 1 and 2. The cross-section for double stripping ( $H^-$  to  $H^+$ ) is about 4% of the cross-section for single stripping ( $H^-$  to  $H^0$ ). The good news is that in a typical accelerator, the residual gases will be mainly  $H_2$  and  $H_2O$ , and then CO and  $CO_2$ . These molecules have relatively low atomic numbers.

This phenomenon is well known and understood, yet it still sometimes requires installation of additional vacuum pumps beyond those specified in the design phase. The vacuum levels can easily be worse than anticipated due to small vacuum leaks, higher-than-expected outgassing rates, etc. In the SNS linac we have measured beam loss due to residual gas stripping in the 87 to 186 MeV coupled cavity linac (CCL) by purposely turning off the vacuum pumps and allowing the gas pressure to rise. As shown in Fig. 3, gas stripping is present to a small degree during normal operations, and it can become significant if there are vacuum problems. There is also unexpected beam loss due to gas stripping in a section of the high-energy beam transport (HEBT) line between the linac and the ring. Additional ion pumps were installed to mitigate this loss. In the J-PARC linac gas stripping was found to cause significant beam loss during the commissioning phase [2]. It was subsequently reduced to acceptable levels by installing additional vacuum pumps in the Separated Drift Tube Linac (S-DTL) and the upstream portion of the linac reserved for future expansion. Also, in the LANSCE linac, residual gas stripping has been estimated [3] to cause about 25% of the  $H^-$  beam loss along the linac. In the ISIS linac, gas stripping is present under nominal conditions, but not at a significant level [4]. However, if the gas pressure increases due to vacuum issues, the ISIS loss can become significant.





**Fig. 1:** Gas stripping cross-section as a function of  $H^-$  beam energy, for various residual gases. Figure reproduced from Ref. [6].

### 3.1.1 Gas stripping example

As an example of how to estimate the amount of beam loss given the reaction cross-section, consider the example of a 100 MeV, 1 mA  $H^-$  beam, in a beam pipe with  $10^{-7}$  Torr of nitrogen gas at 303 K. The following equations are based on Ref. [7].

For our beam energy range of interest, the gas stripping cross-section for nitrogen and oxygen, as shown in Fig. 1, may be written

$$\sigma = \frac{7 \times 10^{-19} \text{ cm}^2}{\beta^2 \text{ atom}}. \quad (1)$$

For later reference, the cross-section for hydrogen for our beam energy range of interest may be written

$$\sigma = \frac{1 \times 10^{-19} \text{ cm}^2}{\beta^2 \text{ atom}}. \quad (2)$$

From the ideal gas law, the gas density is

$$\rho = \left( 2N_A \frac{p}{22410 \times 760} \frac{273}{T} \right) \frac{\text{atoms}}{\text{cm}^3}, \quad (3)$$

where we note that nitrogen is a diatomic gas,  $T$  is the gas temperature in K,  $N_A$  is Avogadro's constant and the pressure  $p$  is in Torr. The beam power lost in a given length of beam line  $l$  is

$$P = E_{\text{beam}} I_{\text{beam}} \frac{d\sigma}{d\Omega} \rho l, \quad (4)$$

where  $E_{\text{beam}}$  is the beam energy and  $I_{\text{beam}}$  is the beam current. Inserting the numbers, we have

$$P = (10^8 \text{ V})(0.001 \text{ A}) \left( \frac{7 \times 10^{-19}}{0.428^2} \text{ cm}^2 \right) \left( 2 \times 6.022 \times 10^{23} \frac{10^{-7} \text{ Torr}}{22410 \times 760 \text{ Torr}} \frac{\text{atoms}}{\text{cm}^3} \right) \left( \frac{273 \text{ K}}{303 \text{ K}} \right) (1 \text{ m}) \quad (5)$$

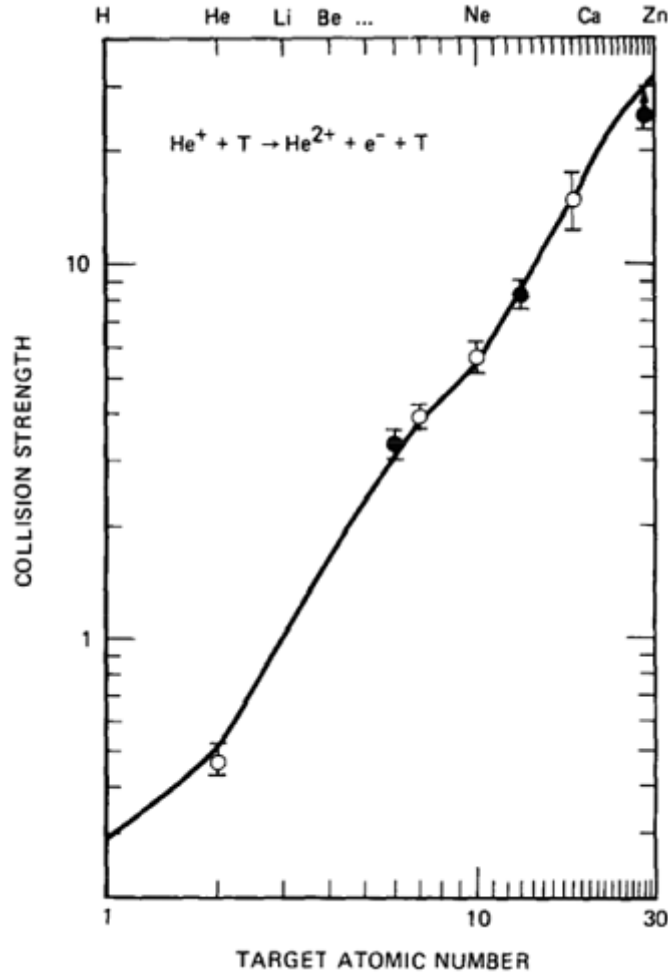
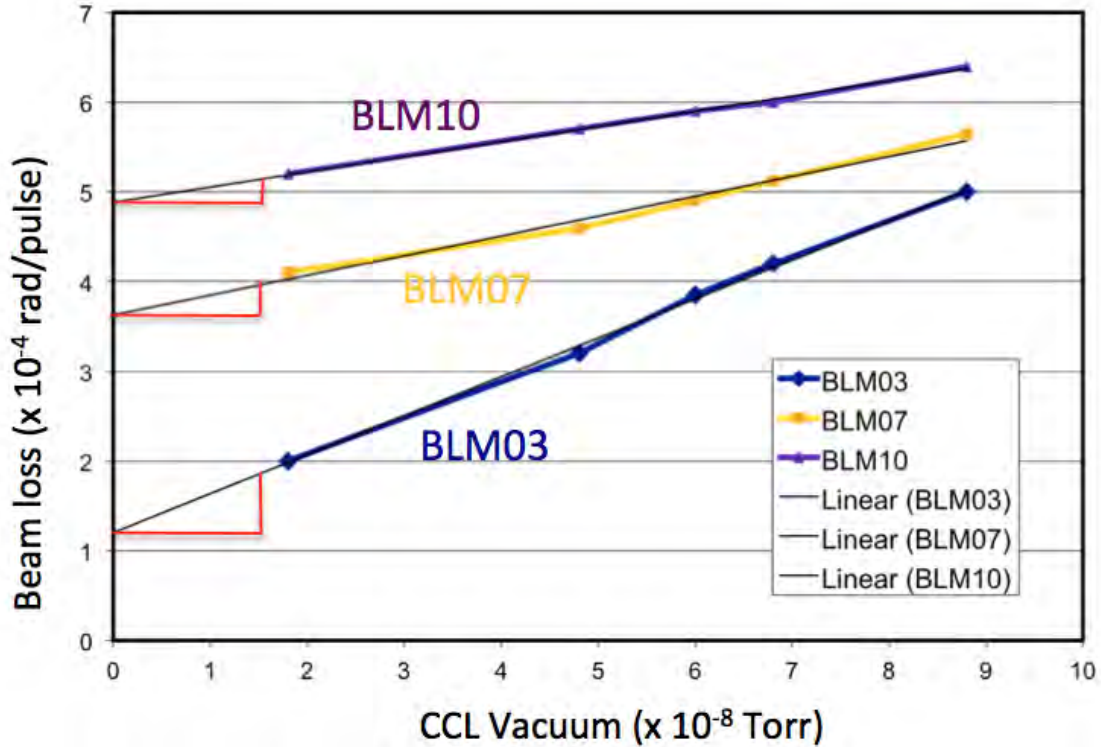


Fig. 2: Gas stripping cross-section as a function of atomic number. Figure reproduced from Ref. [8]

i.e.

$$P = 0.243 \text{ W in 1 m or } 0.243 \text{ W/m.} \quad (6)$$

Figures 4 and 5 show the cross-sections and power lost versus the beam energy for our example case of 1 mA of a  $\text{H}^-$  beam and  $10^{-7}$  Torr of nitrogen gas at 303 K. Note that for a given gas pressure, residual gas stripping causes the beam power lost to increase as the beam energy increases. It is a trade-off between the cross-section decreasing with energy and the beam power increasing with energy. The beam power wins. The next question is 'is this beam loss enough to be concerned about?'. To answer that question, we need a relationship between beam loss and radioactivation. Figure 6 shows one example, in this case for beam loss on copper. Copper is a good example because non-superconducting RF cavities are usually made of copper (e.g. drift tube linacs (DTLs), CCLs, superconducting linacs (SCLs), etc). Given this relationship, we can now produce the plots shown in Figs. 7 and 8. A typical activation limit for hands-on maintenance is 1 mSv/h (100 mrem/h), from all sources. Let us assume that we want to limit the dose from residual gas stripping to 0.1 mSv/h (10 mrem/h). From Fig. 7, we see that we must limit the beam loss to 8.5 W/m at 20 MeV, 0.9 W/m at 100 MeV and 0.12 W/m at 1000 MeV, assuming the residual gas is nitrogen. Note that the allowable beam loss decreases with beam energy. We can also look at this problem from the point of view of maximum allowable gas pressure. From Fig. 8, we see that the maximum gas pressure is  $1.7 \times 10^{-7}$  Torr at 200 MeV or  $2.1 \times 10^{-8}$  Torr at 1000 MeV, assuming nitrogen gas, a copper beam line and a 1 mA beam current.

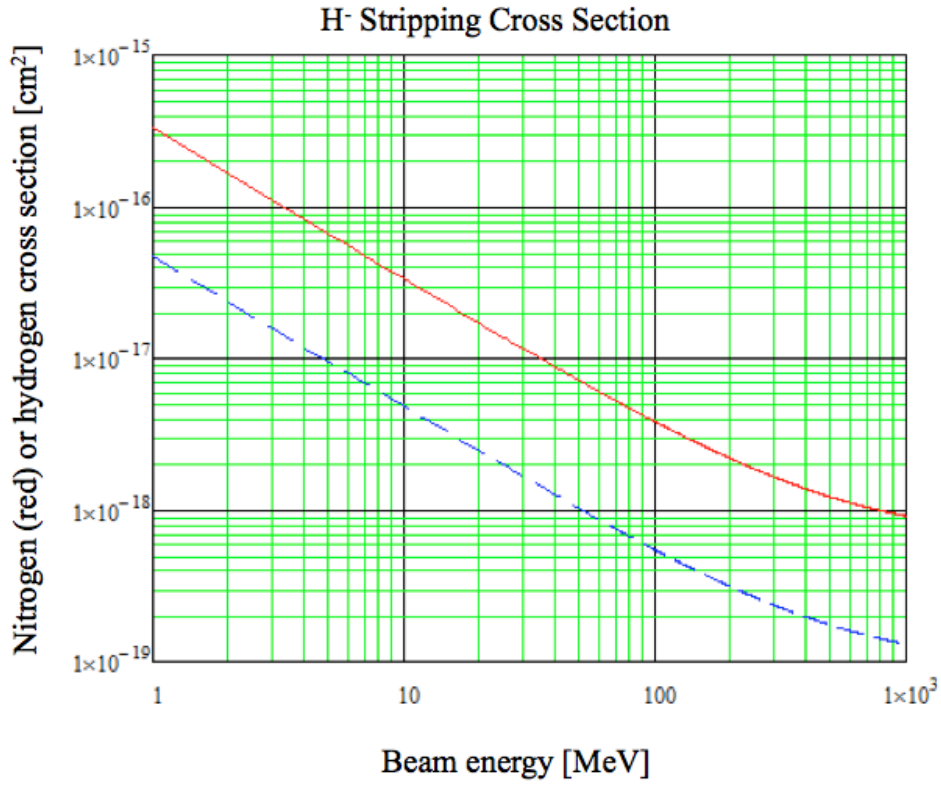


**Fig. 3:** Beam loss in the SNS SCL as a function of gas pressure in the last CCL section. The nominal gas pressure is about  $2 \times 10^{-8}$  Torr, and the red brackets indicate the beam loss due to residual gas stripping during nominal operations. As one moves along the linac toward higher BLM numbers, the slopes of the lines decrease indicating less sensitivity to the upstream phenomenon, due to the BLMs being further away from the location of the loss. The magnitudes of the BLM signals also increase due to the higher beam energies. Figure reproduced from Refs. [9,10].

### 3.2 $H^+$ capture and acceleration

Inadvertent proton acceleration is another interesting source of beam loss in  $H^-$  accelerators. As discussed in Section 3.1, double stripping, where both electrons on the  $H^-$  particle are stripped off, can occur due to residual gas interactions. If these newly created protons are captured into RF buckets, they are accelerated 180 degrees out of phase along with the  $H^-$  particles, and then eventually lost at high beam energies. The most likely place for  $H^+$  capture to occur is at low beam energies, where the double-stripping cross-sections are maximized. The cross-section for double stripping is about 4% of the single-stripping cross-section shown in Fig. 1. For example, recent measurements at LANSCE [11] showed that fully accelerated 800 MeV protons can be easily detected downstream of the linac while only the  $H^-$  ion source is in use. The protons are from double stripping of the  $H^-$  beam in both the 0.75 MeV low-energy beam transport (LEBT) and in the 100–800 MeV CCL. This beam loss mechanism is also observed at J-PARC, where unexpectedly high activation levels were discovered in the beam transport line from the linac to the rapid cycling synchrotron [12]. Adding a chicane bump in the 3 MeV medium-energy beam transport (MEBT) solved the problem by allowing the protons to be intercepted before they could be accelerated to higher beam energies.

One interesting aspect of  $H^+$  capture and acceleration is that the protons are unlikely to survive even RF frequency jumps in the linac, as illustrated in Fig. 9. For example, at the Oak Ridge SNS, where the RF frequency jumps from 402.5 MHz to 805 MHz in the transition from the DTL to the CCL, the protons are unlikely to survive because they are suddenly within decelerating RF buckets after the frequency jump. Alternatively, in the J-PARC linac, where the frequency jumps from 324 MHz to



**Fig. 4:** Gas stripping cross-sections for nitrogen or oxygen (solid red line) and hydrogen (blue dashed line) as a function of beam energy.

972 MHz in the transition from the S-DTL to the Annular Coupled Structure (ACS) linac at 191 MeV, the protons can be accelerated all the way to the end of the linac, only to be lost in the arc leading to the rapid cycling synchrotron.

### 3.3 Intra-beam stripping

In IBSt [1, 13], interactions between the  $H^-$  particles within the beam bunch cause loosely bound electrons to be stripped off, leaving neutral  $H^0$  particles, which are subsequently lost due to lack of focusing, steering and acceleration. The IBSt reaction rate can be written as

$$\frac{dN}{ds} = \frac{N^2 \sigma_{\max} \sqrt{\gamma^2 \theta_x^2 + \gamma^2 \theta_y^2 + \theta_s^2}}{8\pi^2 \sigma_x \sigma_y \sigma_s \gamma^2} F(\gamma \theta_x, \gamma \theta_y, \theta_s), \quad (7)$$

$$F(a, b, c) \approx 1 + \frac{2 - \sqrt{3}}{\sqrt{3}(\sqrt{3} - 1)} \left( \frac{a + b + c}{\sqrt{a^2 + b^2 + c^2}} - 1 \right), \quad (8)$$

where  $N$  is the number of particles in the bunch,  $\gamma$  is the relativistic factor,  $\sigma_{x,y} = \sqrt{\epsilon_{x,y} \beta_{x,y}}$  are the transverse rms bunch sizes,  $\theta_{x,y} = \sqrt{\epsilon_{x,y} / \beta_{x,y}}$  are the transverse local rms angular spreads,  $\epsilon_{x,y}$  and  $\beta_{x,y}$  are the transverse emittance and Twiss parameters, and  $\sigma_s$  and  $\theta_s$  are the rms bunch length and the relative rms momentum spread.  $F$  is weakly dependent on its variables and ranges between 1 and 1.15. This equation assumes a constant cross-section  $\sigma_{\max} \approx 4 \times 10^{-15} \text{ cm}^2$  independent of the relative particle velocity, and this is correct for relative velocities  $2 \times 10^{-4} < \beta_{\text{rel}} < 4 \times 10^{-3}$ . The cross-section drops off for  $\beta_{\text{rel}}$  outside this range.

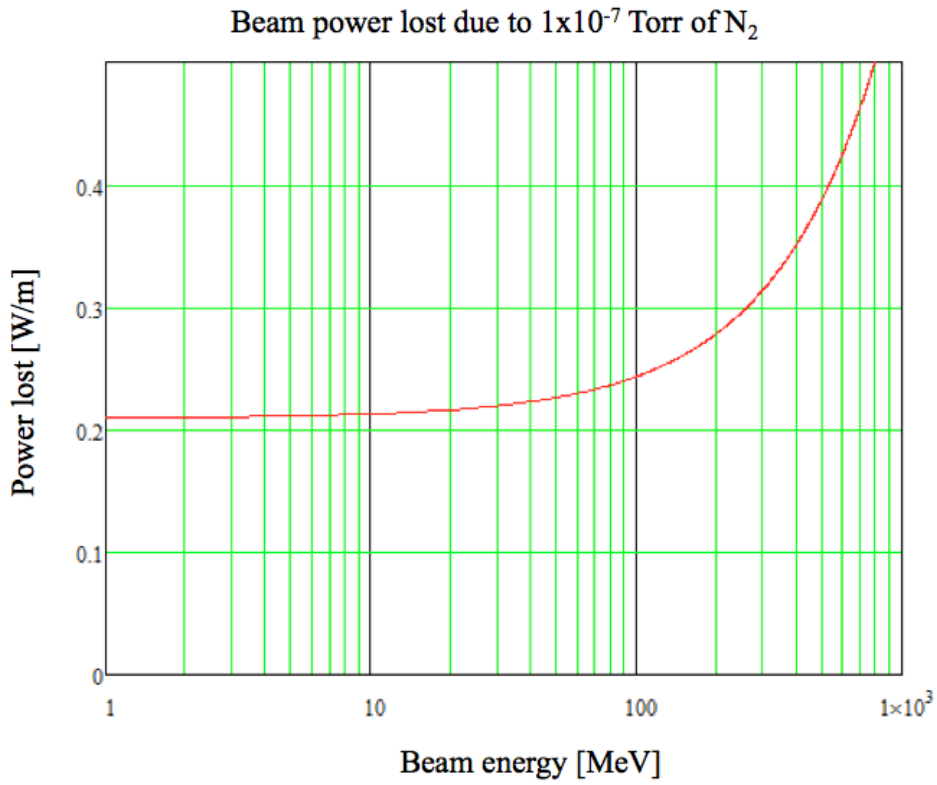


Fig. 5: Beam power lost versus beam energy for our example case

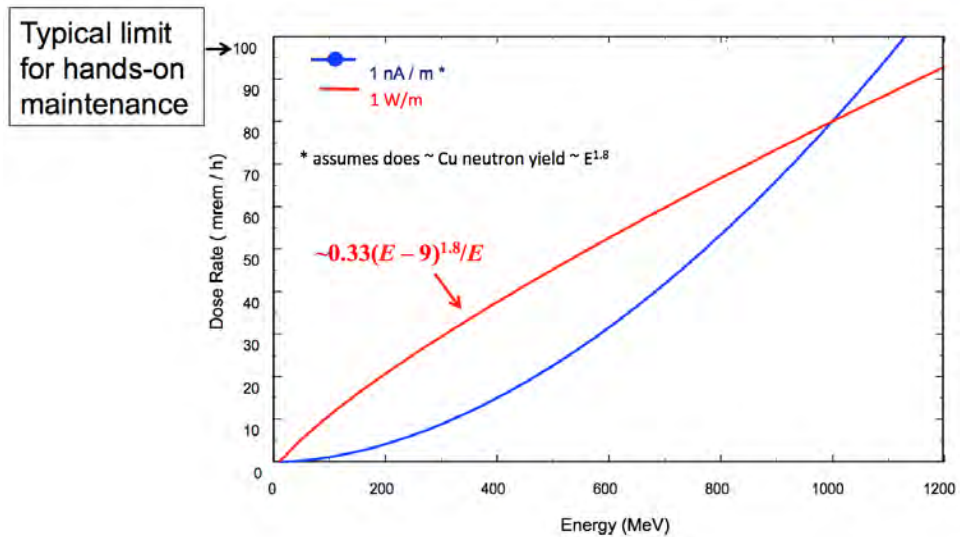
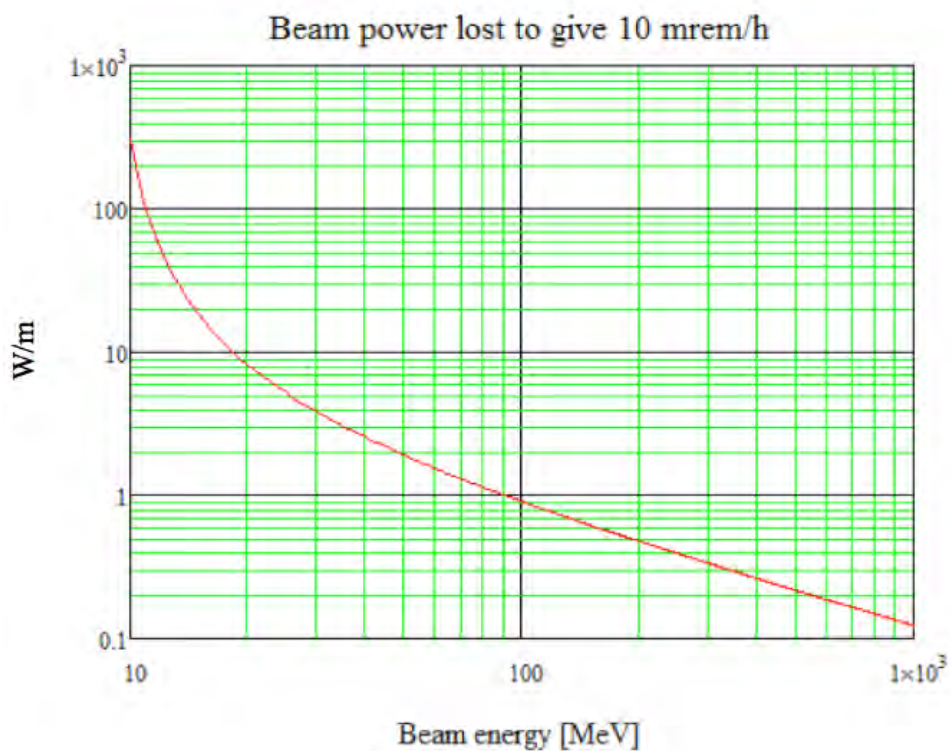
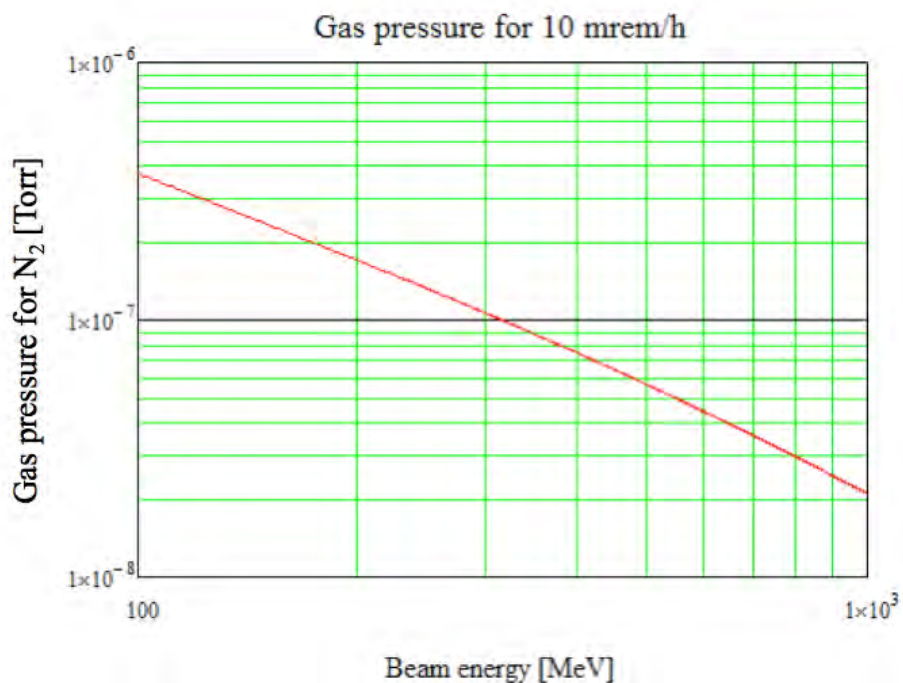


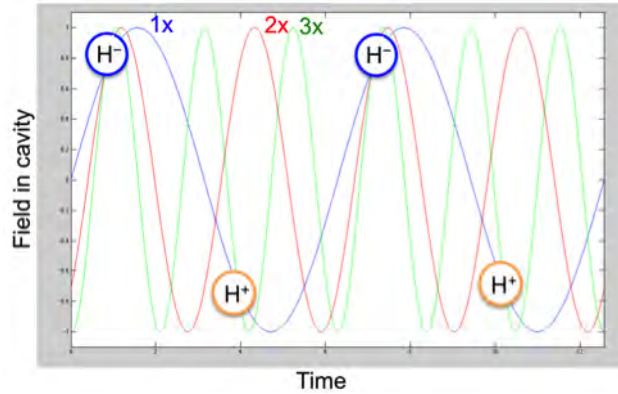
Fig. 6: Dose rate versus beam energy, for proton beam loss of 1 W/m or 1 nA/m, for the case of copper, at 30 cm after 4 h cool down. Figure reproduced from Ref. [14]. Note that  $100 \text{ mrem/h} = 1 \text{ mSv/h}$ .



**Fig. 7:** Allowable beam power loss versus beam energy to produce an activation of 0.1 mSv/h (10 mrem/h) at 30 cm for the case of copper, after 4 h cool down.



**Fig. 8:** Allowable nitrogen gas pressure versus beam energy to produce an activation of 0.1 mSv/h (10 mrem/h) at 30 cm for the case of copper and a 1 mA H<sup>-</sup> beam current, after 4 h cool down.



**Fig. 9:** Electric fields in a linac for three different frequencies.  $H^+$  beams will survive a  $3\times$  frequency jump since the  $3\times$  phase is the same as the  $1\times$  phase. This is not the case for a  $2\times$  frequency jump. Figure reproduced from Ref. [9].

During the design phase of the Oak Ridge SNS, which accelerates  $H^-$  particles to 1 GeV with a design beam power of 1.4 MW, it was believed that the beam loss in the SCL would be negligible, due to the large apertures and low residual gas pressure. Yet, as we discovered during the commissioning phase, the beam loss was much higher than expected, with a measured fractional loss per metre of about  $3 \times 10^{-7}$ . Although this level of beam loss is acceptable for hands-on maintenance, it was nevertheless a puzzle.

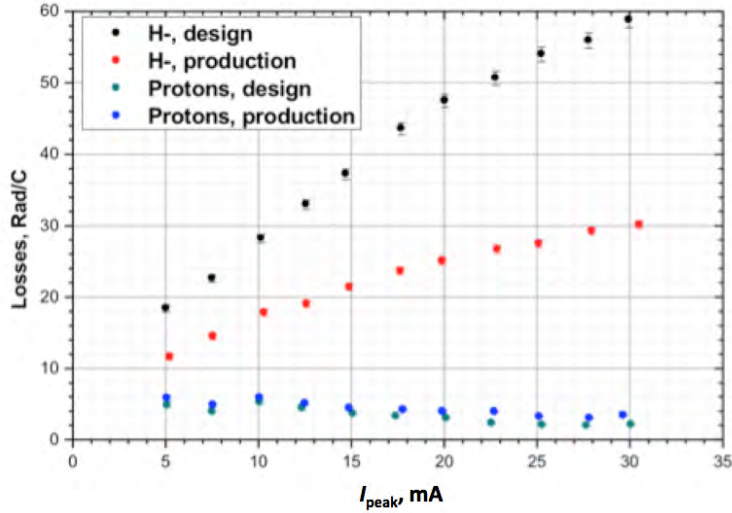
The IBSt reaction rate is proportional to the beam particle density squared, which explains why, before we understood the IBSt loss mechanism at SNS, we were able to empirically reduce the beam loss by lowering the SCL quadrupole focusing strengths by up to about 40%. This is illustrated in Fig. 10, which shows the normalized beam loss as a function of peak beam current, for two different cases: the SCL quadrupoles set for the design gradients and the SCL quadrupole gradients lowered to empirically minimize the beam loss. The lower focusing strengths increase the transverse beam size, which lowers the beam particle density, and in turn lowers the IBSt reaction rate. A demonstration that the IBSt reaction rate depends on the beam density squared (see Eq. (7)) can also be seen in Fig. 10, with the linear relationship between the normalized beam loss and the peak current (or quadratic relationship between the beam density and the un-normalized beam loss rate).

Experiments accelerating protons, rather than  $H^-$  particles, showed that IBSt is by far the dominant loss mechanism in the SNS SCL, as shown in Figs. 10 and 11. Similar measurements at LANSCCE [3] showed that IBSt accounts for about 75% of the  $H^-$  beam loss, with residual gas stripping accounting for the remaining 25%.

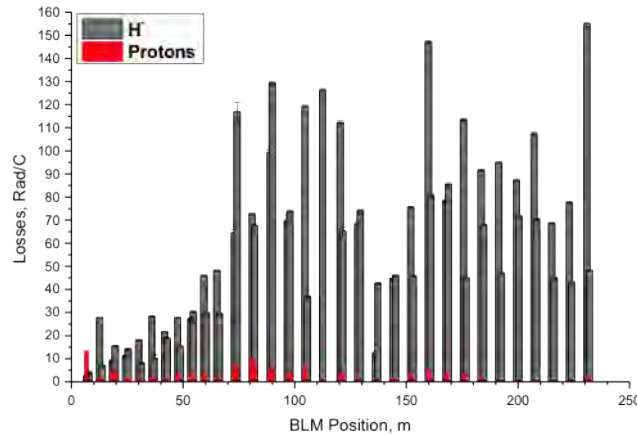
For the case of the SNS SCL, the relativistic  $\beta$  velocities in a given plane in the beam frame, based on particle tracking simulations, vary between  $1 \times 10^{-4}$  and  $7 \times 10^{-4}$ , and thereby mostly satisfy the requirements for Eq. (7). A simple beam loss calculation based on an average stripping cross-section that is on the  $\sigma_{\max}$  plateau, and average transverse and longitudinal beam sizes from model simulations, predicts [13] a total particle loss in the SCL of about  $3 \times 10^{-5}$ . This value is in excellent agreement with the measured [1] value of  $(2-5) \times 10^{-5}$ . A more precise calculation awaits a detailed particle tracking simulation.

### 3.4 Magnetic field stripping

Yet another beam loss mechanism that is important for  $H^-$  ions but not for proton accelerators is magnetic field stripping. This is rarely a problem since the maximum allowable fields are readily calculable and usually avoidable. Magnetic fields are Lorentz transformed to electric fields in the rest frame of the  $H^-$



**Fig. 10:** Normalized beam loss (loss monitor signal divided by the peak beam current) in the SNS SCL for two different optics cases, as a function of ion source current, for both  $H^+$  and  $H^-$  beams. Black:  $H^-$  beam with SCL quadrupole gradients set to design values. Green:  $H^+$  beam with SCL quadrupole gradients set to design values. Red:  $H^-$  beam with SCL quadrupole gradients lowered by up to 40% to minimize the beam loss. Blue:  $H^+$  beam with SCL quadrupole gradients set to the same values as for the  $H^-$  minimum loss case. Figure reproduced from Ref. [16].



**Fig. 11:** BLMs along the SCL, showing the proton (red) versus  $H^-$  (grey) beam loss for the design optics case, for 30 mA beam current. Figure reproduced from Ref. [1].

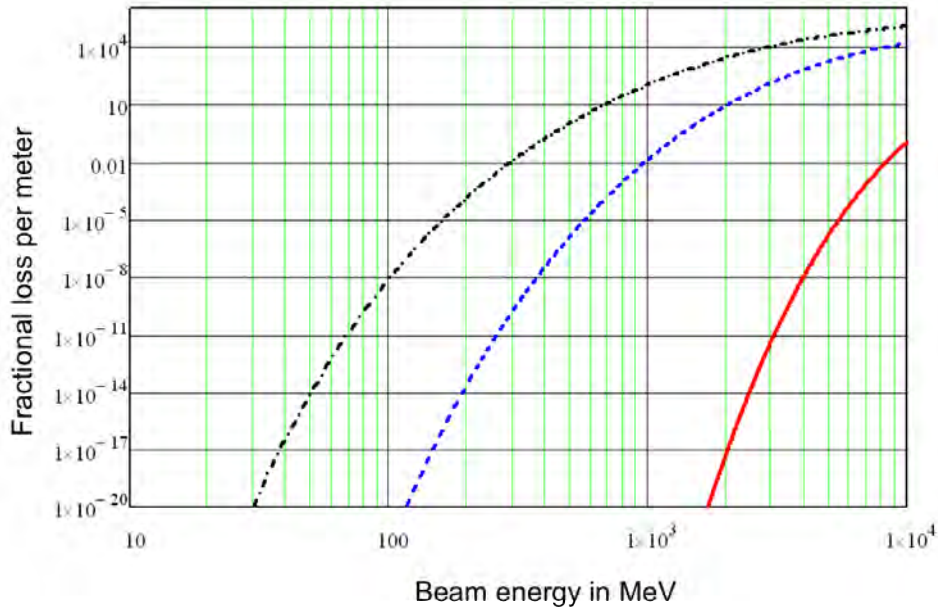
particles and, if the field is high enough, it will strip off some electrons. The fractional loss per unit length is given [15] by

$$\frac{df}{ds} = \frac{B(s)}{A_1} e^{-A_2/\beta\gamma cB(s)}, \quad (9)$$

where  $df/ds$  is the fractional loss per metre,  $B(s)$  is the magnetic field as a function of distance along the beam line,  $A_1 = 2.47 \times 10^{-6}$  V/m,  $A_2 = 4.49 \times 10^9$  V/m and  $\beta$ ,  $\gamma$  and  $c$  are the usual relativistic factors.

The effect is greatest at high beam energies where the Lorentz transform has the greatest effect, as shown in Eq. (9) and plotted in Fig. 12. However, it is easy to overlook the possible scenario where,





**Fig. 12:** Fractional loss per metre due to magnetic field stripping, for magnetic fields of 0.1 (red, solid), 0.5 (blue, dashed) and 1.0 T (black, dot-dash), as a function of  $H^-$  beam energy.

after adjusting quadrupole gradients to minimize the beam loss, the beam size is larger than expected inside quadrupole magnets whose gradients are larger than expected, which could lead to field stripping.

The ISIS facility sees a small amount of field stripping in the 70 MeV transport line between the linac and the ring, at the level of  $<1\%$ , just enough to create some minor hot spots [4]. SNS, J-PARC and LANSCE have not reported any significant beam loss due to this mechanism.

### 3.5 Black-body radiation

Photodetachment using laser beams is a well-developed method to measure  $H^-$  beam profiles and beam emittances, and it is now being developed as a method for charge exchange injection into storage rings and synchrotrons [17]. In these techniques, photons from a laser either strip off the loosely bound electrons from  $H^-$  particles or excite ground-state electrons in  $H^0$  particles so that they can be subsequently stripped completely off by a downstream magnetic field.

Photodetachment can also be caused by black-body radiation, but the stripping rate is minimal for today's  $H^-$  beam energies. The highest-energy  $H^-$  beams to date are produced at the Oak Ridge Spallation Neutron Source, where the beam energy is routinely 940 MeV and the maximum beam energy, demonstrated for short times, is 1.07 GeV. At 1 GeV the beam loss rate due to room-temperature black-body radiation has been estimated [18, 19] to be just  $3 \times 10^{-9}$  per metre or about 100 times less than our maximum allowable loss rate.

However, as the  $H^-$  beam energy increases, the Doppler-shifted black-body photon energies can increase enough to cause significant stripping rates. For example, at 8 GeV, which is a possible charge exchange injection energy for Fermilab's Project X, the stripping rate climbs to  $8 \times 10^{-7}$  per metre [19]. At this level of beam loss photodetachment becomes a serious concern and mitigation methods such as cooling the beam pipe to cryogenic temperatures have been considered.

The probability of beam loss due to black-body photodetachment depends on the overlap of two distributions: the  $H^-$  photodetachment cross-section versus photon energy, which peaks at a photon

energy of about 1.4 eV; and the black-body photon spectral density Doppler shifted to the rest frame of the  $H^-$  ions. For 300 K room-temperature black-body radiation, the probability of stripping is maximum for a beam energy of about 50 GeV [18].

Table 2 shows a summary of the various beam loss mechanisms for some relevant  $H^-$  accelerators.

**Table 2:** Selected beam loss mechanisms observed at various  $H^-$  linacs

| Beam loss mechanism            | SNS                                     | J-PARC   | ISIS  | LANSCE   |
|--------------------------------|---|--|---|--|
| Intra-beam stripping           | Yes, dominant loss in SCL               | Not noted as significant   | Not noted as significant  | Yes, significant, 75% of loss in CCL                   |
| Residual gas stripping         | Yes, moderate stripping in CCL and HEPT | Yes, significant, improved by adding pumping to S-DTL and future ACS section | Yes, not significant when vacuum is good, but can be significant if there are vacuum problems | Yes, significant, 25% of loss in CCL                   |
| $H^+$ capture and acceleration | Possibly, but not significant concern   | Yes, was significant, cured by chicane in MEPT                               | Not noted as significant  | Yes, significant if there is a vacuum leak in the LEPT |
| Field stripping                | Insignificant                           | Insignificant  | Yes, <1% in 70 MeV transport line, some hot spots   | Insignificant  |
| Black-body radiation           | Insignificant                           | Insignificant  | Insignificant   | Insignificant  |

#### 4 Beam loss in both $H^+$ and $H^-$ linacs

The previous sections focused on beam loss in  $H^-$  linacs. We now turn our attention to beam loss that is common to both  $H^+$  and  $H^-$  linacs. The examples we will focus on are:

- beam loss due to beam halo (or beam tails) that can arise from resonances due to magnetic imperfections, collective effects (e.g. space charge) or beam mismatches;
- RF and/or ion source turn on/off transients;
- dark current from the ion source.

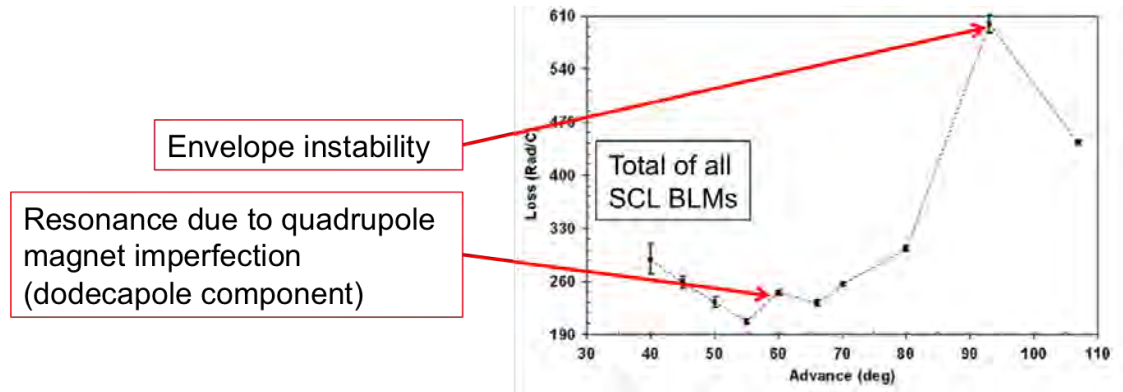
We will discuss each of these one at a time.

##### 4.1 Beam loss due to halo/tails

A well-known source of beam loss, present at the Oak Ridge SNS as well as all other proton and  $H^-$  accelerators, is beam halo or long tails on the beam distribution. When the halo/tails intercept the beam pipe apertures the beam is lost. A certain level of halo/tail formation is inevitable, but it is exacerbated by space charge, mismatched beams, structure resonances, parametric resonances, etc.

Certain phase advances can cause beam loss in linacs and beam transport lines. For example, the  $n\sigma_0 = 180$  or  $360$  degrees resonances drive halo formation, where  $\sigma_0$  is the transverse phase advance per cell for the zero space charge case. The  $\sigma_0 = 90$  degrees resonance, also known as the envelope instability, is commonly avoided in all high-intensity linac designs. Magnet imperfections can also cause

other resonance-driven losses. An example of both of these cases, drawn from the SNS SCL, is shown in Fig. 13. In this figure, the highest beam loss occurs at  $\sigma_0 = 90$  deg, and corresponds to the envelope instability. At  $\sigma_0 = 60$  degrees we see another, much smaller, increase in beam loss due to the unwanted dodecapole component in the SNS SCL quadrupole magnets. This is an example of a magnet imperfection resonance.



**Fig. 13:** Beam loss in the SNS SCL as a function of transverse phase advance per cell, with all RF cavities off. Figure reproduced from Ref. [20].

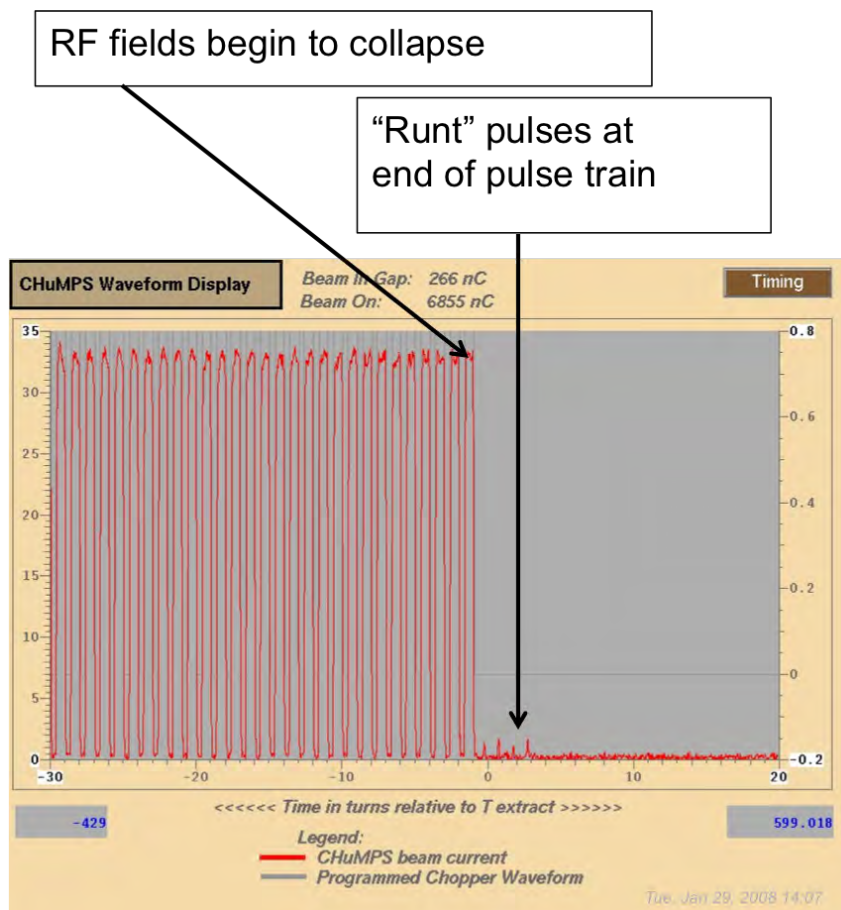
#### 4.2 Beam loss due to RF turn on/turn off transients

This beam loss mechanism is important for accelerators with pulsed RF systems (as opposed to CW linacs). If any beam is accelerated while the RF fields are ramping up or ramping down, it is likely to be lost. This problem is often solved with a chopper system, located at low beam energy, that blanks the beam during these times. The SNS linac has a chopper system located just after the ion source, at a beam energy of 65 keV, but it is not perfect. There is still a small amount of imperfectly chopped beam present at the end of the pulse. This beam is poorly accelerated during the collapse of the RF fields, and it causes beam loss at high beam energy, primarily in the beam transport from the storage ring to the target. We mitigate this loss by purposely ending the Radio Frequency Quadrupole (RFQ) RF pulse about  $3 \mu\text{s}$  before the rest of the RF pulses. The RFQ has a low quality factor, so its field collapses quickly, thus terminating the poorly chopped beam at low beam energy, before it can be accelerated by the rest of the linac. The poorly chopped beam at the beginning of the pulse is mitigated by delaying the ion source turn on until the RF fields have sufficiently ramped up. Figure 14 shows an example of a poorly chopped beam at the end of the pulse. In this figure the RFQ RF pulse does not end early as discussed above.

#### 4.3 Beam loss due to dark current

An unanticipated beam loss mechanism at SNS, discovered during commissioning, is dark current from the ion source. The ion source is pulsed at 60 Hz to create the required 38 mA peak  $\text{H}^-$  ion beam, but it also emits a continuous low-level beam of about  $3 \mu\text{A}$  ('dark current') due to the 13 MHz CW RF transmitter used to help ignite the pulsed plasma. The pulsed RF accelerator systems turn on and off at 60 Hz, and the dark current is only partially accelerated during the on and off transients, which creates beam loss. Even when the dark current is properly accelerated it is undesirable, because it creates beam in the extraction gap of the storage ring.

Figure 15 shows dark current lighting up a view screen located at the injection point of the SNS storage ring, where the stripper foil would normally be located. To mitigate this beam loss mechanism during normal operation, the LEPT chopper was modified to fully blank the head and tail of the beam pulse throughout the entire RFQ pulse length. Also, when the beam is turned off or the machine repetition



**Fig. 14:** A fast BCM in the MEBT shows a poorly chopped beam at the end of the beam pulse [21]. This beam will be accelerated by the downstream RF cavities, only to be lost at high energy.

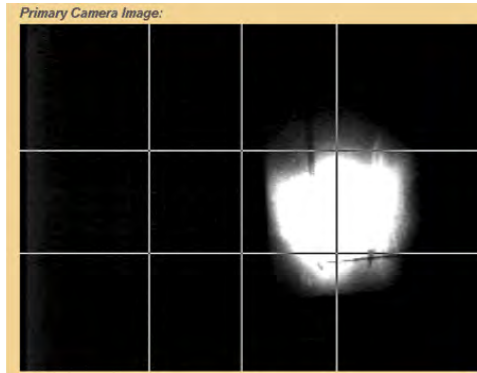
rate is less than 60 Hz, the low-level RF system for the first DTL tank was modified to automatically shift the RF phase 180 degrees to prevent acceleration of the dark current beyond this point. Both  $H^+$  and  $H^-$  linacs, without some kind of beam chopper system, may experience similar beam loss issues due to the ion source turn on/off transients and/or the RF turn on/off transients and/or dark current from the ion source.

#### 4.4 Occasional beam loss

A large amount of beam loss can occasionally occur due to, for example:

- response time for RF feedback and feedforward systems;
- RF trips off due to an interlock;
- fluctuations in the ion source;
- drifts in the RF system (e.g. due to temperature in klystron gallery);
- pulsed magnets miss a pulse or provide only a partial pulse.

The integrated beam power lost may be small compared to the continuous beam loss, but the consequences can be large. For example, occasional but large beam loss can damage superconducting cavities. The energy deposited on the cavity surface desorbs gas or particulates, which creates an environment for arcing or low-level discharge. This can cause the RF cavity performance to degrade over time. At SNS, we have had to lower some cavity fields due to this damage mechanism. Some cavities



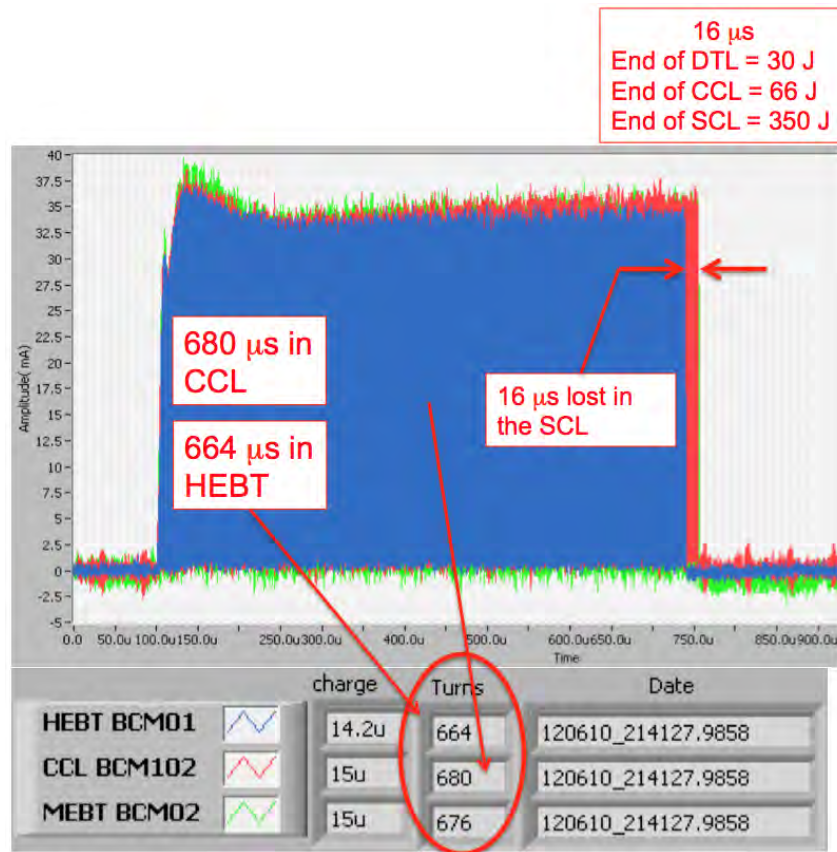
**Fig. 15:** Example of dark current at a view screen located at the SNS ring injection point. The beam is turned off, yet the dark current is present at levels sufficient to light up the view screen. The phase of the first DTL tank is *not* reversed for this image.

have been temporarily turned off. The lower cavity fields cause the final beam energy to be lower. The SNS SCL cavities do not trip off with every errant beam pulse, but the probability for a trip increases with time, and these trips cause downtime. The cavity performance degradation from errant beam can usually be restored by warming up the cavity during a long shutdown and then RF conditioning before resuming beam operation.

RF feedback and feedforward is an important part of beam loss control. The RF system must react to and also anticipate the beam loading caused by high-intensity beams. Otherwise there will not be a constant accelerating field in the cavity for the duration of the beam pulse, which can cause beam loss. Also, when the beam is turned back on after a trip, the RF system may have to re-optimize the feedback and feedforward parameters, and beam loss can be higher than normal during this time. For example, at SNS, after a latched beam-off trip, we slowly increase the average beam current over a period of about 1 min, by ramping both the peak current and the repetition rate, to give the RF system time to adapt. Beam losses are elevated during this time. Also, sudden changes in the beam pulse structure can cause the beam loading to change too rapidly for the RF system to compensate, and this can also cause beam loss. Similarly, if a RF system trips off in mid-pulse the collapsing field in the cavity will only partially accelerate the beam and cause beam loss in the downstream portion of the linac or beam transport lines. Due to the response time of the machine protection system (15–20  $\mu\text{s}$  at SNS), the ion source will continue to inject beam into the linac, only to be lost downstream of the affected cavity.

## 5 Errant beam capture at SNS

At SNS we have assembled a system to capture signals due to errant beam events (i.e. events due to sudden occasional beam loss). It is based on a combination of beam current monitors (BCMs), beam position monitors and beam loss monitors (BLMs). BCMs upstream and downstream of the SCL provide a differential measurement of how much beam is lost. The SCL BLM system, comprising 76 ion chamber detectors along the SCL, is used to identify the location of the loss. An automated report system assembles the data for each event for offline analysis. An example of one such analysis is shown in Fig. 16. In this case the event occurred about 664  $\mu\text{s}$  after the start of the pulse. Due to the response time of the machine protection system an additional 16  $\mu\text{s}$  of beam was injected into the linac, but it was lost before reaching the exit of the linac. Based on this image alone it is not possible to say where the beam was lost and what the beam energy was, but the amount of loss corresponds to about 30 J if it is lost in the DTL, 66 J in the CCL and 350 J in the SCL. Clearly, 350 J is enough to be concerned about for superconducting cavities. Another example of an errant beam event in the SNS linac is shown in Fig. 17. In this case about 12  $\mu\text{s}$  of beam is lost in the SCL, and it is due to the sudden collapse of the RF field in



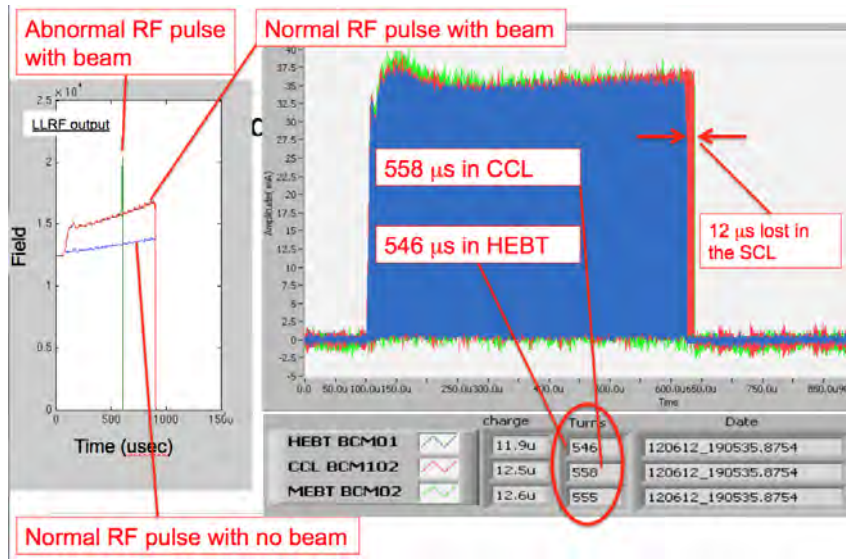
**Fig. 16:** Example of an errant beam event in the SNS linac [22, 23]

one of the warm linac RF cavities. Yet another example of an errant beam capture is shown in Fig. 18. A fast BCM in the MEBT, just downstream of the RFQ, shows a drop in the peak beam current toward the end of the beam pulse, most likely due to a fault in the ion source. The drop in peak current will cause a sudden change in beam loading in the RF cavities that will be too fast for the RF system to respond, and it will cause beam loss.

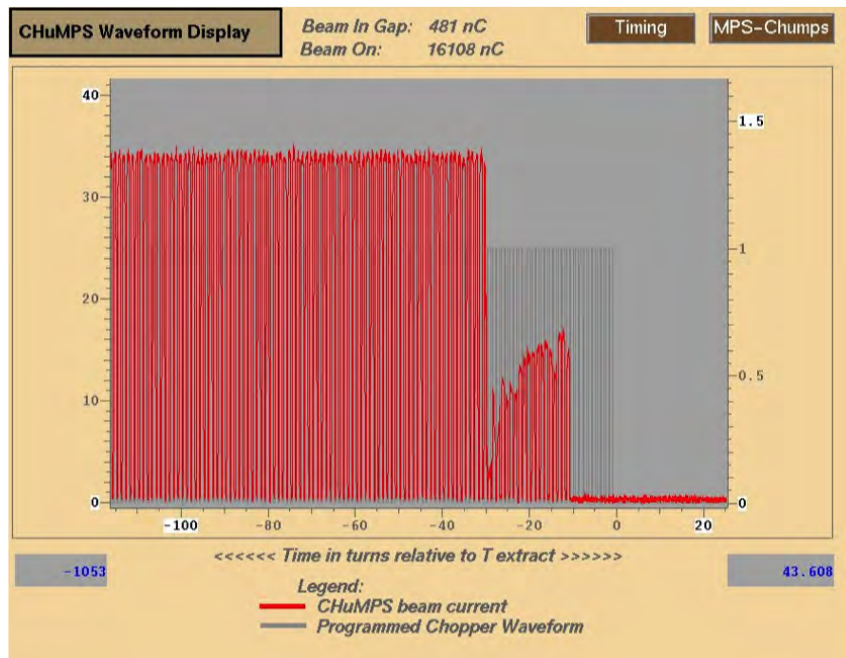
At the time of these measurements we estimated that <10% of the errant beams were due to the ion source and the LEBT, and that most of these occurred during the first week of a new source installation, most likely due to high-voltage arcing. More than 90% of BLM trips were due to warm linac RF faults. The RF faults occurred at various times during the pulse. Those that occurred during the RF fill had reproducible times, and those that occurred during the RF flat top were at random times. Based on these results we focused on reducing the warm linac RF faults. We did this through a combination of empirical adjustments to the RF field amplitudes, the RF fill times, the RF resonant frequencies and by frequent cryopump reconditioning to optimize the vacuum.

## 6 Beam loss mitigation

We have discussed a variety of causes of beam loss in linacs. We will now turn our attention to methods of mitigating beam loss. Table 3 shows some of the more common methods and the following sections discuss some of these in more detail.



**Fig. 17:** Example of an errant beam event in the SNS linac due to the sudden collapse of the RF field in one of the warm linac RF cavities [22, 23].



**Fig. 18:** A fast BCM just downstream of the RFQ shows a sudden drop in beam current toward the end of the beam pulse [22].

**Table 3:** Some methods of beam loss mitigation

| Cause of beam loss                                | Mitigation  |
|---|---|
| Beam halo—both transverse and longitudinal        | Scraping, collimation, better matching from one lattice to the next, magnet and RF adjustments    |
| Intra-beam stripping                              | Increase beam size (both transverse and longitudinal)   |
| Residual gas stripping                            | Improve vacuum  |
| H <sup>+</sup> capture and acceleration           | Improve vacuum, add chicane at low energy   |
| Magnetic field stripping                          | Avoid by design   |
| Dark current from ion source                      | Deflect at low energy, reverse (phase shift) RF cavity field when beam is turned off              |
| Off-normal beams (sudden, occasional beam losses) | Turn off beam as fast as possible, track down troublesome equipment and modify to trip less often |

### 6.1 Low-energy scraping

A common cause of beam loss is beam tails or beam halo striking the beam pipe. The halo/tails can be present from the very beginning (at the exit of the ion source) or they can be created by mismatched beams, space charge, parametric resonances, etc. Many beam loss mitigation efforts focus on reducing the beam halo/tails. At SNS we have two main methods to control halo/tails: low-energy scraping and matching.

We have found that beam scraping at low beam energy is an effective method to reduce beam loss at SNS due to halo/tails. Any scraping is best done at low beam energies where the scraper system can be more compact and where the consequent radioactivation is less. In 2004–2005, left and right scrapers were added to the MEBT section of the warm linac, between the RFQ and the first DTL tank, where the beam energy is 2.5 MeV. Figure 19 shows a typical scraping result. The magnitude of the beam loss reduction varies from case to case and location to location. It is sometimes more than a factor of 10.

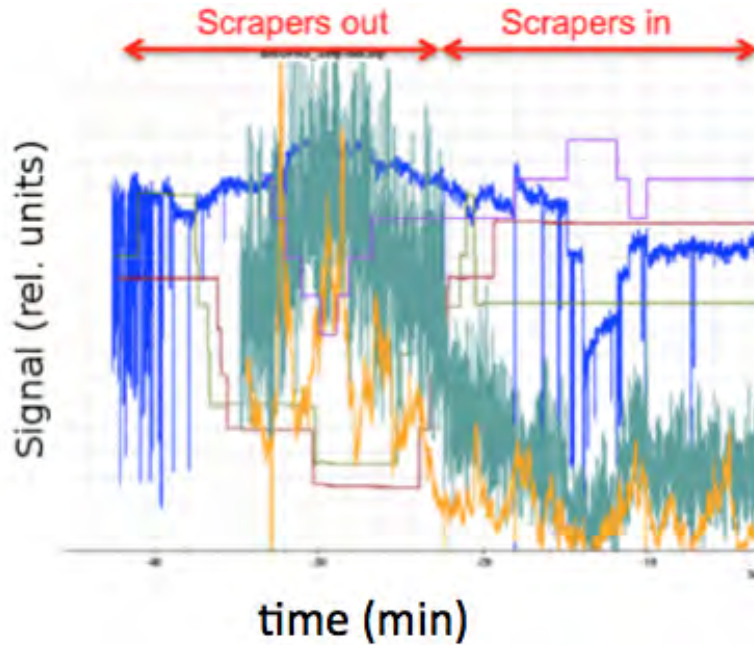
These scrapers produce a measurable reduction in the beam halo/tails, as can be seen on the MEBT slit-and-collector emittance scanner results shown in Fig. 20. The halo/tail reduction continues into the DTL, as can be seen in the non-Gaussian tails on a wire-scanner measurement shown in Fig. 21. However, by the time the beam reaches the HEFT at the exit of the linac, there is very little measurable difference in the beam profile with and without MEBT scraping, as shown by the halo/tail measurement in Fig. 22. Nevertheless, as shown in Fig. 19, there are still large loss reductions at certain points downstream of the HEFT due to MEBT scraping, especially in the ring injection dump beam line.

Based on the success of these scrapers, we added top/bottom scrapers in summer 2013. Unfortunately there is not sufficient space in the SNS linac to add scrapers at other locations, for example between the warm and cold linac sections, but this may be desirable in the next-generation SCL designs.

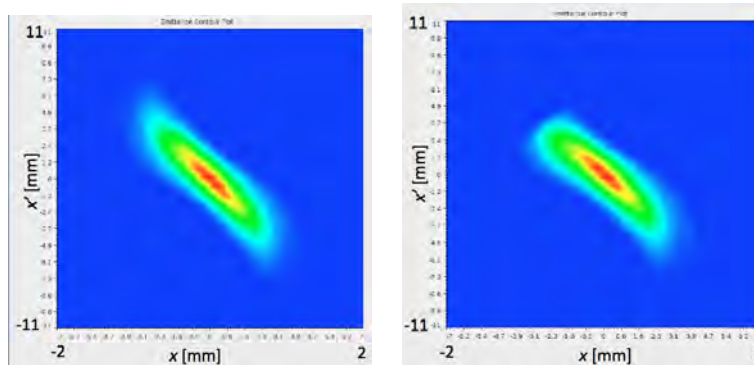
### 6.2 Beam loss reduction by increasing the H<sup>-</sup> beam size

In linacs that accelerate H<sup>-</sup> beams, IBSt can be the dominant source of beam loss. The SNS and LAN-SCE linacs are two such examples. Because the IBSt reaction rate is proportional to the beam density squared, a large reduction in beam loss can be achieved by increasing the beam size, in the longitudinal, transverse, or both, dimensions. Figure 23 shows the design quadrupole strengths for the SNS SCL, and also a set of strengths that were empirically determined to reduce the beam loss by about 50%. The lower quadrupole strengths increase the transverse beam size, thereby lowering the beam density and lowering the beam loss, as shown in Fig. 10. Further reduction in quadrupole strengths will cause the beam size to increase so much that the beam halo/tails will scrape on the beam pipe and increase the beam loss.





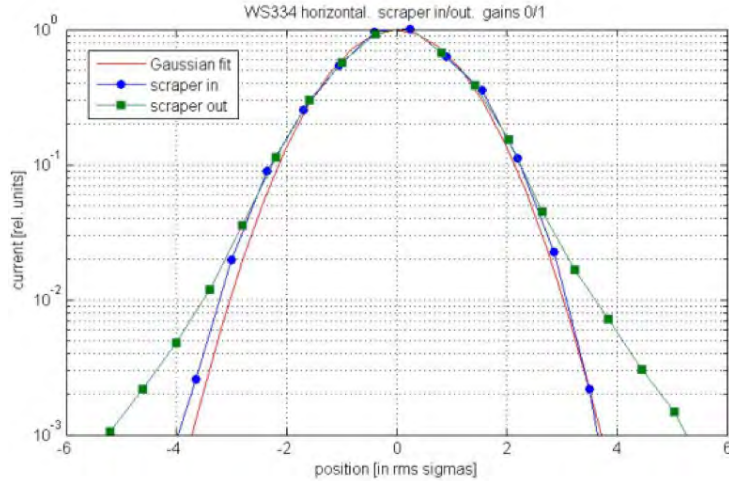
**Fig. 19:** Beam loss reduction due to low-energy scraping. Magenta, light green and violet show the scraper positions. Blue shows the beam charge, and orange and dark green show BLM signals. The beam loss is reduced by up to 50–60% at certain locations by scraping about 3% of the beam, primarily in the CCL, the beginning of the SCL and the ring injection dump beam line. Figure reproduced from Ref. [24].



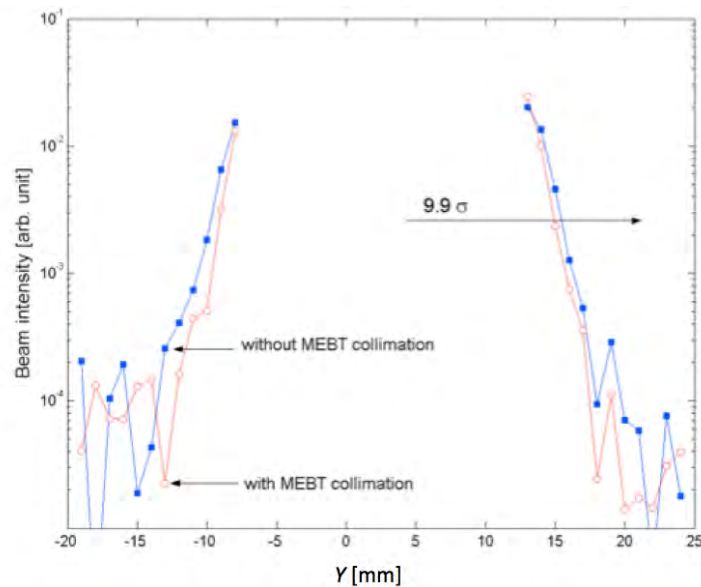
**Fig. 20:** Horizontal emittance scanner without (left) and with (right) scraping. This emittance station is located in the MEBT, downstream of the scrapers, where the beam energy is still 2.5 MeV. Figure reproduced from Ref. [25].

### 6.3 Matching

To minimize beam loss, conventional wisdom dictates that the Twiss parameters of the beam should be matched when the beam passes from one lattice section to the next, e.g. from one FODO lattice to another FODO lattice. For a perfect beam distribution this makes good sense because it minimizes the required aperture and prevents phase-space dilution. However, what if the beam distribution is not perfect and the Twiss parameters of the core of the beam are different from the tails of the distribution? Perhaps it is better to mismatch the core of the beam to allow better transmission (lower beam loss) for the part of the distribution that causes beam loss (i.e. the tails or halo of the beam). Figure 24, for example, shows measured rms beam sizes in the beginning of the SNS HEBT for the low-loss production tune, fitted with an envelope model with starting parameters varied to give the best fit. Figure 24 also shows the design



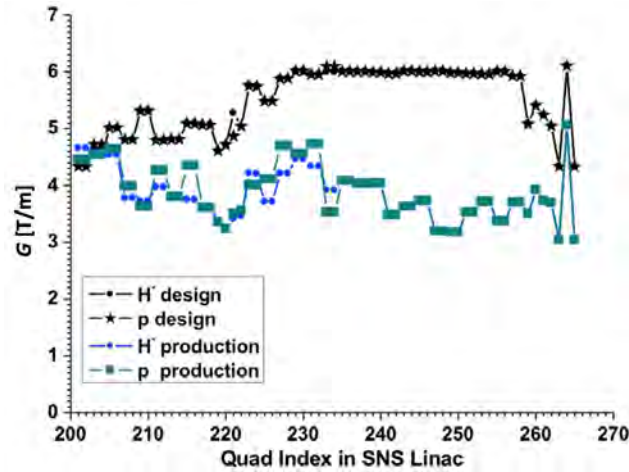
**Fig. 21:** A wire-scanner beam profile measurement in the DTL, before and after inserting the MEBT scrapers. The beam energy is 39.7 MeV. Blue circles—scraper is in; green squares—scraper is out; solid red line—Gaussian fit. Figure reproduced from Ref. [26].



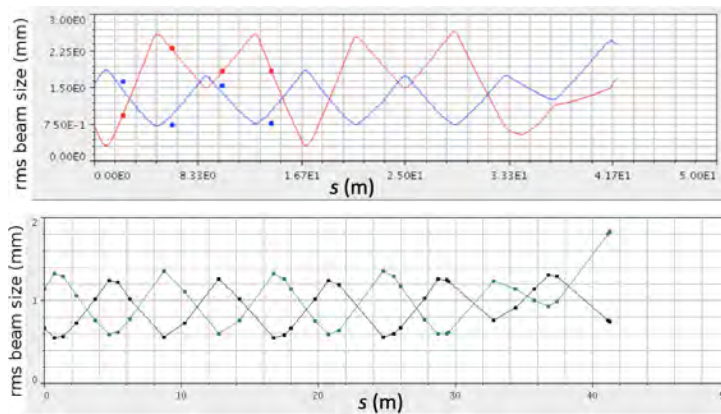
**Fig. 22:** A wire-scanner beam profile measurement in the HEBT, before and after inserting the MEBT scrapers. Figure reproduced from Ref. [27].

case, which assumes a well-matched beam. The low-loss tune clearly does not have a well-matched core of the beam in the horizontal plane.

Another example of beam matching versus beam tails comes from a series of large-dynamic-range wire-scanner beam profile measurements in the SNS DTL [28]. Figure 25 shows two cases: the low-loss production tune and after matching the Twiss parameters at the DTL entrance. The low-loss tune shows non-Gaussian tails as large as 30% of the peak, starting from the first wire scanner in the DTL. The well-matched case shows much improved non-Gaussian beam tails at the beginning of the DTL, but the losses are higher and the tails start to re-form by the end of the DTL, and then persist throughout the linac and transport lines.



**Fig. 23:** Quadrupole strengths in the SNS SCL. Black shows the design strengths and green shows the empirically determined strengths, approximately 40% lower than the design strengths, that reduce the beam loss by about 50%. Figure reproduced from Ref. [16].

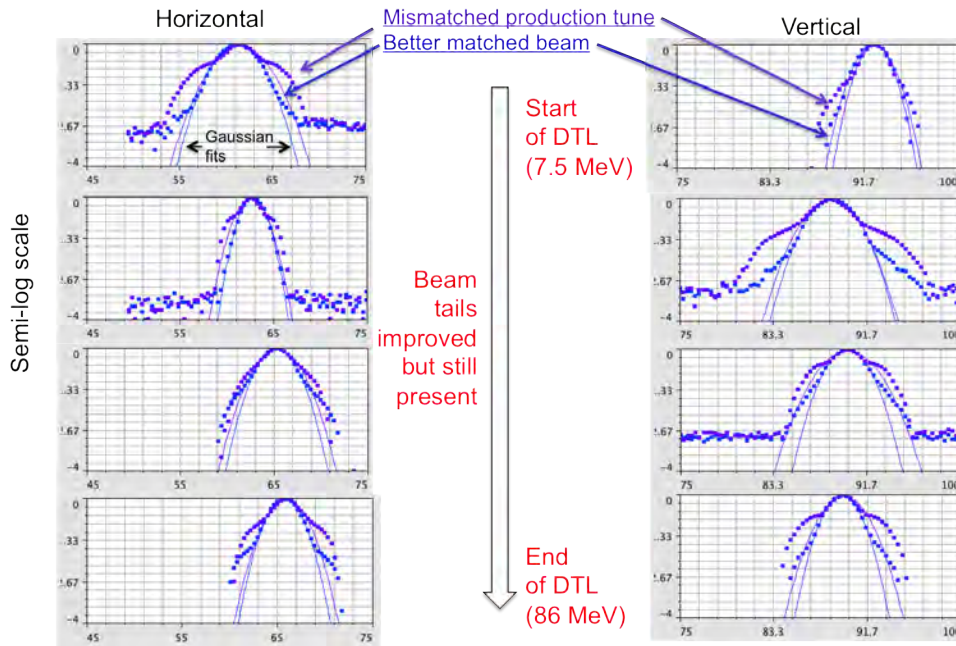


**Fig. 24:** Beam sizes (rms) in the HEBT. Top: the low-loss production case, showing beam sizes corresponding to a mismatched beam. Dots show the measured rms beam sizes and the solid lines show the results of an envelope model adjusted to fit the data. Red is horizontal, blue is vertical. Bottom: the corresponding envelope model for the design case. In this plot the dots indicate lattice element locations. Upper plot reproduced from Ref. [29].

At SNS, in the HEBT, the DTL and other areas of the accelerator, the low-loss tune does not have well-matched Twiss parameters for the core of the beam. Setting up the matched core case is a good place to start the beam loss minimization process, but the minimum-loss case can only be found by empirical adjustments to quadrupole gradients and the RF phases and amplitudes. With our present suite of beam instrumentation it is not possible to accurately characterize the parameters of the beam tails/halo, and then use that data to better match the beam, because of the limited dynamic ranges of the measurements.

#### 6.4 Beam loss reduction by magnet and RF phase adjustments

As discussed in the previous section, when the SNS accelerator complex is operated with the quadrupole gradients set to the design values, the beam loss is high and the same is true for some of the RF cavity phases. We have empirically found that the loss can be reduced by at least a factor of two by making small adjustments to the matching quadrupole gradients and to certain RF phases. The minimum beam loss case is very sensitive to certain RF phases, such as those in the DTL. Just 1 degree of phase change



**Fig. 25:** Normalized horizontal and vertical beam profiles in the SNS DTL showing beam tail formation in the DTL. Purple is the low-loss production tune and blue shows the result of matching the beam at the entrance to the DTL. The solid lines show Gaussian fits to the data. Figure reproduced from Ref. [9].

can double the beam loss at certain locations. As the SNS accelerator complex has matured, we have worked to bring the quadrupole gradient set points in line with design values from simulation codes. In some cases the simulation codes have been updated to more accurately reflect the actual accelerator, and in other cases low-loss tunes have been found that minimize the discrepancies between the accelerator and the model. The concept is that model-based adjustments that have a physics basis are better than random adjustments that may lead to a localized minimum in beam loss or to an operating point that is difficult to sustain.

The SNS DTL quadrupole gradients are by definition operated according to design, since they are permanent magnets. The CCL gradients are now also operated very close their design values, starting in August 2012. The SCL quadrupoles are up to 40% below their design values, to increase the beam size as discussed above. The HEBT quadrupole gradients are mostly operated according to their design values, with the biggest exceptions being at the beginning and end of the HEBT, where the lattice changes from the SCL to the HEBT, and from the HEBT to the ring. The beam loss in the injection dump beam line is very sensitive to changes in the HEBT quadrupole gradients.

Similar experiences are seen at other accelerator facilities, including LANSCE [30], ISIS [31] and the HIPA cyclotron at PSI. At HIPA the last 20–50% of beam loss reduction comes from empirical tuning [32].

## 7 Summary

Beam loss is a very important driver to the cost, design and operation of high-intensity  $H^+$  and  $H^-$  accelerators, and will become even more important as next-generation accelerators come online. Beam loss in  $H^-$  accelerators is much more complex than  $H^+$  accelerators due to the many more beam loss mechanisms available to  $H^-$  beams. Modern design and simulation codes cannot accurately predict beam loss and so prudent designs must include flexible machine lattices and mitigation measures such as scrapers and collimators. However, the future is bright for the next-generation accelerators that can

take advantage of and expand on the observations and experiences of today's high-intensity accelerator facilities.

### Acknowledgements

ORNL is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy.

Notice: This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide licence to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

### References

- [1] A. Shishlo *et al.*, *Phys. Rev. Lett.* **108** (2012) 114801. <http://dx.doi.org/10.1103/PhysRevLett.108.114801>
- [2] A. Miura *et al.*, LINAC2010, Tsukuba, Japan, 2010, p. 587. <http://www.jacow.org>.
- [3] L. Rybarczyk *et al.*, IPAC2012, New Orleans, LA, USA, 2012, p. 3892. <http://www.jacow.org>.
- [4] A. Lechford, Private communication, August 2012.
- [5] G. Gillespie, *Phys. Rev. A* **15** (1977) 563. <http://dx.doi.org/10.1103/PhysRevA.15.563>
- [6] G. Gillespie, *Phys. Rev. A* **16** (1977) 943. <http://dx.doi.org/10.1103/PhysRevA.16.943>
- [7] R. Shafer, Tech. Note TN:LANSCE-1:99-085, May 1, 1999.
- [8] G. Gillespie, *NIM B* **2** (1984) 231. [http://dx.doi.org/10.1016/0168-583X\(84\)90196-4](http://dx.doi.org/10.1016/0168-583X(84)90196-4)
- [9] M. Plum, Proc. HB2012 Workshop, Beijing, China, 2012. <http://www.jacow.org>.
- [10] J. Galambos, Private communication.
- [11] R. McCrady *et al.*, LINAC2010, Tsukuba, Japan, 2010, p. 908. <http://www.jacow.org>.
- [12] K. Hasegawa, LINAC2010, Tsukuba, Japan, 2010, p. 449. <http://www.jacow.org>.
- [13] V. Lebedev *et al.*, LINAC2010, Tsukuba, Japan, 2010, p. 929. <http://www.jacow.org>.
- [14] J. Galambos *et al.*, Snowmass 2001, Snowmass Village, June 30 - July 21, 2001.
- [15] A.J. Jason, D.W. Hudgins and O.B. VanDyke, *IEEE Trans. Nucl. Sci.* **NS-28**(3) 2703 (1981) 2703; <http://dx.doi.org/10.1109/TNS.1981.4331890>
- [16] A. Shishlo *et al.*, IPAC2012, New Orleans, LA, USA, 2012, p. 1074. <http://www.jacow.org>.
- [17] Y. Liu *et al.*, Proc. 2011 Particle Accelerator Conf., New York, 2011, p. 1433. <http://www.jacow.org>.
- [18] H.C. Bryant and G.H. Herling, *J. Mod. Opt.* **53**(1–2) (2006) 45; <http://dx.doi.org/10.1080/09500340500280835>
- [19] D. Johnson, Proc. HB2008 Workshop, Nashville, TN, USA, 2008, p. 290. <http://www.jacow.org>.
- [20] Y. Zhang *et al.*, *Phys. Rev. Spec. Top. Accel. Beams* **14** (2010) 044401. <http://dx.doi.org/10.1103/PhysRevSTAB.13.044401>
- [21] L. Longcoy, SNS electronic logbook entry, 29 January 2008.
- [22] C. Peters, Errant beam update, SNS Accelerator Advisory Committee Meeting, 7 May 2013.
- [23] W. Blokland, Private communication, 2012–2013.
- [24] J. Galambos, Ramp up progress, SNS Accelerator Advisory Committee Meeting, 3 February 2010.
- [25] A. Zhukov, A. Aleksandrov and A. Shishlo, Transverse emittance measurements in MEBT at SNS, Proc. Linear Accelerator Conf., LINAC2010, Tsukuba, Japan, 2010, p. 614. <http://www.jacow.org>.

- [26] A. Aleksandrov *et al.*, Beam dynamics studies and beam quality in the SNS normal-conducting linac, Proc. 2005 Particle Accelerator Conf., Knoxville, TN, USA, 2005, p. 3381. <http://dx.doi.org/10.1109/PAC.2005.1591478>. <http://www.jacow.org>.
- [27] D. Jeon, J. Galambos, J. Tang and Y. Zhang, Collimation in the SNS linac and HEBT, Accelerator Performance Seminar, Oak Ridge, TN, USA, 3 June 2010.
- [28] C.K. Allen, Private communication, 2012.
- [29] S. Cousineau, SNS electronic logbook, 5 August 2011.
- [30] L. Rybarczyk, Proc. HB2012 Workshop, Beijing, China, 2012, p. 324. <http://www.jacow.org>.
- [31] D. Adams, Proc. HB2012 Workshop, Beijing, China, 2012, p. 560. <http://www.jacow.org>.
- [32] M. Seidel, J. Grillenberger and A.C. Mezger, Proc. HB2012 Workshop, Beijing, China, 2012, p. 555. <http://www.jacow.org>.

# High-Intensity Synchrotron Radiation Effects

*Y. Suetsugu*

High Energy Accelerator Research Organization, Tsukuba, Japan

## Abstract

Various effects of intense synchrotron radiation on the performance of particle accelerators, especially for storage rings, are discussed. Following a brief introduction to synchrotron radiation, the basic concepts of heat load, gas load, electron emission, and the countermeasures against these effects are discussed.

## Keywords

Accelerator vacuum system; synchrotron radiation; photon stimulated gas desorption; photoelectron; heat load.

## 1 Introduction

Recent high-power (that is, high-current and high-energy) particle accelerators generate intense synchrotron radiation (SR). This is a good photon source. However, it has the following potentially harmful effects on accelerator performance:

- i) heat load: damage to beam pipes or instruments,
- ii) gas load: short lifetime, noise to particle detectors,
- iii) electron emission: beam instabilities, gas load,
- iv) radiation: radiation damage.

The first three effects are directly related to the beam and the vacuum system. In this paper, basic and practical concepts to understand the three effects are presented, along with measures to treat these problems, that is, to protect the machine in a broad sense. These problems affect accelerator vacuum systems, but they have widespread effects upon overall machine performances as well. The understanding of these problems is also useful in designing and constructing accelerators.

## 2 Synchrotron radiation

Synchrotron radiation comprises electromagnetic waves emitted when a high-energy charged particle is accelerated in a direction orthogonal to its velocity, such as in a magnetic field (Fig. 1) [1]. The SR is useful as a photon source. The main features of SR compared to other photon sources are:

- high intensity and high photon flux,
- wide range of wavelengths, from infrared to hard X-ray,
- well understood spectrum intensity,
- high brightness,
- high polarization ratio.

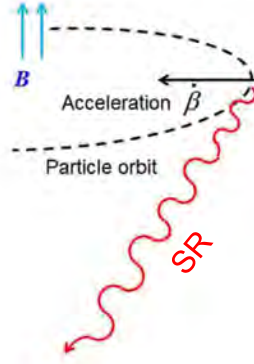


Fig. 1: Synchrotron radiation

An accelerated charged particle emits electromagnetic radiation. The radiation fields (electric field  $\vec{E}$  and magnetic field  $\vec{B}$ ) are given by using electromagnetic potentials:

$$\vec{E} = -\frac{\partial}{\partial t} \vec{A} - \nabla \phi, \quad \vec{B} = \nabla \times \vec{A} . \quad (1)$$

Here,  $\phi$  and  $\vec{A}$  are the Lienard-Wiechert scalar and vector potentials, which are given by

$$\vec{A}(t) = \frac{e}{4\pi\epsilon_0 c} \left[ \frac{\vec{\beta}}{R(1-\vec{n}\cdot\vec{\beta})} \right]_{\text{ret}}, \quad \phi(t) = \frac{e}{4\pi\epsilon_0} \left[ \frac{1}{R(1-\vec{n}\cdot\vec{\beta})} \right]_{\text{ret}}, \quad (2)$$

where  $\vec{R}(t_{\text{ret}})$  is the distance vector from the source to the observer (see Fig. 2), and  $t_{\text{ret}}$  is the retarded time,  $ct_{\text{ret}} = ct - R(t_{\text{ret}})$ , and  $\vec{\beta}$  is the ratio of the velocity  $\vec{v}$  to the speed of light  $c$  (that is,  $\vec{\beta} = \vec{v}/c$ ). Hence the electric and magnetic fields are obtained by

$$\vec{B} = \frac{1}{c} [\vec{n} \times \vec{E}]_{\text{ret}}, \quad (3)$$

$$\vec{E} = \frac{e}{4\pi\epsilon_0} \left[ \frac{(1-\beta^2)(\vec{n}-\vec{\beta})}{R^2(1-\vec{n}\cdot\vec{\beta})^3} \right]_{\text{ret}} + \frac{e}{4\pi\epsilon_0 c} \left[ \frac{\vec{n} \times (\vec{n}-\vec{\beta}) \times \dot{\vec{\beta}}}{R(1-\vec{n}\cdot\vec{\beta})^3} \right]_{\text{ret}} . \quad (4)$$

At observing points far from the emitting point, the radiation field of the latter term of  $\vec{E}$  ( $\propto 1/R$ ) is more important, and the former term can be neglected.

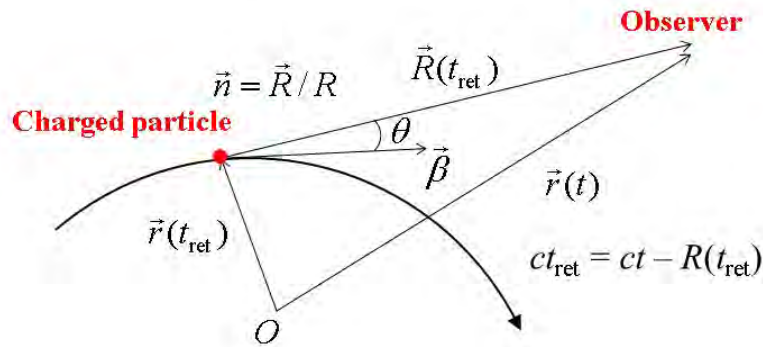


Fig. 2: Coordinate system



The pointing vector, that is, the radiation energy flow toward  $R$  per unit area, is given by

$$\vec{S}_r(t) = \frac{1}{\mu_0} \vec{E} \times \vec{B} = \frac{1}{\mu_0 c} E^2 (1 - \vec{\beta} \cdot \vec{n}) \vec{n} \Big|_{\text{ret}} = \varepsilon_0 c E^2 (1 - \vec{\beta} \cdot \vec{n}) \vec{n} \Big|_{\text{ret}} . \quad (5)$$

Then, the instantaneous differential radiation per unit solid angle  $d\Omega$  becomes

$$\frac{dP}{d\Omega} = \vec{n} \cdot \vec{S} R^2 \Big|_{\text{ret}} = \varepsilon_0 c E^2 (1 - \vec{n} \cdot \vec{\beta}) R^2 \Big|_{\text{ret}} = \frac{e^2}{16\pi^2 \varepsilon_0 c} \frac{\left| \vec{n} \times \left\{ (\vec{n} - \vec{\beta}) \times \dot{\vec{\beta}} \right\} \right|^2}{(1 - \vec{n} \cdot \vec{\beta})^5} \Big|_{\text{ret}} . \quad (6)$$

If  $\dot{\vec{\beta}}$  is parallel to  $\vec{\beta}$ ,

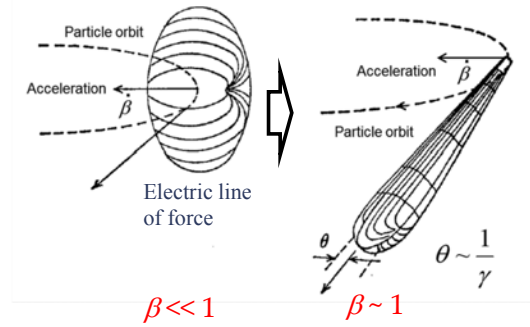
$$\frac{dP}{d\Omega} = \frac{e^2 \dot{\beta}^2}{16\pi^2 \varepsilon_0 c} \frac{\sin^2 \theta}{(1 - \beta \cos \theta)^5} . \quad (7)$$

On the other hand, if  $\dot{\vec{\beta}}$  is orthogonal to  $\vec{\beta}$ ,

$$\frac{dP}{d\Omega} = \frac{e^2 \dot{\beta}^2}{16\pi^2 \varepsilon_0 c} \frac{(1 - \beta \cos \theta)^2 - (1 - \beta^2) \sin^2 \theta}{(1 - \beta \cos \theta)^5} . \quad (8)$$

For both cases, when  $\beta \approx 1$ , the term  $(1 - \beta \cos \theta)^5$  approaches zero if  $\theta$  approaches to zero (see Fig. 3). This means that the power beams to the front of the orbit. This is called ‘beaming’. The angle of beaming  $\theta$  is given by

$$\theta = \frac{1}{\gamma}, \quad \gamma = \frac{1}{\sqrt{1 - \beta^2}} = \frac{E_e}{m_0 c^2} = \frac{E_e [\text{MeV}]}{0.511} \quad (\text{for electron}) . \quad (9)$$



**Fig. 3:** Beaming of SR

Now, consider a charged particle in a homogeneous magnetic field  $B$ . The acceleration in  $B$  is orthogonal to  $\beta$ , and is given by

$$\dot{\vec{\beta}}_{\perp} = \frac{\beta^2 c}{\rho} , \quad (10)$$

where the bending radius of charge particle  $\rho$  at the electron energy  $E_e$  is given by

$$\frac{1}{\rho [\text{m}]} = \frac{eBc}{\beta E_e} = 0.2998 \frac{B [\text{T}]}{\beta E_e [\text{GeV}]} . \quad (11)$$

Then, the instantaneous radiation power is obtained by integrating Eq. (8) over  $\psi$  (the angle on the plane of  $\dot{\vec{\beta}}$ ) and  $\theta$  (the angle orthogonal to the plane of  $\dot{\vec{\beta}}$ ),

$$P = \frac{2cr_e m_e c^2}{3} \frac{\beta^4 \gamma^4}{\rho^2} = \frac{cC_\gamma}{2\pi} \frac{E_e^4}{\rho^2}, \quad C_\gamma \equiv \frac{4\pi}{3} \frac{r_e}{(m_e c^2)^3} = 8.85 \times 10^{-5} \frac{\text{m}}{\text{GeV}^3}. \quad (12)$$

Here,  $r_e$  is the classical electron radius.

Note that the radiation power depends on the mass of the radiating particle as  $1/m^4$ . Synchrotron radiation is, therefore, much more important for electron and positron rings. For accelerators utilizing a superconducting system, however, such as the LHC, SR is also important for the proton beams, because the heating might transfer a significant heat load to the cryogenics system.

Hereafter, we consider the case of an electron or a positron deflected by a dipole magnet of a ring with a circumference of  $C$  (see Fig. 4). The radiation energy along the ring per electron is

$$U_0 = \oint P dt = \frac{C_\gamma}{2\pi} E_e^4 \oint \left( \frac{1}{\rho_x^2} + \frac{1}{\rho_y^2} \right) ds, \quad c dt = ds. \quad (13)$$

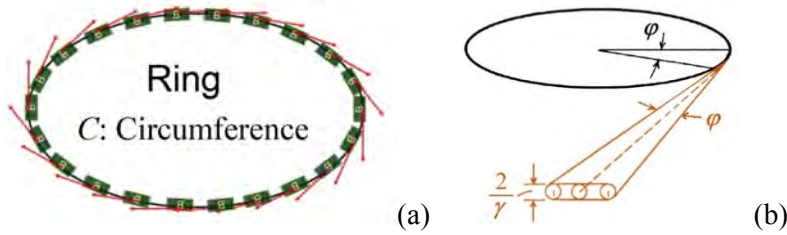
For an isomagnetic field (i.e.  $\rho = \text{constant}$ ),

$$U_0 = C_\gamma \frac{E_e^4}{\rho}. \quad (14)$$

For a circulating beam current  $I_e$ , the total radiation power  $P_{\text{le}}$  becomes

$$P_{\text{le}} = U_0 \times \frac{I_e}{e} = C_\gamma \frac{E_e^4}{\rho} \times \frac{I_e}{e}, \quad (15)$$

$$P_{\text{le}} [\text{W}] = 8.85 \times 10^4 \frac{E_e [\text{GeV}]^4}{\rho [\text{m}]} I_e [\text{A}] = 2.65 \times 10^4 E_e [\text{GeV}]^3 B [\text{T}] I_e [\text{A}]. \quad (16)$$



**Fig. 4:** (a) Schematic configuration of a ring with a circumference of  $C$ , where “B” means a dipole magnet, and (b) synchrotron radiation emitted within a deflection angle of  $\phi$ .

The average power line density in the ring is obtained by

$$\langle P_{\text{le, line}} \rangle = P_{\text{le}} / C. \quad (17)$$

The power in an angle of  $\phi$  is

$$P_{\text{le}}(\phi) = P_{\text{le}} \frac{\phi}{2\pi}. \quad (18)$$

Until now, we have only considered SR in a time domain regime. Sometimes it is also useful and important to consider SR in the frequency domain. The frequency spectrum of the electric field is obtained by Fourier transform of  $E(t)$ :

$$\tilde{E}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} E(t) e^{i\omega t} dt . \quad (19)$$

From Eq. (5),

$$\frac{dP}{d\Omega} = \vec{n} \cdot \vec{S} R^2 \Big|_{\text{ret}} = \varepsilon_0 c E^2 R^2 \Big|_{\text{ret}} , \quad (20)$$

$$\frac{dW}{d\Omega} = \int \frac{dP(t)}{d\Omega} dt = \frac{1}{\mu_0 c} \int_{-\infty}^{+\infty} (RE)^2 dt = \frac{1}{\mu_0 c} \int_{-\infty}^{+\infty} |R\tilde{E}(\omega)|^2 d\omega . \quad (21)$$

Then, the frequency spectrum of power is given by

$$\begin{aligned} \frac{d^2W}{d\Omega d\omega} &= \frac{1}{\mu_0 c} (R\tilde{E}(\omega))^2 = \frac{1}{2\pi\mu_0 c} \left| \int_{-\infty}^{+\infty} (RE) e^{i\omega t} dt \right|^2 \\ &= \frac{e^2}{16\pi^3 \varepsilon_0 c} \left| \int_{-\infty}^{+\infty} \left[ \frac{|\vec{n} \times \{ (\vec{n} - \vec{\beta}) \times \dot{\vec{\beta}} \}|^2}{(1 - \vec{n} \cdot \vec{\beta})^5} e^{i\omega(t + \frac{R(t)}{c})} dt \right] \right|^2 . \end{aligned} \quad (22)$$

Finally, the spatial and spectral energy distribution of SR per unit frequency and solid angle is given by

$$\frac{d^2W}{d\Omega d\omega} = \frac{e^2}{16\pi^3 \varepsilon_0 c} \gamma^2 \frac{\omega^2}{\omega_c^2} K_{2/3}^2(\xi) F(\xi, \theta) , \quad (23)$$

$$\xi \equiv \frac{1}{2} \frac{\omega}{\omega_c} (1 + \gamma^2 \theta^2)^{3/2} , \quad (24)$$

$$F(\xi, \theta) \equiv (1 + \gamma^2 \theta^2)^2 \left[ 1 + \frac{\gamma^2 \theta^2}{1 + \gamma^2 \theta^2} \frac{K_{1/3}^2(\xi)}{K_{2/3}^2(\xi)} \right] . \quad (25)$$

where  $K_i(\zeta)$  is the modified Bessel function. The former term in Eq. (25) is the  $\sigma$  mode, where the electrical field is orthogonal to the deflecting field ( $B$ ). The latter term is the  $\pi$  mode, where the electrical field is in the plane of the deflecting field and the line of observation. For high energies, the  $\sigma$  mode is dominant. Here,

$$\omega_c \equiv \frac{3}{2} \frac{c\gamma^3}{\rho} \quad (26)$$

is the critical frequency. This is the frequency that halves the total energy.

The photon number (photon flux) with a beam current  $I_e$  per unit solid angle and frequency is given by

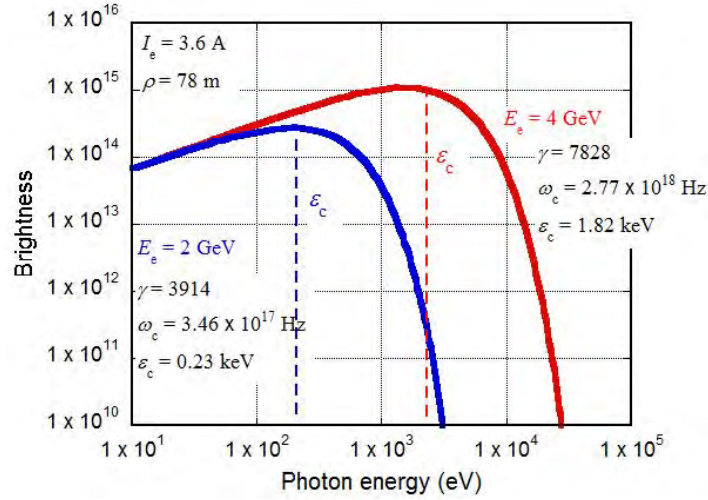
$$\frac{d^2\dot{N}_{\text{ph},I_e}}{d\Omega(d\omega/\omega)} = \frac{d^2P_{I_e}}{d\Omega d\omega} \frac{1}{\hbar} = \frac{d^2W}{d\Omega d\omega} \frac{I_e}{e} \frac{1}{\hbar} , \quad (27)$$

where  $\hbar \equiv h/2\pi$ . The spatial and spectral photon flux distribution per unit solid angle  $d\theta d\psi$  and the band width  $d\omega/\omega$  that is, the brightness, is given by

$$\frac{d^3 \dot{N}_{\text{ph,le}}}{d\theta d\psi (d\omega / \omega)} = C_\omega E^2 I_e \frac{\omega^2}{\omega_c^2} K_{2/3}^2(\xi) F(\xi, \theta) , \quad (28)$$

$$\begin{aligned} C_\omega &\equiv \frac{3\alpha}{4\pi^2 e (m_e c^2)^2} = 1.3255 \times 10^{22} \frac{\text{photons}}{\text{s rad}^2 \text{ GeV}^2 \text{ A}} , \\ &= 1.3255 \times 10^{13} \frac{\text{photons}}{\text{s mrad}^2 \text{ GeV}^2 \text{ A } 0.1\% \text{ bandwidth}} \end{aligned} \quad (29)$$

where  $\alpha$  is the fine-structure constant. The brightness is a key parameter in considering the performance of light (photon) sources. Figure 5 shows the typical brightness as a function of photon energies and the critical energies.



**Fig. 5:** Examples of brightness.  $\varepsilon_c$  is also indicated

The critical energy of photons is given by:

$$\varepsilon_c = \frac{3}{2} \frac{\hbar c \gamma^3}{\rho} \equiv \hbar \omega_c , \quad (30)$$

$$\varepsilon_c [\text{eV}] = 2218 \times 10^3 \times \frac{E_e [\text{GeV}]^3}{\rho [\text{m}]} = 0.665 \times 10^3 \times E_e [\text{GeV}]^2 B [\text{T}] . \quad (31)$$

Using the critical energy, the mean photon energy is expressed as

$$\langle \varepsilon \rangle = \frac{8}{15\sqrt{3}} \varepsilon_c , \quad (32)$$

and the total photon flux is

$$\dot{N}_{\text{ph}} = \frac{15\sqrt{3}}{8} \frac{P_{\text{tot}}}{\varepsilon_c} . \quad (33)$$

The critical energy is a key parameter characterizing SR.

The total photon number is given by integrating Eq. (28) with respect to  $\theta$ ,  $\psi$  (that is, the whole ring), and  $\omega$ ,

$$\dot{N}_{\text{ph,le}} = \frac{15\sqrt{3}}{4} C_{\psi} I_e E_e = 8.08 \times 10^{20} I_e [\text{A}] E_e [\text{GeV}] [\text{photons s}^{-1}], \quad (34)$$

$$C_{\psi} \equiv \frac{4\alpha}{9em_e c^2} = 3.9614 \times 10^{19} \frac{\text{photons}}{\text{s rad GeV A}}. \quad (35)$$

The average photon number per unit length along the ring is obtained by

$$\langle \dot{N}_{\text{ph,le,line}} \rangle = \dot{N}_{\text{ph,le}} / C. \quad (36)$$

The photon number for an angle of  $\phi$  is

$$\dot{N}_{\text{ph,le}}(\phi) = \dot{N}_{\text{ph,le}} \frac{\phi}{2\pi}. \quad (37)$$

### 3 Effects of synchrotron radiation

The three main effects of SR on an accelerator (especially on the vacuum system) (see Fig. 6) are given below.

- i) Heat load: when SR hits a surface, it transfers the energy to the surface. The SR heats up the beam pipe, and sometimes damages it by excess heating and thermal stress.
- ii) Gas load: when SR hits a surface, it desorbs gas molecules on the surface. The gas desorption increases vacuum pressure. The pressure rise reduces the beam lifetime and increases background noise in the detector.
- iii) Emission of electrons: when SR hits a surface, the surface emits electrons (photoelectrons). The emitted photoelectrons enhance the formation of the electron cloud, which leads to electron cloud instabilities (for positive beams).

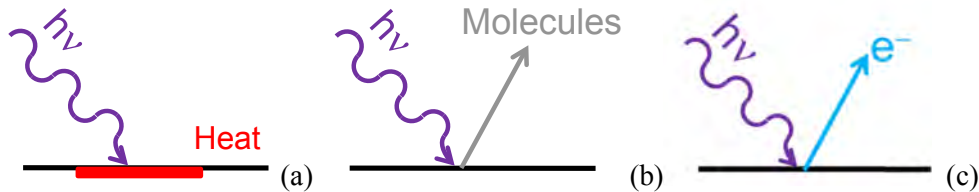


Fig. 6: Three major effects of SR. (a) heat load; (b) gas load; (c) electron emission

#### 3.1 Heat load

##### 3.1.1 General

When SR hits the inner wall of a beam pipe, it transfers energy to the surface, resulting in heating of the surface. As described above, the SR ‘beams’ (concentrates) in the front direction. If the irradiated area is not properly cooled, the surface is easily excessively heated and damaged. For example, if  $E_e = 4$  GeV,  $I_e = 3.6$  A,  $\rho = 74$  m, and  $C = 2000$  m (parameters for SuperKEKB [2]), from Eq. (16),

$$\langle P_{\text{le,line}} \rangle = 88.5 \times 10^3 \times 4^4 \times 2.6 / 74 / 2000 = 550 \text{ W m}^{-1}. \quad (38)$$

The power density is sufficiently high to melt metals if the irradiation area is not cooled. The heat load is actually distributed along the ring. The sources (emitting points) are in the bending magnets. For a uniform beam pipe, the heat load is maximum in the bending magnet, and decreases gradually on the

downstream side, as shown in Fig. 7 for the SuperKEKB [2]. In this case the average power line density is  $\sim 0.6 \text{ kW m}^{-1}$ , but the peak power line density is  $2.3 \text{ kW m}^{-1}$ . When considering heating by SR the maximum power density is more important than the average.

Dependencies of the SR power line and area densities on the distance from the emitting point to the irradiated point  $R$  and the incident angle  $\theta_i$  are shown in Fig. 8. The line power density in the magnet is proportional to  $1/R$ , and is almost constant, because  $R$  and  $\theta_i$  are constant. The SR power line density outside the magnet, on the other hand, is proportional to  $1/R \times \theta_i \propto 1/R^2$ , because the incident angle  $\theta_i$  is also almost proportional to  $1/R$ . For the power area density, on the other hand, the vertical spread angle of  $2/\gamma$  should be taken into account. Power area density is key in evaluating the thermal stress.

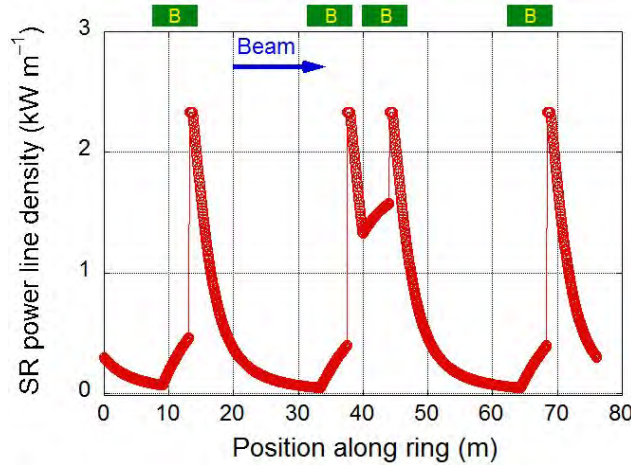


Fig. 7: Example of the distribution of SR power line density (SuperKEKB [2])

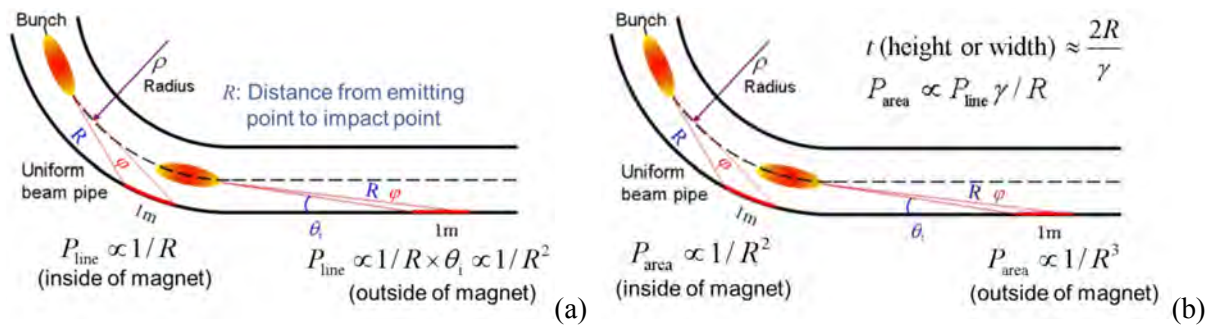


Fig. 8: Typical dependencies of SR power (a) line and (b) area densities upon  $R$  and  $\theta_i$

### 3.1.2 Countermeasures

The basic principle of the heat load countermeasure is to receive SR at specific places (photon stops) with a proper cooling system at large  $R$  and small  $\theta_i$ . In general, this can be achieved in two ways: using distributed photon stops (photon masks) and localized photon stops.

#### 3.1.2.1 Distributed photon stops

In this case, small photon stops are placed upstream of bellows chambers or flanges to make short SR shadows, as shown in Figs. 9 and 10 [3–6]. The photon stops have a relatively low height ( $H$ ) of  $\sim 10 \text{ mm}$ . The typical shadow length, given by  $H/\tan \theta_i$ , is  $200\text{--}400 \text{ mm}$ . Most of the heat load is distributed along the ring, and the heat load at the photon stops is relatively small ( $\theta_i$  is also small). The structure of the beam pipes is simple. Distributed photon stops must be used if the power density at the localized photon stop (described below) is too high.

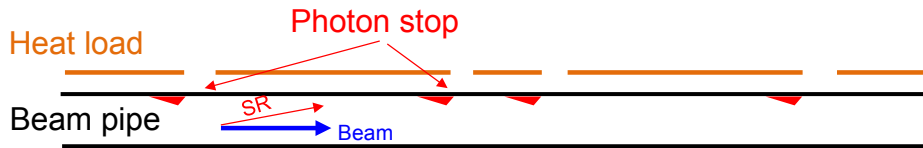


Fig. 9: Schematic configuration of distributed photon stops

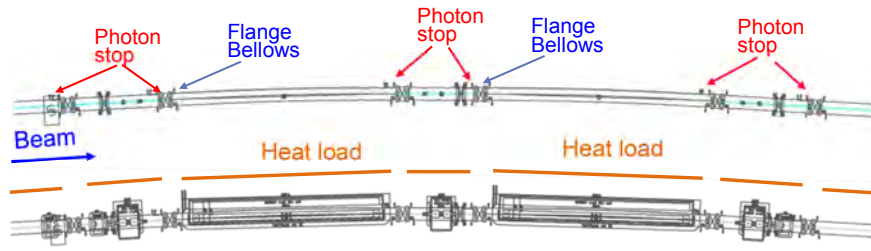


Fig. 10: Example of the distributed photon stop scheme (SuperKEKB [2])

3.1.2.2 Localized photon stops

In this case, large photon stops that result in long SR shadows are placed locally, as shown in Figs. 11 and 12 [7–10]. The typical mask height is 100–200 mm, and the typical shadow length is a few metres of SR, i.e. the photon stops receive the SR power corresponding to the power of a few metres. This means that most of the heat load concentrates in the photon stops, usually at a much higher power density than in the case of the distributed photon stops. One of the criteria used to decide on a particular photon stop scheme, distributed or localized, is the SR power density. Sometimes, in light sources, the photon stop is called a ‘crotch absorber’. The structure of the beam pipe is likely to be complicated here. Effective pumping is realized by putting pumps at the same places as the photon stops (see below).

Various types of photon stops (masks) have been designed in various accelerators [7–12]. In designing the photon stops, simulations using finite element methods (FEM) are very useful in evaluating the temperature and stress distribution. Key design points are:

- i) to obtain a slanting irradiated surface (i.e. make  $\theta_i$  as small as possible) to reduce the power density in the horizontal direction as well as in the vertical direction;
- ii) to prepare sufficient and effective cooling paths;
- iii) to use materials with high thermal conductivity and high thermal strength.

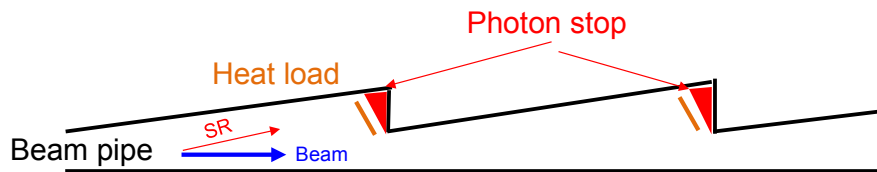


Fig. 11: Schematic configuration of the localized photon stops

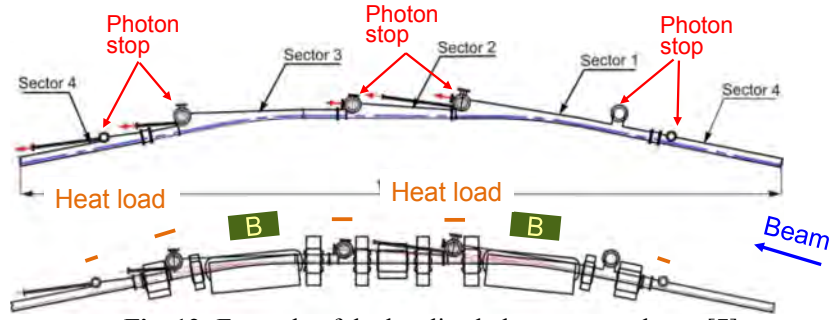


Fig. 12: Example of the localized photon stop scheme [7]

### 3.1.2.3 Other countermeasures

- i) Use materials with a high thermal conductivity and a high thermal strength, e.g. copper, copper-chromium alloys, and GlidCop [13].
- ii) Use beam pipes with an antechamber, where the SR hits a point far from the emission point. The power area density decreases as  $R$  increases.

## 3.2 Gas load

### 3.2.1 General

When SR hit the inner surface of the beam pipe, it desorbs gas molecules adsorbed on the surface. This is called photon stimulated gas desorption (PSD) [9, 14–16]. Residual gases in beam pipes during the operation are the result of PSD. For example, if  $E_e = 4$  GeV and  $\rho = 74$  m, the critical energy  $\epsilon_c$  is 1.9 keV. Because the temperature corresponding to 1 eV is approximately 12 000°C, a 1 keV photon is enough to destroy the chemical bonding between adsorbed molecules and surface molecules (a few electron volts). PSD is also much more effective than baking. Considerable gas desorption is expected with PSD compared to with thermal gas desorption for large photon numbers.

The main effects of the gas load are given below.

- i) Energy loss due to scattering with the residual gases. This leads to particle loss and then a shorter beam lifetime. The loss of particles increases the background noise of the detectors, and can also cause radio activation.
- ii) Generation of ions by ionization of residual gases. Ion instabilities can be excited in beams with negative charges, such as electrons. Furthermore, ions generated by ionization of residual gases hit the beam pipe wall and desorb gases from the surface (this is called ion stimulated gas desorption (ISD)). This sometimes leads to a pressure instability.

Here, we briefly discuss the beam lifetime  $\tau$ , which is defined as

$$I_e = I_{e0} e^{-\frac{t}{\tau}} . \quad (39)$$

where  $I_e$  and  $I_{e0}$  are the beam current and the initial beam current, respectively. For sufficiently large apertures,  $\tau$  can be usually expressed by

$$\frac{1}{\tau} = \sum_i \{ \sigma_B(Z_i) + Z_i \sigma_M + \sigma_R(Z_i) \} p_i . \quad (40)$$

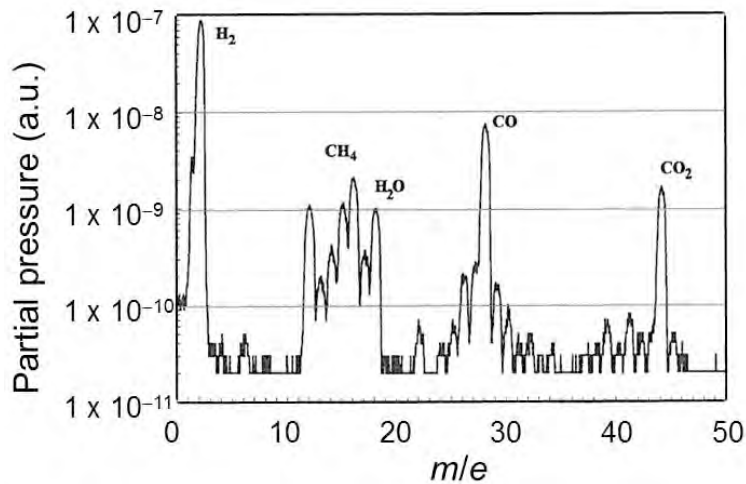
Here,  $\sigma_B$ ,  $\sigma_M$  and  $\sigma_R$  are the cross-sections of three major interaction processes with gas molecules, i.e. a Bremsstrahlung with nuclei, Moller scattering with electrons outside nuclei, and Rutherford scattering with nuclei, respectively. As indicated by this equation, lifetime is inversely proportional to the pressure.



**3.2.2 Photon stimulated gas desorption**

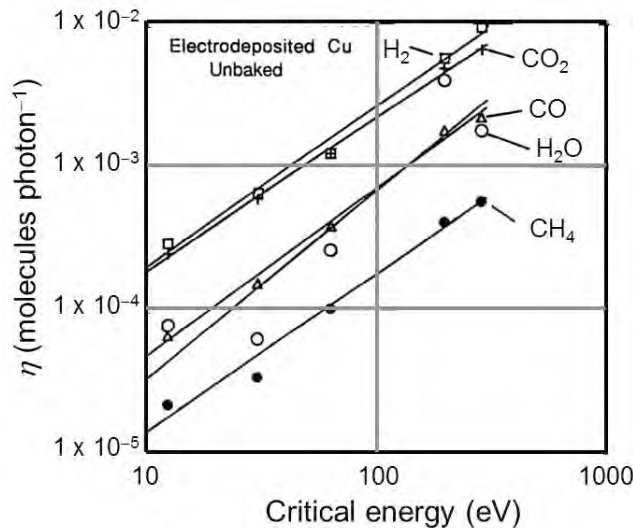
The SR irradiation on the inner surface results in the emission of photoelectrons. The photon energy is high enough to cause electron emission (photoelectrons) from material surfaces where the work functions are a few electron volts. The quantum efficiency  $\eta_e$ , the yield of photoelectrons per photon, is  $\sim 0.1$  electrons photon<sup>-1</sup>. We will discuss electron emission from the surface in the section below. These electrons hitting the surface desorb molecules from the surface, because they also have sufficiently high energies. This is called electron stimulated gas desorption (ESD). It is believed that most PSD comes from ESD.

The number of gas molecules emitted by one photon is the PSD rate. It is usually expressed by  $\eta$  [molecules photon<sup>-1</sup>]. After the usual baking, the major desorbed gases are hydrogen (H<sub>2</sub>), carbon monoxide (CO), and carbon dioxide (CO<sub>2</sub>) (Fig. 13) [9]. Water (H<sub>2</sub>O) is the main gas for a non-baked system.



**Fig. 13:** Typical mass spectrum of the residual gases due to PSD [9]

The  $\eta$  increases with the incident photon energy (critical energy) because the deposit energy increases, as shown in Fig. 14 [14].



**Fig. 14:** Dependence of  $\eta$  on the critical energy of SR [14]

The  $\eta$  increases as the incident angle decreases [15]. A rough surface, therefore, can result in a decrease in  $\eta$ . Note that if the surface is smooth and the incident angle is small, the reflection of SR should be taken into account.

Another important property of PSD is aging (scrubbing). The  $\eta$  decreases with the integrated photon number (photon dose,  $D$ ), as shown in Fig. 15. This phenomenon is called beam aging or scrubbing. Typical values of  $\eta$  before SR irradiation are  $10^{-3}$  to  $10^{-2}$  molecules photon $^{-1}$ . The  $\eta$  decreases to  $10^{-6}$  to  $10^{-7}$  after sufficient aging. Usually,  $\eta$  varies according to the function

$$\eta = D^{-1-0.6} \quad (41)$$

Practically, when designing the vacuum system, an  $\eta$  of  $1 \times 10^{-5}$  to  $1 \times 10^{-6}$  molecules photon $^{-1}$  is assumed considering the aging effect.  $\eta$  also strongly depends on the surface condition, the degree of contamination, the thickness of the oxide layer, the roughness of the surface, etc. [16].

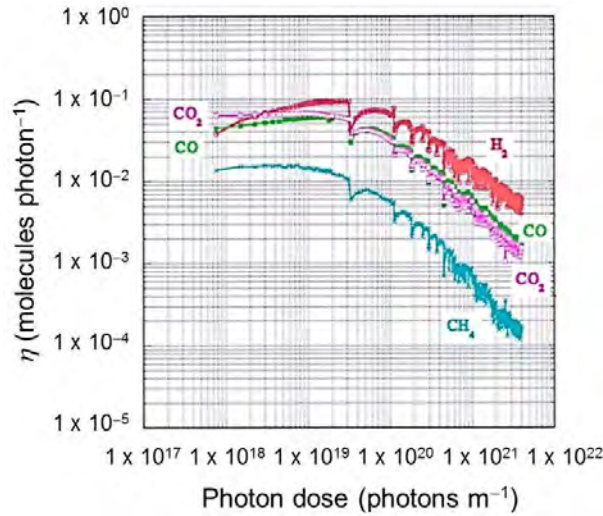


Fig. 15: Typical beam scrubbing (aging) of  $\eta$

Here we estimate the gas load. For example, if  $E_e = 4$  GeV,  $I_e = 2.6$  A and  $C = 3000$  m, from Eq. (36),

$$\langle \dot{N}_{\text{ph,le,line}} \rangle = 8.08 \times 10^{20} \times 4 \times 2.6 / 3000 = 2.8 \times 10^{18} \text{ photons s}^{-1} \text{ m}^{-1} \quad (42)$$

If  $\eta = 1 \times 10^{-6}$  molecules photon $^{-1}$ ,

$$\langle \dot{N}_{\text{ph,le,line}} \rangle = 2.8 \times 10^{18} \times 1 \times 10^{-6} = 2.8 \times 10^{12} \text{ molecules s}^{-1} \text{ m}^{-1} \quad (43)$$

The average line gas desorption rate (gas load) along the ring  $Q_{\text{av,line}}$  (at  $T = 25^\circ\text{C} = 298$  K) is

$$\begin{aligned} Q_{\text{av,line}} &= \langle \dot{N}_{\text{mol,le,line}} \rangle \times k_B T = 2.8 \times 10^{12} \times 1.38 \times 10^{-23} \times 298 \\ &= 1.1 \times 10^{-8} \text{ Pa m}^3 \text{ s}^{-1} \text{ m}^{-1} \end{aligned} \quad (44)$$

Here we use the ideal gas law equation

$$pV = N_{\text{mol}} k_B T \quad (45)$$

Here,  $V$  and  $k_B$  are the volume and Boltzmann constant, respectively. The expression is convenient in designing a vacuum system, because the pumping speed is usually expressed in cubic metres per second ( $\text{m}^3 \text{s}^{-1}$ ). If the average linear pumping speed is  $S_{\text{av,line}}$  ( $\text{m}^3 \text{s}^{-1} \text{ m}^{-1}$ ) in the ring, the obtained average pressure  $p_{\text{av}}$  (Pa) is

$$p_{av} = \frac{Q_{av,line}}{S_{av,line}} \text{ Pa} \quad (46)$$

The distribution of gas load is almost the same as that of photons. Basically, the gas load is high downstream of the bending magnets, as in the case of heat load (Fig. 16). In this case, the average photon line density is  $\sim 5.5 \times 10^{18} \text{ photons s}^{-1} \text{ m}^{-1}$ . The maximum photon line density is  $\sim 3 \times 10^{19} \text{ photons s}^{-1} \text{ m}^{-1}$ .

Note here that the distribution of gas load is not exactly the same as for photons, because the PSD depends on the beam dose and  $\theta^i$ . Actually, the difference between the maximum and the minimum values decreases with time.

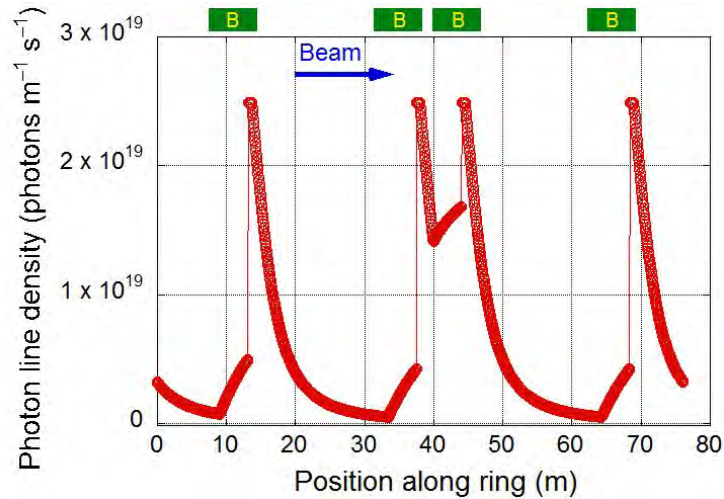


Fig. 16: Distribution of photon line density along the ring (SuperKEKB) [2]

### 3.2.3 Countermeasures

The basic principle of countermeasures is to prepare proper pumps at places where photons irradiate the beam pipe. There are two ways to treat the gas load: distributed pumping and localized pumping.

#### 3.2.3.1 Distributed pumping

Distributed pumping works well with the distributed photon stops described above, as shown in Fig. 17 [3–6]. Beam pipes are usually very narrow and long; hence their conductance is small, typically  $< 0.1 \text{ m}^3 \text{ s}^{-1} \text{ m}^{-1}$ . In the distributed pumping scheme, pumps are located along the beam pipe, just alongside the beam channel, and the beam pipe is then effectively evacuated. Uniform pumping speed along the ring is realized. Distributed pumping is effective where the gas loads are distributed evenly along the beam pipe. In a distributed pumping scheme, the structure of the beam pipes is relatively simple.

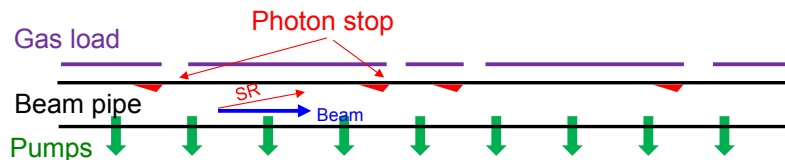
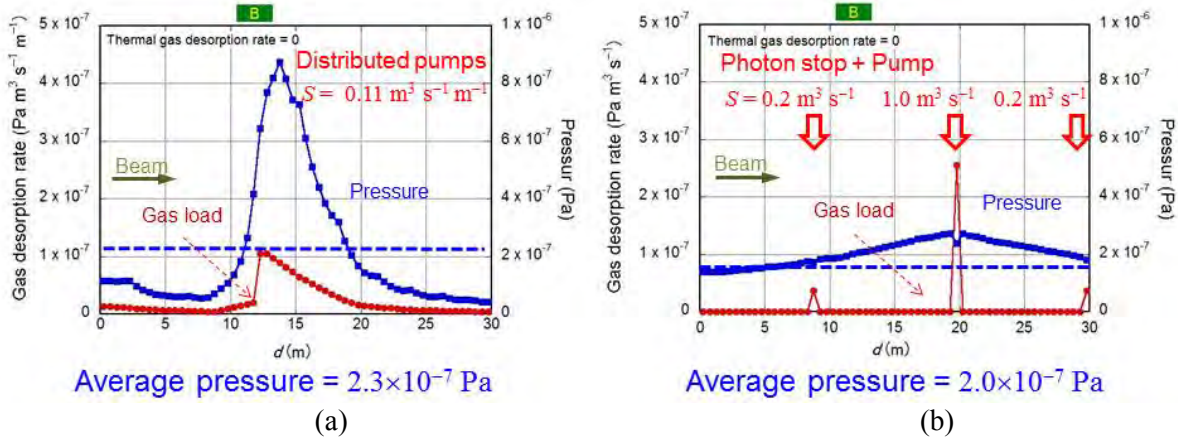


Fig. 17: Distributed pumping scheme

A commonly used pump in the distributed pumping scheme is the distributed sputter-ion pump (DIP). This is a sputter-ion pump operating in the bending magnets’ magnetic field. The DIP was popular until ca. 1990. Another commonly used pump is the non-evaporable getter (NEG) pump. The

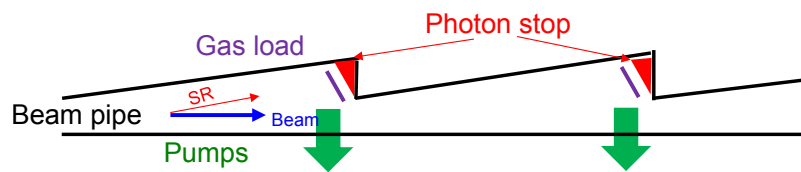
NEG strips are placed along the beam pipe. A thin film of the NEG ingredient coating inside the beam pipe has recently been used in various facilities. In the case of the previous example, if we use a distributed pumping system with an average pumping speed of  $\sim 0.11 \text{ m}^3 \text{ s}^{-1} \text{ m}^{-1}$ , an average pressure of  $2.3 \times 10^{-7} \text{ Pa}$  is obtained as shown in Fig. 18, assuming  $\eta = 1 \times 10^{-6} \text{ molecules photon}^{-1}$ . A similar pressure profile to that of the photon line density is obtained.



**Fig. 18:** Comparison of distributed and localized pumping scheme, where the photon line density in Fig. 15 and  $\eta = 1 \times 10^{-6} \text{ molecules photon}^{-1}$  are assumed. (a) Distributed pumping; (b) localized Pumping.

### 3.2.3.2 Localized pumping

The localized pumping scheme works well with the localized photon stops, as shown in Fig. 19 [7–12]. In this case, the pumps are placed near the localized photon stops, usually downstream of the bending magnets. The localized photon scheme is related to a localized gas load. The pumps are concentrated at the locations where the gas loads are large. This is a reasonable approach to achieve ultra-high vacuum, and is adopted in many recent photon sources. The widely used pumps are turbo-molecular pumps, sputter ion pumps, Ti-sublimation pumps, and NEG cartridges, etc. Here the structure of the beam pipes is more than in the case of the distributed pumping scheme. As indicated in Fig. 18, if localized pumps are used and the thermal gas desorption is ignored, a lower average pressure is obtained compared to that obtained with distributed pumping, even with smaller total pumping speeds. Note, however, that low thermal gas desorption is essential. Otherwise the pressure between the adjacent pumps will be high owing to the limited conductance of the beam pipe.



**Fig. 19:** Localized pumping scheme

### 3.2.3.3 Other countermeasures

An important measure is to avoid contamination during the beam pipe's manufacturing and assembly process. A clean assembly environment should be ensured. Surface treatments, such as chemical cleaning, argon glow discharge, and pre-baking, are effective in reducing thermal gas desorption. Using an antechamber scheme is also effective, because photons hit photon stops in the antechamber, which is separated from the beam channel. Desorbed gas is confined within the antechamber. The antechamber structure is usually adopted for a localized photon stop scheme. The antechamber scheme also provides a relatively smooth beam channel, which contributes to a lower beam impedance.

### 3.3 Electron emission

#### 3.3.1 General

When SR hits the surface, photoelectrons are emitted from the surface, as described above. The yield of photoelectrons from one incident photon, the quantum efficiency  $\eta_e$ , is  $\sim 0.1$  electrons  $\text{photon}^{-1}$ . If the beams are positively charged (i.e. positrons or protons), they attract the electrons. The electrons that are accelerated by the next bunch's electric fields hit the surface and emit electrons, which are called secondary electrons. If the secondary electron yield (SEY), the number of electrons emitted by one incident electron, is larger than 1, the enhancement of electrons (multipactoring) can occur. This positive feedback leads to the accumulation of electrons around the beams. This group of electrons is called the electron cloud [17].

#### 3.3.2 Secondary electron yield

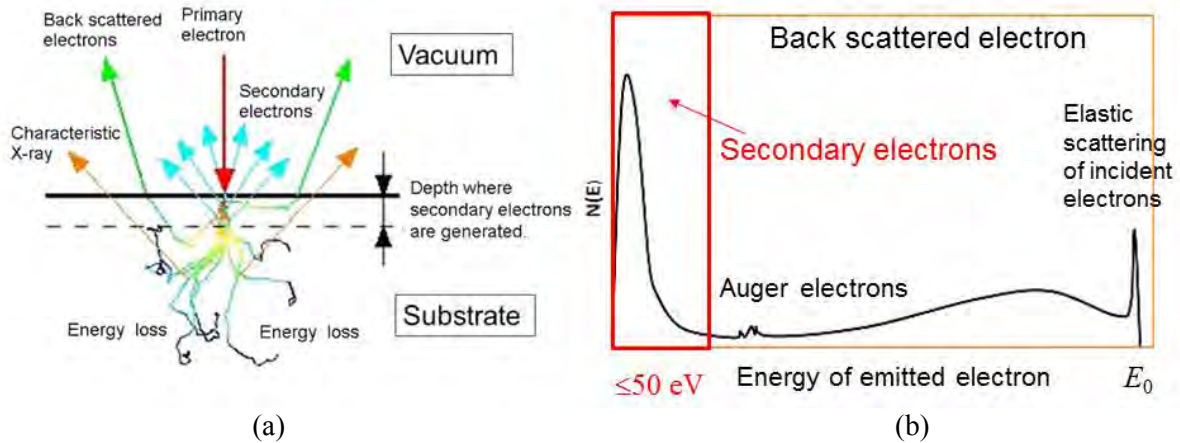
Figure 20 shows the creation process and the typical energy spectrum of secondary electrons [18]. The secondary electrons are emitted from the surface following the cosine law, i.e., uniformly. The energy of secondary electrons is less than 50 eV. The SEY depends on the incident angle of electrons as in the case of the photoelectron yield's dependence on the incident angle of the photons. The SEY  $\delta$  increases with the incident angle  $\theta$ , which is the angle between the direction of the incident electron and the normal to the surface [19]. The dependence can be explained as follows: for shallow incidence (large  $\theta$ ), electrons generated along the path of the incident electron can easily escape to vacuum (see Fig. 20). The following two formulae are commonly used in simulations. For  $\theta \sim 0^\circ$ ,

$$\delta \approx \frac{\delta_0}{\cos \theta} \quad (47)$$

Here  $\delta_0$  is the SEY at normal incidence of the primary electron. For  $\theta \rightarrow 90^\circ$ ,

$$\delta \approx \delta_0 e^{\alpha X_m (1 - \cos \theta)} \quad (48)$$

Here,  $X_m$  is the depth at which secondary electrons are generated at normal incidence and  $\alpha$  is the absorption rate. Usually,  $\alpha X_m \sim 0.4$  is used.



**Fig. 20:** (a) Process of generating secondary electrons; (b) typical energy spectrum of secondary electrons [18]

Figure 21 shows the dependence of SEY on the energy of the incident electron (primary electron). The SEY has a maximum at an incident electron energy of 200–400 eV and decreases gradually with increasing energy. Two formulae for  $\delta$  are usually used for the simulation [20–24]. One of these is

$$\delta(E_r) \approx \delta_{\max} 1.11 E_r^{-0.35} \left( 1 - e^{-2.3 E_r^{1.35}} \right), \quad (49)$$

where  $\delta_{\max}$  is the maximum yield for perpendicular incident,  $E_r \equiv E_p/E_{\text{pm}}$ , where  $E_p$  is the energy of the incident electron, and  $E_{\text{pm}}$  is the primary electron energy at which the yield is at a maximum.

The other expression is:

$$\delta(E_r) \approx \delta_{\max} \frac{sE_r}{s-1+E_r^s} \tag{50}$$

where  $s \sim 1.4$ .

A decrease in SEY with electron dose (integrated electrons per unit area) is observed, as in the case for the PSD rate ( $\eta$ ) [25]. The decrease is also called as the aging or conditioning. The SEY also strongly depends on the surface conditions and materials.

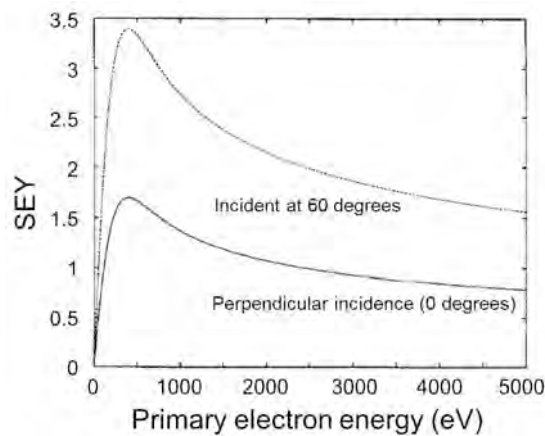


Fig. 21: Dependence of SEY on the energy of the incident electrons [20]

### 3.3.3 Electron cloud instability

This section briefly introduces the electron cloud effect (ECE) [17, 26–28]. If the electron density around the beam exceeds a threshold value, the electron cloud causes a beam instability, called the electron cloud instability. First, the displacement of the top bunch affects the following bunches via the electron cloud. Then, the perturbation of the electron cloud (a kind of wake field) affects the following bunches or the electrons in the same bunch. The former is called a coupled bunch instability, and the latter is called a head–tail instability, as explained in Fig. 22. The electron cloud instability leads to the blow-up of beam size, which increases the emittance of the beam, and decreases in the luminosity in the colliders. The electron cloud instability is a critical issue in recent high-intensity proton and positron storage rings. Many theoretical and experimental studies have been conducted about the formation of the electron cloud, the simulation of beam instability, the measurement of beam emittance, and countermeasures.

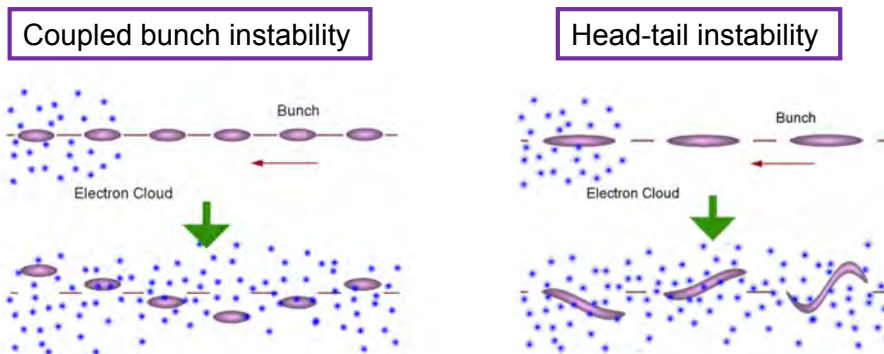


Fig. 22: Two types of electron cloud instabilities

Here we roughly estimate the number of generated photoelectrons. For the SuperKEKB positron ring, where  $E_e = 4 \text{ GeV}$ ,  $I_e = 3.6 \text{ A}$  and  $C = 3000 \text{ m}$ , the average photon linear density along the ring from Eq. (36) is

$$\langle \dot{N}_{\text{ph,le,line}} \rangle = 8.08 \times 10^{20} \times 4 \times 3.6 / 3000 = 3.9 \times 10^{18} \text{ photons s}^{-1} \text{ m}^{-1} . \quad (51)$$

If the quantum efficiency ( $\eta_e$ ) is 0.1, the number of emitted photoelectrons is

$$\langle \dot{N}_{\text{ele,le,line}} \rangle = \eta_e \times \langle \dot{N}_{\text{ph,le,line}} \rangle = 3.9 \times 10^{17} \text{ electrons s}^{-1} \text{ m}^{-1} . \quad (52)$$

In the case of the SuperKEKB, the threshold of the electron density,  $\rho_{e,\text{th}}$ , to result in the head–tail instability is  $2 \times 10^{11} \text{ electrons m}^{-3}$ . It has been found that  $\rho_{e,\text{th}}$  is easily achieved if no countermeasures are adopted.

### 3.3.4 Countermeasures

The basic principles of the countermeasures are to suppress electron emissions and remove electrons around the beams. Various countermeasures have been proposed and studied, and some have been applied in practice.

#### 3.3.4.1 Beam pipe with antechambers

The SR irradiates the side wall of the antechamber, far from the beam (Fig. 23) [3, 29]. Hence, the photoelectrons do not easily interface with the beam. Note that some photons may hit the outside of the antechamber at points far from the photon emitting point owing to the vertical spread of  $\sim 2/\gamma$ . Furthermore, multipactoring of secondary electrons is more significant for a large beam current. The antechamber structure is, therefore, effective at low beam currents.

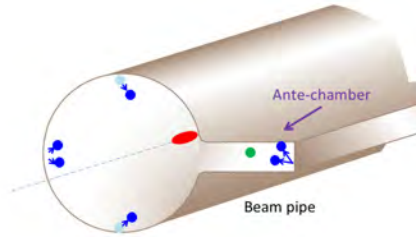


Fig. 23: Beam pipe antechamber

#### 3.3.4.2 Inner coating with a low SEY

At high beam currents, the main mechanism forming the electron cloud is the multipactoring of secondary electrons (Fig. 24) [30–34]. In this situation, some inner coatings with a low SEY are effective in suppressing the electron cloud formation. Possible candidates are TiN, graphite, and NEG ingredients.

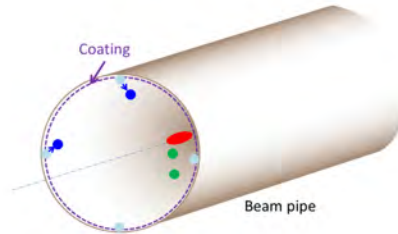
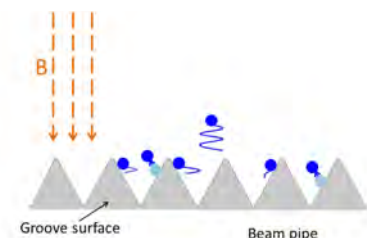


Fig. 24: Beam pipe inner coating

### 3.3.4.3 Grooved surface

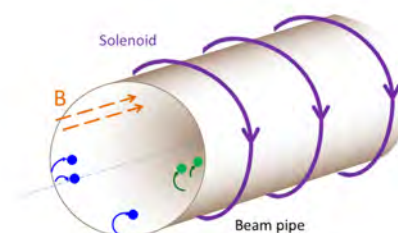
A surface with a grooved structure is found to have a low SEY (Fig. 25) [35–37]. The SEY is structurally reduced, especially in a magnetic field. A coating of material with a low SEY on the groove enhances the reduction of SEY. Beam impedance may be a concern.



**Fig. 25:** Grooved surface

### 3.3.4.4 Solenoid field

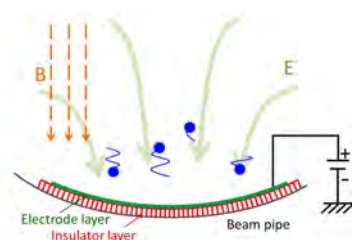
With a magnetic field along the beam pipe, the electrons emitted from the surface return to the surface due to Larmor motion (Fig. 26) [38, 39]. Emitted photoelectrons or secondary electrons have an energy of several tens of electron volts. Hence, a magnetic field of several tens of gauss is sufficient. Drastic effects were observed in PEP-II and KEKB.



**Fig. 26:** Solenoid field

### 3.3.4.5 Clearing electrode

An electrode in a beam pipe with a high positive potential attracts the electrons around the beam orbit (Fig. 27) [23, 37, 40, 41]. A drastic effect in reducing electron density is expected and has been confirmed in experiments. The effect was also demonstrated at DAFNE. Beam impedance is also a concern with this countermeasure.



**Fig. 27:** Clearing electrode

### 3.3.4.6 Other countermeasures

Alternative effective countermeasures are making a wide bunch gap and changing the bunch filling patterns. Experience with KEKB, however, suggests that these techniques are not such effective countermeasures.



## References

- [1] A.W. Chao and M. Tigner, *Handbook of Accelerator Physics and Engineering* (World Scientific, Singapore, 1999). <http://dx.doi.org/10.1142/3818>
- [2] N. Ohuchi, Proc. IPAC2014, Dresden, 2014, p. 1877.
- [3] Y. Suetsugu *et al.*, *Vacuum* **84** (2010) 694. <http://dx.doi.org/10.1016/j.vacuum.2009.06.027>
- [4] J. Heim *et al.*, Proc. PAC1995, Dallas, 1995, p. 527.
- [5] C. Benvenuti *et al.*, *Vacuum* **44** (1993) 507. [http://dx.doi.org/10.1016/0042-207X\(93\)90083-M](http://dx.doi.org/10.1016/0042-207X(93)90083-M)
- [6] LEP Vacuum Group, *Vacuum* **41** (1990) 1882. [http://dx.doi.org/10.1016/0042-207X\(90\)94121-6](http://dx.doi.org/10.1016/0042-207X(90)94121-6)
- [7] V. Avagyan, H. Gagiyan, S. Nagdalyan and A. Petrosyan, Proc. EPAC2002, Paris, 2002 p. 2532.
- [8] D. Hunt, K. Kennedy and T. Stevens, Proc. PAC1995, Dallas, 1995, p. 2067.
- [9] G.Y. Hsiung *et al.*, *J. Vac. Sci. Technol.* **A12** (1994) 1639. <http://dx.doi.org/10.1116/1.579029>
- [10] C. Hetzel *et al.*, Proc. IPAC2012, New Orleans, 2012, p. 2558.
- [11] K. Watanabe *et al.*, Proc. PAC1993, Washington DC, 1993, p. 3845.
- [12] Y.T. Cheng *et al.*, Proc. IPAC2011, San Sebastian, 2011, p. 1692.
- [13] SCM Metal Products, Inc. Available at:  
[http://www.aps.anl.gov/APS\\_Engineering\\_Support\\_Division/Mechanical\\_Operations\\_and\\_Maintenance/Miscellaneous/tech\\_info/Glidcop/SCM\\_Glidcop\\_product\\_info.pdf](http://www.aps.anl.gov/APS_Engineering_Support_Division/Mechanical_Operations_and_Maintenance/Miscellaneous/tech_info/Glidcop/SCM_Glidcop_product_info.pdf).
- [14] J. Gomez-Goni *et al.*, CERN VT Note 93-1 (1993).
- [15] B.A. Trickett *et al.*, *J. Vac. Sci. Technol.* **A10** (1992) 217. <http://dx.doi.org/10.1116/1.578139>
- [16] A.G. Mathewson, *Vacuum* **44** (1993) 479. [http://dx.doi.org/10.1016/0042-207X\(93\)90077-N](http://dx.doi.org/10.1016/0042-207X(93)90077-N)
- [17] For example, reports presented at E-CLOUD'02, CERN, 15–18 April 2002; E-CLOUD'04, Napa, 19–23 April 2004; E-CLOUD'07, Daegu, 9–12 April 2007; ECL2 Workshop, CERN, 28 February–2 March 2007; E-CLOUD'10, Cornell University, 8–12 October 2010.
- [18] A.J. Hatch, *Nucl. Instrum. Methods PR-A* **41** (1966) 261. [http://dx.doi.org/10.1016/0029-554X\(66\)90010-3](http://dx.doi.org/10.1016/0029-554X(66)90010-3)
- [19] R.E. Kirby and F.K. King, *Nucl. Instrum. Methods PR-A* **469** (2001) 1.  
[http://dx.doi.org/10.1016/S0168-9002\(01\)00704-5](http://dx.doi.org/10.1016/S0168-9002(01)00704-5)
- [20] F. Zimmermann, SLAC-PUB-7664 (1997).
- [21] V. Baglin *et al.*, Proc. EPAC1998, Stockholm, 1998, p. 359.
- [22] L.F. Wang *et al.*, *Phys. Rev. Special Topics – Acc. Beams* **5** (2002) 124402.  
<http://dx.doi.org/10.1103/PhysRevSTAB.5.124402>
- [23] L.F. Wang, D. Raparia, J. Wei and S.Y. Zhang, *Phys. Rev. Special Topics – Acc. Beams* **7** (2004) 034401-1. <http://dx.doi.org/10.1103/PhysRevSTAB.7.034401>
- [24] U. Iriso-Ariz *et al.*, Proc. PAC2003, Portland, 2003, p. 797.
- [25] K. Shibata *et al.*, Proc. EPAC2008, Genoa, 2008, p. 1700.
- [26] K. Ohmi, *Phys. Rev. Lett.* **75** (1995) 1526. <http://dx.doi.org/10.1103/PhysRevLett.75.1526>
- [27] K. Ohmi and F. Zimmermann, *Phys. Rev. Lett.* **85** (2000) 3821.  
<http://dx.doi.org/10.1103/PhysRevLett.85.3821>
- [28] Y. Susaki and K. Ohmi, Proc. IPAC2010, Kyoto, 2010, p.1545.
- [29] M.T.F. Pivi *et al.*, Proc. IPAC2011, San Sebastian, 2011, p. 1063.
- [30] K. Kennedy *et al.*, Proc. PAC1997, Vancouver, 1997, p. 3568.
- [31] B. Henrist *et al.*, *App. Surface Sci.* **172** (2001) 95.  
[http://dx.doi.org/10.1016/S0169-4332\(00\)00838-2](http://dx.doi.org/10.1016/S0169-4332(00)00838-2)
- [32] C.Y. Vallgren *et al.*, Proc. IPA2011, San Sebastian, 2011, p. 1587.

- [33] J.R. Calvey *et al.*, Proc. IPAC2011, San Sebastian, 2011, p. 796.
- [34] Y. Suetsugu *et al.*, *Nucl. Instrum. Methods PR-A* **578** (2007) 470.  
<http://dx.doi.org/10.1016/j.nima.2007.06.015>
- [35] M. Pivi *et al.*, *J. Appl. Phys.* **104** (2008) 104904. <http://dx.doi.org/10.1063/1.3021149>
- [36] M. Pivi *et al.*, Proc., PAC2007, Albuquerque, 2007, p. 1997.
- [37] J. Conway *et al.*, Proc. PAC2011, New York, 2011, p. 1250.
- [38] Y. Funakoshi *et al.*, Proc. EPAC2006, Edinburgh, 2006, p. 610.
- [39] Y. Cai *et al.*, Proc. PAC2003, Portland, 2003, p. 350.
- [40] D. Alesini *et al.*, Proc. IPAC2012, New Orleans, 2012, p. 1107.
- [41] Y. Suetsugu *et al.*, *Nucl. Instrum. Methods PR-A* **598** (2008) 372.  
<http://dx.doi.org/10.1016/j.nima.2008.08.154>

## Beam–Material Interactions

*N.V. Mokhov<sup>1</sup> and F. Cerutti<sup>2</sup>*

<sup>1</sup>Fermilab, Batavia, IL 60510, USA

<sup>2</sup>CERN, Geneva, Switzerland

### Abstract

This paper is motivated by the growing importance of better understanding of the phenomena and consequences of high-intensity energetic particle beam interactions with accelerator, generic target, and detector components. It reviews the principal physical processes of fast-particle interactions with matter, effects in materials under irradiation, materials response, related to component lifetime and performance, simulation techniques, and methods of mitigating the impact of radiation on the components and environment in challenging current and future applications.

### Keywords

Particle physics simulation; material irradiation effects; accelerator design.

## 1 Introduction

The next generation of medium- and high-energy accelerators for megawatt proton, electron, and heavy-ion beams moves us into a completely new domain of extreme energy deposition density up to 0.1 MJ/g and power density up to 1 TW/g in beam interactions with matter [1, 2]. The consequences of controlled and uncontrolled impacts of such high-intensity beams on components of accelerators, beamlines, target stations, beam collimators and absorbers, detectors, shielding, and the environment can range from minor to catastrophic. Challenges also arise from the increasing complexity of accelerators and experimental set-ups, as well as from design, engineering, and performance constraints.

All these factors put unprecedented requirements on the accuracy of particle production predictions, the capability and reliability of the codes used in planning new accelerator facilities and experiments, the design of machine, target, and collimation systems, new materials and technologies, detectors, and radiation shielding and the minimization of radiation impact on the environment. Particle transport simulation tools and the physics models and calculations required in developing relevant codes, such as FLUKA [3–5], GEANT4 [6–8], MARS15 [9–12], MCNP6 [13, 14], and PHITS [15, 16], are all driven by application. The most demanding applications are the high-power accelerators (e.g., spallation neutron sources, heavy-ion machines, and neutrino factories), accelerator driven systems, high-energy colliders, and medical facilities [2].

This paper gives a brief overview of the principal issues in the field. It is divided into two main sections. The first section is devoted to specific details of interactions of fast particles with matter. The second section characterizes the behaviour of materials under irradiation and highlights related applications at particle accelerators.

## 2 Interactions of fast particles with matter

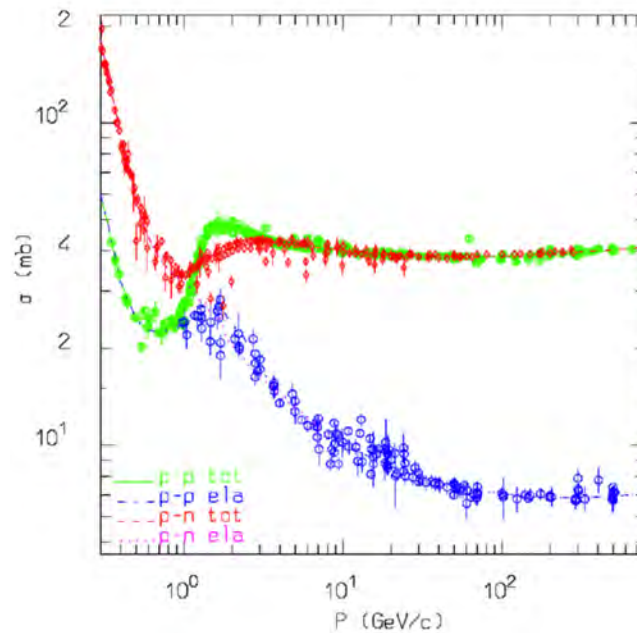
Electromagnetic interactions, decays of unstable particles, and strong inelastic and elastic nuclear interactions all affect the passage of high-energy particles through matter. The physics of these processes is described in detail in numerous books, handbooks, and reviews (see, for example, Refs. [2, 17–19]).

At high energies, the characteristic feature of the phenomenon is the creation of hadronic cascades and electromagnetic showers in matter due to multiparticle production in electromagnetic and strong nuclear interactions. Because of consecutive multiplication, the interaction avalanche rapidly accrues, passes the maximum and then dies as a result of energy dissipation between the cascade particles and due to ionization energy loss. Energetic particles are concentrated around the projectile axis forming the shower core. Neutral particles (mainly neutrons) and photons dominate with a cascade development when energy drops below a few hundred megaelectronvolts.

The length scale in hadronic cascades is a nuclear interaction length,  $\lambda_I$ , (16.8 cm in iron), while in electromagnetic showers it is a radiation length,  $X_0$ , (1.76 cm in iron); see Refs. [17, 18] for definitions and values of these quantities in other materials. The hadronic cascade longitudinal dimension is  $(5 \div 10)\lambda_I$ , while in electromagnetic showers it is  $(10 \div 30)X_0$ . It increases logarithmically with primary energy in both cases. Transversely, the effective radius (95% of energy deposited) for a hadronic cascade is about  $\lambda_I$ , while for electromagnetic showers it is about  $2R_M$ , where  $R_M$  is the Molière radius equal to  $0.0265X_0(Z + 1.2)$ . Low-energy neutrons coupled to photons propagate for much larger distances in matter around the cascade core, both longitudinally and transversely, until they thermalize down to an energy of the order of a fraction of an electronvolt and possibly undergo radiative capture, still implying the emission of photons of several megaelectronvolts. Muons—created predominantly in pion and kaon decays during cascade development—can travel hundreds and thousands of metres in matter along the cascade axis. Neutrinos—usual muon partners in such decays—propagate even farther, hundreds and thousands of kilometres, until they exit the Earth’s surface.

## 2.1 Nuclear reactions: particle and residue production

Hadron production is ruled by non-elastic nuclear reactions. For a sound description of hadron–nucleus (h–A) and nucleus–nucleus (A–A) interactions, one has to rely on a comprehensive understanding of hadron–nucleon (h–N) interactions over a wide energy range as a basic ingredient. Figure 1 shows the total and elastic N–N cross-sections. Below 1 GeV/c, the two cross-sections (total and elastic) tend to coincide both for p–p (n–n) and p–n, rapidly increasing with decreasing energy and with about a factor of three difference between p–p and p–n at the low-energy end, as expected on the basis of symmetry and isospin considerations. At high energies, the isospin dependence disappears and the reaction cross-section, given by the difference between total and elastic cross-sections, becomes predominant.

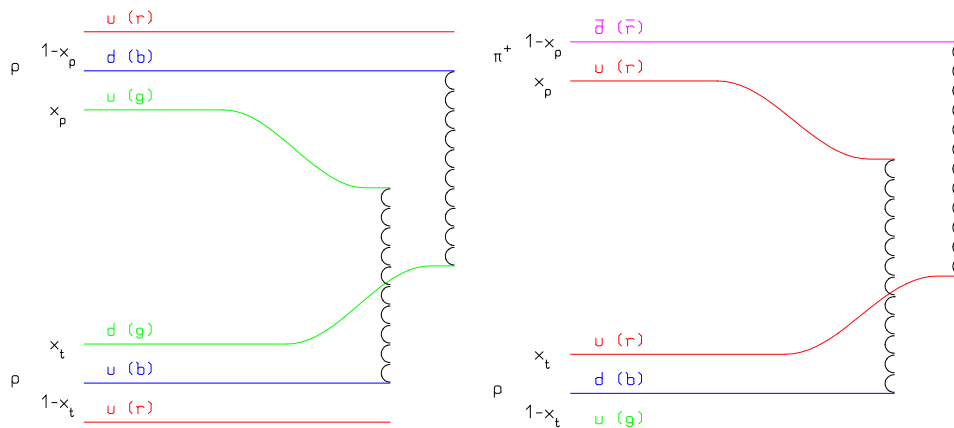


**Fig. 1:** Total (tot) and elastic (ela) proton–proton and proton–neutron cross-sections as a function of proton momentum. Points are experimental data, dashed lines are adopted parameterizations [19].

The non-elastic channel with the lowest threshold, i.e., single pion production, in N–N interactions ( $N_1 + N_2 \rightarrow N_1' + N_2' + \pi$ ) opens at a projectile kinetic energy of 290 MeV and becomes important above 700 MeV. In pion–nucleon interactions ( $\pi + N \rightarrow \pi' + \pi'' + N'$ ), the threshold is as low as 170 MeV. Both reactions are normally described in the framework of the isobar model, assuming that they proceed through an intermediate state containing at least one resonance. There are two main classes of reaction, those in which the intermediate state coincides with a single resonance (possible in  $\pi$ –N reactions) and those in which it initially contains two particles. The former exhibits a bump in the cross-section, corresponding to the mass of the formed resonance. Resonance masses, widths, cross-sections, and branching ratios are extracted from data and conservation laws whenever possible, making explicit the use of spin and isospin relations. They can be also inferred from inclusive cross-sections when needed. For a discussion of resonance production, see, for example, Refs. [20–22].

As soon as the projectile energy exceeds a few gigaelectronvolts, the description in terms of resonance production and decay becomes increasingly difficult. The number of resonances that should be considered grows exponentially and their properties are often poorly known. Furthermore, the assumption of one or two resonance creations is unable to reproduce an experimental feature of high-energy strong interactions, i.e., the large yield of secondary particles that belong neither to the projectile nor to the target fragmentation region but rather to the central region, at small Feynman  $x$  values. Different models, based directly on quark degrees of freedom, must be introduced.

Models based on interacting strings have emerged as a powerful tool in understanding quantum chromodynamics at the soft hadronic scale (low transverse momentum), that is in the non-perturbative regime. The dual parton model [23] is one of these models and is built by introducing partonic ideas into a topological expansion, which explicitly incorporates the constraints of duality and unitarity, typical of Regge theory. In this context, hadrons are considered as open strings with quarks, antiquarks, or diquarks sitting at the ends. For instance, mesons (colourless combinations of a quark and an antiquark) are strings with their valence quark and antiquark at the two opposite ends. At sufficiently high energies, the leading term in the interaction corresponds to a pomeron exchange (a closed string exchange), which has a cylindrical topology. When a unitarity cut is applied to the cylindrical pomeron, two hadronic chains are left as the sources of particle production. As a consequence of colour exchange in the interaction, each colliding hadron splits into two coloured partons, one carrying colour charge  $c$  and the other  $\bar{c}$ . The parton with colour charge  $c$  (or  $\bar{c}$ ) of one hadron combines with the parton of the complementary colour of the other hadron, to form two colour-neutral chains. These chains appear as two back-to-back jets in their own centre-of-mass systems. The exact method of building up these chains depends on the nature of the projectile–target combination; examples are shown in Fig. 2.



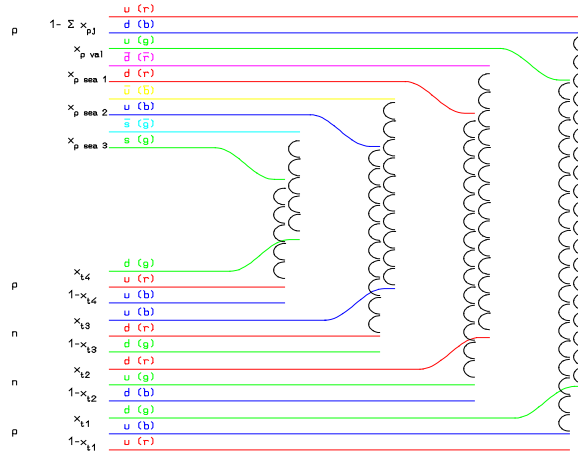
**Fig. 2:** Dual parton model leading two-chain diagram for (left) p–p and (right)  $\pi^+$ –p scattering. The respective colour and quark combination shown in the figure is just one of the allowed possibilities. Momentum fractions are also indicated.

The chains produced in an interaction are then hadronized. The dual parton model gives no prescriptions for this stage of the reaction. All the available chain hadronization models, however, rely on the same basic assumptions, the most important one being chain universality; that is, chain hadronization does not depend on the particular process that originated the chain, and until the chain energy is much larger than the mass of the hadrons to be produced, the fragmentation functions (which describe the momentum fraction carried by each hadron) are the same. As a consequence, fragmentation functions can, in principle, be derived from hard processes and  $e^+e^-$  data, with (few) parameters valid for all reactions and energies. In fact, mass and threshold effects are non-negligible at typical chain energies and require a suitable treatment. Examples of  $h-N$  particle production can be found, for instance, in Ref. [19].

When moving to nucleus interactions ( $h-A$  and  $A-A$ ), the increased complexity of the problem is usually schematized into a sequence of three stages, discussed next.

### 2.1.1 Cascade

In the Glauber formalism [24, 25], the inelastic interaction of a hadron with a nucleus is described through multiple contemporary interactions with  $\nu$  target nucleons. The Glauber–Gribov model [26–28] represents the diagram interpretation of the Glauber cascade. The  $\nu$  interactions of the hadron projectile originate  $2\nu$  chains; two of them are formed by the projectile valence quarks and the valence quarks of one target nucleon (valence–valence chains), while the remaining  $2(\nu - 1)$  chains involve projectile sea quarks and valence quarks of the other struck nucleons (sea–valence chains). The chain-building process is illustrated in Fig. 3 for a proton–nucleus interaction; analogous diagrams apply for other hadron projectiles.

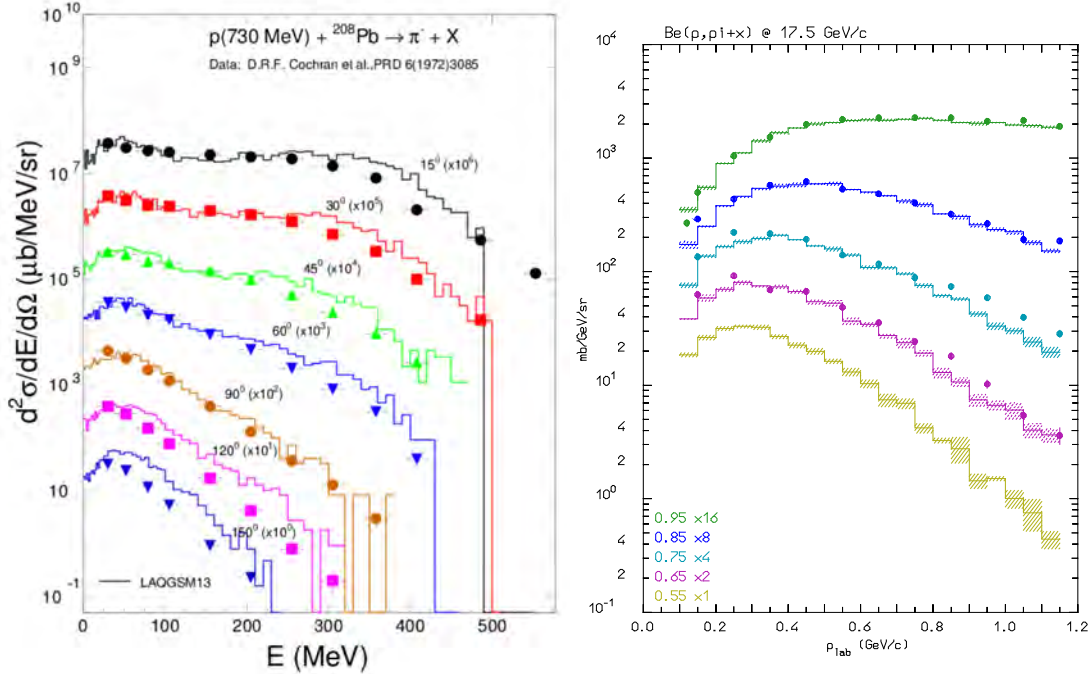


**Fig. 3:** Dual parton model leading two-chain diagram for  $p-A$  Glauber scattering with four collisions. The colour and quark combination shown in the figure is just one of the allowed possibilities. Momentum fractions are also indicated.

This sharing of the projectile energy among many chains naturally softens the energy distributions of the reaction products and boosts the multiplicity with respect to  $h-N$  interactions. In this way, the model accounts for the major  $A$ -dependent features without any degrees of freedom, except in the treatment of mass effects at low energies.

The Fermi motion of the target nucleons must be included to obtain the correct kinematics, in particular, the smearing of transverse momentum ( $p_T$ ) distributions. All nuclear effects on the generated hadrons (‘secondaries’) are accounted for by the subsequent intranuclear cascade. The formation zone concept is essential to explain the observed reduction of the re-interaction probability with respect to the naive free cross-section assumption. It can be understood as a ‘materialization’ time. At high

energies, the ‘fast’ particles produced in the Glauber cascade have a high probability of materializing outside the nucleus without triggering a secondary cascade. Further cascading only involves the slow fragments produced in the target fragmentation region of each primary interaction, and therefore the re-interaction probability tends quickly to saturate with energy as the Glauber cascade reaches its asymptotic regime. Only a small fraction of the projectile energy is thus left available for the intranuclear cascade and the following stages. Examples of pion production at different energies are shown in Fig. 4.



**Fig. 4:** Left: Double differential spectra of negative pions generated by 730 MeV protons on  $^{208}\text{Pb}$ . Symbols are experimental data [29] and histograms are MARS15 results, both scaled according to the angle as indicated. Right: Double differential spectra of positive pions generated by 17.5 GeV/c protons on  $^9\text{Be}$ . Symbols are experimental data [30] and histograms are FLUKA results, both scaled according to the angle as indicated (cosine values are given).

The Glauber cascade and the formation zone act together in reaching a regime where the ‘slow’ part of the interaction is almost independent of the projectile energy. Owing to the very slow variation of the  $h$ – $N$  cross-section from a few gigaelectronvolts to a few teraelectronvolts, the Glauber cascade is almost energy-independent and the rise in the multiplicity of ‘fast’ particles is related only to the increased multiplicity of the elementary  $h$ – $N$  interactions. At the end of the cascading process, the residual excitation energy is directly related to the number of primary and secondary collisions that have taken place. Each collision does indeed leave a ‘hole’ in the Fermi sea, which carries an excitation energy related to its depth in the Fermi sea.

### 2.1.2 Pre-equilibrium

At energies lower than the  $\pi$  production threshold, a variety of pre-equilibrium models have been developed [31], following two leading approaches: the quantum-mechanical multistep model and the exciton model. The former has a very good theoretical background but is quite complex, while the latter relies on statistical assumptions, and is simple and fast. Exciton-based models are often used in Monte Carlo codes to link the intranuclear cascade stage of the reaction to the equilibrium one.

Typically, the intranuclear cascade stage continues until all nucleons in the nucleus are below a smooth threshold of a few tens of megaelectronvolts and all particles except nucleons (e.g., pions) have been emitted or absorbed. The input configuration for the pre-equilibrium stage is characterized by the total number of remaining protons and neutrons, by the number of particle-like excitons (nucleons

excited above the Fermi level) and of hole-like excitons (holes created in the Fermi sea by the intranuclear cascade interactions), and by the excitation energy and momentum of the resulting nucleus. All these quantities can be derived by properly recording what occurred during the intranuclear cascade stage.

The pre-equilibrium stage, while distributing the excitation energy among all degrees of freedom through N–N elastic scattering, accounts for intermediate energy emissions of single nucleons and light particles formed by nucleon coalescence.

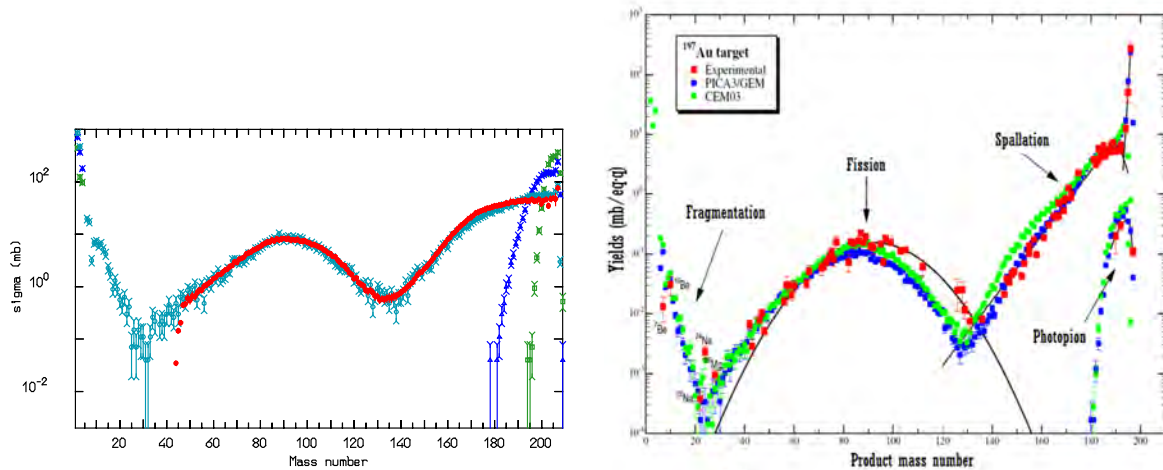
### 2.1.3 Final de-excitation

The last stage of the reaction chain assumes that the remaining nucleus (typically more than one in the case of A–A interactions, featuring both a projectile- and a target-like residual system) is a thermally equilibrated system, characterized by its mass and atomic numbers and a given excitation energy. The latter is dissipated through the ‘evaporation’ of single nucleons, light particles, and intermediate mass fragments, or by fission. The neutron evaporation is favoured over charged particle emission, owing to the Coulomb barrier, especially for medium-heavy nuclei, whose excitation energy is higher, owing to the larger cascading chances and to the larger number of primary collisions in the Glauber cascade at high energies.

Many evaporation or fission models are based on the standard Weisskopf–Ewing formalism [32]. For light residual nuclei, where the excitation energy may overwhelm the total binding energy, statistical fragmentation models (Fermi break-up) are more appropriate.

The end of the de-excitation process is characterized by the emission of  $\gamma$ -rays, corresponding to the transition between specific levels of the residual nucleus.

Although the reaction may originate from a particularly high-energy projectile, its evaporation, fission, or break-up stage is a low-energy phenomenon, much slower than the previous stages and sensitive to nuclear physics ingredients. In fact, it determines what is left after the interaction, yielding the distribution of residues, which, in the case of an energetic nuclear reaction on a high-Z material, fills the whole mass (and charge) range of the nuclide chart, as demonstrated in Fig. 5.



**Fig. 5:** Left: Mass distribution of the nuclei generated by 1 GeV/n  $^{208}\text{Pb}$  ions on hydrogen. Data [33] (red) are compared with FLUKA results (light blue). The distributions obtained after the cascade (green) and pre-equilibrium (dark blue) stages are also shown. Right: Mass distribution of the nuclei generated by bremsstrahlung photons (up to 1 GeV) on  $^{197}\text{Au}$ . Data (see Ref. [34] and references therein) in red are compared with the results of the CEM03 model used in MARS15 (green). The distribution obtained using a different model (blue) is also shown. Reproduced with permission from S. Mashnik.



It is worth mentioning that photonuclear (see Fig. 5 right) and electronuclear interactions can also be coherently described in this framework, through an appropriate definition of the initial state. Electronuclear interactions are, in fact, nuclear interactions by virtual photons.

## 2.2 Radionuclides

Among the residual nuclei generated in non-elastic nuclear reactions, one can probably find radionuclides, which are subject to further decay into other nuclear species and so are responsible for continuous delayed radiation, mostly of an electromagnetic nature (electrons, positrons, and photons), owing to  $\beta$  and  $\gamma$  decays. This initiates decay chains, governed by the radioactivity laws, where the time evolution of isotope populations is given by the Bateman equations:

$$dN_n/dt = P_n + \sum_i (b_{i,n} \cdot \lambda_i \cdot N_i) - \lambda_n \cdot N_n . \quad (1)$$

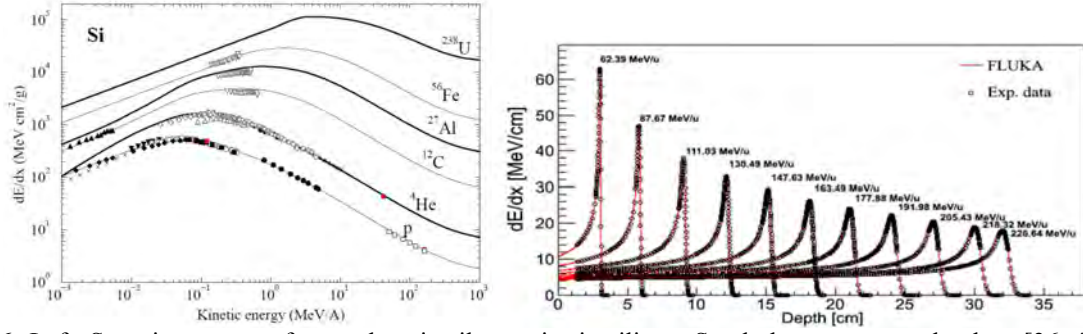
In Eq. (1),  $N_n$  represents the population of a certain isotope,  $n$ , varying as a function of its direct production rate by nuclear reactions,  $P_n$ , its decay constant,  $\lambda_n$ , and the growth rate by parent decay expressed by the sum extended to all parent isotopes,  $i$ . The latter takes into account the respective branching ratios  $b_{i,n}$  for the relevant channel.

A Monte Carlo code simulating the *prompt radiation* propagation in a given geometry, and thereby calculating the radionuclide production rates,  $P_n$ , can also be provided with the capability of solving, on-line, the system of Bateman equations for an irradiation profile (sequence of time intervals and corresponding beam intensities) and several cooling times (time distances from the irradiation end). This enables it to calculate specific activities as well as residual dose rates owing to the *decay radiation*, whose spatial propagation can be simulated at the same time.

## 2.3 Ionization energy loss

Tracking the transport of charged particles in matter involves accounting for the Coulomb interactions they experience with the electrons and the atomic nuclei of the medium. While the incoming particle energy loss is largely dominated by the first interactions, the other interactions are the main responsible for the particle's trajectory changes. In most cases, multipurpose Monte Carlo codes treat these processes as *continuous*, contrary to *discrete* events, such as nuclear reactions, bremsstrahlung emissions, Compton scattering, or particle decay. This means that the particle proceeds through steps, at the end of each of which its energy and direction is changed to take into account the cumulative effect of all Coulomb interactions (inducing atom ionization and excitation) experienced along the track. Actually, the generation of energetic knock-on electrons ( $\delta$  rays) can also be treated as a discrete event (above the electron transport energy limit, of the order of 1 keV), paying a significant penalty in computation time, where it is justified by the need not to assume the particle energy loss as translating into a local energy deposition, but to consider the range of the electron carrying part of that energy loss elsewhere. Analogously, as far as the particle trajectory is concerned, single scattering with an atomic nucleus can also be explicitly simulated when needed (implying much shorter steps), instead of utilizing multiple scattering algorithms, providing the trajectory alteration as a cumulative outcome.

The well-known Bethe–Bloch formula [35] gives the *mean* loss rate (stopping power) as a function of the particle speed and charge and of the relevant material properties. Nevertheless, several corrections (such as shell, Barkas, Bloch, Mott, and Lindhard–Sørensen corrections) have to be included, to ensure suitable accuracy. Moreover, with high- $Z$  projectiles it is necessary to evaluate their effective charge, since electron capture becomes important at low energies. In addition, actual energy losses feature significant fluctuations with respect to the mean value, making the latter far from being exhaustive, but requiring the implementation of a proper distribution function.



**Fig. 6:** Left: Stopping power of several projectile species in silicon. Symbols are measured values [36–41] and curves are MARS15 predictions. Right: Profiles of energy deposition in water by protons at different energies in the hadron therapy range. Circles are experimental data and curves are FLUKA results [42].

Examples of the accuracy achieved in the description of the ionization process are shown in Fig. 6, where, in addition to the reproduction of measured stopping power values for a notable variety of radiation types, the study of many Bragg peaks of clinical use in proton therapy is reported.

## 2.4 Displacements of atoms

The dominant mechanism of structural damage of inorganic materials is displacement of atoms from their equilibrium position in a crystalline lattice as a result of irradiation, with the formation of interstitial atoms and vacancies in the lattice. The resulting deterioration of material critical properties is characterized—in the most universal way—as a function of the number of displacements per target atom; this number is a strong function of projectile type, energy, and charge, as well as material properties, including temperature.

Three major codes (FLUKA, MARS15, and PHITS) use very similar implementations of the Norgett-Robinson-Torrens (NRT) model [43, 44] to calculate the number of displacements per target atom [2]. A primary knock-on atom created in nuclear collisions can generate a cascade of atomic displacements. This is taken into account via the damage function  $\nu(T)$ . The number of displacements per target atom is expressed in terms of the damage cross-section  $\sigma_d$ :

$$\sigma_d(E) = \int_{T_d}^{T_{\max}} \frac{d\sigma(E,T)}{dT} \nu(T) dT, \quad (2)$$

where  $E$  is the kinetic energy of the projectile,  $T$  is the kinetic energy transferred to the recoil atom,  $T_d$  is the displacement energy, and  $T_{\max}$  is the highest recoil energy according to kinematics. In a modified Kinchin–Pease model [43],  $\nu(T)$  is zero at  $T < T_d$ , unity at  $T_d < T < 2.5T_d$ , and  $k(T)E_d/2T_d$  at  $2.5T_d < T$ , where  $E_d$  is ‘damage’ energy available to generate atomic displacements by elastic collisions.  $T_d$  is an irregular function of atomic number ( $\sim 40$  eV). The displacement efficiency,  $k(T)$ , introduced as a result of simulation studies on evolution of atomic displacement cascades [45], drops from 1.4 to 0.3 once the primary knock-on atom energy is increased from 0.1 to 100 keV, and exhibits a weak dependence on target material and temperature.

The implementation of this model in MARS15 [46] and FLUKA [47] includes electromagnetic elastic (Coulomb) scattering, the Rutherford cross-section with Mott corrections, and nuclear form-factors (a factor of two effect). Resulting displacement cross-sections due to Coulomb scattering are shown in Fig. 7 (left) for various projectiles on silicon and carbon targets. For elementary particles, the energy dependence of  $\sigma_d$  disappears above 2–3 GeV, while it continues to higher energies for heavy ions. For projectiles heavier than a proton,  $\sigma_d$  grows with the projectile charge  $z$  as  $z^2/\beta^2$  at  $\gamma\beta > 0.01$ , where  $\beta$  is the projectile velocity. All products of elastic and inelastic nuclear interactions, as well as Coulomb elastic scattering of transported charged particles (hadrons, electrons, muons, and heavy ions) from 1 keV to 10 TeV, contribute to the number of displacements per target atom in the model. The

number of displacements per target atom for neutrons from  $10^{-5}$  eV to 20–150 MeV is described in MARS15 using the NJOY99+ENDF-VII database [48, 49] for 393 nuclides [50]. A corresponding output is shown in Fig. 7 (right). FLUKA adopts database information for neutrons up to 20 MeV, while at higher energies, where many reaction channels are open, it describes neutron elastic and inelastic interactions through its models and determines the number of displacements per target atom explicitly by calculating non-ionizing energy losses of the products.

Such results are then corrected using the experimental defect production efficiency  $\eta$ , where  $\eta$  is a ratio of a number of single interstitial atom vacancy pairs (Frenkel pairs) produced in a material to the number of defects calculated using the NRT model. The values of  $\eta$  have been measured [51] for many important materials in the reactor energy range.

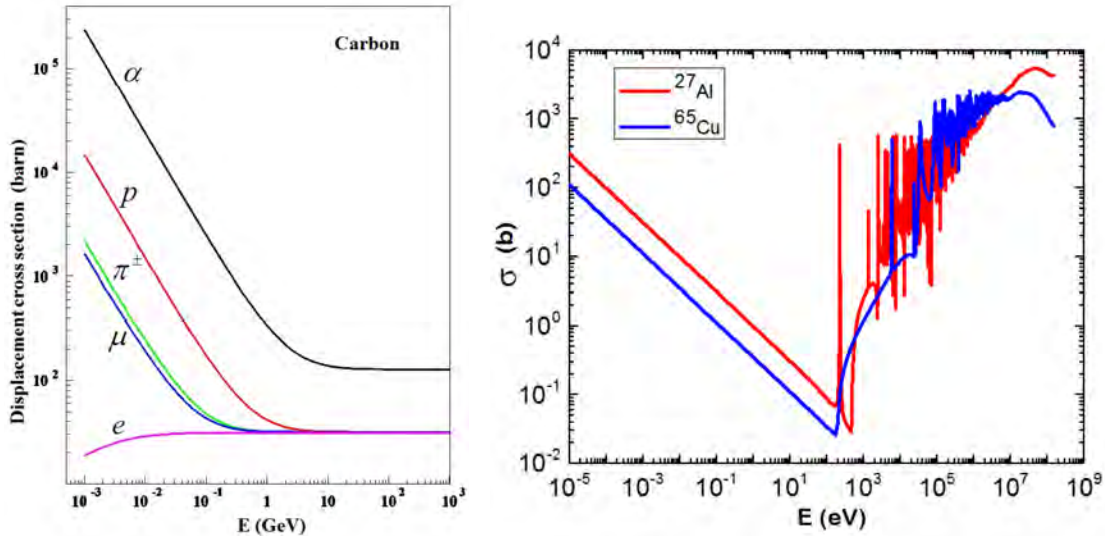


Fig. 7: Left: Displacement cross-section in carbon for various charged particles (MARS15). Right: NRT neutron defect production cross-sections on aluminium and copper.

### 3 Materials under irradiation

Depending on the material, the level of energy deposition density, and the time structure, one can face a variety of effects in materials under irradiation. The two categories of materials response are related to the component lifetime and performance [2]:

1. component damage (lifetime):
  - thermal shocks and quasi-instantaneous damage (see also A. Bertarelli’s contribution in these proceedings);
  - insulation property deterioration due to dose build-up;
  - radiation damage to inorganic materials due to atomic displacements, as well as helium and hydrogen production;
  - detector component radiation aging and damage.
2. operational (performance):
  - superconducting magnet quenching;
  - single-event effects in electronics;
  - detector performance deterioration;
  - radioactivation, prompt dose and impact on environment.

### 3.1 Thermal shock

Short pulses with energy deposition density ranging from 200 J/g (tungsten) and 600 J/g (copper) to  $\sim 1$  kJ/g (nickel, Inconel) and  $\sim 15$  kJ/g cause thermal shocks, resulting in fast ablation and slower structural changes, or melting. Two outstanding examples are the Fermilab antiproton ( $\bar{p}$ ) production target damages by a 120 GeV proton beam in the period from 1993 to 2011, and a 980 GeV proton beam-induced Tevatron collimator damage.

#### 3.1.1 Antiproton production target

Significant effects in the evolution of the Fermilab antiproton production target have been observed [52]. This 10 cm diameter target stack is made up of six target discs 0.95 cm thick separated—in early days—by two 0.32 cm thick copper cooling discs, later replaced with copper mini-cylinders to provide better airflow. The entire assembly slowly rotates, distributing the primary beam, with time, over a cylindrical section of the target. In Tevatron Run I at Fermilab, evidence of external target damage, sustained when the rotation mechanism failed for several months with only vertical motion available, was discovered. Figure 8 shows damage at the exit of a nickel target chord. Ejection of nickel pieces has also led to a contamination incident.



**Fig. 8:** Tevatron Run I antiproton production target damage in 1994. Courtesy of A. Leveling and J. Morgan

When the target was rotated properly, it was not damaged, although the outer titanium sleeve showed signs of swelling. Nickel was used in the first year of Run II. After several months of operation at  $4.5 \times 10^{12}$  protons per pulse with root mean square beam spot sizes of  $\sigma_x = 0.22$  mm and  $\sigma_y = 0.16$  mm, a region of damage about 2.5 mm wide developed on the target; the titanium jacket evaporated in that region and there was a 15% drop in  $\bar{p}$  yield.

In September 2002, the targets were replaced with Inconel targets, which have an excellent high-temperature tensile strength, although a relatively low thermal conductivity compared with copper and nickel. The switch to Inconel-600 extended the service life of each target from weeks to months ( $\times 6$  for the entire stack), with practically no decrease in  $\bar{p}$  yield. The 11.43 cm outer diameter Inconel target disc with the copper mini-cylinders, providing best airflow for cooling, is shown in Fig. 9 (left).

In 2007, the target was used in a ‘consumable mode’ to maximize the  $\bar{p}$  yield. Figure 9 (right) shows the Inconel-600 disc remnants after 4 months of operation with a total  $2.65 \times 10^{19}$  protons on

target and a root mean square beam spot size of  $\sigma_x \sim 0.18$  mm,  $\sigma_y \sim 0.22$  mm. The observed phenomenon was attributed to chemical oxidation of the damaged—by thermal shock—target surface at the target chord beam exit. Because of tight limits established for vertical target positioning, the copper cooling cylinders between the target discs emerged relatively undamaged, though distorted in some places, as seen in the figure.

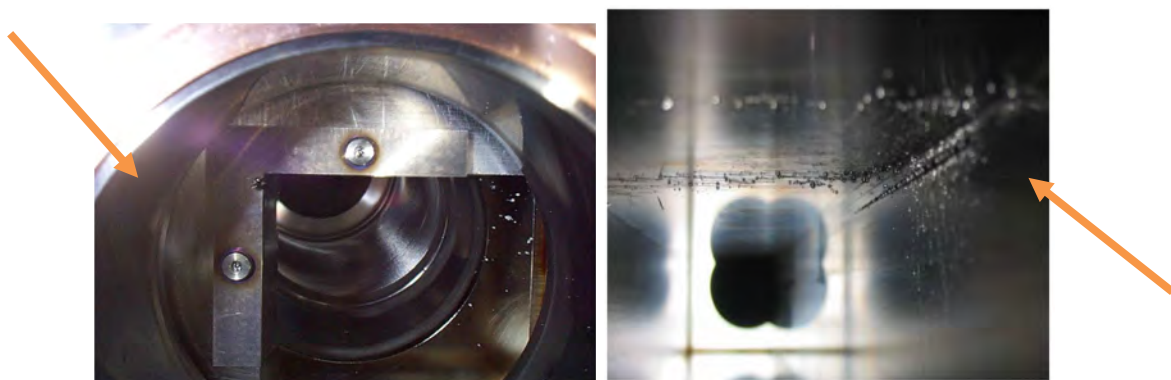
Radioactive particles from a damaged target were also a problem amplified by radioactive titanium pieces ejected from the damaged jacket. In Run II, the titanium jacket was replaced by a thin-walled cylinder of graphite, and then by a 6.35 mm thick beryllium jacket, both being nearly transparent to the primary beam and products of its interactions with the target.



**Fig. 9:** Tevatron Run II antiproton production target. Left: Top view. Right: Target number 7 in 2007 with damaged Inconel discs and distorted copper cooling cylinders in between. Courtesy of A. Leveling and J. Morgan.

### 3.1.2 *Tevatron collimator damage*

Another example of the fast material ablation at accelerators is the destruction of the Tevatron primary (Fig. 10, left) and secondary (Fig. 10, right) collimators caused by an accidental loss of the 980 GeV beam in 2003 [53]. The damage was induced by a failure in the Collider Detector Facility Roman pot detector positioning at the end of a  $980 \times 980$  GeV proton–antiproton colliding beam store.



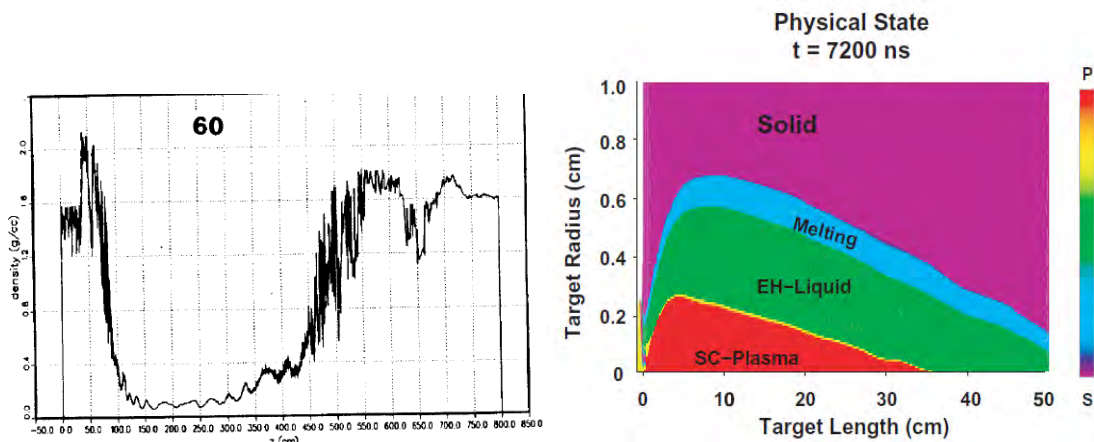
**Fig. 10:** Left: The hole indicated was created in the Tevatron 5 mm thick primary tungsten collimator. Right: The 25 cm long groove indicated was created in the Tevatron secondary stainless steel collimator [2, 53].

The dynamics of this failure over the first 1.6 ms, including excessive halo generation and superconducting magnet quenching, were studied via realistic simulations using the MARS [9–12] and STRUCT [54] codes. It was shown that the misbehaved beam-induced ablation of the tungsten primary

collimator resulted in the creation of the hole seen in it, while the simulated parameters of the groove in the stainless steel secondary collimator jaw surface fully agreed with the post-mortem observations [2, 53].

### 3.2 Hydrodynamic regime

Beam pulses with energy deposition density in excess of 15 kJ/g bring materials to the hydrodynamic regime [2]. This was demonstrated in studies [55] for the SSC 20 TeV proton beam (400 MJ, 300  $\mu$ s spill), first on a graphite beam dump and later for the collider superconducting magnets, steel collimators, and tunnel-surrounding Austin chalk. Since the beam duration was comparable to the characteristic time of expected hydrodynamic motions, the static energy deposition capability of the MARS code has been combined with the 2D and 3D hydrodynamics of the LANL's MESA and SPHINX codes. It was found, in simulations, that a hole was drilled by the beam in the graphite dump at a rate of 7 cm/ $\mu$ s with generated pressures of a few kbar (Fig. 11, left). Later these effects were studied in detail for the Super Proton Synchrotron (SPS) and Large Hadron Collider (LHC) targets, beam dumps, and collimators using coupling of the FLUKA code (energy deposition) with BIG2 [56, 57] and LS-DYNA [58] (hydrodynamics) codes. Figure 11 (right) shows the calculated physical state of the solid tungsten target at the end of the SPS proton pulse (root mean square beam spot size of 0.088 mm) at 7.2  $\mu$ s. It can be seen that within the inner 2 mm radius, a strongly coupled plasma state exists, which is followed by an expanded hot liquid. The melting front is seen propagating outwards.



**Fig. 11:** Left: Axial density of graphite beam dump in 60  $\mu$ s after the 20 TeV beam spill start [55]. Right: Tungsten target physical state after the SPS beam pulse [56]. EH, expanded hot; SC, strongly coupled. Reproduced from [56] with permission from N. Tahir.

### 3.3 Hydrogen and helium gas production

At accelerators, radiation damage to inorganic structural materials—being primarily driven by displacement of atoms in a crystalline lattice (see Section 2.4)—is amplified by increased hydrogen and helium gas production for high-energy beams. In the Spallation Neutron Source (SNS) beam windows, the ratio of He atoms to the number of displacements per target atom is about 500 times that in fission reactors. These gases can lead to grain boundary embrittlement and accelerated swelling. In the simulation codes analysed here, uncertainties on production of hydrogen are about 20%, while for helium uncertainties could be as high as 50%.

### 3.4 Dose in organic materials

Unlike megaelectronvolt-type accelerators, which have insulators made mostly of ceramics or glasses, the majority of insulators in high-energy accelerator equipment are made of organic materials: epoxy resin, G11, polymers, etc. [2]. Apart from electronics and optical devices, the organic materials are the most sensitive to radiation. A large number of radiation tests have been made on these materials and the

results are extensively documented [59]. The impact of radiation on organic materials is a three-step process [60]:

1. production of free radicals by radiation;
2. reaction of free radicals: crosslinking, chain scission, formation of unsaturated bonds (C=C, etc.), oxidation, and gas evolution;
3. change of molecular structure: modification and degradation affected by irradiation temperature and atmosphere as well as by presence of additives.

The findings for organic materials under irradiation are [60]:

- degradation is enhanced at high temperatures;
- radiation oxidation in the presence of oxygen accelerates degradation;
- radiation oxidation is promoted in the case of low dose rates;
- additives can improve radiation resistance; for example, 1% by weight of antioxidant in polyethylene can prolong its lifetime 5 to 10 times.

Dose limits on insulators are usually defined for a certain level of change in the material properties critical to the application. For example, 10% degradation in ultimate tensile strength is a typical criterion for epoxy, CE epoxy resins and G11. Similar changes in electrical resistivity are often used as a criterion. For the given insulator and irradiation conditions, radiation damage is proportional to the peak energy deposition density or dose accumulated in the hottest region. Radiation damage thresholds, based on the results of dedicated radiation tests [59], experience, or indirect evidence, are used worldwide as a basis for design and an estimate of component lifetime or operation time prior to replacement. For example, the dose limit used for the LHC superconducting magnet insulators is 25 to 40 MGy [61]. Other projects utilizing superconducting magnet technologies assume a lower limit, of 7–10 MGy.

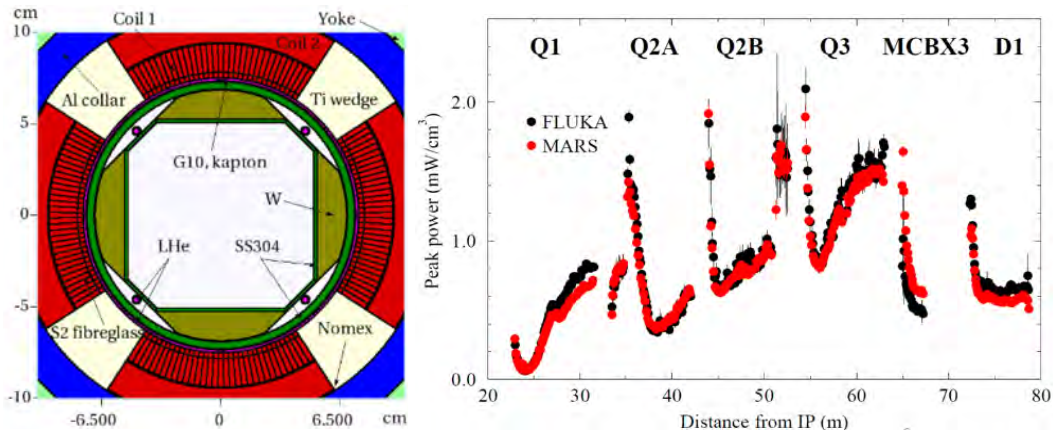
It is worth noting here that energy deposition—responsible for damage in insulators and, e.g., for cable quench stability in superconducting magnets—is modelled in accelerator applications quite accurately. In the majority of cases, FLUKA and MARS15 results on energy deposition coincide within 10% and agree with data.

### 3.5 Quenching

A magnet quench is a dramatic yet fairly routine event within a particle accelerator. It occurs when one of the superconducting magnets that steer and focus the particle beams warms above a critical temperature, bringing operations to an abrupt halt. A quench often starts when stray particles from the beam enter a magnet's coils, producing an initial burst of heat. Within a fraction of a second, parts of the superconducting wire in the magnet lose their ability to conduct electricity without resistance, generating more heat that quickly spreads throughout the entire magnet. The coolant surrounding the magnet begins to boil.

As with the other types of the beam impact on materials, the beam-induced quench creation and propagation in the superconducting coils depend on the level and profile of energy deposition density, its time structure, operational current and—for pulses longer than ~1 ms—on the cooling efficiency of the superconducting cable. The accelerator class superconducting magnet will quench if the peak power density in the innermost cable exceeds 40–60 mW/cm<sup>3</sup> in the Nb<sub>3</sub>Sn quadrupoles at  $I_{op}/I_c = 0.8$  [62–64]. Here,  $I_{op}$  is the magnet operational current, and  $I_c$  is the magnet critical current. The quench limit in NbTi based coils is 13 mW/cm<sup>3</sup>, again at  $I_{op}/I_c = 0.8$ . Studies of beam-induced effects in accelerator superconducting magnets are described in Refs. [61, 64], which consider the high-luminosity upgrade of the LHC inner triplet magnets. A tiny fraction of the 7 TeV proton beams or products of their interactions lost on the superconducting magnets would induce hadronic and electromagnetic showers with energy deposition levels that could easily exceed these quench limits. Optimized using thorough

FLUKA and MARS15 studies, absorbers and high-Z magnet bore inserts (related to the electromagnetic nature of energy deposition at that location) are to be incorporated in the high-luminosity LHC inner triplet region to mitigate this problem safely. Figure 12 from Ref. [61] shows the details of the protection system (left) and the resulting peak power density profile—well below the quench limits—on the innermost superconducting cable by the collision debris at the luminosity of  $5 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$  (right).



**Fig. 12:** Left: Details of the FLUKA-MARS15 model in the innermost region of the high-luminosity LHC inner triplet first quadrupole, with 16 mm thick tungsten inserts (olive) in the mid-planes. Right: Longitudinal peak power density profile on the innermost superconducting cable of the inner triplet, orbit correctors (MCBX) and separation dipole (D1) coils, as calculated by FLUKA (black) and MARS15 (red) for 14 TeV centre-of-mass collision debris at  $5 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$  luminosity.

In the case of a large superconducting magnet, which can be several metres long and carry currents of 10,000 A or more, the quench creates a loud roar as the coolant—liquid helium with a temperature close to absolute zero—turns into gas and vents through pressure relief valves, like steam escaping a tea kettle [65]. Such a quench generates as much force as an exploding stick of dynamite. A magnet usually withstands this force and is operational again in a few hours after cooling back down. If repair is required, it takes valuable time to warm up, fix, and then cool down the magnet, i.e., days or weeks in which no particle beams can be circulated, and no science can be done.

During CERN’s LHC start-up operations in 2008, with current ramping up and no beam circulating, an electrical fault in a dipole–quadrupole interconnection was responsible for the development of an electrical arc puncturing the helium enclosure [66]. As a consequence, several superconducting magnets quenched and, despite helium relieving to the tunnel, large pressure forces displaced dipoles in a few subsectors. Eventually, the replacement of a number of magnets was necessary. To mitigate potentially destructive quenches, the superconducting magnets that form the LHC are equipped with fast-ramping heaters, which are activated once a quench event is detected by a complex quench protection system. Since the dipole bending magnets are connected in series, each power circuit includes 154 individual magnets; should a quench event occur, the entire combined stored energy of these magnets must be dumped at once. This energy is transferred into dumps that are massive blocks of metal, which heat up to several hundreds of degrees Celsius, through resistive heating, in a matter of seconds.

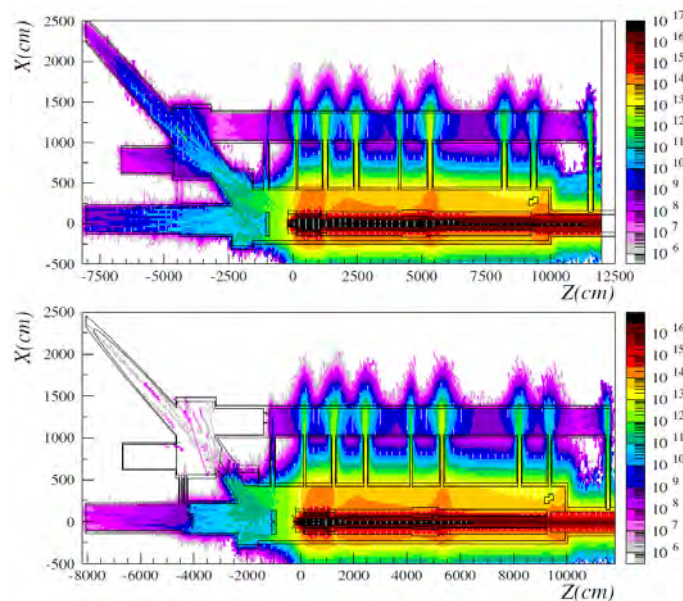
### 3.6 Radiation to electronics

Electronic components and systems exposed to a mixed radiation field experience three different types of radiation damage: damage from the total ionizing dose, displacement damage, and so-called single-event effects. The latter events range from single event or multiple bit upsets and single-event transients to possible destructive latch-ups, destructive gate ruptures or burn-outs (single-event gate ruptures and burn-outs).



The first two types of damage refer to the steady accumulation of defects causing measurable effects, which can ultimately lead to device failure. They are evaluated through total ionizing dose (in grays) and non-ionizing energy deposition, respectively. The latter is generally quantified by accumulated silicon 1 MeV equivalent neutron fluence, which requires the use of conversion factors to weight the effect of other energies and particle types with respect to one of the 1 MeV neutrons in silicon, as electronic device material. As for stochastic single-event-effect failures, these form an entirely different group, since they result from the ionization by a single particle, which is able to deposit enough energy to perturb the operation of the device. They can only be characterized in terms of their probability of occurrence as a function of accumulated high-energy hadron fluence, not overlooking the dependence on device type as well as on particle nature. Actually, the hadron energy threshold is usually intended as 20 MeV, but unstable hadrons of lower energies must also be counted. Concerning neutrons of lower energies, one has to weight them according to the ratio of their single-event upset cross-section to that of hadrons above 20 MeV, substantially reflecting the  $(n, \alpha)$  cross-section behaviour in representative microchip materials.

Such failures can lead to serious consequences: for instance, single-event effects were responsible for the shut-down of the CERN Neutrinos to Gran Sasso (CNGS) facility in 2007. The CNGS facility was designed to produce an intense muon neutrino beam directed towards the Gran Sasso National Laboratory (LNGS) in Italy, 732 km from CERN. The physics program extended over 5 yr with up to  $4.5 \times 10^{19}$  protons impinging on the primary target per year, extracted from the CERN SPS with two nominal extractions of  $2.4 \times 10^{13}$  protons at 400 GeV/c each, in a CNGS cycle of 6 s, corresponding to an average power at the target of 510 kW. The facility was started with gradually increasing intensity in 2007, but had to be shut down after only  $8 \times 10^{17}$  protons on target, owing to successive failures in the ventilation system. After detailed analysis, it was found that the microcontrollers that failed were placed in relatively high radiation areas, i.e., near to the ducts connecting the target and service galleries (see Fig. 13). The failure was due to single-event effects induced by high-energy hadrons.

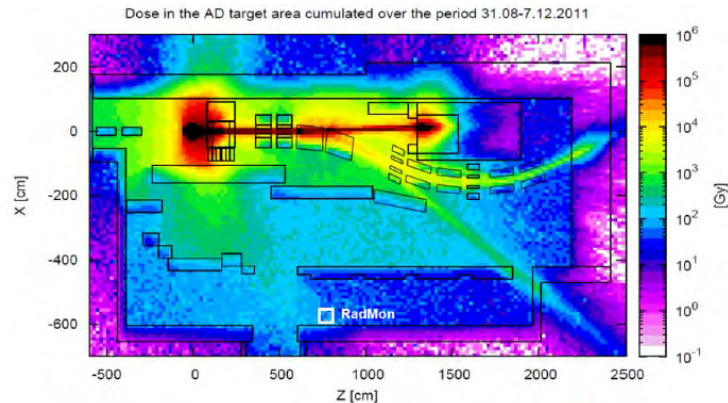


**Fig. 13:** Annual high-energy hadron fluence (in  $\text{cm}^{-2}$ ) in the CNGS facility, as predicted by FLUKA calculations, before and after the installation of the shielding aimed to create a protected area (the rectangle centred at about  $Z = -20$  m,  $X = 12$  m) for the control electronics. The 400 GeV/c proton beam impinges from the left on the carbon target at the reference frame origin. In the mentioned area, the radiation levels, initially in the range of  $10^7$  to  $10^9 \text{ cm}^{-2} \text{ yr}^{-1}$ , are reduced by the shielding below  $10^6 \text{ cm}^{-2} \text{ yr}^{-1}$ .

Enhancement of the electronics protection was mandatory, therefore, a radiation safe area was created with the introduction of fixed and mobile shielding. In total,  $53 \text{ m}^3$  of concrete was poured *in situ*, the ventilation system had to be completely reconfigured, and all the electronics had to be moved

to the new safe area. According to calculations (as in Fig. 13), the new shielding ensures considerable attenuation with respect to the former layout, for all quantities of interest, i.e., total ionizing dose, silicon 1 MeV neutron equivalent (1 MeV n-eq) fluence, and high-energy hadron fluence. In particular, the latter quantity is reduced to only at most one order of magnitude higher than the rate from cosmic rays at sea level ( $10^5 \text{ cm}^{-2}$  per year, roughly corresponding, in a radiation environment generated by a primary hadron source, to a 1 MeV n-eq fluence ten times larger and to a total ionizing dose below  $1 \text{ mGy yr}^{-1}$ ). Radiation monitors deployed at various points in the service galleries were used to benchmark the FLUKA calculations [67], obtaining remarkable agreement. In particular, high-energy hadron fluence values measured in the service gallery by two RadMon detectors [68], located in line-of-sight positions at the duct exits at about 50 and 80 m from the target (see Fig. 13), respectively, were reproduced within 10%, with an estimated experimental uncertainty of 20%.

Besides further instances at SPS and LHC energies [69, 70], another example of successful calculation of relevant radiation levels is given by the study of the antiproton decelerator target area [71], where the antiproton beam to be injected into the antiproton decelerator ring is generated by the impact of the Proton Synchrotron proton beam at 26 GeV/c onto an iridium target. Dose values measured on the beam line 10 m downstream of the target (at the station of the PS-ACOL irradiation facility [72]) as well as by a RadMon detector in a quite peripheral position, as indicated in Fig. 14, matched FLUKA results very well. In the first location, 66–80 mGy per pulse of  $1.4 \times 10^{13}$  protons compared with a 68–70 mGy prediction. In the RadMon location, a reading of 50 ( $\pm 15$ ) Gy over 14 weeks compared with a 60 Gy prediction (see Fig. 14). The same RadMon yields a thermal neutron to high-energy hadron fluence ratio of 5 ( $\pm 40\%$ ), in full agreement with the simulation outcome of 5 ( $\pm 10\%$ ).



**Fig. 14:** Dose map of the antiproton decelerator target area (top view), as calculated by FLUKA and normalized to the proton beam intensity accumulated over the 14 weeks of the indicated period [71]. Along the missing axis ( $Y$ ), values are averaged over a 10 cm interval, at the height corresponding to the indicated RadMon position. The 26 GeV/c proton beam impinges the iridium target at the reference frame origin from the left.

The CNGS incident led to the careful evaluation of all electronic systems located in the LHC underground areas, typically either fully commercial or based on ‘commercial-off-the-shelf’ components, and of the respective radiation levels. An extensive mitigation strategy, consisting of relocation to safe areas, as well as suitable shielding design and installation, allowed minimization of the single-event-effect impact on the accelerator operation. In fact, single-event-effect-induced downtime decreased from an initial value of 400 h in 2011 to 250 h in 2012, reducing the single-event-effect dump rate referred to cumulated luminosity by a factor of four (from 12 to 3 dumps per inverse femtobarn) [70]. This remarkable achievement has still to be dramatically improved during the new LHC run (Run II) and especially the high-luminosity era. To this purpose, and in the context of any other project implying a challenging radiation environment, a prevention strategy has to be implemented from the early stage onward, entailing the availability of protected areas, possibly relying on a dedicated shielding solution, for electronic equipment not validated by radiation testing, together with the development and adoption of radiation-tolerant and radiation-hardened electronics.

### 3.7 Shielding

In an accelerator context, typical radiation sources are represented by regular and accidental beam impacts on beam-intercepting devices, such as collimators, dumps, targets, or even unexpected obstacles, for instance plastic and metallic dust [73]. In the case of rings, nuclear reactions between beam particles and residual gas nuclei inside the vacuum chamber play an additional role. With electron and positron beams, synchrotron radiation becomes a main concern (it can carry an important power with hadron beams too, but the photon spectrum is much softer and is absorbed by the first material layers). Finally, colliders are affected with beam–beam collision debris around experimental insertions.

For a given source term, the induced radiation levels first depend on the relative position, namely on the radial and longitudinal distance from the shower generation. Clearly, the geometrical attenuation (proportional to the square of the distance in the case of an isotropic source) is often insufficient and, to allow the integration of a radiation facility in the environment, to guarantee accessible areas, or to ensure the correct operation of the electronic equipment (see Section 3.6), a specific shielding has to be designed.

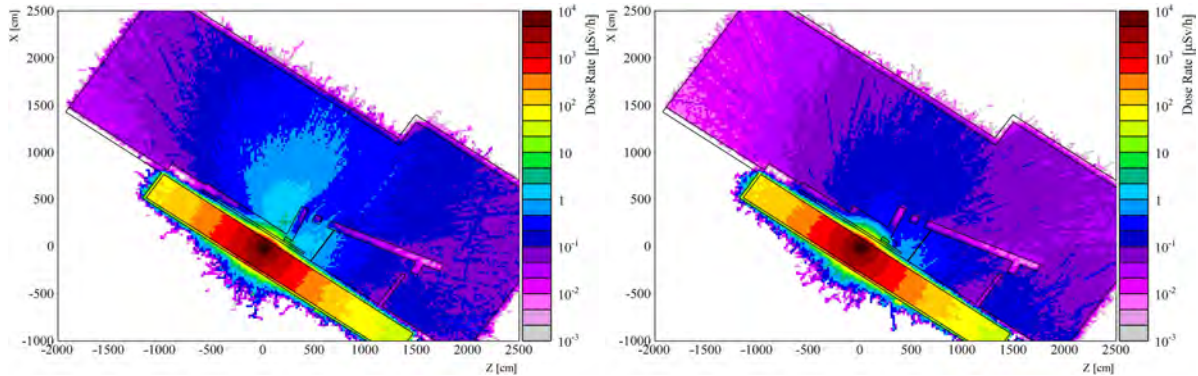
*High energy hadron* propagation is limited (in addition to ionization losses affecting charged species) by non-elastic reactions, replacing the primary particle with lower-energy products. Their occurrence is proportional to the inverse of the nuclear interaction length, i.e., to the density of the traversed material. Therefore, such a radiation field is effectively attenuated by dense materials. When coupled to cost considerations favouring cheap options, this typically suggests the use of iron or bigger volumes of concrete for massive shielding.

At large radial distances from the primary source, beyond a considerable material amount, the radiation field starts to be dominated by *low-energy neutrons*. These are further slowed down by nuclear elastic scattering, more effectively on light nuclei, as in hydrogen-rich materials. Approaching thermal energies, they are removed in the presence of particular isotopes with pronounced capture cross-sections, such as  $^{10}\text{B}$  or  $^{113}\text{Cd}$ . In this context, borated polyethylene is a common solution to wash out the neutron field.

Even in the case of energetic hadron sources, energy deposition is ruled by electromagnetic showers initiated by high-energy *photons* from neutral pion decay. In peripheral areas, photons accompany low-energy neutron propagation, being produced in non-elastic reactions, e.g., radiative capture. With electron beams, bremsstrahlung and synchrotron radiation make photons play a crucial role. At energies above a certain threshold (of the order of 10 MeV in lead and 100 MeV in carbon), photon interaction consists of electron and positron pair production and hence electromagnetic shower development, with further photon generation by lepton bremsstrahlung. Going below that threshold, photons mainly lose energy through Compton scattering, and eventually are absorbed by the photoelectric effect, which is strongly favoured in high- $Z$  materials.

Concerning *high-energy muons*, typically produced in pion and kaon decays, as mentioned at the beginning of this paper, they cannot be stopped within limited distances, since they are affected predominantly by ionization and multiple Coulomb scattering. Bremsstrahlung and direct  $e^+e^-$  pair production rule their transport at energies larger than 1 TeV.

For radioprotection purposes, depending on the aspects to be considered, particle fluence in a given location is transformed into *effective dose* or *ambient dose equivalent* (both expressed in sieverts), through the use of respective sets of conversion coefficients, which are a function of particle type and energy [74, 75]. *Prompt* dose equivalent outside a radiation facility, reflecting the relevant radiation level in a public space during normal or accidental operation of the facility, is the quantity to minimize below acceptable limits through the facility shielding design.

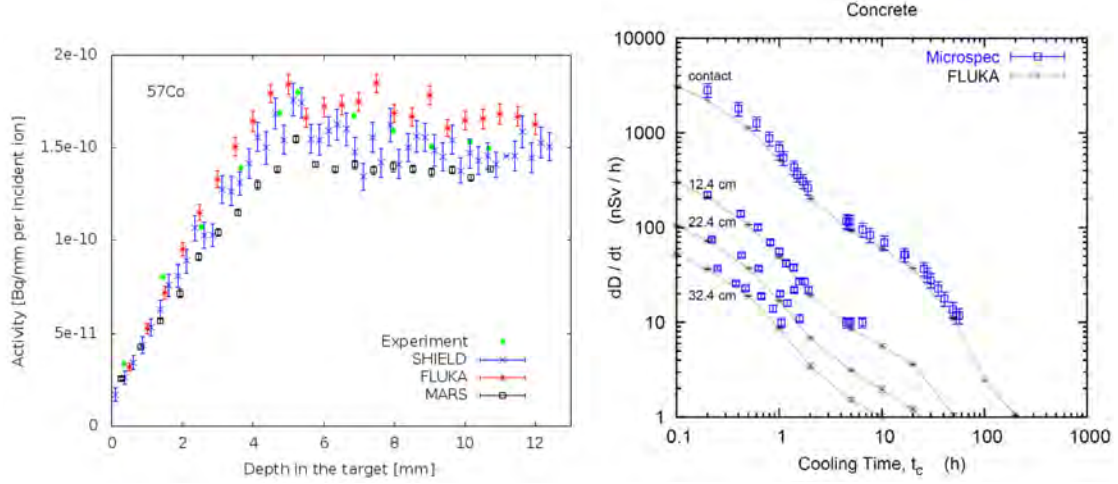


**Fig. 15:** Prompt ambient dose-equivalent maps (top view) in the intersecting storage rings tunnel on the side of the n\_TOF EAR-2 neutron beam line, directed along the missing  $Y$  axis at the origin of the reference frame (courtesy of J. Vollaire). Left: with default shielding, as described in the text. Right: with improved shielding, as described in the text.

As an example, Fig. 15 shows the simulated effect of the shielding optimization performed during the construction of the second experimental area (EAR-2) of the n\_TOF facility at CERN [76]. The Proton Synchrotron proton beam impinges at 20 GeV/c, with  $7 \times 10^{12}$  protons per pulse and an average current of  $1.6 \times 10^{12}$  p/s, on the n\_TOF spallation target (a massive lead block), generating two neutron beams, which reach EAR-1 and EAR-2, respectively. The first is located after a 185 m horizontal flight path along the proton beam direction, whereas the new area was built at  $90^\circ$ , 20 m above the lead target, to provide a significantly higher neutron flux. Along the path towards EAR-2, the neutron beam line runs contiguous to a tunnel hosting a workplace, just behind a 60 cm thick concrete wall, where strict limits apply in terms of prompt ambient dose equivalent. With default shielding consisting of a concrete block 80 cm thick and 2 m high, neutron streaming induces a dose-equivalent rate exceeding few  $\mu\text{Sv/h}$  (Fig. 15 left), implying that the area should be classified, from the radioprotection standpoint [77]. Nevertheless, with the extension of the concrete protection, whose volume was increased almost three times, and the addition of two iron plates, a quite significant improvement was achieved, enabling the  $0.5 \mu\text{Sv/h}$  limit to be complied with (Fig. 15 right), as nicely confirmed by later radiation monitor measurements (J. Vollaire, private communication).

### 3.8 Activation

As anticipated in Section 2.2, material activation is responsible for continuous delayed emissions, defining the radiation conditions of a facility during its shut-down periods. This also limits access and intervention possibilities during beam absence, and affects equipment handling, e.g., waste disposal. Therefore, activation levels have to be evaluated since the design stage. Calculation reliability mainly depends on accuracy in radionuclide production (see Fig. 5) and on knowledge of actual material compositions. Various activation benchmark experiments have been performed [78–80]. As an example, Fig. 16 (left) shows the activity profile of  $^{57}\text{Co}$  generated along a copper target by a 500 MeV/n  $^{238}\text{U}$  beam [80], penetrating a distance of about 5 mm. As suggested by the measured shape, which is well reproduced by simulation codes, the considered nuclide is mainly produced in secondary neutron re-interactions. Figure 16 (right) illustrates the role of isotopes of different lifetimes. In this case, various samples were put in the vicinity of a copper target irradiated by a 120 GeV/c proton and positive pion beam and then transferred to a low-background laboratory, to measure the time evolution of residual dose rates at several distances from the sample [79]. Values scale down with increasing distance, as predicted, and, for the concrete sample considered here, the time profile is shaped by  $^{11}\text{C}$  decay (positron emission with  $t_{1/2} = 20$  min) for the very first few hours and by  $^{24}\text{Na}$  decay ( $\beta^-$  transmutation into  $^{24}\text{Mg}$  generating two  $\gamma$  lines with  $t_{1/2} = 15$  h) later on.



**Fig. 16:** Left: Activity profile of  $^{57}\text{Co}$  in a copper target hit by a 500 MeV/n  $^{238}\text{U}$  beam [80]. Experimental data are compared with the predictions of the indicated codes. Right: Time evolution of residual ambient dose-equivalent rates at different distances from a concrete sample, previously exposed to prompt radiation emerging from a nearby copper target hit by 120 GeV/c protons and positive pions [79]. Blue symbols are experimental data and black symbols connected by lines are FLUKA results.

The calculation of 3D residual dose maps, coupled with an intervention plan detailing the position of the workers and the duration of their actions, allows for the evaluation of individual and collective doses, to be compared with legal limits, design limits (required as facility design criteria not to surpass a given dose per intervention and per year), and optimization thresholds (imposing, if exceeded, optimization of the intervention plan, to minimize the dose to personnel according to the ‘as low as reasonably achievable’ principle). In this regard, a few guidelines concerning material choice, material amount, and equipment handling, should be considered at the beginning of every new project: whenever possible, lower-activation, radiation-resistant, more easily disposable materials must be preferred; only essential components should be installed, in particular in high-loss areas, and they must be easily accessible and enable fast installation, maintenance, repair, and dismantling.

A special aspect is represented by air activation, which has to be considered from the points of view of release in the environment and accessibility delay of an irradiated area. The activity of a certain air radioisotope inside the latter at the end of the irradiation period  $T$  is given by

$$A_T = A_S (1 - \exp(-(\lambda + m_{\text{on}})T)), \quad (3)$$

where  $\lambda$  is the radioisotope’s decay probability per unit time and  $m_{\text{on}}$  is the relative air exchange rate during irradiation, i.e., the fraction of the total air volume in the area that is renewed per unit time. The quite low interaction probability of particles in air might limit the Monte Carlo statistical accuracy of air radioisotope production; hence, an alternative two-step method involves scoring the energy distribution of hadron fluence in air and then folding it with the cross-sections for radioisotope production from the air target nuclei. In this way, the saturation activity  $A_S$  can be calculated as [81]

$$A_S = \frac{V\lambda}{\lambda + m_{\text{on}}} \sum_{P,T,j} \phi_P(E_j) \sigma_{P,T}(E_j) N_T (\Delta E)_{j,P}, \quad (4)$$

where the summation has to be performed over the produced hadron species  $P$  (mainly neutrons, protons, and charged pions), the air nuclear species  $T$  ( $^{12}\text{C}$ ,  $^{14}\text{N}$ ,  $^{16}\text{O}$ , and  $^{40}\text{Ar}$ ), and all the bins  $j$  of width  $\Delta E$  into which the hadron energy range has been divided.  $V$  is the irradiated air volume,  $\phi$  is the differential fluence rate,  $\sigma$  is the production cross-section for the considered radioisotope, and  $N_T$  is the number of target nuclei per unit volume, calculated from the air composition.

For each air radioisotope, the total amount of activity released into the atmosphere during the irradiation period,  $T$ , is

$$A_{\text{on}} = m_{\text{on}} A_S \left( T - \frac{1 - \exp(-(\lambda + m_{\text{on}})T)}{\lambda + m_{\text{on}}} \right) \exp(-\lambda t_{\text{on}}) , \quad (5)$$

where  $t_{\text{on}}$  is the time taken by the air flux to reach the release point from the irradiated area where air is being activated. Finally, the total amount of activity released into the atmosphere after shut-down can be obtained by

$$A_{\text{off}} = A_T \frac{m_{\text{off}}}{\lambda + m_{\text{off}}} \exp(-\lambda t_{\text{off}}) , \quad (6)$$

with  $m_{\text{off}}$  and  $t_{\text{off}}$  representing the same quantities as  $m_{\text{on}}$  and  $t_{\text{on}}$ , respectively, but referred to the period following the irradiation end.

### 3.9 Simulation tools in challenging accelerator applications

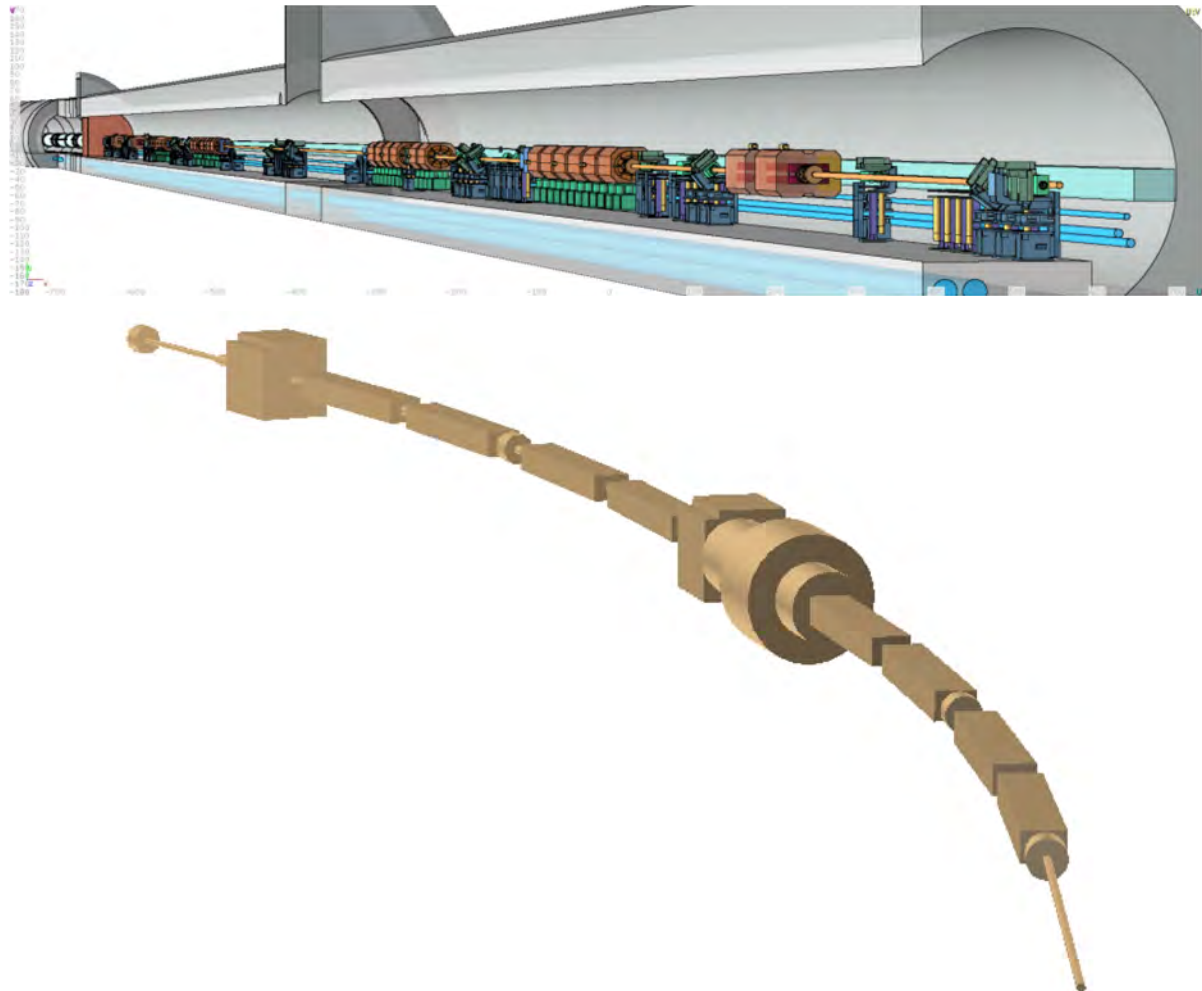
The challenge here is to produce a detailed and accurate (to a percent level) model of all particle interactions with 3D system components (up to tens of kilometres of the accelerator lattice in some cases) in the energy region spanning up to 20 decades, as a basis of accelerator, detector, and shielding designs and their performance evaluation, for both short-term and long-term effects.

The current versions of five general-purpose all-particle codes are capable of this: FLUKA, GEANT4, MARS15, MCNP6, and PHITS. These are used extensively worldwide for accelerator applications in conjunction with such accelerator tracking codes as STRUCT [54] and SixTrack [82–84]. A substantial amount of effort (up to several hundreds of person-years) has been put into development of these codes over the last few decades. The user communities for the codes reach several thousands of people worldwide. The five codes listed above can handle a very complex geometry, have powerful user-friendly built-in graphical user interfaces with magnetic field and tally viewers, and variance reduction capabilities. Tallies include volume and surface distributions (1D to 3D) of particle flux, energy, reaction rate, energy deposition, residual nuclide inventory, prompt and residual dose equivalent, number of displacements per target atom for radiation damage, event logs, intermediate source terms, etc. All the aspects of beam interactions with accelerator system components are addressed in sophisticated Monte Carlo simulations benchmarked—wherever possible—with dedicated beam tests.

In accelerator applications, particle shower simulations lie in a multidisciplinary field, in which particle dynamics in accelerators and radiation-matter interaction play together. In fact, their source term is often provided through multiturn tracking in accelerator rings, which requires dedicated codes, eventually dumping, in static loss files, a beam particle sample characterized in the phase space at a certain interface. Conversely, tracking codes happen to be faced with the problem of dealing with particle scattering in beam-intercepting devices, such as collimators. Innovative solutions adopt different types of on-line coupling between tracking and interaction codes, which exchange particle run times to perform their respective tasks. There has been a quantum leap in coupling these general-purpose codes with tracking codes for accelerators:

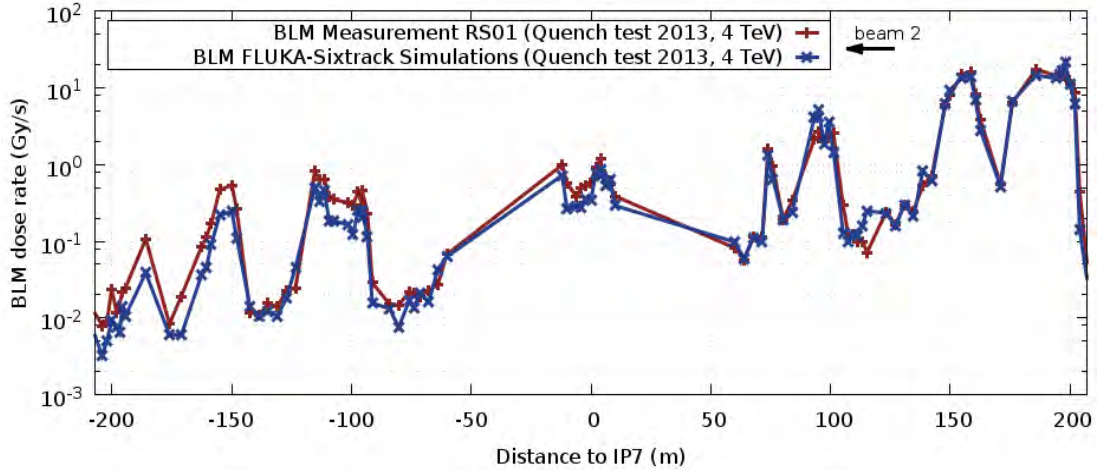
- MMBLB = MAD-MARS Beam Line Builder (since 2000) [85]. Earlier used the STRUCT code [54] tracking; currently coupled with MAD-X [86, 87], and available in the ROOT-based version [2, 50].
- BDSIM combines C++ in-vacuum accelerator style particle tracking and GEANT4 physics (since the mid-2000s) [88].
- FLUKA LineBuilder and Element Database [89] and active coupling to SixTrack; the two codes communicate with each other through a network port [90].

Figure 17 shows two examples of the complexity affordable nowadays in accelerator line modelling for beam-machine interaction studies, which is coupled to the required accuracy in geometry detail implementation (from vacuum chamber and collimator aperture to magnet coil structure and radiation monitor positioning) as well as in magnetic field treatment and scoring resolution.



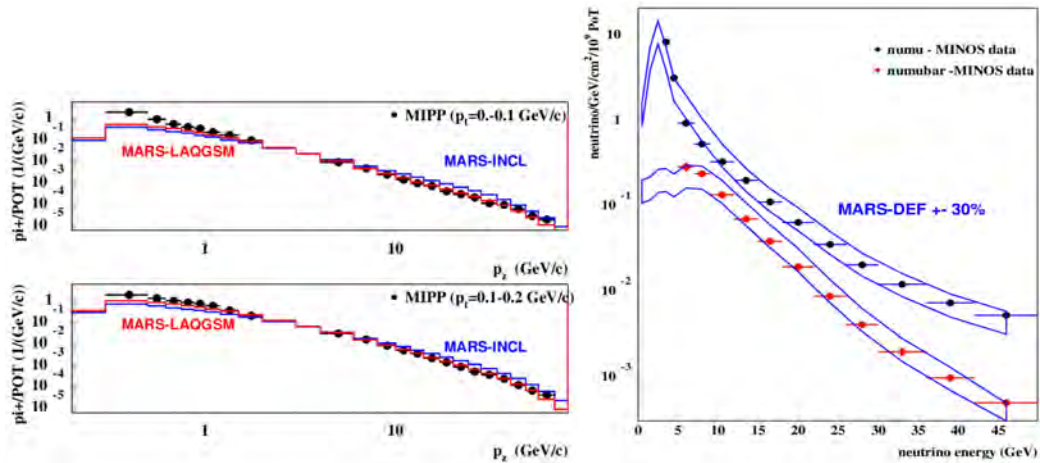
**Fig. 17:** Top: FLUKA geometry of the LHC betatron cleaning insertion (interaction region 7) by the FLUKA LineBuilder and Element DataBase [89]. Bottom: MARS15 geometry of the Fermilab Booster by the ROOT-Based MMBLB [2, 50].

Towards the end of the LHC run 1 (early 2013), several quench tests were performed, to explore the actual quench limits of the superconducting magnets and verify the quality of theoretical calculations [91]. In particular, during a so-called collimation quench test [92, 93] at a beam energy of 4 TeV, the horizontal primary collimator of the betatron cleaning insertion was impacted by a peak proton loss rate equivalent to about 1 MW for 1 s, with no quench occurring in the downstream dispersion suppressor. The propagation of the induced particle shower was measured by the beam loss monitor system [94], giving a picture of the energy deposition profile over several hundred metres, as shown in Fig. 18. This provided a very challenging opportunity to benchmark the adopted SixTrack-FLUKA simulation chain [95], which yielded the impressive agreement reported in the figure, both in terms of pattern and absolute signal comparison, spanning a few orders of magnitude.



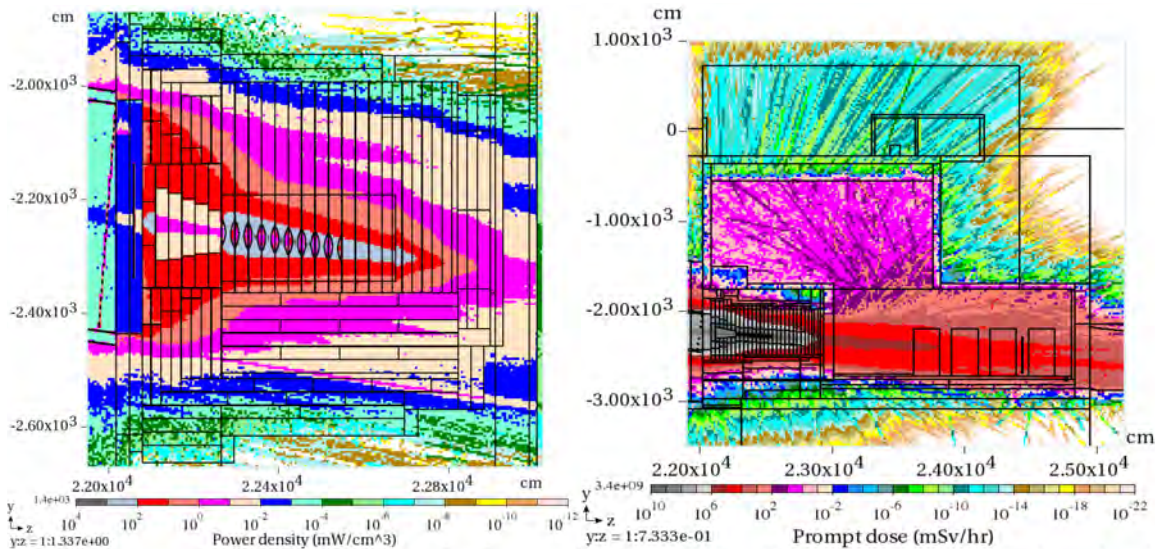
**Fig. 18:** Absolute beam loss monitor signal pattern at the peak loss rate of the 2013 LHC collimation quench test, averaged over the shortest available time interval of  $40 \mu\text{s}$  (RS01): data (red) are compared with predictions (blue) by the SixTrack-FLUKA coupling according to the simulation strategy discussed in Ref. [95].

All the power and the described features of the current version of the MARS15 code are fully exploited in the new neutrino and precision fixed-target experiments for the Intensity Frontier under design and construction in the USA, such as Mu2e and the Long Baseline Neutrino Facility/DUNE. The code was thoroughly benchmarked against the data at corresponding proton beam energies of 8 GeV (for Mu2e) and 120 GeV (for the Long Baseline Neutrino Facility/DUNE). As an example, Fig. 19 shows that MARS15 calculations agree very well with experimental data for a 120 GeV beam on a thick graphite target for pion spectra (left) and neutrino and antineutrino spectra (right) at the MINOS (Main Injector Neutrino Oscillation Search) near detector. Results of the MARS15 calculations for a 2.4 MW beam were used to optimize the design of the Long Baseline Neutrino Facility hadron absorber at the end of the 200 m long decay channel, as well as complicated radiation shielding around the absorber, along with the configuration of all the conventional facilities, as shown in Fig. 20. Note that the calculated power density profiles span 12 orders of magnitude in the region of interest, while for the prompt dose it was necessary to cover 24 decades, which would be impossible without applying the corresponding variance reduction and other sophisticated Monte Carlo techniques.



**Fig. 19:** MARS15 results versus recent MIPP (main injector particle production) NuMI target data on (left) pion spectra and (right) MINOS (Main Injector Neutrino Oscillation Search) neutrino and antineutrino spectra at the near detector. Results are normalized per proton on target indicated as POT and PoT.





**Fig. 20:** Left: MARS15-calculated power density map in the Long Baseline Neutrino Facility hadron absorber. Right: MARS15-calculated prompt dose distribution in the entire hadron absorber complex.

A very challenging, and at the same time exciting, application was a muon collider project in which the design energy of muon beams varied over several years from 62.5 GeV to 3 TeV. As described in Ref. [96] for the Higgs Factory muon collider, a detailed 3D model was built using MARS15 for the entire collider ring, including the interaction region, the chromaticity correction and matching sections, the arc, the machine-detector interface, and the SiD-like collider detector, with the silicon vertex detector and tracker based on the design proposed for the Compact Muon Solenoid detector upgrade.

Figure 21 shows the 3D model, while Fig. 22 shows the components in the machine-detector interface region. At a muon energy of 62.5 GeV with  $2 \times 10^{12}$  muons per bunch, the electrons from muon decays deposit more than 300 kW in the superconducting magnets of the Higgs Factory interaction region and storage ring. This heat deposition corresponds to an unprecedented average dynamic heat load of 1 kW/m around the 300 m long ring, or a multimewatt room temperature equivalent, if the heat is deposited at helium temperature. That is about one hundred times above acceptable levels. The detector backgrounds in such a project are also much too excessive. The suppression needed on both the fronts has been achieved through substantial effort. First, the lattice and magnets—with a dipole component in the interaction region quadrupole magnets and large apertures varying along the lattice—were designed appropriately. To further protect the collider, thick tungsten masks and liners (with tight elliptical apertures varying according to the beam envelope) in the magnet interconnect regions and inside each magnet have been optimized using massive iterative MARS15 studies. The configuration and composition of a sophisticated tungsten nozzle in the vicinity of the interaction point and other details of the machine-detector interface were optimized simultaneously. As a result, the average dynamic heat load on the superconducting coils of  $\sim 1$  kW/m was reduced to the allowable value of 10 W/m at 4.5 K, with the peak power density in the coils being reduced to below the quench limit, with a necessary safety margin [96]. The detector backgrounds were also reduced adequately [97].

Various examples of design and operation of ambitious research facilities have been discussed in this paper. All of them required a thorough consideration of the specifics of particle-matter interactions and corresponding physics processes in the phase space regions of interest, understanding of the beam-induced microscopic and macroscopic effects in the components, close iterative work with the lattice, magnet and detector designers, and use of the modern state-of-the-art simulation tools.

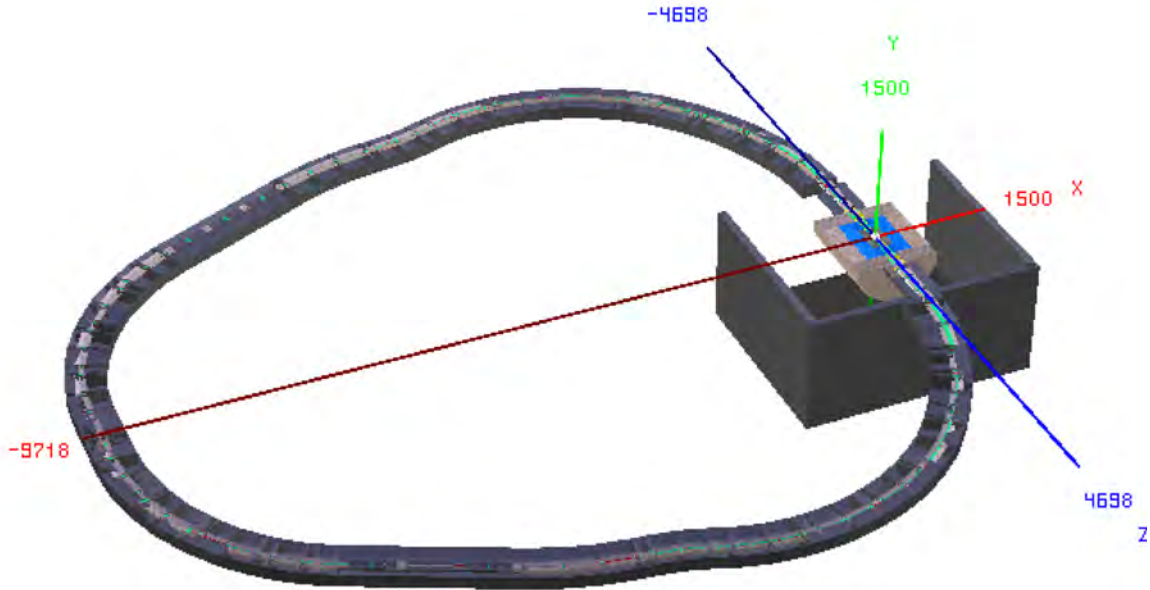


Fig. 21: MARS15 3D model of the Higgs Factory muon collider

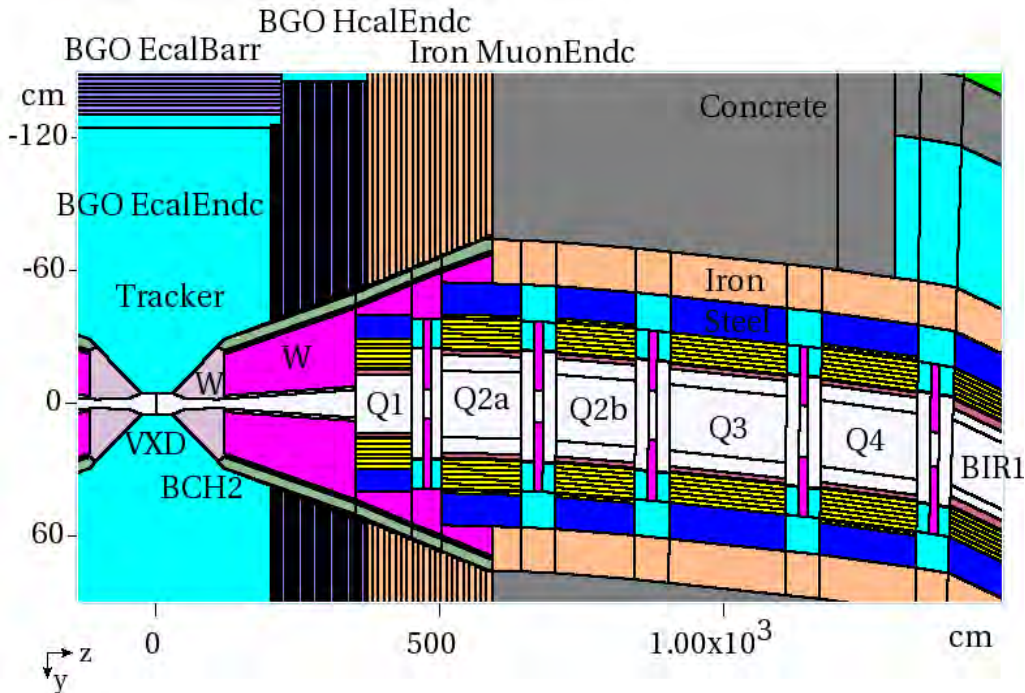


Fig. 22: Higgs Factory machine-detector interface MARS15 model with tungsten nozzles on each side of the interaction point, tungsten masks in interconnect regions and tungsten liners inside each magnet [96]. BCH2, borated polyethylene layer; BGO HcalEndc, Bismuth Germanate hadron endcap calorimeter; BGO EcalBarr, Bismuth Germanate electromagnetic barrel calorimeter; BGO EcalEndc, Bismuth Germanate electromagnetic endcap calorimeter; BIR1, first dipole magnet in the interaction region; MuonEndc, muon endcap detector; Q, quadrupole magnets; W, tungsten; VXD, vertex detector.

### Acknowledgements

We wish to thank many CERN and Fermilab colleagues for their very valuable contributions, in particular M. Brugger, A. Ferrari, K. Gudima, I. Rakhno, S. Roesler, S. Striganov, I. Tropin, and V. Vlachoudis.

## References

- [1] N.V. Mokhov, Proc. Workshop on Shielding Aspects of Accelerators, Targets and Irradiation Facilities (SATIF-10), Geneva, Switzerland, OECD NEA No. 6898, 2010, p. 105.
- [2] N.V. Mokhov, *Rev. Accel. Sci. Technol.* **6** (2013) 275.  
<http://dx.doi.org/10.1142/S1793626813300132>
- [3] A. Ferrari *et al.*, FLUKA: a multi-particle transport code, CERN-2005-010 (2005).
- [4] G. Battistoni *et al.*, *Ann. Nucl. Energy* **82** (2015) 10–18.  
<http://dx.doi.org/10.1016/j.anucene.2014.11.007>
- [5] FLUKA, <http://www.fluka.org>
- [6] S. Agostinelli *et al.*, *Nucl. Instr. Methods A* **506**(3) (2003) 250–303.  
[http://dx.doi.org/10.1016/S0168-9002\(03\)01368-8](http://dx.doi.org/10.1016/S0168-9002(03)01368-8)
- [7] J. Allison *et al.*, *IEEE Trans. Nucl. Sci.* **53**(1) (2006) 270–278.  
<http://dx.doi.org/10.1109/TNS.2006.869826>
- [8] Geant4, <http://geant4.cern.ch>
- [9] N.V. Mokhov, The MARS Code System User’s Guide, Fermilab-FN-628 (1995).
- [10] N.V. Mokhov and S.I. Striganov, *AIP Conf. Proc.* **896** (2007) 50.  
<http://dx.doi.org/10.1063/1.2720456>
- [11] N. Mokhov *et al.*, *Prog. Nucl. Sci. Technol.* **4** (2014) 496–501.  
<http://dx.doi.org/10.15669/pnst.4.496>
- [12] MARS Code System, <http://www-ap.fnal.gov/MARS/>
- [13] T. Goorley *et al.*, *Nucl. Technol.* **180**(3) (2012) 298–315. <http://dx.doi.org/10.13182/NT11-135>
- [14] Monte Carlo Code Group, A General Monte Carlo N-Particle (MCNP) Transport Code, <http://mcnp.lanl.gov>
- [15] K. Niita *et al.*, *Radiat. Meas.* **41**(9–10) (2006) 1080–1090.  
<http://dx.doi.org/10.1016/j.radmeas.2006.07.013>
- [16] PHITS, Particle and Heavy Ion Transport Code System, <http://phits.jaea.go.jp/>
- [17] A.N. Kalinovskii *et al.*, *Passage of High-Energy Particles through Matter* (American Institute of Physics, New York, 1989).
- [18] J. Beringer *et al.* (Particle Data Group), *Phys. Rev. D* **86**(1) (2012) 010001.  
<http://dx.doi.org/10.1103/PhysRevD.86.010001>
- [19] A. Ferrari and P.R. Sala, The Physics of High Energy Reactions, Proc. Workshop on Nuclear Reaction Data and Nuclear Reactors Physics, Design and Safety, Trieste, April 1996, Ed. A. Gandini and G. Reffo (1998), p. 424.
- [20] S. Huber and J. Aichelin, *Nucl. Phys. A* **573**(4) (1994) 587–625. [http://dx.doi.org/10.1016/0375-9474\(94\)90232-1](http://dx.doi.org/10.1016/0375-9474(94)90232-1)
- [21] A. Engel *et al.*, *Nucl. Phys. A* **572**(3–4) (1994) 657–681. [http://dx.doi.org/10.1016/0375-9474\(94\)90405-7](http://dx.doi.org/10.1016/0375-9474(94)90405-7)
- [22] S. Tei *et al.*, *Z. Phys. A* **356** (1997) 421. <http://dx.doi.org/10.1007/s002180050198>
- [23] A. Capella *et al.*, *Phys. Rep.* **236**(4–5) (1994) 225–329. [http://dx.doi.org/10.1016/0370-1573\(94\)90064-7](http://dx.doi.org/10.1016/0370-1573(94)90064-7)
- [24] R.J. Glauber and G. Matthiae, *Nucl. Phys. B* **21**(2) (1970) 135–157.  
[http://dx.doi.org/10.1016/0550-3213\(70\)90511-0](http://dx.doi.org/10.1016/0550-3213(70)90511-0)
- [25] R.J. Glauber, in *Lectures in Theoretical Physics*, Eds. A.O. Barut, and W.E. Brittin (Interscience, New York, 1959), Vol. 1, p.315.

- [26] V.N. Gribov, *Sov. Phys. JETP* **29** (1969) 483.
- [27] V.N. Gribov, *Sov. Phys. JETP* **30** (1970) 709.
- [28] L. Bertocchi, *Nuovo Cimento* **11A**(1) (1972) 45–62. <http://dx.doi.org/10.1007/BF02722777>
- [29] D.R.F. Cochran *et al.*, *Phys. Rev. D* **6**(11) (1972) 3085.  
<http://dx.doi.org/10.1103/PhysRevD.6.3085>
- [30] I. Chemakin *et al.*, *Phys. Rev. C* **65**(2) (2002) 024904.  
<http://dx.doi.org/10.1103/PhysRevC.65.024904>
- [31] E. Gadioli and P.E. Hodgson, *Pre-equilibrium Nuclear Reactions* (Clarendon Press, Oxford, 1992).
- [32] V.F. Weisskopf, *Phys. Rev.* **52**(4) (1937) 295. <http://dx.doi.org/10.1103/PhysRev.52.295>
- [33] T. Enqvist *et al.*, *Nucl. Phys. A* **686**(1–4) (2001) 481–524. [http://dx.doi.org/10.1016/S0375-9474\(00\)00563-7](http://dx.doi.org/10.1016/S0375-9474(00)00563-7)
- [34] S.G. Mashnik *et al.*, *J. Nucl. Radiochem. Sci* **6** (2005) A1–A19.  
[http://dx.doi.org/10.14494/jnrs2000.6.2\\_A1](http://dx.doi.org/10.14494/jnrs2000.6.2_A1)
- [35] K.A. Olive *et al.* (Particle Data Group), *Chin. Phys. C* **38**(9) (2014) 090001.  
<http://dx.doi.org/10.1088/1674-1137/38/9/090001>
- [36] M. Fama *et al.*, *Nucl. Instr. Methods B* **193**(1–4) (2002) 91–96. [http://dx.doi.org/10.1016/S0168-583X\(02\)00732-2](http://dx.doi.org/10.1016/S0168-583X(02)00732-2)
- [37] H. Whitlow *et al.*, *Nucl. Instr. Methods B* **195**(1–2) (2002) 133–146.  
[http://dx.doi.org/10.1016/S0168-583X\(02\)00946-1](http://dx.doi.org/10.1016/S0168-583X(02)00946-1)
- [38] G. Konac *et al.*, *Nucl. Instr. Methods B* **136–138** (1998) 159–165.  
[http://dx.doi.org/10.1016/S0168-583X\(98\)80016-5](http://dx.doi.org/10.1016/S0168-583X(98)80016-5)
- [39] H. Ikeda *et al.*, *Nucl. Instr. Methods B* **115**(1–4) (1996) 34–38. [http://dx.doi.org/10.1016/0168-583X\(95\)01511-6](http://dx.doi.org/10.1016/0168-583X(95)01511-6)
- [40] P. Mertens and P. Bauer, *Nucl. Instr. Methods B* **33**(1–4) (1988) 133–137.  
[http://dx.doi.org/10.1016/0168-583X\(88\)90530-7](http://dx.doi.org/10.1016/0168-583X(88)90530-7)
- [41] J. Ziegler, *J. Appl. Phys.* **85**(3) (1999) 1249. <http://dx.doi.org/10.1063/1.369844>
- [42] S. Molinelli *et al.*, *Phys. Med. Biol.* **58**(11) (2013) 3837. <http://dx.doi.org/10.1088/0031-9155/58/11/3837>
- [43] G.H. Kinchin and R.S. Pease, *Rep. Prog. Phys.* **18** (1955) 1. <http://dx.doi.org/10.1088/0034-4885/18/1/301>
- [44] M.J. Norgett *et al.*, *Nucl. Eng. Design* **33**(1) (1975) 50–54. [http://dx.doi.org/10.1016/0029-5493\(75\)90035-7](http://dx.doi.org/10.1016/0029-5493(75)90035-7)
- [45] R.E. Stoller, *J. Nucl. Mat.* **276**(1–3) (2000) 22–32. [http://dx.doi.org/10.1016/S0022-3115\(99\)00204-4](http://dx.doi.org/10.1016/S0022-3115(99)00204-4)
- [46] N.V. Mokhov *et al.*, Simulation and verification of DPA in materials, Proc. Workshop on Appl. High Intensity Proton Accel. (World Scientific, Singapore, 2010) pp. 128–131.  
[http://dx.doi.org/10.1142/9789814317290\\_0015](http://dx.doi.org/10.1142/9789814317290_0015)
- [47] F. Cerutti, Radiation damage calculation by FLUKA, 2nd Workshop on Radiation Effects in Superconducting Magnet Materials, KEK, Tsukuba (2013).
- [48] M.B. Chadwick *et al.*, *Nucl. Data Sheets* **107**(12) (2006) 2931–3060.  
<http://dx.doi.org/10.1016/j.nds.2006.11.001>
- [49] R.E. MacFarlane *et al.*, LANL Report LA-12740-M (1994).
- [50] N. Mokhov *et al.*, *Progress Nucl. Sci. Technol.* **4** (2014) 496–501.  
<http://dx.doi.org/10.15669/pnst.4.496>

- [51] C.H.M. Broeders and A.Y. Konobeyev, *J. Nucl. Mat.* **328**(2–3) (2004) 197–214.  
<http://dx.doi.org/10.1016/j.jnucmat.2004.05.002>
- [52] J. Morgan, P-bar Production Target Performance in Run II, Fermilab P-bar Technical Note No. 683 (2003).
- [53] A.I. Drozhdin *et al.*, Beam-induced damage to the tevatron collimators: analysis and dynamic modeling of beam loss, energy deposition and ablation, Fermilab-FN-751 (2004).
- [54] A.I. Drozhdin and N.V. Mokhov, STRUCT User’s Reference Manual, SSCL-MAN-0034 (1994),  
<http://www-ap.fnal.gov/users/drozhdin/STRUCT/>
- [55] D.C. Wilson *et al.*, *IEEE Proc. Part. Accel. Conf.* (1993) 3090.
- [56] N.A. Tahir *et al.*, *New J. Phys.* **10** (2008) 073028.  
<http://dx.doi.org/10.1088/1367-2630/10/7/073028>
- [57] N.A. Tahir *et al.*, *Phys. Rev. ST Accel. Beams* **15** (2012) 051003.  
<http://dx.doi.org/10.1103/PhysRevSTAB.15.051003>
- [58] M. Scapin *et al.*, *Comput. Structures* **141** (2014) 74–83.  
<http://dx.doi.org/10.1016/j.compstruc.2014.05.008>
- [59] H. Schonbacher, in *Handbook of Accelerator Physics and Engineering*, 2nd edn, Eds. A.W. Chao *et al.* (World Scientific, Singapore, 2013), p. 793.
- [60] A. Idesaki, Irradiation effects of gamma-rays on cyanate ester/epoxy resins, Second Workshop on Radiation Effects in Superconducting Magnet Materials, KEK, Tsukuba (2013).
- [61] N.V. Mokhov *et al.*, *Phys. Rev. ST Accel. Beams* **18** (2015) 051001.  
<http://dx.doi.org/10.1103/PhysRevSTAB.18.051001>
- [62] A. Zlobin *et al.*, Proc. European Particle Accelerator Conf. 2002, MOPLE017 (2002) 2451.
- [63] I. Novitski and A.V. Zlobin, *IEEE Trans. Appl. Supercond.* **17**(2) (2007) 1059–1062.  
<http://dx.doi.org/10.1109/TASC.2007.898532>
- [64] P.P. Granieri and R. van Weelden, *IEEE Trans. Appl. Supercond.* **24** (2014) 4802806.
- [65] T. Peterson, Magnet quench, *Symmetry Magazine*, Fermilab/SLAC, November 2008.  
<http://www.symmetrymagazine.org/article/november-2008/explain-it-in-60-seconds-magnet-quench>
- [66] Interim Summary Report on the Analysis of the 19 September 2008 Incident at the LHC, CERN/AT/PhL, EDMS 973073 (2008).
- [67] A. Ferrari *et al.*, CERN-AB-Note-2006-038, EDMS 745389 (2006).
- [68] G. Spiezia *et al.*, *Proc. Sci.* **024** (2011) 1–12.
- [69] K. Røed *et al.*, *IEEE Trans. Nucl. Sci.* **58**(3) (2011) 932–938.  
<http://dx.doi.org/10.1109/TNS.2010.2097605>
- [70] C. Adorisio *et al.*, HL-LHC Preliminary Design Report, Ed. G. Apollinari *et al.*, CERN-ACC-2014-0300 (2014).
- [71] M. Calviani and E. Nowak, TS-Note-2012-069 (TECH), EDMS 1235761 (2012).
- [72] M. Tavlet and M.E. Leon Florian, 1st European Conference on Radiation and its Effects on Devices and Systems, RADECS 91 (IEEE, 1991) p. 582–585,  
<http://dx.doi.org/10.1109/RADECS.1991.213533>  
<http://dx.doi.org/10.1109/RADECS.1991.213533>
- [73] T. Baer *et al.*, Proc. IPAC 2012, New Orleans, 2012, p. 3936.
- [74] M. Pelliccioni, *Radiat. Prot. Dosim.* **88**(4) (2000) 279–297.  
<http://dx.doi.org/10.1093/oxfordjournals.rpd.a033046>
- [75] S. Roesler and G.R. Stevenson, CERN-SC-2006-070-RP-TN, EDMS 809389 (2006).

- [76] E. Chiaveri *et al.*, CERN-INTC-2012-029 (2012).
- [77] I. Bergström, M.Sc. thesis, Luleå tekniska universitet, 2009.
- [78] M. Brugger *et al.*, *Radiat. Prot. Dosim.* **116**(1–4) (2005) 6–11.  
<http://dx.doi.org/10.1093/rpd/nci051>
- [79] M. Brugger *et al.*, *Radiat. Prot. Dosim.* **116**(1–4) (2005) 12–15.  
<http://dx.doi.org/10.1093/rpd/nci052>
- [80] E. Mustafin *et al.*, Proc. EPAC 2006, Edinburgh, 2006, p. 1834.
- [81] C. Birattari *et al.*, *Radiat. Prot. Dosim.* **14** (1986) 311.
- [82] F. Schmidt, SixTrack Version 4.2.16: single particle tracking code treating transverse motion with synchrotron oscillations in a symplectic manner, CERN-SL-94-56 (1994).
- [83] G. Ripken and F. Schmidt, CERN-SL-95-12 (1995).
- [84] CERN BE/ABP Accelerator Beam Physics Group, SixTrack–6D Tracking Code,  
<http://sixtrack.web.cern.ch/SixTrack>.
- [85] M.A. Kostin *et al.*, An improved MAD-MARS beam line builder: user’s guide, Fermilab-FN-0738-rev (2004).
- [86] H. Grote *et al.*, The MAD-X Program, User’s Reference Manual, CERN (2015).
- [87] CERN BE/ABP Accelerator Beam Physics Group, MAD–Methodical Accelerator Design,  
<http://madx.web.cern.ch/madx/>.
- [88] L. Nevay *et al.*, Proc. IPAC 2014, Dresden, 2014, p. 182.
- [89] A. Mereghetti *et al.*, Proc. IPAC 2012, New Orleans, 2012, p. 2687.
- [90] A. Mereghetti *et al.*, Proc. IPAC 2013, Shanghai, 2013, p. 2657.
- [91] B. Auchmann *et al.*, *Phys. Rev. ST Accel. Beams* **18** (2015) 061002.  
<http://dx.doi.org/10.1103/PhysRevSTAB.18.061002>
- [92] B. Salvachua *et al.*, CERN-ACC-NOTE-2014-0036 (2014).
- [93] B. Salvachua *et al.*, Proc. IPAC 2014, Dresden, 2014, p. 174.
- [94] B. Dehning *et al.*, *AIP Conf. Proc.* **648** (2002) 229. <http://dx.doi.org/10.1063/1.1524405>
- [95] E. Skordis *et al.*, Proc. IPAC 2015, Richmond, 2015, p. 2116.
- [96] N.V. Mokhov *et al.*, Proc. IPAC 2014, Dresden, 2014, Fermilab-CONF-14-175-APC-TD (2014).
- [97] N.V. Mokhov *et al.*, Proc. IPAC 2014, Dresden, 2014, Fermilab-CONF-14-18 4-APC (2014).

## Reliability Considerations for the Operation of Large Accelerator User Facilities

*F. J. Willeke*

Brookhaven National Laboratory, Upton, NY

### Abstract

The lecture provides an overview of considerations relevant for achieving highly reliable operation of accelerator based user facilities. The article starts with an overview of statistical reliability formalism which is followed by high reliability design considerations with examples. The article closes with operational aspects of high reliability such as preventive maintenance and spares inventory.

### Keywords

Accelerator reliability; spare inventory, Weibull distribution, preventive maintenance, redundant components, single point failure.

## 1 Introduction

In previous decades, accelerators were developed and optimized as tools to explore the energy frontier for studying sub nuclear particles. However, more recently, another aspect of accelerator optimization has become more important, which is highly reliable operations to produce a large quantity of particle collisions ('particle factories') or photons (light sources), serving a large and diverse user community. The reliability aspect is particularly relevant for light sources. Light sources have large user communities of several thousand users organized in small independent research teams, each of which uses only a small fraction of the beam time. Even small operational inefficiencies due to frequent failures and interruptions might cause the total loss of allocated beam time of some research teams, with significant disruption of their science programmes. For these reasons, an increasing emphasis has been put on highly reliable operations. Reliability is usually defined as the total relative amount of beam time made available to users within the scheduled time period. A reliability of 95% is considered a tolerable lower limit for modern light sources. Reliability values of the order of 98% are reported frequently and are not an unusual achievement. This means that for a scheduled yearly beam time of for example 5000 h, only 250 h or less of user operations may be lost due to failures. Assuming that, on average, full recovery from a failure requires two hours, the time between interruptions must be larger than 40 h (assuming 24 h/d and 7 d/week of operation) on average. Science with synchrotron radiation has become very sophisticated and the delivery of a beam is not a sufficient criterion for reliability any more. Users need a beam of the planned beam energy and with nearly constant intensity, high spatial stability and high reproducibility of all beam parameters after changes of operational mode, such as changes in photon energy by changing the field strengths of undulator magnets. Accelerators consist of a large number of active components, many of them with high power consumption, which must function simultaneously to enable beam operation. They are connected and coordinated by sophisticated digital controls, and precision timing is usually a condition for proper functioning. For a facility with 100,000 of such components, any component may fail only after  $4 \times 10^6$  h of operation.

In the past, an accelerator facility needed several years or even a decade of operations to mature operations and develop the hardware system before such demanding operating goals could be realized. What appears to be desirable is to develop requirements to be taken into account in the design of the accelerator and in planning for its operations. Thus, accelerators must be designed for high reliability. When operating the facilities, all components must be maintained carefully, based on a comprehensive

preventive maintenance programme to minimize unscheduled downtime. This involves the monitoring of all components over time to identify any deviation from normal functioning, so as to prevent a failure during operations by timely repair or replacement.

These topics are discussed in this paper, which is organized as follows:

- introduction to reliability theory and definition of relevant parameters and properties;
- aspects and examples for high-reliability design;
- maintenance programmes;
- spares management.

## **2 Short summary of reliability definitions and relationships**

### **2.1 General remarks**

The purpose of this section is to show how reliability relevant parameters and functions that are used to analyse failure statistics and to make reliability predictions are related to observable and measurable quantities. To provide some understanding of the underlying statistical nature of the relationships, a short derivation of the most important formulae from basic principles and assumptions is presented.

In many areas of physics and engineering, the behaviour of complex, though deterministic, systems is described successfully by a statistical model in which events are considered random and independent, as insufficient detailed information about these systems is available. Random events are considered independent. This implies that the probability for a failure is independent of the history of previous failures or the failure of other components. While this is a rather practical and successful approach in most cases, we must keep in mind that we are dealing with deterministic systems and we are using the concept of statistics as a model. In particular, it should be noted at this point that the number of components in the systems considered is small, so the uncertainty of statistical prediction is considerable. Moreover, one needs to be aware that failures and different times or in different subsystems are not independent, adding another element of uncertainty to the outcome of statistical modelling.

Let us consider the following example:

A circuit breaker trips due to external overvoltage and a large number of magnet power supplies lose power and trip. During recovery from the event, one supply is found to be damaged from that occurrence and needs to be repaired. The two failures, the circuit breaker trip and the power supply failure, are clearly not independent. The power supply would most likely not have failed and continued to function for a while. However, it is also quite likely that the power supply would have failed at a somewhat later time during normal operating procedures, such as turn-off–turn-on cycles because there must have been a hidden defect, as power supplies are designed to survive power failures. Such events might be considered quasi-independent. However, there are strongly dependent failures, such as a large cooling-water leak causing water to penetrate a power supply, which may lead to corrosion and subsequent failure. Such events are usually fairly rare and do not have a large statistical significance. For now, we will assume that there are no significant dependencies of failures and that the statistical failure model is adequate.

### **2.2 Mean time between failure**

The mean time between failures (MTBF) is an important observable parameter, which can be related to other statistical functions and parameters related to failure occurrence and system reliability. For a single component, it is simply defined as the average time between two failure events, which is the number of



failures in a certain time period. This assumes that the system can be restored or repaired after a failure event. For non-repairable components one just averages the time to failure for a number of components. This is quite intuitive and, as we will see, the equivalency of these two definitions can be shown. To assess the impact of failure, another quantity is quite relevant; the time it takes on average to return a failed system to service, the mean time to repair (MTTR). The availability of a repairable system is defined as

$$\text{Availability} = 1 - \frac{\text{MTTR}}{\text{MTBF} + \text{MTTR}}.$$

Where a system consists of different constituents or subsystems (labelled here by index  $i$ ), the availability can be written as

$$\text{Availability} = \prod_i \left[ 1 - \frac{\text{MTTR}_i}{\text{MTBF}_i + \text{MTTR}_i} \right] \cong 1 - \sum_i \frac{\text{MTTR}_i}{\text{MTBF}_i + \text{MTTR}_i}, \text{ for } \text{MTBF} \gg \text{MTTR}.$$

To predict statistical failure behaviour, it is useful to introduce the concept of an instantaneous failure probability, which is closely related to the MTBF. We define  $p$  as the probability for a failure to occur in a small interval of time  $\Delta t$ . For the time being, we assume that  $p$  is the same for any interval of time,

$$p = \lambda \cdot \Delta t,$$

where  $\lambda$  is called the instantaneous failure rate. If  $\lambda$  is constant in time, the failure density distribution,  $f$ , or the probability of surviving a certain number  $n - 1$  of time intervals but failing in the  $n$ th time interval is

$$f_n \Delta t = [1 - p]^{n-1} p,$$

and  $f_n$  is a normalized distribution with

$$\sum_{n=1}^{\infty} f_n = 1,$$

using

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1 - q} \text{ for } |q| < 1.$$

The MTBF can be interpreted as the expectation value of the time to failure for the density distribution, thus

$$\begin{aligned} MTTBF = \langle n \rangle \Delta t &= \sum_{n=1}^{\infty} n \cdot [1 - p]^{n-1} p \cdot \Delta t = \\ &= -p \cdot \Delta t \cdot \frac{d}{dp} \sum_{n=1}^{\infty} [1 - p]^{n-1} p \cdot \Delta t = -p \cdot \Delta t \frac{d}{dp} \frac{1}{p} = \frac{\Delta t}{p} = \frac{1}{\lambda}. \end{aligned}$$

This is an important relationship, which allows us to relate the parameters of statistical functions to observations:  $\text{MTBF} = 1/\lambda$ . Note that this simple relationship holds only for time-independent instantaneous failure rates but that an equivalent relationship can be given for more complicated cases of time dependent  $\lambda(t)$  (called the hazard function) as well.

### 2.3 Failure and survival functions

The failure and survival functions are important tools to predict failures. The failure function  $F_N$  gives the probability that failure occurs within a time  $N \cdot \Delta t$ . It is given as the sum over the failure distribution density up to a time  $N \cdot \Delta t$ :

$$F_N = \sum_{n=1}^N f_n = 1 - (1 - p)^N.$$

The survival function  $S_N$  is the complement of the failure function and gives the probability that the system will survive for a time  $N \cdot \Delta t$ .

$$S_N = 1 - F_N.$$

## 2.4 Systems with identical components

Accelerators are built with subsystems which contain identical components. The considerations of the previous section can be generalized to include multicomponent systems. Assume a system of  $N$  identical components, each component having an instantaneous failure probability of  $\lambda \cdot \Delta t$ ; ( $\lambda$  is also called the hazard function). The probability of  $m$  components failing during a time  $\Delta t$  is then:

$$P_{Nm} = \binom{N}{m} \cdot (1 - p)^{N-m} p^m.$$

Note that  $P_{Nm}$  is a normalized distribution function with

$$\sum_{m=1}^N P_{Nm} = (q + p)^N |_{q \rightarrow 1-p} = 1.$$

The average number of failed components within a time interval  $\Delta t$  is, as might have been expected:

$$\langle m \rangle = \sum_{m=1}^N P_{Nm} \cdot m = p \frac{d}{dp} (q + p)^N |_{q \rightarrow 1-p} = N \cdot p.$$

The likelihood of having no failure in the time interval is  $P_{N0} = (1 - p)^N$ . When computing the MTBF for the system with  $N$  components,  $1 - p$  in the equation for the MTBF for one component in Section 2.2 has to be replaced by  $(1 - p)^N$ :

$$\text{MTBF} = \frac{\Delta t}{1 - (1 - p)^N} \cong \frac{\Delta t}{N \cdot p} = \frac{1}{N \cdot \lambda}.$$

Thus the MTBF for the system with  $N$  components is  $N$  times smaller than for a single-component system as one might have assumed intuitively.

## 2.5 Non-constant failure rates

So far, we have assumed that the probability for failure within  $\Delta t$  is constant in time. However, there are many reasons for a non-constant instantaneous failure rate. New systems have a certain fraction of components with hidden defects, leading to enhanced failure rates early in the life cycle. Many components wear-out or age as a result of other effects (damage due to repetitive high temperature, accumulation of dust and aggressive chemicals, change of material properties, such as elasticity, with time and so on). Another reason for time dependence of failures is changing external conditions, such as temperature and humidity. Systems with high voltage, for example, tend to develop arcs if the humidity of the air changes. Another parameter is the time since the last maintenance, during which mechanical clamps might loosen or dust might have accumulated.

For these reasons, to describe real systems, we must develop the formalism under the assumption that  $\lambda$  is not constant but may depend on time.

We start with the assumption that  $p$  is the probability for a failure within a short interval of time  $\Delta t$ , but that the probability may vary for different time intervals labelled  $n$ . Thus  $p \rightarrow p_n = \lambda_n \cdot \Delta t$ . The failure density distribution then becomes

$$f_N \cdot \Delta t = \lambda_N \cdot \Delta t \cdot \prod_{n=1}^N (1 - \lambda_n \cdot \Delta t).$$

The expression for  $f$  is more conveniently represented by a continuous function by making the time steps infinitesimally small. To make this transition we rewrite the probability density distribution  $f$  as:

$$f_N = \lambda_N \cdot \exp \left[ \sum_{n=1}^N \ln(1 - \lambda_n \Delta t) \right].$$

At this point, we can make the transition  $\Delta t \rightarrow 0$  and write, correspondingly,

$$f_N = \lim_{N \rightarrow \infty} \left\{ \lambda_N \cdot \exp \left[ \sum_{n=1}^N \ln(1 - \lambda_n \Delta t) \right] \right\}.$$

Since  $\lambda_n \Delta t$  is approaching zero, the logarithm can be expressed in terms of its Taylor expansion,  $\ln(1 - \lambda_n \Delta t) \rightarrow -\lambda_n \Delta t$ ,

resulting in

$$f_N = \lim_{N \rightarrow 0} \left\{ \lambda_N \cdot \exp \left[ \sum_{n=1}^N -\lambda_n \Delta t \right] \right\},$$

and this leads to

$$f(t) = \lambda(t) \cdot \exp \left[ - \int_0^t \lambda(\tau) d\tau \right].$$

The failure function  $F(t)$ , which gives the probability of failure within the interval  $[0, t]$  is the integral over the probability density distribution  $f(t)$ :

$$F(t) = \int_0^t d\theta \lambda(\theta) \cdot \exp \left[ - \int_0^\theta \lambda(\tau) d\tau \right] = 1 - \exp \left[ - \int_0^t \lambda(\tau) d\tau \right].$$

The complement of the failure function is the survival function  $S(t)$ , which gives the probability of surviving a time  $t$  without failure:

$$S(t) = \exp \left[ - \int_0^t \lambda(\tau) d\tau \right].$$

The instantaneous failure rate may be expressed by  $F(t)$  and  $S(t)$ :

$$\lambda(t) = \frac{\frac{d}{dt} F(t)}{S(t)}.$$

Now we can express the MTBF in terms of the continuous functions that we have derived. The MTBF is the expectation value of the time until failure using the probability density distribution,

$$\begin{aligned}
\text{MTBF} &= \int_0^{\infty} dt t \cdot f(t) \\
&= \int_0^{\infty} dt t \cdot \lambda(t) \cdot \exp\left[-\int_0^t \lambda(\tau) d\tau\right] \\
&= \left[ t \cdot \exp\left[-\int_0^t \lambda(\tau) d\tau\right] \right]_0^{\infty} + \int_0^{\infty} dt \exp\left[-\int_0^t \lambda(\tau) d\tau\right] \\
&= \int_0^{\infty} dt S(t) .
\end{aligned}$$

For example, if  $\lambda$  constant:

$$\text{MFBF} = \int_0^{\infty} dt \exp\left[-\int_0^t \lambda d\tau\right] = \int_0^{\infty} dt \exp[-\lambda t] = \frac{1}{\lambda}.$$

With these functions, we are now able to calculate an expectation value for the medium residual lifetime  $\text{MRL}(t)$  of a system that has survived a certain time  $t$  without failure. We use a similar expression to that for calculating the MTBF, except that the integral now extends from time  $t$  to infinity and the expression is divided by the probability of survival up to the time  $t$ , since only cases that have survived up to time  $= t$  are being considered

$$\text{MRL}(t) = \frac{\int_0^{\infty} f(t + \tau) \cdot \tau \cdot d\tau}{S(t)},$$

$$\text{MRL}(t) = \frac{\int_0^{\infty} d\tau S(t + \tau)}{S(t)}.$$

Note that for statistical failures, i.e.,  $S(t) = \exp(-\lambda t)$ ,  $\text{MRL} = 1/\lambda$ , which is identical to MTBF.

## 2.6 Statistical modelling of real systems

A useful parameterization for describing the failure statistics of real systems is the Weibull parameterization. The parameters of the Weibull model can be chosen so as to describe any of the three failure modes discussed so far: premature failure, statistical failure, and wear-out or ageing-related failure. Weibull was a Swedish engineer who introduced his model in the 1930s [1] in the context of describing fatigue and wear-out and this model has been applied to many use-cases. Since then, many modified or alternative parameterizations have been proposed and successfully applied (see, for example, Ref. [2], and quotations therein). However, it would stray too far from the purpose of this lecture to discuss them all. In the Weibull model, the instantaneous failure probability function  $\lambda(t)$  is described as

$$\lambda(t) = \frac{a}{b} \left(\frac{t}{b}\right)^{a-1}$$

The probability density distribution for failures in the Weibull model is

$$f(t) = \frac{a}{b} \left(\frac{t}{b}\right)^{a-1} \cdot \exp\left[-\left(\frac{t}{b}\right)^a\right].$$

The probability for failure is then

$$F(t) = 1 - \exp\left[-\left(\frac{t}{b}\right)^a\right].$$

The parameter  $b$  is a lifetime parameter, which for  $a = 1$  equals the MTBF. The parameter  $a$  describes the nature of the failure statistics. If  $a < 1$ , the model describes early, premature, failure, where the failure probability decreases with increasing time;  $a = 0$  describes the case where the failure probability function is constant, which is referred to as the statistical failure mode;  $a > 1$  describes an increasing failure rate with increasing time, as expected for wear-out and ageing. If  $a > 1$ , the stronger the deviation from  $a = 1$ , the sharper the probability density distribution  $f(t)$  peaks around the lifetime value  $t = b$ . For  $a < 1$ , the closer  $a$  is to zero, the faster the decay but the slower the approach of probability to failure to  $F = 1$ . A real system has elements of all three phases of failure. The hazard function for various values of  $a$  is depicted in Fig. 1.

Given an inventory of identical components, a certain fraction,  $c_1$ , will fail prematurely, another fraction,  $c_2$ , will fail statistically and the majority of the components,  $c_3$ , will fail due to wear-out and ageing. The total probability for failure is a weighted sum of the three components,

$$F(t) = \sum_{i=1}^3 c_i \cdot \left\{ 1 - \exp \left[ - \left( \frac{t}{b_i} \right)^{a_i} \right] \right\}.$$

The survival function is then written as

$$S(t) = \sum_{i=1}^3 c_i \cdot \exp \left[ - \left( \frac{t}{b_i} \right)^{a_i} \right].$$

The MTBF in the Weibull model is more complex than in the simple case of purely statistical failure, and is  $MTBF = b \cdot \Gamma \cdot \left( 1 + \frac{1}{a} \right)$ .

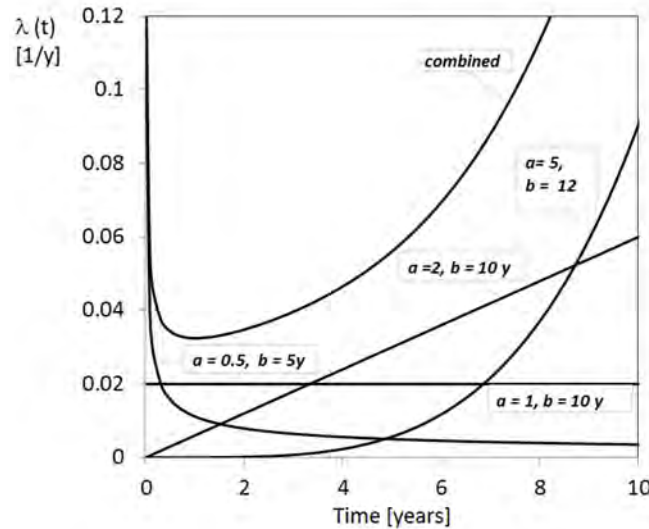


Fig. 1: Instantaneous failure rate in the Weibull model for different form factors  $a$  and lifetime  $b$

### 3 Accelerator design for high reliability

#### 3.1 General remarks

In Section 2, we discussed the three modes of failure: premature, statistical, wear-out, and ageing. Combining the three modes of failure in one graph, one obtains the so-called bathtub curve (see Fig. 2) with an initially enhanced failure rate, a steady but lower failure rate in the middle and an enhanced rate at the end of the components' life cycle. The way the accelerator complex is designed, constructed, and operated has a large impact on the failure rate in all three phases.

Premature failure can be reduced by careful quality assurance and inspection of purchased components. Suppliers should be chosen based on proven reliability records and good workmanship. Often, hidden damage occurs during transport. The purchasing contract should include requirements for shock, temperature, and humidity detectors as part of packaging the components for transport. Considerable effort should be invested in acceptance testing the equipment. The tests should be comprehensive but safe, as damage of sensitive equipment during acceptance testing might occur.

The wear-out and ageing phase can be strongly influenced by regular maintenance, preventive maintenance, operating the equipment below maximum power rating, controlling the installed equipment's environment in terms of temperature, humidity, and dust, and exposure to aggressive chemicals.

All three phases of failure, however, are influenced by how the accelerator, and its subsystems and components, are designed. The following sections will discuss various aspects of high-reliability accelerator design.

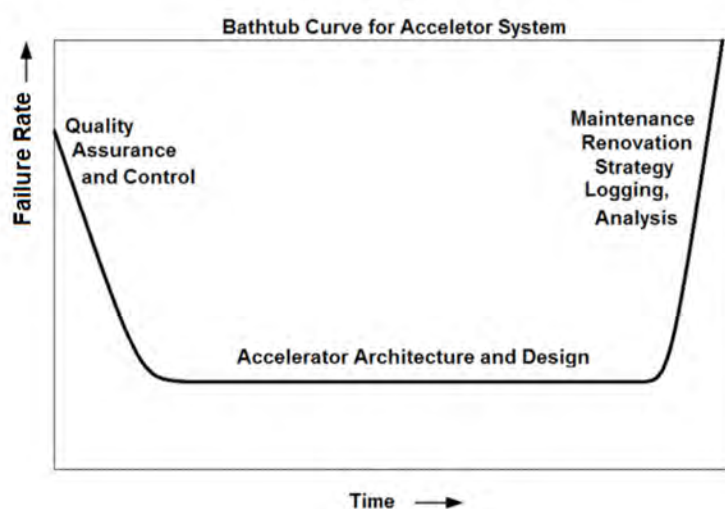


Fig. 2: The 'bathtub' curve of system failures over the entire life cycle

The overall optimization of an accelerator system is a compromise between three major considerations, which may lead to conflicting requirements. These are cost, performance, and reliability. The challenge of building a new accelerator facility is to find solutions that support all three requirements satisfactorily.

## 3.2 High-reliability design considerations

Next, we discuss some of the main considerations of high-reliability design. This will be followed by a more detailed discussion of examples.

### 3.2.1 Overall complexity

The more complex a system is, the higher the probability for hidden errors, or wrong or incompatible sets of parameters. These errors may 'sleep' within the system up to the point where special parameters in particular configurations are required. Troubleshooting in complex systems can be very time consuming. So, the complexity and interdependence of systems, which are, of course, unavoidable, should be minimized in a high-reliability design.

### **3.2.2 Unavoidable weakness**

Some weakness in the design with respect to high reliability is thus unavoidable. For example, the accelerator beam will most probably be lost if there is a failure of a magnet power supply or if the RF system trips. The designers should be aware of these weaknesses and mitigate the risk of failure, for example, by reducing the number of individual supplies or installing spares as hot spares.

### **3.2.3 Subsystem architecture**

The architecture of the subsystems, which defines their interdependence and their common failure modes, plays an important role in mitigating reliability risks. The system should be designed and configured such that failure in one component does not start a chain of failures in other components.

### **3.2.4 Fail-safe design**

Provision for component failure is a part of every good design, to avoid collateral damage in case a failure occurs.

### **3.2.5 Overrated design**

An effective way to achieve good reliability is overrated design. High-power components, such as magnet power supplies, RF transmitters, RF cavities, and pulsed magnet systems, are particularly susceptible to failure, with a potential for collateral damage. Operating these systems at the highest performance implies that operating at the temperature limit or with the maximum tolerated mechanical vibrations will shorten the components' lifetimes and increase the probability of failure. However, overrating goes along with increased cost and may not always be affordable.

### **3.2.6 Environmental impact**

Changing environmental conditions, for example, by varying temperature, increasing humidity, changing from a dry and dusty environment to a humid environment, or exposure to aggressive chemicals (for example, created by synchrotron radiation) are important factors of failures. Protecting the equipment from such influences will be a major contribution to high reliability.

### **3.2.7 Built-in redundancy and hot spares**

Built-in redundancy is an effective way of minimizing downtime, even if a failure does occur. Redundancy, however, may be quite costly. Hot spares are spares that need to be available anyway, to avoid long downtimes during repair or replacement but that are already installed place, ready to be used if the primary component fails. The combination of the two concepts is very effective, to avoid large downtimes and the additional cost may be relatively moderate.

### **3.2.8 Built-in diagnostics**

Built-in diagnostics, preferably with post-mortem analysis capabilities, is a strong tool to detect potential failures before they occur or to identify quickly the root cause of failure, thereby speeding up repair and recovery from failure. Built-in diagnostics comes with a cost and must be compromised in a cost-effective design. In each case, the designs should have provisions for integrating built-in diagnostics at a later date without the necessity for major design changes.

### **3.2.9 Repair and maintenance-friendly design**

Failures do happen in any complex technical system. This fact needs to be taken into account when designing the system. A modular concept makes it easy to diagnose and isolate the source of an error. The mechanical layout should take into account that systems need to be easily accessible for quick

repair. The system must include well-documented and easily assessable measurement points for quick diagnosis.

### **3.2.10 Documentation**

The effort to produce and maintain complete, up-to-date and readily available comprehensive documentation will pay for itself when errors occur in complex systems. While large documentation is usually needed for procuring and building components, the maintenance of documentation in the operational phase is often neglected, owing to lack of resources. This might be the cause of long repair and recovery.

### **3.3 Subsystem architecture**

A basic choice is to choose either a compact design, which combines many functions and features implemented in the same hardware, thereby saving on redundant components, or a modular design. The modular design is friendlier for troubleshooting and repair. It also allows more easily for the incorporation of hot spares. When coupling the two types of approach, attention needs to be paid to avoid accumulating disadvantages with respect to reliable operations.

Consider, for example, a number of switched-mode power supplies. These devices are supplied with a constant d.c. voltage from individual supplies or from a common supply. The supply turns the d.c. input into a pulsed voltage with a rectangular pulse shape via a fast switch, operating at a fixed frequency in the kilohertz range. The pulse length is varied to achieve the desired output current. The output filter turns the rectangular waveform into a d.c. current.

The first, though expensive, solution is to provide a d.c. voltage supply for each individual switched-mode supply.

The second, more economical, solution is to provide a common d.c. voltage source. One can even go one step further by implementing a multichannel power supply, which has, besides the common d.c. voltage input, other common components, such as an auxiliary voltage supply for power supply electronics, and interlock and alarm features. This approach turns the switched-mode power supply system into a cost-effective compact system. The disadvantage is that most of the failures in one supply or channel will cause the common voltage supply to be shut off. All the switched-mode units supplied by the voltage source will be tripped as a consequence. Recovery involves restoring a large number of power supplies. Often, a few supplies need some extra effort to return them to service. The impact of a simple trip can be considerable. The lifetime of all the supplies turned off unnecessarily is affected.

A design developed at SLAC overcomes this disadvantage [3] by adding two isolation switches, which separate the failing switched-mode supply from its voltage source. The voltage supply thus remains turned on and all the other switched-mode supplies are decoupled. While this adds both cost and complexity, it appears to be a good compromise and will achieve considerably higher reliability performance.

### **3.4 Fail-safe design**

Fail-safe design is good engineering practice, to protect a device from its own failures. However fail-safe designs that might provide optimum protection of the device might not be favourable for high reliability. Perfect protection of a device implies many trips, with many of them not being necessary to protect the device. Figure 3 illustrates the dilemma. Consider two redundant sensors for the protection of a device. Combining the two signals via a logical AND gate establishes an enhanced reliability system with little chance of false trips, but the protective function is not perfect and a faulty sensor might lead to damage of the device. This case is not fail-safe. In the opposing case, where the two signals are combined by a logical OR gate, the device is fairly well protected but false signals will lead to



unnecessary trips. Reliability is compromised in this case. A system with three sensors will mitigate the shortcoming if combined as shown in Fig. 4.

Variable trip thresholds are also a good way to overcome the fail-safe dilemma. Early in operation, when there is little or no experience with the device, the trip threshold can be set low. After operational experience has been gained and the entire operation is matured, the thresholds can be raised, thereby reducing the probability of unnecessary trips. These must be part of the design from the beginning. It is also important to design a system to administer and safeguard parameters, to avoid equipment damage resulting from incorrect parameters.

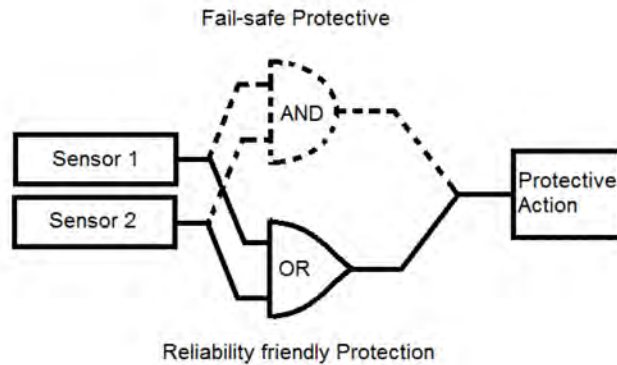


Fig. 3: Fail-safe versus high-reliability protection

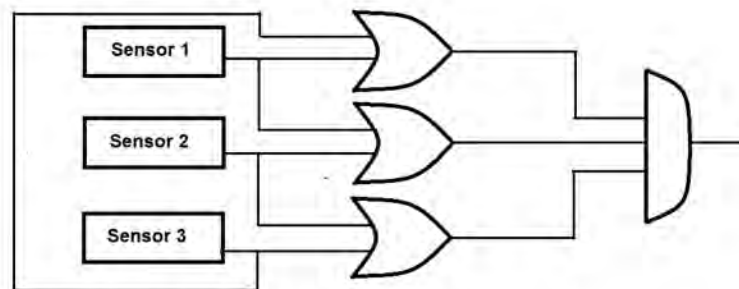


Fig. 4: Fail-safe and high-reliability arrangement of three sensors

### 3.5 Built-in redundancy

Built-in redundancy has the potential for improving the reliability of a system considerably. It will not reduce the failure rate but it will reduce the MTTR significantly. It might also offer a convenient way to perform preventive maintenance while keeping the system running. Closely related to built-in redundancy are hot spares. Hot spares are spare components that are installed in the system and can be switched into service without much effort and in minimum time.

Consider a system with two redundant components, labelled 1 and 2, with failure functions  $F_1(t)$ ,  $F_2(t)$ . The system fails if the two components fail in the same period of time. The components will be redundant only if the failed redundant components are repaired immediately after discovery of failure. This implies that all inactive components of a redundant system need to be checked for full functionality at regular intervals. Let the time between functional tests be  $t_c$ . Another assumption needed for full redundancy is that failures of the redundant components would be completely independent and uncorrelated (which in a real-use case would have to be verified carefully). The redundant system will fail if the two systems fail simultaneously within a time  $t_c$ . The corresponding failure function  $F(t_c)$  is

$$F(t_c) = F_1(t_c) \cdot F_2(t_c) = \left( 1 - \exp \left[ - \int_0^{t_c} \lambda_1(t') dt' \right] \right) \cdot \left( 1 - \exp \left[ - \int_0^{t_c} \lambda_2(t') dt' \right] \right).$$

Let us consider a well-matured system with constant (statistical) failure rate,

$$\int_0^{t_c} \lambda_1(t') dt' = \lambda_1 \cdot t_c,$$

and let us assume that  $t_c \ll \text{MTBF}$  or  $\lambda_1 \cdot t_c \ll 1$ . In this case, the failure function is approximated as

$$F(t_c) = \lambda_1 \lambda_2 t_c^2.$$

Thus by the choice of  $t_c$ , the failure probability of the redundant system can be made arbitrarily small under the assumption of uncorrelated failures. The  $\text{MTBF}_r$  of the redundant system is then

$$\text{MTBF}_r = \left( \frac{\lambda_1 \lambda_2 t_c^2}{t_c} \right)^{-1} = \frac{\text{MTBF}^2}{t_c}.$$

For, example, consider a system of two redundant components with a MTBF of 480 h for each component. If the system is checked daily, the probability of failure will be  $(24/480)^2 = 1/400$  and the  $\text{MTBF}_r$  of the redundant system is 9600 h.

If there are  $n$  parallel components that need to be active simultaneously, the  $\text{MTBF}_s$  for the system is  $n$  times the  $\text{MTBF}_c$  of a single component, according to the considerations in Section 2. However, if there is one built-in (hot) spare, the  $\text{MTBF}_s$  of the system is given by

$$\text{MTBF}_s = \frac{\text{MTBF}_c^2}{t_c \cdot n}.$$

Figure 5 shows as an example a crate with five switched-mode power supplies developed at DESY in 2000 (J. Eckoldt, private communication), one of which is redundant. Assuming a MTBF of 100,000 h for each supply, and assuming that the function of the redundant supply is checked once per month, the MTBF is improved to 347,000 h by using the redundant component. Thus, with an increase of less than 25% in cost, the reliability of the device is increased by a factor of 3.5.



**Fig. 5:** Crate with five switched-mode power supplies, one of them being redundant (courtesy of J. Eckoldt, DESY)

### 3.6 Redundant safeguards

The case of pairs of redundant components can be easily generalized for a system of  $n$  safeguards. Each individual safeguard (labelled  $i$ ) has a mean time between failure of  $\text{MTBF}_i$ . Each safeguard is checked at regular intervals  $\Delta t$ . The  $\text{MTBF}_s$  for the redundant system of safeguards is then

$$\text{MTBF}_s = \frac{\prod_{i=1}^n \text{MTBF}_i}{\Delta t^{n-1}}.$$

This shows that by increasing the number of safeguards, the reliability of the redundant system is increasing enormously. However, this is only true if  $\Delta t < \text{MTBF}_i$ , that is, if the components are checked at regular, sufficiently small, intervals. It is also important that the redundant components are as diverse as possible to reduce the probability of correlated failures.

### 3.7 Overrated design

Overrating of high-power components has a number of positive effects on the lifetime and probability of failure of the components. These have mainly to do with operating temperature. The operating temperature will be less in an overrated device as the cooling is designed for larger power dissipation than the power dissipated in normal operating conditions. The change in temperature when the device is turned on and off will also be reduced. This leads to reduced mechanical stress. All these factors will increase lifetime and reduce failure rate. In the electronics industry, the following Arrhenius-based expression is used to describe the impact of temperature and temperature changes on the failure rate [4]:

$$\frac{\lambda}{\lambda_0} = \left(\frac{\Delta T}{\Delta T_0}\right)^2 \cdot \exp\left[-\frac{E}{kT} \cdot \left(1 - \frac{T}{T_0}\right)\right].$$

The first factor describes the effect of thermal cycling and fatigue; the second factor describes the thermal stress due to high temperature. The index ‘0’ indicates a reference temperature level for comparison. The parameter  $k$  is Boltzmann’s constant, and  $E$  is a typical excitation energy level that leads to changes in the material under consideration

Thus, overrating leads to a reduction in the failure probability. Another positive effect is that the system can be operated further away from critical limits and trip thresholds, thereby reducing the number of false trips. On the negative site is increased cost and, in many cases, increased installation space requirements.

### 3.8 Environmental impact: dust, humidity, temperature

In Section 3.7, we discussed the impact of temperature and temperature changes on the lifetime and on the probability of failure. The impact of temperature on the lifetime of electrolytic temperature is pronounced and well understood. Manufacturers quote the following expression for the lifetime, based on Arrhenius law (see for example, Ref. [5]):

$$MTBF(T) = MTBF(T_{ref}) \cdot 2^{-\left(\frac{T-T_{ref}}{10^{\circ}\text{C}}\right)}.$$

Figure 6 shows how the lifetime of film capacitors is affected by internal temperature, which is determined by ambient temperature and internal heat production. Similar behaviour is observed for electrolytic capacitors.

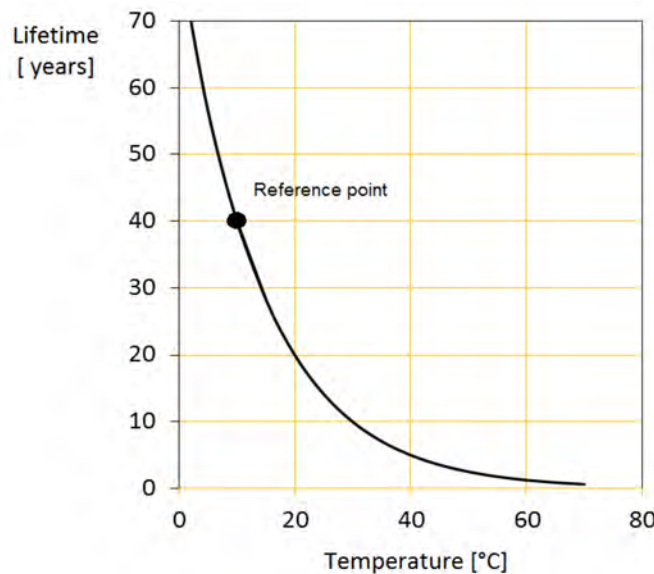
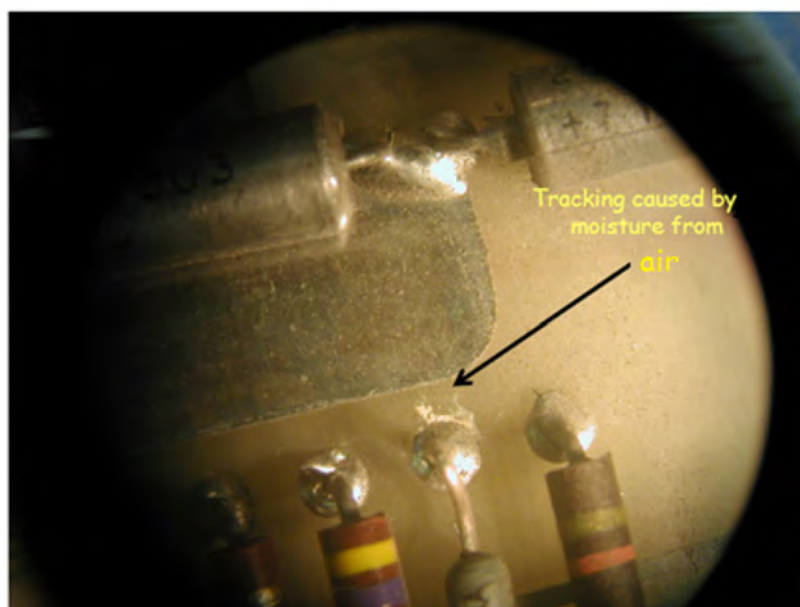


Fig. 6: Lifetime of film capacitors as a function of internal temperature

Exposure to changing humidity and dust are other environmental factors that can significantly compromise the reliability of technical components.

Dust may have constituents that are electrically conducting and could cause electrical shorts between connectors on electrical printed circuit boards. Once a current is flowing, the printed circuit board material can become carbonized, which aggravates the short to the point that the electronic component becomes non-functional. Needless to say, such a process can be accelerated by the presence of humidity or of chemical radicals in the air. Such radicals may be produced by synchrotron radiation, which is produced by high-energy electron beams in the accelerator tunnel. The malfunction in such cases may develop only gradually and slowly and may initially only cause occasional failures. The cause of such failures can be very difficult to find and repeated failures may cause significant downtime. Figure 7 shows a board in the quench protection system of the Tevatron (Fermilab) superconducting magnet system that was damaged by this effect (H. Edwards and P. Czarapata, Groemitz Miniworkshop on Accelerator Reliability (Groemitz) 2005, unpublished data). The combination of dust and humidity caused some mysterious modulator trips in the RF system of the HERA electron-positron collider at DESY, Hamburg. In the late 1990s, a high frequency of arcing on the modulator was reported. In an attempt to explain the events, they were analysed as a function of air humidity and dust particles in the air. There was no clear correlation. However, after further investigation, it was found that the arcs always occurred when the outside weather conditions changed from dry to humid periods. A small amount of outside air had been added to the circulating, well-conditioned, internal airflow. During dry periods, small amounts of dust from the outside air accumulated on the surface of high voltage ( $\approx 70$  kV) carrying components. When the humidity increased, the dust particles acted as launch points for arcs into the slightly more humid air in the modulator room. Thus, even a tiny amount of exposure to environmental conditions can have surprisingly large effects.



**Fig. 7:** Damage to quench protection board as a result of dust and humidity

An effective, though not quite inexpensive method, was chosen to mitigate the environmental impact on electronics in the new synchrotron light source, NSLS-II, at Brookhaven National Laboratory. All electronics components are enclosed in a sealed rack system, in which air is circulated around the equipment for cooling and is then cooled with water-to-air heat exchangers. This method has the advantage of keeping cooling water away from the magnet power supply and other electrical equipment and maintaining air cleanliness and low humidity. As power dissipation is low in all devices, air cooling is sufficient to maintain favourable operating temperatures. Figure 8 shows the NSLS-II rack system.

### 3.9 Error-prone solutions

In reviewing the cause of failure in accelerator equipment, there are two outstanding items. The first one is leaking cooling water, which can cause electrical shorts, damage of electrical equipment, or the formation of acidic liquid, which will lead to corrosion. While large leaks are detected easily and the recovery from a large leak is straightforward, small leaks may remain unnoticed and a large amount of damage might result before the problem is noticed. For this reason, water cooling systems have to be designed and manufactured with great care. Wherever air cooling can be used instead of water cooling, air cooling should be given preference. Water piping should, if possible, be installed underneath delicate equipment rather than overhead. Double floors that provide a space to accommodate water piping are an expensive but effective solution to avoid damage by cooling-water accidents. An example of a water-free cooling design is the air cooled sealed NSLS-II power supply enclosure mentioned previously (see Fig. 8), which keeps cooling water away from electrical components. Water cooling circuits should be regularly checked for pressure drop and the design should foresee an efficient way of performing such tests frequently.

Another frequent source of failure is cable terminations and connectors. The list of potential issues is long and spans from assembly errors, insufficient ground connections, corrosion, and mechanical damage, cable damage at the fitting, miswiring, and confusion of connectors after repair or test. The root cause is that cable connections often have to be mounted in the field under sometimes difficult conditions, such as limited space, visibility, accessibility, dust, etc. For this reason, it should be checked from case to case whether analogue hard-wiring can be replaced by digital data connections, which offer the ability to replace many critical electromechanical connectors and switches. To save cost on cable connections by using cheap components and inexpensive labour might not be an optimum decision in view of the loss in reliability caused by low-quality connectors and poor workmanship. It is advisable to design a comprehensive quality-control programme to assure the adequacy of cable connections. To reduce the risk of slowly developing poor connectivity, the use of gold-plated connectors is encouraged. In the case of high-current bolted cable connectors, the materials have to be carefully chosen to match in thermal expansivity, to avoid premature wear-out of the bolting elements.



**Fig. 8:** NSLS-II sealed racks (equipment enclosures)

### 3.10 Built-in diagnostics

The next two topics are strongly related to operation and maintenance of high reliability, but have to be considered during the design and construction phase of a facility. The first topic is built-in diagnostics. Usually, a failure announces itself by more or less significant changes in some of the analogue variables inside a device. Therefore, it is helpful in both preventing and analysing a failure if the internal analogue variables of a technical device are measured, recorded, logged, distributed, and analysed. Thus, built-in diagnostic tools have the potential to increase the MTBF and reduce the MTTR of a component or system.

Comprehensive capturing of relevant internal data has to be well integrated in the design of a device, component, or system. With increasing complexity of a technical system, the need of internal diagnostics for supporting preventive maintenance and troubleshooting becomes more crucial. There are many examples of good implementation of built-in diagnostics. The cost of the additional design and construction effort is not negligible. However, it will be extremely difficult to achieve reliabilities above 95% without built-in diagnostics. For these reasons, built-in diagnostics is a standard component of modern technical designs.

### 3.11 Repair- and maintenance-friendly design

Accessibility for troubleshooting and repair is an important aspect of service-friendly design in support of high operational reliability by quick recovery from failure. Another service-friendly design feature is modularity. Modularity refers to the ability of independent testing for proper functioning of a part of a system without major rearrangement, partial disassembly, or reconfiguration. Modularity also refers to the physical arrangement of the constituents in a way that allows a larger subsystem to be exchanged with a minimum of effort. A thoughtful design foresees easy access to internal components for taking measurements during troubleshooting. This includes the provision of circuit board extensions to easily access measurement points, or the provision of a connector that can be connected to a test module for automated troubleshooting.

## 4 High-reliability operations

### 4.1 General remarks

Reliable operations that are based on a high-reliability design and implementation are achieved by a process of continuous improvements. Many subsystems of an accelerator facility are custom-designed systems with few components. They will mature during operation by small improvements and partial replacement of weak components. While this process is expected to require a significant effort in the start-up years of a facility, it will settle to a lower level of effort after a few years but will persist up to the time when larger investments and refurbishment become necessary, when the components arrive at the end of their life cycle.

An important tool for organizing efficient and reliable operation is comprehensive data logging and analysis of the performance of all components. The analysis toolkit will usually not be provided as part of the construction but its creation is an iterative process, which also requires a certain amount of operation time to reach a sufficiently high level of maturity.

The logged data need to have timing information and the systems should be equipped with circular buffers to allow post-mortem analysis. Data should be easily accessible from off-site to allow experts to perform analysis and troubleshooting without their having to be physically on-site.

Root cause analysis is helpful in understanding larger incidents, to prevent them from happening again. Commercial software tools are available to support such activities.

High operational reliability requires well thought-through operational strategies to mitigate the impact of failure, in particular, in view of the always-limited operational resources. Important elements of such strategies are scheduled maintenance and a preventive maintenance programme based on comprehensive monitoring of components and analysis of the data, as discussed previously.

One way of optimizing the facility output is to develop a figure of merit for operational performance and use this to relate any component failure to reductions in performance. This enables one to decide rationally whether to interrupt operations for an unscheduled intervention or to run with reduced performance until the next scheduled intervention. Part of such strategy is the inclusion of back-up plans for operating with reduced performance, such as accelerator studies or special operation modes.

## 4.2 Preventive maintenance

With the large number of components and the large diversity of equipment in a large accelerator facility, the opportunity for preventive maintenance to replace, repair, or adjust equipment before failure occurs is large. Systematic preventive maintenance of each piece of equipment is, in most cases, unrealistically expensive, and the effectiveness of such activity is very poor. Preventive maintenance, therefore, needs to be properly focused on equipment and use-cases where it will have a high probability of being effective in preventing failure. In this section, we will discuss preventive maintenance opportunities that have proved effective.

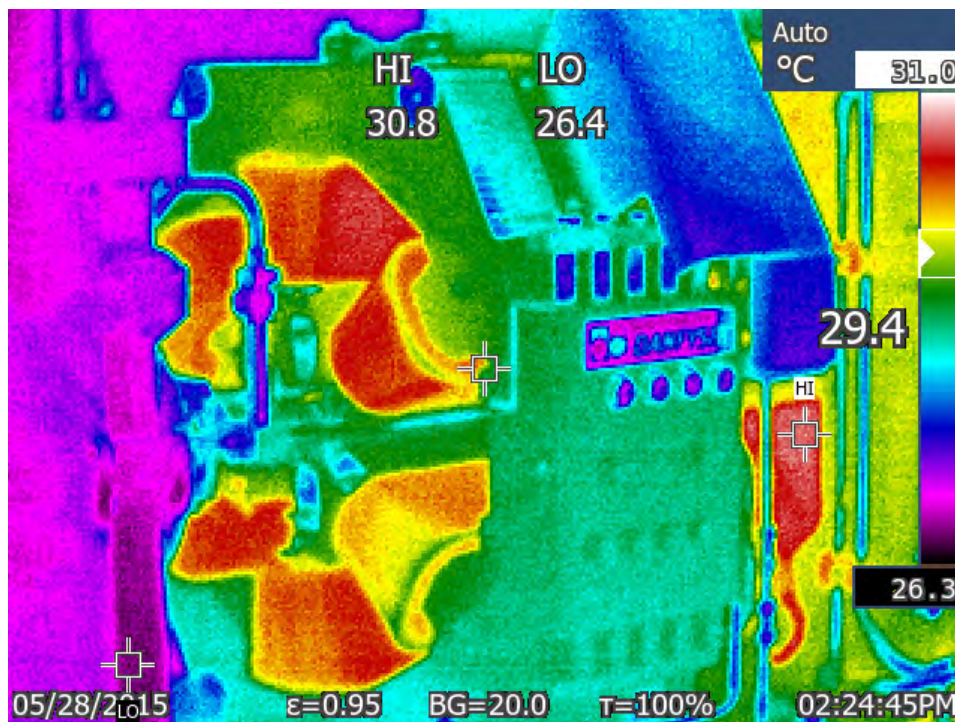
Mechanical rotating equipment is an obvious candidate for preventive maintenance. Examples of such equipment are water pumps, mechanical vacuum pumps, fan systems, air compressors, chilled water compressor systems, and turbines in cryogenic cold-engines. Such equipment is needed in many industrial installations. Such systems usually come with detailed maintenance plans from the manufacturer. Often, preventive maintenance is offered as part of a service package by the manufacturer. The preventive maintenance of such systems has been standardized and standards published, for example the Society of Automotive Engineers JA1011 standard *Evaluation Criteria for Reliability-Centered Maintenance* [6].

Another obvious class of equipment are battery-based devices, such as uninterruptable power supplies. Batteries have a well-known lifetime in terms of operating hours. Thus, preventive maintenance is very effective in such cases.

It is also obvious that filter systems need cleaning or filter replacement at regular intervals, which depend somewhat on the environment in which they have to work. Maintenance plans can be optimized after short operation periods.

Much equipment is cooled by air driven by internal fan systems. Fans should be interlocked so that in case of failure, there is no consequent damage to high-power equipment. The lifetime of fans should be known, in principle. However, the range of achieved operating hours is quite large. The performance and lifetime of fans depends on humidity, dust, and ambient temperature. Therefore, some judgement and experience is required to know when to exchange fans preventively, to avoid excessive costs or to prevent shutdown of equipment during operations.

The preventive maintenance that can be carried out on electrical equipment, such as magnet power supplies, is somewhat less obvious. Often, these systems have clamps or bolts to hold equipment in place. These mechanical fixtures need to be checked from time to time to ensure that all equipment is tightly connected, to avoid arcing and other damage to high-power equipment. An efficient method to check the integrity of connections that carry a high current is to use thermal imaging (see Fig. 9), which allows a large amount of equipment to be monitored in a very short time. Critical equipment can also be monitored continuously by a fixed installation; the cost of thermal imaging has reduced dramatically in the recent decade.



**Fig. 9:** Screenshot of thermal imaging of high-current connectors and magnet coils which is a very efficient method to check for weak connectors, obstructed cooling channels on magnet coils, etc.

Water cooling piping at or inside equipment should be regularly checked, since, as pointed out, cooling-water leaks are often the root cause of malfunction. A static pressure test will easily detect cooling-water leaks that can cause significant damage or downtime.

Some equipment exhibits sign of wear-out and fatigue. Some thyatron tubes in pulsed power equipment announce the end of their lifetime by requiring higher voltages to maintain the discharge current. Poor contacts also show increased transition resistance. Latent winding faults on magnet coils can be recognized by carefully examining inductive voltages and comparing them with electrical current changes.

The lifetime of piping systems and cooling channels inside equipment is very difficult to estimate. Here, it is important to check the water quality—its resistance, oxygen content, and pH—on a regular basis. It is also very difficult to estimate the impact of synchrotron radiation on the cooling water and the cooling channels. Variable thermal loads may lead to repeated thermal stress, which can cause fatigue and subsequent cooling-water leaks, which can be detected by static pressure testing on isolated cooling channels.

Maintenance is labour-intensive and is one of the highest cost elements in operating an accelerator. It is important that precious resources are used in the most effective way. This requires that maintenance programmes to focus on components with a high-failure probability. Error and failure analysis, supported by modelling, can be helpful tools to develop effective maintenance programmes.

During the start-up of operations of a new facility, it is usual for a large number of teething issues on new systems to be addressed; it is difficult to introduce a systematic preventive maintenance programme at the same time, owing to limited resources and a lack of data on failure events. Preventive maintenance is thus most effective in the mature operation phase. Preventive refurbishment is naturally most reasonable if the systems enter the wear-out phase. However the time constants for the subsystems are probably different, so that not all operational reliability phases occur simultaneously.

Next, we will demonstrate, with an example, how the formalism of reliability engineering can be applied to decision-making in maintenance activities. Consider a system of 200 components that are



subject to wear-out. Figure 10 shows the available data on failures that have occurred in components within a few years. The failures start to appear after about 200 weeks and the failure rate afterwards is accelerated. Do these observations suggest that preventive refurbishment will avoid considerable downtime in the future? The data are well described by a two-parameter Weibull failure function

$$F(t) = 1 - \exp(-(t/\tau)^\alpha),$$

the parameters  $\alpha$  and  $\tau$  being obtained by a least square fit. We then calculate the mean residual life:

$$\text{MRL} = \int_0^\infty dt \cdot \frac{S(t + t_0)}{S(t_0)}.$$

This allows us to project future failure rates and enables us to make a rational decision on preventive refurbishment, to prevent unscheduled downtime (see Fig. 11). The MTBF is about 494 weeks and the form factor is  $\alpha = 6.1$ . Thus, the system is described well by failures in the wear-out phase. The mean residual life at the end of the observation period is reduced to only one-third of the original system and is expected to become very small after another operation time of ~200 weeks, which constitutes a high risk for unscheduled downtime. In this situation, preventive refurbishment might be considered an overall optimum measure.

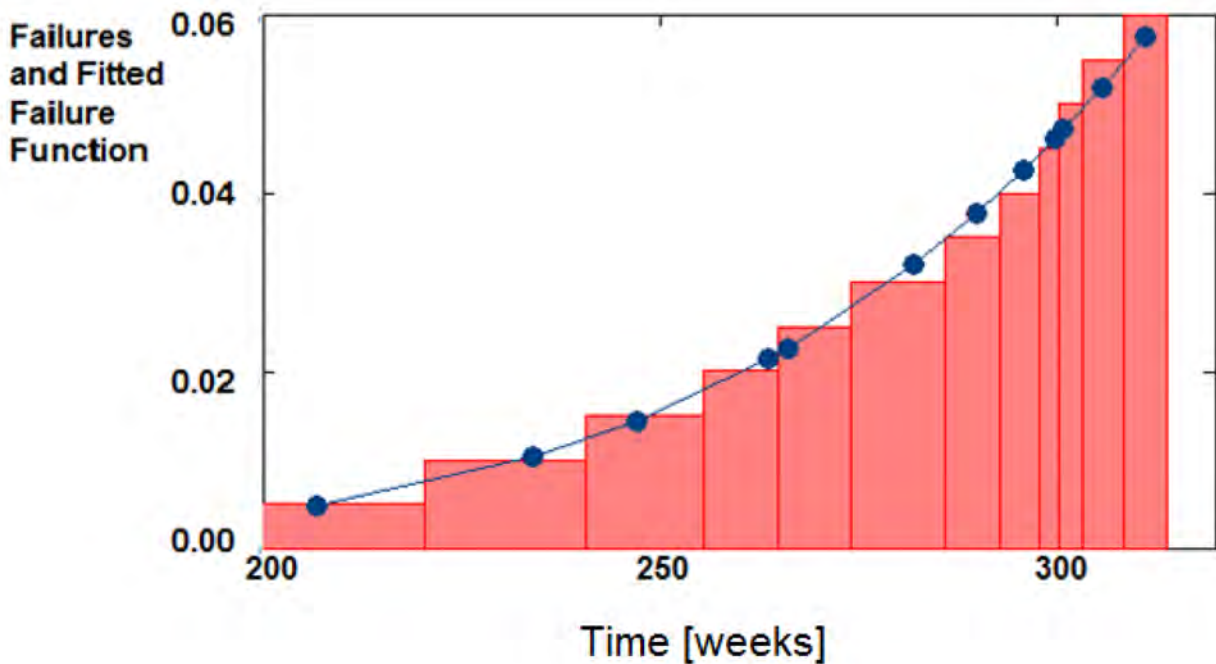
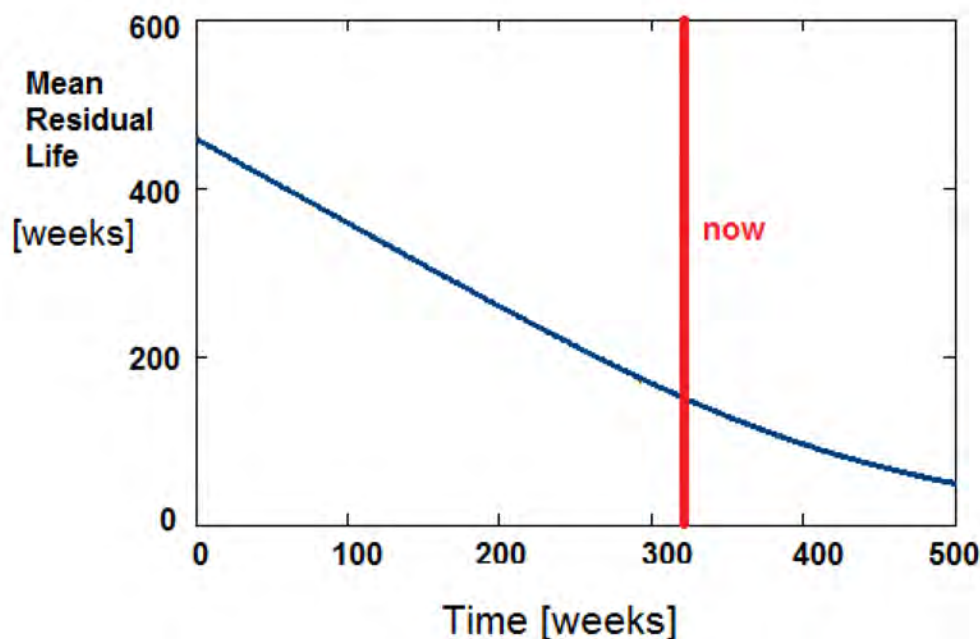


Fig. 10: Example of accumulated failures (columns) and Weibull Failure function (two-parameter fit, dots)



**Fig. 11:** Mean residual life as a function of operating time. The mean residual life is expected to drop significantly after additional operation time.

### 4.3 Speed-up of repair

The MTTR includes the time to identify a malfunctioning component. This can be quite time consuming and cumbersome, as trips and faults that are observed are often only the consequence of a hidden primary malfunction. This time may exceed the time for repairing or exchanging a component. The information provided about a failure event is thus an important factor in the overall system reliability. The following measures have proven to be very helpful in speeding up the troubleshooting and investigation process:

- **Transient recording** is based on continuous monitoring with an appropriate data rate, which may have arbitrarily short time-scales. Data storage is limited, and this defines a data cycle in which the oldest data are continuously overwritten by new data (also called a ‘circular buffer’). The failure event needs to provide a trigger that stops the circular buffer and data for the time period that precedes the event that can be retrieved (post-mortem data).
- An **asset management** database with life cycle data on each component may be very helpful in finding the root cause of a failure event. Knowledge of a component’s history (previous failures, unusual operating conditions, time in operation, problems with similar equipment) provides, in some cases, a clue to understanding a malfunction.
- **Remote access** to the process data generated by a system or an item of equipment is important, as it tends to save time and cost, since experts may then not need to come on-site to perform troubleshooting or to direct repairs, so long as they have access to the Internet somewhere on the planet.
- Having the data from a failure available is sometimes not sufficient. **Analysis tools** are likely to be required if a large amount of data needs to be scanned for anomaly behaviour of the equipment.
- It is worthwhile to maintain a **database with information on failures**, so as not to be solely dependent on the memory of experts involved in resolving past problems.
- **Start-up checklists** are helpful, to find malfunction early and to avoid having to repeat a start-up procedure due to late discovery.

#### 4.4 Spares inventory

The availability of spares is crucial for high-reliability operations, as many spare parts require a long time for replacement or repair. Therefore, for all breakable equipment of an accelerator facility, there should be at least one spare, which is best procured as part of the construction project. For equipment with a purchasing lead time of the order of the system MTBF, (which is  $n$  times smaller than the single component MTBF,  $n$  being the number of pieces of equipment in the system) more than one spare is necessary to avoid extended downtimes.

To plan highly reliable operations, it is necessary to determine the rate of consumption of spares. The replacement of used spares is a major cost factor in an operations budget. For a system with identical components, the failure density distribution of which may be described by the Weibull model, the probability of failure of each component in the interval  $[0, t]$  is described by

$$F(t) = 1 - \exp \left[ - \left( \frac{t}{b} \right)^a \right]$$

and the probability of surviving longer than  $t$  is  $S(t) = 1 - F(t)$ . The lifetime  $\tau$  of a component is defined as the time when  $S(\tau) = 1/e$ . It is identical to parameter  $b$  in the distribution:

$$1 - \frac{1}{e} = 1 - \exp \left[ - \left( \frac{\tau}{b} \right)^a \right] \rightarrow 1 = \left( \frac{\tau}{b} \right)^a \rightarrow \tau = b.$$

Where  $a = 1$  (statistical failures),  $\tau$  is identical to the MTBF and the lifetime of a system of  $n$  identical components is  $\tau/n$ , as shown in Section 2. The annual replacement of spares of a system of  $n$  identical components is then  $n/(\tau/1 \text{ year})$ .

However, this consideration is only correct where there is an equilibrium in decay and replacement. If we have a component with a long lifetime, let us say 20 years, the probability for failure in the first few years is very small. Thus the yearly reinvestment of cost per unit  $\times n/\tau$  is unnecessary.

Therefore, for adequate but economical repurchases of spares, we need to take into account that we are starting with a complete spare inventory and that it might take years before the available spare is used and needs to be replaced. A compact formula for the replacement of spares over time is given by Hoffstaetter and Willeke (unpublished data) and is briefly presented in the following.

Let us consider a system with  $n$  identical components. The failure of components of the system is described by the function  $F(t)$ , which may be modelled using a Weibull distribution. This assumption is not necessary for the assessment we are going to make, but it will allow us to calculate examples easily, to emphasize the importance of the considerations that will follow.

We remember that the failure density distribution function  $f(t)$  is the time derivative of  $F(t)$ . We will further define the rate of component replacement as  $R(t)$

$$f(t) = \frac{d}{dt} F(t).$$

We now consider replacements of failed components. The number of initially installed components to be replaced after some step in time  $\delta t$  is  $R_0(t) \cdot \delta t = f(t) \cdot \delta t$ . However, the replaced parts  $f(t) \cdot \delta t$  with  $0 < t' < t$  are also subject to failure, at a rate  $f(t - t')$  at time  $t$ , and we have to add these to the list of replacements (the infinitesimally small factor  $\delta t$  will be dropped from now on):

$$R_1(t) = f(t) + \int_0^t f(t') \cdot f(t - t') dt'.$$

The replacements of the already replaced parts, however, are also subject to failure, so we are left with an infinite number of nested integrals:

$$R(t) = f(t) + \int_0^t dt' f(t') \left[ f(t-t') + \int_0^{t-t'} dt'' f(t'') \left[ f(t-t'') \int_0^{t-t'-t''} dt''' f(t''') f(t-t''') \dots \right] \right]$$

The replacement at time  $t$  replaces initially installed components with failure rate  $f(t)$ , and components that were replaced at any earlier time  $t'$  are replaced at the rate  $f(t-t')$ . This consideration leads to the compact integral equation

$$R(t) = f(t) + \int_0^t dt' R(t') \cdot f(t-t').$$

Note that the function  $f(t)$  can be assumed to be zero for  $t < 0$ . Any other value does not make sense, since  $f(t)$  describes the failures of parts that did not yet exist at  $t < 0$ . Parts that do not yet exist also do not have to be replaced, so that  $R(t < 0) = 0$  as well. For this reason, the integrand is

$$R(t') \cdot f(t-t') = 0 \text{ for } t' < 0 \text{ and for } t' > t.$$

Thus, we can rewrite the equation as

$$R(t) = f(t) + \int_{-\infty}^{\infty} dt' R(t') \cdot f(t-t').$$

We use the fact that the Fourier transform of a convolution

$$\int_{-\infty}^{\infty} dt' f(t') \cdot R(t-t')$$

in the time domain equals the product of the Fourier transform of  $f(t)$  and  $R(t)$ . With

$$(\tilde{f}(\omega), \tilde{R}(\omega)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (f(t), R(t)) \cdot \exp[-i\omega t] dt.$$

We arrive at

$$\tilde{R}(\omega) = \tilde{f}(\omega) + \sqrt{2\pi} \cdot \tilde{f}(\omega) \tilde{R}(\omega),$$

with the solution

$$\tilde{R}(\omega) = \frac{\tilde{f}(\omega)}{1 - \sqrt{2\pi} \cdot \tilde{f}(\omega)}.$$

The replacement function is thus

$$R(t) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \tilde{R}(\omega) \cdot \exp[i\omega t] d\omega = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \frac{\tilde{f}(\omega)}{1 - \sqrt{2\pi} \cdot \tilde{f}(\omega)} \cdot \exp[i\omega t] d\omega.$$

Note that

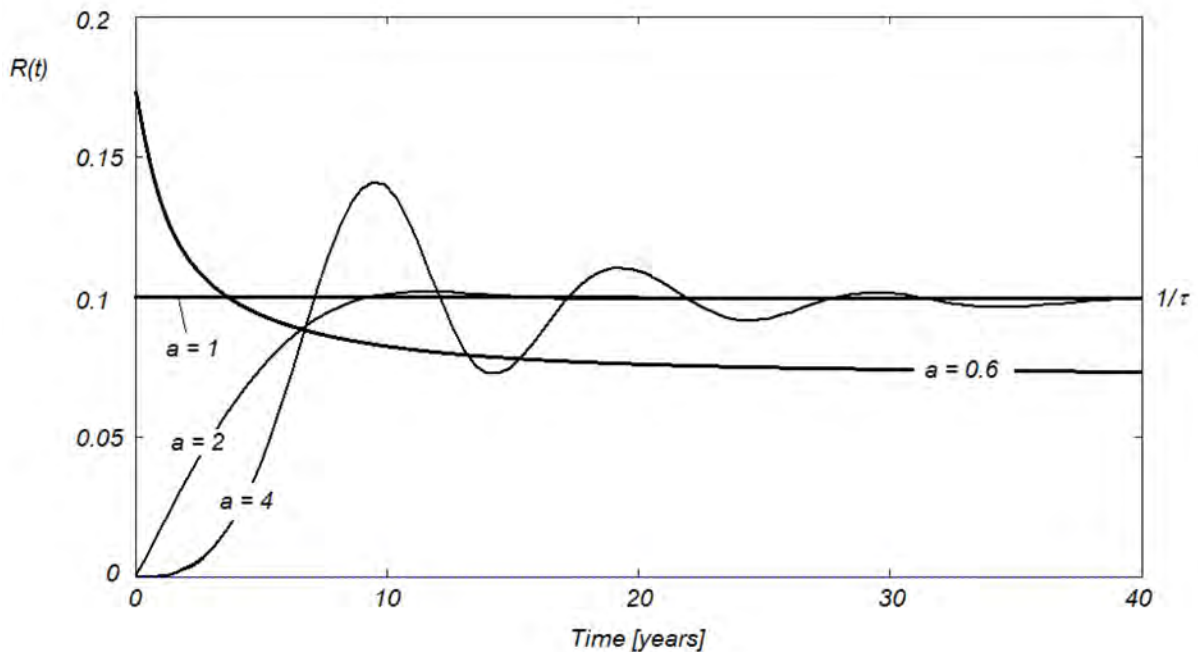
$$\tilde{f}(0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) dt = \frac{1}{2\pi},$$

and the integrand in the expression has a pole at  $\omega = 0$ , since

$$\sqrt{2\pi} \cdot \tilde{f}(\omega)_{(\omega=0)} = \int_0^{\infty} f(t) dt = F(\infty) = 1,$$

which requires some care in the integration. This pole reflects the fact that for times much longer than the lifetime  $\tau$ , the replacement function approaches the constant  $1/\tau$ , so that the time integral of  $R$ ,  $\bar{R}(0) = \int_{-\infty}^{\infty} R(t')dt'$ , tends to infinity.

The replacement function for a system with Weibull parameter  $a = 1$  is the constant  $1/t$  over the entire range from zero to infinity. For  $a > 1$ , the replacement function is zero for  $t = 0$ , around  $t = \tau$ , the rate is larger than  $1/\tau$  and  $t$  will oscillate around  $1/\tau$  with a frequency  $1/\tau$  and decreasing amplitude. For  $a < 1$ , the replacement function departs quickly from its start value and approaches zero very slowly. So in this case, the replacement will not approach the limit of  $1/\tau$ . This is due to the fact that for long times, the failure rate is approaching zero. Figure 12 shows the replacement as a function of time for various values of  $a$  ( $a = 3, 2, 1$ , and  $0.6$ ).



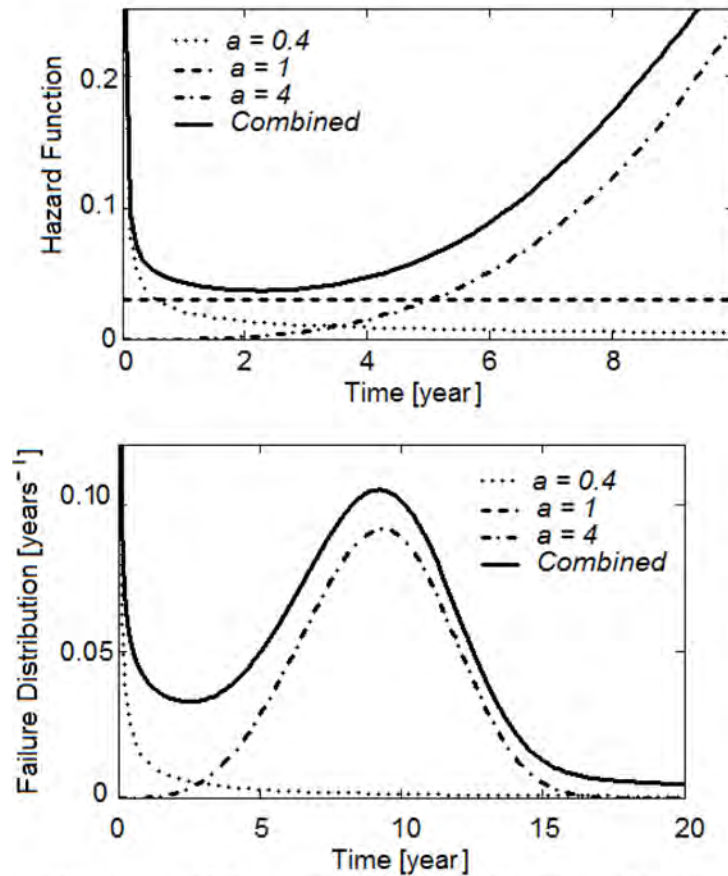
**Fig. 12:** Replacement of used spares as a function of time for various system parameters (Weibull parameter  $a$ ) for  $a = 1$  (constant failure rate),  $a = 2, 3$  (failure due to ageing and wear-out),  $a = 0.5$  (early failure).

Let us consider an example:

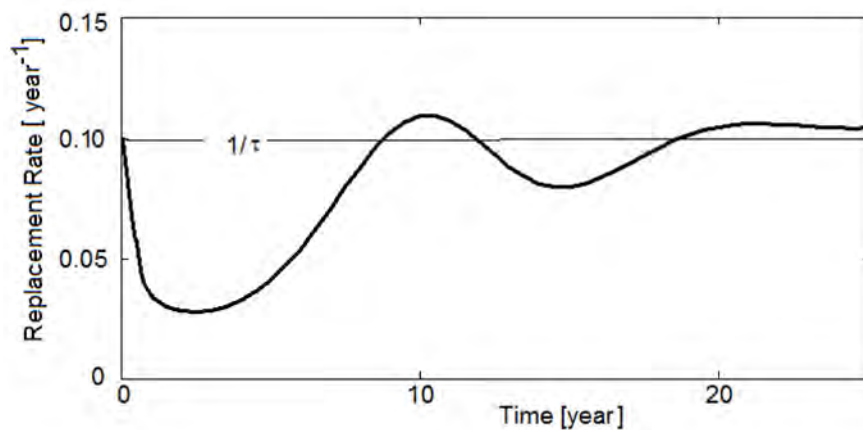
Consider a system that has a certain number of components  $N$ . This number  $N$  is assumed sufficiently large that statistical fluctuations are smaller than the systematic trends. A fraction  $c_1 = 10\%$ <sup>1</sup> fail prematurely, with a lifetime parameter of  $\tau_1 = 1$  year and a form factor  $a = 0.4$ . A fraction  $c_2 = 30\%$  has statistical failure characteristics with a lifetime of 20 years and the remainder of the components fail because of wear-out, with a lifetime parameter of 20 years as well and a form factor of  $a = 4$ .

The corresponding hazard function  $\lambda(t)$  and the failure probability distribution function (called here  $f(t)$ ) are shown in Fig. 12; the replacement function  $R(t)$  is shown in Fig. 13. We see that the estimated expenses per unit of time for spare replacement start out much lower than the average  $1/\tau$ .

<sup>1</sup> This value is intentionally chosen unrealistically large to make the effect more clearly visible in the results.



**Fig. 13:** Example hazard functions and failure probability distribution. The combined hazard function is calculated by  $\Lambda(t) = f(t)/S(t)$  with  $f(t)$  and  $S(t)$  being the sum of the corresponding partial functions.



**Fig. 14:** Replacement of spares over time for a system characterized by the hazard function in Figure 13

The expected expenditure rate on spares in the first few years of operation (up to about half the lifetime) is significantly lower than the average spare consumption rate of  $1/\tau$  which is obtained asymptotically for a well matured system

$$\lim_{t \rightarrow \infty} R(t) = \frac{1}{\tau}.$$

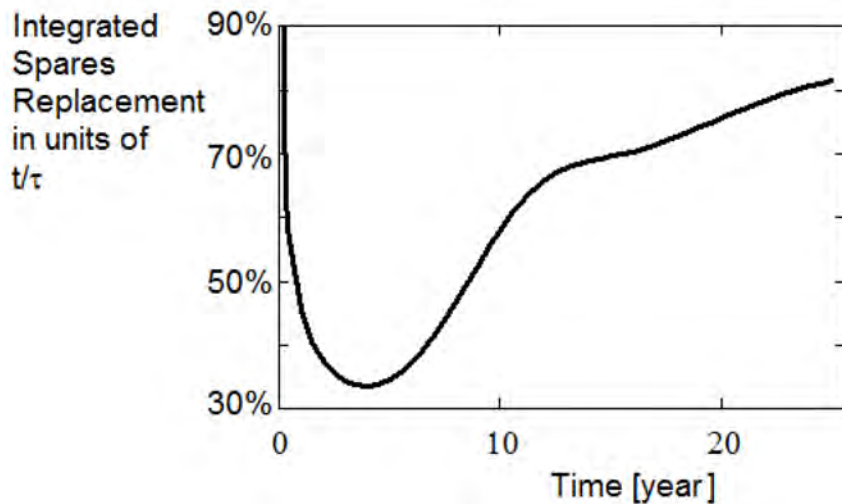
The consumption rate will exceed average spare consumption once the equipment reaches the end of its life; after this, the cost will approach the average cost while oscillating around the average.

The integrated spare replacement  $I_s$  of a large period of  $t \approx \tau$ ,

$$I_s(t) = \int_0^t R(t') dt',$$

will remain significantly below the integral  $t/\tau$  which is based on the asymptotic value of the spare consumption rate  $I/\tau$  (see Fig. 15).

Thus taking the more advanced spare cost analysis into account will release substantial funds for planning and addressing other reliability issues as they occur in the early phase of a facility's life.



**Fig. 15:** Integrated spare replacement over time is considerably lower than the average spare replacement for a mature system with a replacement rate of  $I/\tau$ .

#### 4.5 Human performance factors

Human performance issues are a topic that greatly exceeds the scope of this lecture. While it should not be completely omitted, it is impossible to do full justice to this topic, given its importance. Here, we list the most important aspects of the impact of human error on the reliability of a facility. A facility with a nearly perfect hardware system and nearly perfect controls is fairly insensitive to human error. However, the implementation of nearly perfect systems is most certainly not the most cost-efficient implementation of operations. For this reason, human intervention is required to keep systems running and to handle exceptional events. These interventions are prone to human error.

Human errors are difficult to eliminate completely but there are measures that can be reasonably taken to minimize them.

- A clear definition of the line of command is mandatory, to handle exceptional operations. These need to be communicated clearly and all those involved in operations need to understand and accept these definitions. The line of command may not be static but may change for different phases and modes of operation. Particular care has to be taken to describe the transition from one mode to the other. In particular, the return to normal service needs to be addressed.
- The operational roles, responsibilities, and accountabilities also need to be defined, spelt out, and communicated clearly.
- Operational information needs to be distributed regularly in briefings, and shift turnover meetings, to avoid information gaps or misunderstandings, which may lead to failure and operational inefficiencies.

- It is important to set all procedure and operational rules in writing. Care should be taken to prepare such write-ups and the consensus of all those involved should be secured.
- Automation of operational procedures is advisable wherever feasible, affordable, and safe, to avoid overloading operators, especially in emergency situations.
- Facility operators should have a solid base training, which should emphasize the handling of exceptional situations.
- Comprehensive system information should be available online or in the form of hard copies.
- Operational software should take ergonomic considerations into account. The use of colour to convey information is not recognizable by everybody and should be minimized.
- Obviously a comprehensive, well optimized alarm system is essential, to recover quickly from faults and trips.
- Access to equipment and controls should be carefully optimized. While access limitations might stand in the way of quick recovery from a failure, the absence of access limitations can be a cause of failures through false settings and misunderstandings.
- Ambiguous naming is a frequent source of misunderstanding, uncoordinated action, and loss of operational efficiency. Names and labels need to be unambiguous. The information on naming needs to be communicated well; the wrong use of names may not be ignored or considered insignificant.

## 5 Closing remarks

This report intends to provide an overview of considerations that are related to high-reliability operations of accelerator-based science user facilities. It would have been beyond the scope of this report to cover each of the topics comprehensively. For this reason, this write-up should be considered as an encouragement to study the areas discussed in this report in more detail. Quite a number of textbooks and scientific and technical publications on reliability engineering, human performance issues, and system maintenance are available for further study.

## References

- [1] W. Weibull, Trans. R. Inst. Technol., Stockholm, Sweden **27** (1949).
- [2] F. Bayle and A. Mettas, Acceleration models in reliability predictions justification and improvements, 2010 Reliability and Maintainability Symposium, San Jose, CA, USA (2010), <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5443868>
- [3] P. Bellomo and D. MacNair, SLAC next-generation high availability (HA) magnet power system, ARW2009, Vancouver 2009, <http://indico.triumf.ca/contributionDisplay.py?contribId=5&sessionId=7&confId=749>
- [4] C. Chen *et al.*, IEEE Power Electron. Specialists Conference, Aachen (2004), Proceedings: Power Electronics Specialists Conference, 2004. PESC 04. 2004 IEEE 35th Annual (Volume:6)
- [5] <http://www.rubycon.co.jp/de/products/alumi/pdf/Life.pdf>
- [6] [http://standards.sae.org/ja1011\\_200908/](http://standards.sae.org/ja1011_200908/)



# Beam Transfer and Machine Protection

V. Kain

CERN, Geneva, Switzerland

## Abstract

Beam transfer, such as injection into or extraction from an accelerator, is one of the most critical moments in terms of machine protection in a high-intensity machine. Special equipment is used and machine protection aspects have to be taken into account in the design of the beam transfer concepts. A brief introduction of the principles of beam transfer and the equipment involved will be given in this lecture. The main concepts of machine protection for injection and extraction will be presented, with examples from the CERN SPS and LHC.

## Keywords

Beam Transfer; machine protection; kicker systems; interlocking system; single-turn failures.

## 1 Introduction

Injection and extraction from an accelerator are critical operations on short time-scales with a high risk of significant beam loss in the case of equipment malfunction or badly adjusted parameters. Figure 1, which is taken from a presentation at an LHC workshop on beam-induced quenches, highlights how prominent losses induced by beam transfer actions are in the day-to-day operation of accelerators. The example is from the Relativistic Heavy Ion Collider accelerator at Brookhaven National Laboratory. While in the first part of this lecture some of the typical injection, extraction, and transfer line configurations, along with associated equipment, are discussed, the second part is dedicated to a presentation of the machine protection systems developed for the LHC injection and beam dump system, as examples for machine protection systems for beam transfer.

## 2 Injection

Various different injection techniques exist in the different machines around the world. The chosen technique depends on the requirements in terms of loss acceptance, brightness, and particle type used. This section, as well as Sections 3 and 4, is based on the lectures by B. Goddard, W. Bartmann, M. Barnes and M. Meddahi at the CERN Accelerator School [2].

### 2.0.1 *Single-turn injection for hadrons*

This type of injection is typically used for transfer between machines in an accelerator chain in association with *boxcar stacking*, where a larger ring is filled sequentially by a smaller one. Angle and position errors at the injection point lead to injection oscillations, while optical errors lead to betatron mismatch. Both can cause emittance blow-up. As most of the machine protection concepts will be introduced in connection with this type of injection, it will be described in detail in Section 2.1.

### 2.0.2 *Multiturn injection for hadrons*

In the case of multiturn injection for hadrons, phase-space painting is used to increase the total circulating intensity. The variant of  $H^-$ -injection allows injection into the same phase-space area after stripping the  $H^-$  ions to protons. In this way the intensity is increased while keeping the emittance small.

## Beam losses at RHIC

- Injection mismatch
- Poor beam lifetime
  - Bad orbit, working point, chromaticity
  - Emittance growth due to weak resonances, beam–beam, intra-beam, beam–gas interactions
- Beam instability
  - Transition crossing, strong orbital resonance, etc.
  - Can be fast
- system failure
  - Injection kicker mistiming, injection damper misphasing
  - Oscillation of a magnet power supply
  - Abort kicker dysfunction
  - RF cavity failure
    - Cause debunched beam

Brookhaven Science Associates

BROOKHAVEN  
NATIONAL LABORATORY

**Fig. 1:** Most frequent causes of beam loss at the Relativistic Heavy Ion Collider accelerator at Brookhaven National Laboratory. Issues during beam transfer, e.g., injection and beam abort, are prominent. Presentation by M. Bei at the Workshop on Beam-induced Quenches at CERN [1].

### 2.0.3 Lepton injection

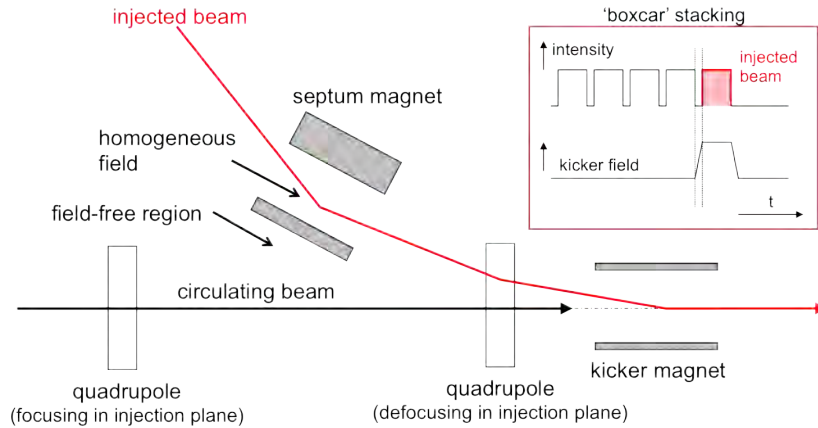
In the case of high-energy leptons, as were used for the LEP accelerator, injection precision and matching is of less concern for emittance blow-up, owing to damping with synchrotron radiation. This fact can be used to inject beams off-momentum or at an angle and increase the intensity in the same phase-space volume after damping. These techniques are called betatron and synchrotron accumulation, respectively.

## 2.1 Single-turn injection and injection equipment

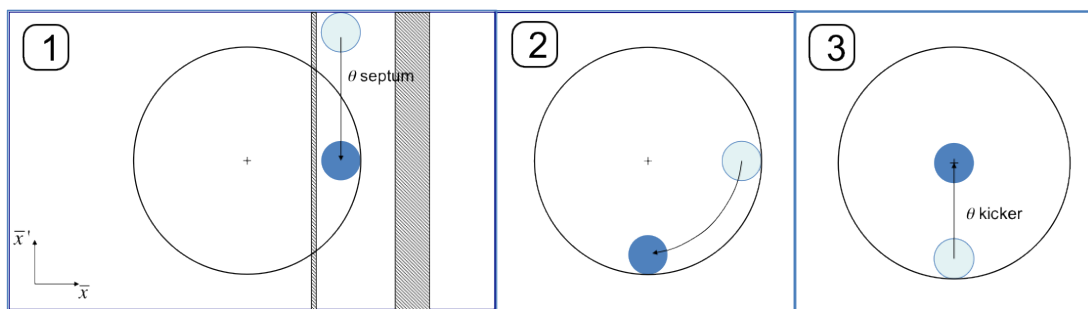
Figure 2 shows the principle of single-turn injection. A transfer line brings the beam close to the orbit of the circular machine. The last two magnetic elements are special injection magnets. The first of these is called a septum magnet. It is already so close to the circulating beam that the iron yoke and coils of a conventional magnet would no longer fit between the trajectory of the injected beam and the orbit of the circulating beam. Instead, a septum magnet has an aperture with a magnetic field for the injected beam and another aperture without a magnetic field through which the circulating beam passes. Between the two regions, there is a thin septum. The septum deflects the beam onto the closed orbit at the centre of a kicker magnet. The kicker magnet compensates for the remaining angle. The kicker magnet, however, has only one aperture, through which both the injected and circulating beams pass. Kicker magnets are pulsed magnets with very fast rise and fall times, such that they provide the required magnetic field at the moment of the passage of the injected beam and no field during the passage of the circulating beam. The septum and kicker magnets are typically installed on either side of a defocusing quadrupole in the plane of the kicker magnet to minimize the kick strength. The actions of the two magnet systems in normalized phase-space are shown in Fig. 3.

### 2.1.1 Septum magnet

A general introduction to septa can be found in Ref. [3]. Two main types of septa exist – electrostatic septa and magnetic septa. Electrostatic septa have a very thin (typically  $\leq 100 \mu\text{m}$ ) separation between



**Fig. 2:** Single-turn injection



**Fig. 3:** The single-turn injection process in normalized phase-space. The beam indicated by the blue circle undergoes a large deflection as it passes through the septum magnet (plot 1), then travels through a  $\pi/2$  phase advance to meet the kicker magnet (plot 2). The kicker deflection aligns the beam on the central orbit (plot 3).

the zero-field and high-field regions. They are often used for extraction systems, to minimize losses at extraction and the required kick strength or strength of the following septa.

Magnetic septa are pulsed or DC dipole magnets with a 2–20 mm separation between the zero-field and high-field regions. Figure 4 shows a direct-drive pulsed septum magnet. This type of magnet is often used under vacuum to minimize the distance between the circulating and deflected beams. To minimize the self-inductance and for mechanical reasons, the coil generally consists of a single turn, resulting in large required currents of typically 5–25 kA. Single-turn injection uses magnetic septa.

**2.1.2 Kicker magnet**

A full lecture on kicker magnets can be found in Ref. [4]. Kicker magnets typically produce rectangular field pulses with fast rise and fall times. The field strength is lower than can be achieved with septum magnets. For comparison, the LHC injection septum provides roughly 12 mrad of bending angle in five magnets, whereas the injection kicker magnets provide 0.8 mrad in four magnets. The injection energy in the LHC is 450 GeV. To achieve fast rise and fall times, a voltage pulse is pre-charged in a so-called pulse forming line, a coaxial cable, or a pulse forming network, all of which are lumped elements. This voltage pulse is launched towards the kicker when the *main switch* is closed. The main sub-systems of a typical kicker system are shown in Fig. 5. Pulse forming networks and lines accumulate electric energy over a comparatively long time from a power supply (a few milliseconds in the case of a fast resonant charging power supply) and then release it in the form of a square pulse of relatively short duration (a few microseconds). Figure 6 shows the LHC injection kicker magnet.

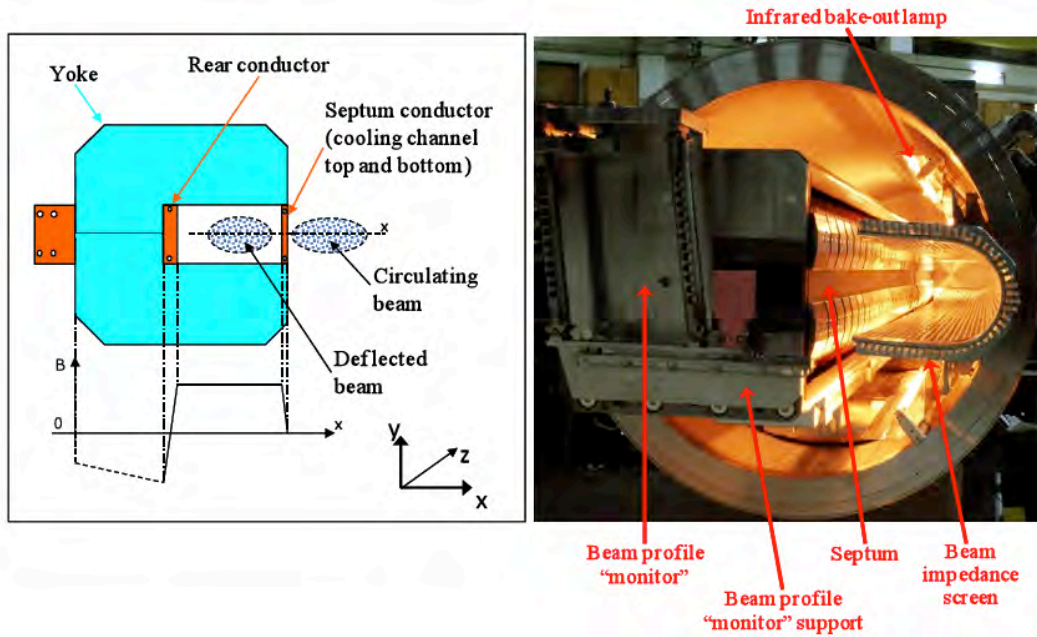


Fig. 4: Direct-drive pulsed septum

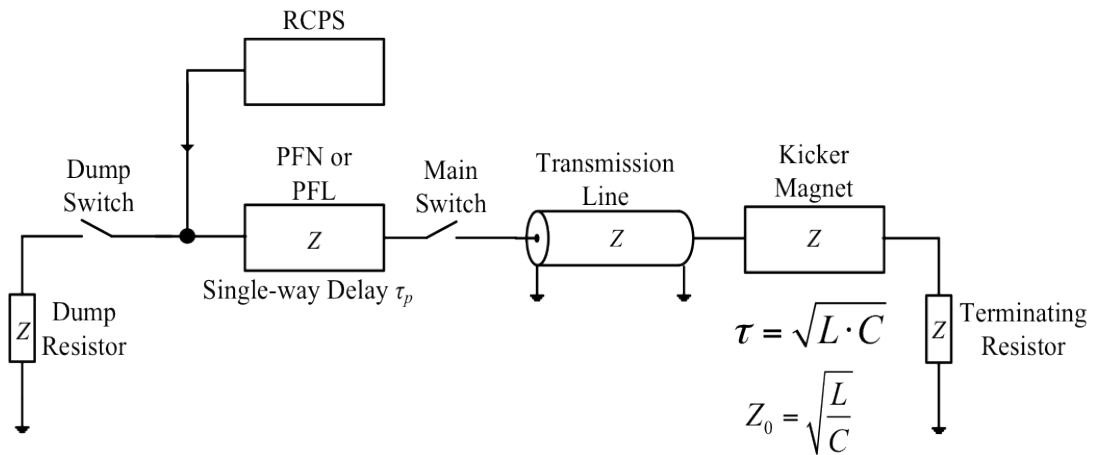


Fig. 5: The main sub-systems of a kicker system: The pulse forming network (PFN) or pulse forming line (PFL) is charged to a voltage  $V_p$  by the resonant charging power supply (RCPS). When the *Main Switch* closes a pulse of magnitude  $V_p/2$  is launched towards the magnet through the transmission line (coaxial cable). The impedances in the system should be matched, to avoid reflections. The length of the pulse in the magnet can be controlled between 0 and  $2\tau = \sqrt{L \cdot C}$  by adjusting the timing of the dump switch relative to the main switch.

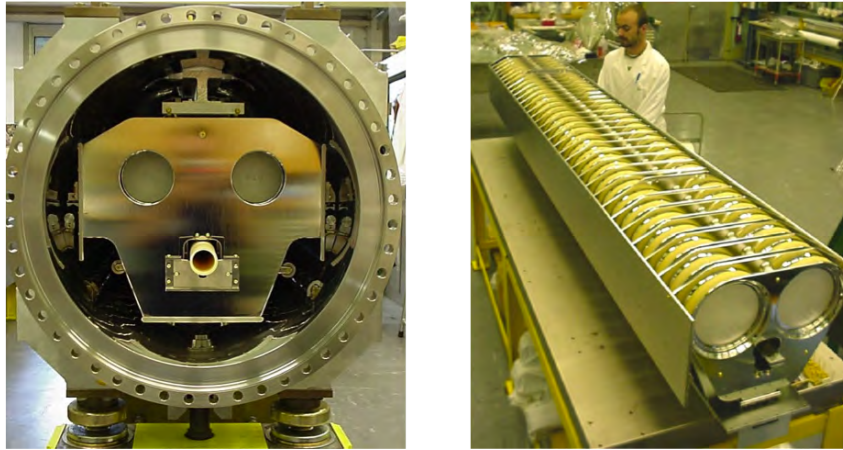
A typical kicker cross-section is shown in Fig. 7. The characteristics of the field in the kicker gap are:

$$B = \frac{\mu_0 I}{g}, \tag{1}$$

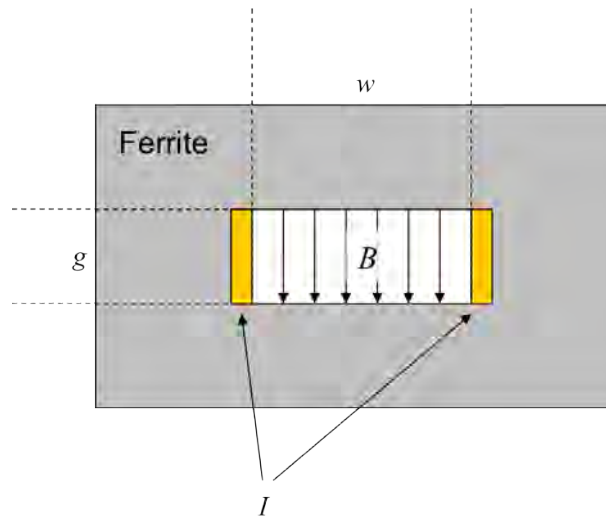
$$L = \frac{\mu_0 w l}{g}, \tag{2}$$

$$\frac{dI}{dt} = \frac{V}{L}, \tag{3}$$

where  $B$  is the magnetic field,  $I$  the current,  $g$  the gap height,  $l$  the length,  $w$  the gap width,  $L$  the



**Fig. 6:** The LHC injection kicker MKI



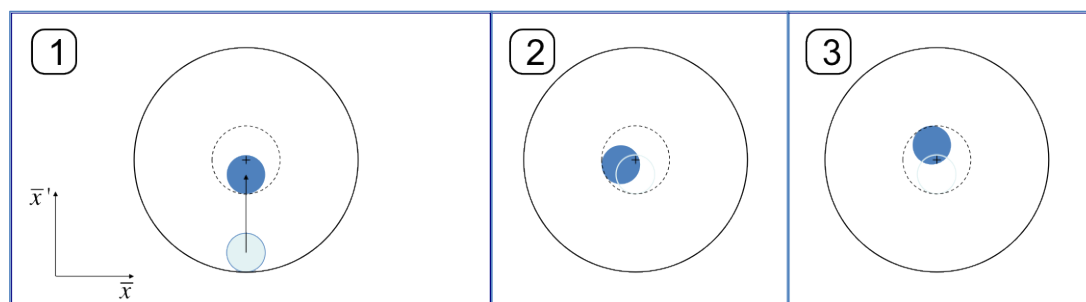
**Fig. 7:** Simplified cross-section of kicker magnet:  $B$ , magnetic field;  $g$ , gap height;  $I$ , current;  $w$ , gap width

inductance, and  $V$  the kicker voltage. A typical kicker  $dI/dt$  would be 3 kA in 1  $\mu$ s. In general, kickers have a low inductance. Nevertheless, for short rise times, very high voltages of the order of several tens of kilovolts are required. Kicker switches, as indicated in Fig. 5, therefore need to hold high voltages and to switch high currents within short rise times. *Thyratron* switches (gas tubes) or solid-state switches are frequently used for this purpose. As can be anticipated, kicker switches are associated with a number of possible failure scenarios. There are two main categories.

- **Erratic turn-on:** The switches turn on spontaneously and the kickers fire asynchronously, e.g., during the circulating beam passage instead of the injected beam.
- **Missing:** The switch does not fire.

So-called magnet *flash-overs* are another kicker failure category, and are independent of the switches. A flash-over is a discharge between kicker electrodes while the kicker has high field. Depending on which cell, and hence where in the kicker, the flash-over occurs, it results in a reduced or even increased kick strength, owing to reflection.

As the bending angle exerted by the kicker magnets is generally large, kicker failures belong to the group of very fast beam loss mechanisms. If the kicker magnets are being used to inject a beam, the



**Fig. 8:** Normalized phase-space at the injection point in the horizontal plane over three turns. The blue beam is injected with an angle error, leading to an oscillation of the particle distribution around the central orbit turn after turn.

resulting *injection oscillations* would be too large to establish a circulating beam. The beam would be lost on the first turn. In case of an asynchronous kick on the circulating beam, the resulting oscillations around the ring would be so large that the beam would be lost within a single turn. Not all kicker failures can be prevented by design but the risk for some can be minimized. For example, fast resonant charging power supplies are used to charge the pulse forming networks quickly. The charging then only takes place a few milliseconds before the kick is required. The probability of an error is reduced in this way. Many kicker failures require *passive protection*, see Section 5.2.

## 2.2 Mismatch at injection

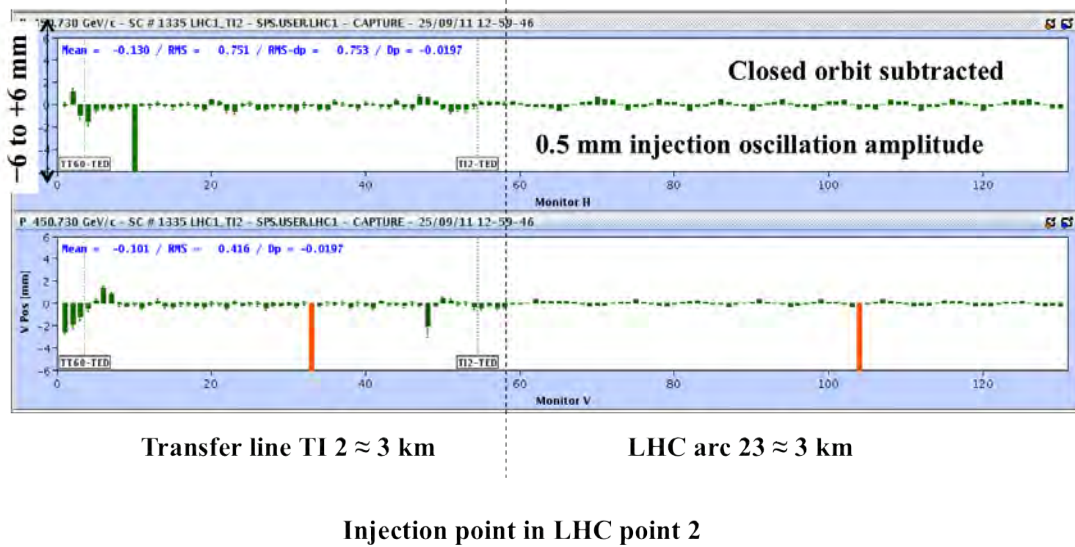
Errors during injection can lead to beam loss due to *mismatch*.

### 2.2.1 Steering errors

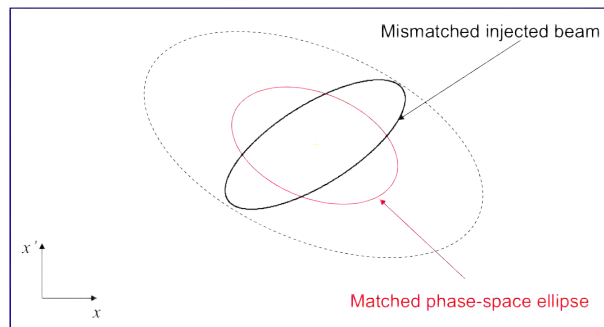
Steering errors, a difference in position and angle (e.g.,  $x$  and  $x'$ ) between the circulating orbit and the injected trajectory at the injection point, lead to *injection oscillations*. The injected beam will oscillate around the closed orbit. With the presence of non-linearities in the machine, this oscillation will decohere after a number of turns. As a result, the beam emittance and hence the beam sizes around the ring will increase. If the initial oscillation amplitude is large enough, it can lead to beam loss as early as the first turn. The mechanism of generating injection oscillations is illustrated in Fig. 8. Figure 9 shows the LHC injection oscillation display in the LHC control room. Instead of showing the trajectory turn after turn at one beam position monitor, the injection oscillation display shows the trajectory along many beam position monitors, covering many oscillation phases.

### 2.2.2 Optical mismatch

The Twiss parameters at the injection point define the shape and orientation of the particles' phase-space ellipse. One speaks of optical mismatch if the shape and orientation of the ellipse of the injected beam does not match the ellipse of the circulating beam at the injection point. An example of such a situation is shown in Fig. 10. The black mismatched ellipse will change its orientation turn after turn, with its extremities following the contour of the dashed ellipse. This causes beam size beating at the injection point or any other point of observation. Non-linear effects, e.g., magnetic field multipoles, however, will introduce amplitude dependent differences in particle motion and, over many turns, a phase-space oscillation is transformed into an emittance increase. This is illustrated in Fig. 11. So-called filamentation eventually fills a larger ellipse. The larger ellipse has the same shape and orientation as the matched one.



**Fig. 9:** Injection oscillation display from the LHC control room, showing the trajectory with respect to the reference in the transfer line and the first turn minus the closed orbit in the first 3 km of the LHC.

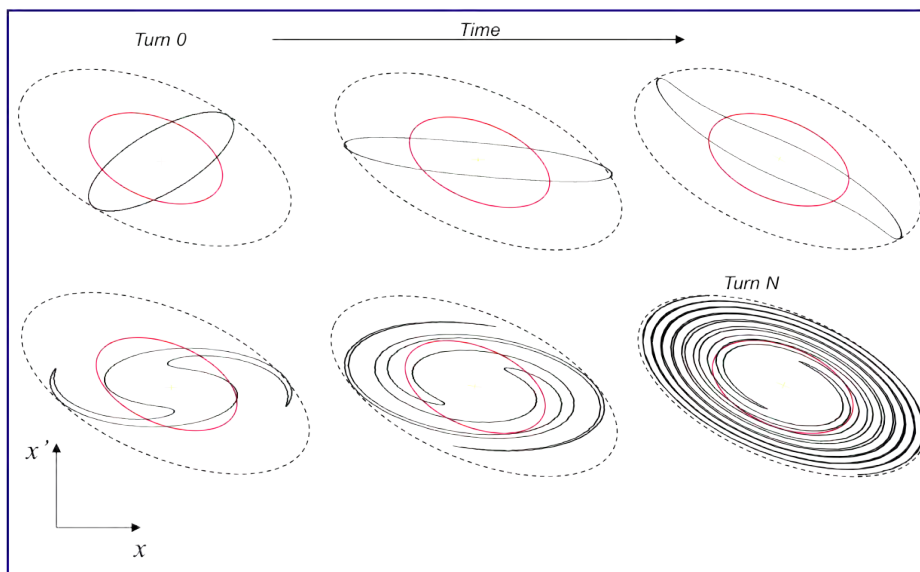


**Fig. 10:** Phase-space plot showing an optical mismatch at the injection point

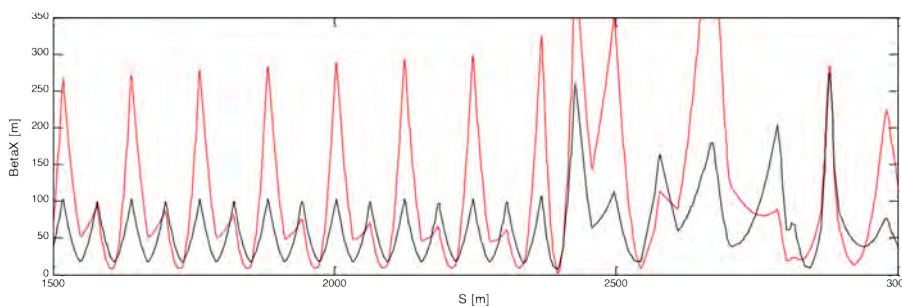
### 3 Transfer lines

Transfer lines transport the beam between accelerators and onto targets, dumps, and instruments. The beam dynamics in a transfer line are not only determined by the strength of the elements in the transfer line. Instead, the Twiss parameters at any point in the line,  $\alpha(s)$ ,  $\beta(s)$ , are also a function of the initial functions,  $\alpha_1$  and  $\beta_1$ . This is illustrated in Fig. 12. The strengths of the line are kept the same but the initial conditions are modified. As a result, the beta functions change. The same is true for the dispersion function. Another difference with respect to circular machines is the effect of an error at a particular location in the line. Whereas in a circular machine the error will introduce a perturbation around the entire ring, in a transfer line the error will affect only the part of the line downstream of the error location, see Fig. 13.

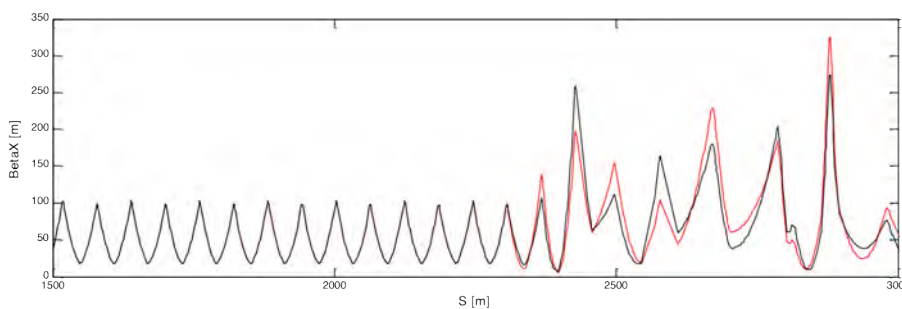
If transfer lines are used to link two machines, the Twiss parameters of the extraction point have to be propagated through the line and matched to the Twiss parameters at the injection point of the next machine by adjusting the quadrupole strengths in the line accordingly, see Fig. 14. From the previous discussion, we know that the Twiss parameters, and hence the orientation of the phase-space ellipse of the injected beam, have to be the same as the Twiss parameters, and orientation, of the phase-space ellipse of the circulating beam at the injection point, to avoid optical mismatch. The Twiss parameters can be propagated if the transfer matrix  $M_{1 \rightarrow 2}$  is known:



**Fig. 11:** Owing to non-linearities, optical mismatch will lead to emittance blow-up over many turns

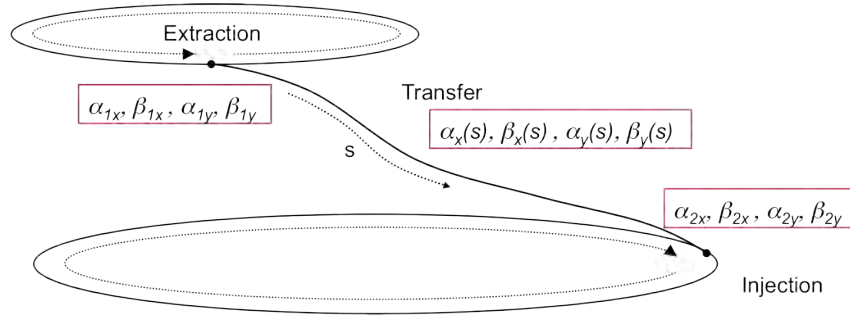


**Fig. 12:** The horizontal beta function in a transfer line as a function of the longitudinal position  $s$ : The strengths of the magnetic elements in the line are kept the same, but the initial conditions are modified to produce the red and black optics.

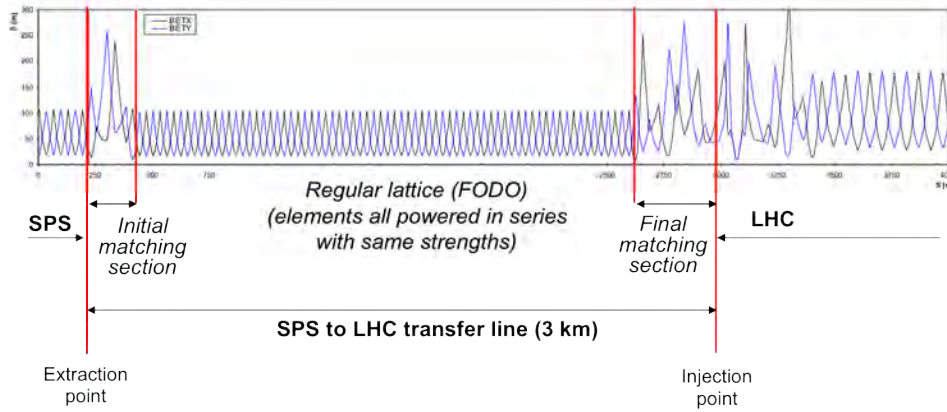


**Fig. 13:** The effect of an error in a transfer line is only seen downstream of the error location. For the red beta function, a quadrupole error was introduced at location  $s \approx 2300$  m.





**Fig. 14:** If transfer lines are used to link two machines, the Twiss parameters of the extraction point have to be propagated through the line and matched to the Twiss parameters at the injection point of the next machine.



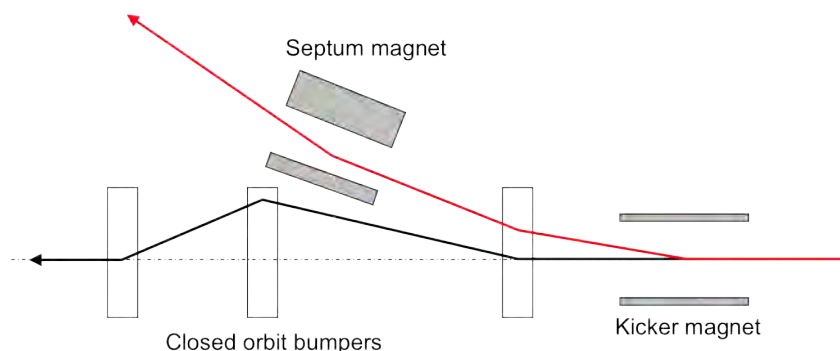
**Fig. 15:** Layout of the SPS to LHC transfer line with a long central section with a regular lattice and initial and final matching sections.

$$\begin{pmatrix} x \\ x' \end{pmatrix}_{s_2} = M_{1 \rightarrow 2} \begin{pmatrix} x \\ x' \end{pmatrix}_{s_1} = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix} \cdot \begin{pmatrix} x \\ x' \end{pmatrix}_{s_1}, \quad (4)$$

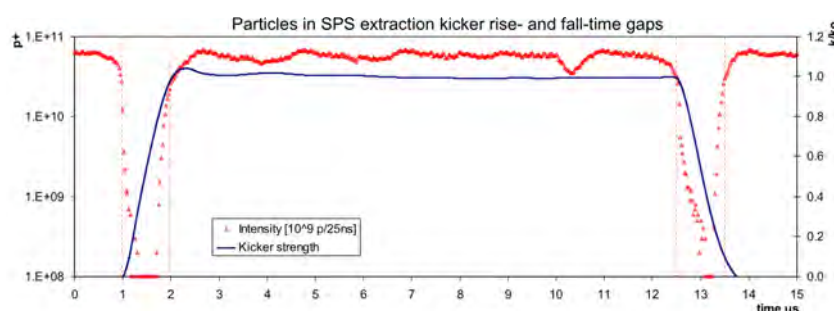
$$\begin{pmatrix} \beta_2 \\ \alpha_2 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} C^2 & -2CS & S^2 \\ -CC' & CS' + SC' & -SS' \\ C'^2 & -2C'S' & S'^2 \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \alpha_1 \\ \gamma_1 \end{pmatrix}. \quad (5)$$

Typically, eight variables need to be matched ( $\alpha$ ,  $\beta$ , the dispersion  $D$  and its derivative  $D'$  in both planes). Frequently, constraints such as phase advance requirements for transfer line collimators or insertions for special equipment such as stripping foils have to be respected. Independently powered quadrupole magnets with adjustable strength are used for this purpose. If the transfer lines are long, the problem can be simplified by designing the lines in separate sections. The main part of the line then consists of a regular central section with a regular lattice, e.g., a FODO or doublet with the quadrupoles at a regular spacing and powered in series. At the end and the beginning of the line, initial and final matching sections with independently powered quadrupoles are installed. An example of this type of layout is the LHC transfer line shown in Fig. 15.

To counteract trajectory deviations due to magnet misalignments and field and powering errors, transfer lines are usually also equipped with beam position monitors and independently powered small dipole corrector magnets. Horizontal correctors and beam position monitors are located at large  $\beta_x$  and vertical correctors and beam position monitors are located at large  $\beta_y$ .



**Fig. 16:** Travelling from right to left, the beam is deflected by the kicker magnet into the septum, which bends the beam into the transfer line.



**Fig. 17:** One turn takes  $23 \mu\text{s}$  for the SPS beam at  $400 \text{ GeV}/c$ . For the CERN Neutrino to Gran Sasso beam, two batches, each of  $10 \mu\text{s}$ , fill the SPS. The batches are extracted one after the other at intervals of  $50 \text{ ms}$ . The extraction kickers have to rise between the batches. The beam in the gaps between the batches is lost on the extraction septum blade or in the SPS.

## 4 Extraction

As in the case of injection, there are various different techniques to extract the beam from a circular machine. A few methods will be briefly summarized.

### 4.1 Single-turn fast extraction

As for single-turn injection, kicker magnets and septa are used for single-turn fast extraction. To reduce the required kicker magnet strength, the beam is sometimes moved close to the septum by a closed orbit bump. An example of a typical fast extraction layout is shown in Fig. 16. The kicker magnet deflects the entire beam into the septum in a single turn. The septum then bends the beam into the transfer line. The smallest deflection angles are required for a phase advance of  $\pi/2$  between the kicker and the septum. The septum deflection may also be in the other plane from the kicker deflection. Losses around the ring or in the extraction channel at the moment of the extraction process can originate from orbit errors (e.g., closed orbit bump errors), kicker failures, kick synchronization errors or particles in the extraction gap during which the extraction kicker field rises (e.g., uncaptured beam). An example of a kicker rising in an extraction gap populated with particles is shown in Fig. 17 for the CERN Neutrino to Gran Sasso fast extraction from the CERN SPS.

### 4.2 Non-resonant multiturn extraction

This type of extraction is used to fill a larger machine from a smaller one over several turns. The beam is sliced into equal parts by a fast orbit bump that puts the whole beam onto the septum. The beam is

**Table 1:** LHC nominal beam parameters

|                | <b>Energy<br/>[GeV]</b> | <b>Number of bunches</b> | <b>Bunch intensity</b> | <b>Stored energy<br/>[MJ]</b> |
|----------------|-------------------------|--------------------------|------------------------|-------------------------------|
| SPS extraction | 450                     | 288                      | $1.7 \times 10^{11}$   | 2.4                           |
| LHC top energy | 7000                    | 2808                     | $1.7 \times 10^{11}$   | 360                           |

extracted in a few turns with the machine tune rotating the beam. This type of extraction is an intrinsically high-loss process.

### 4.3 Resonant multiturn extraction

Multipole fields from sextupoles or octupoles are used to distort the circular normalized phase-space and define a stable area delimited by unstable fixed points. The beam is slowly extracted by approaching the respective resonance increasingly closely and going across it. By adjusting the tune change speed, the extracted intensity can be controlled and a constant particle spill can be achieved.

### 4.4 Resonant low-loss multiturn extraction

An alternative to the non-resonant multiturn extraction with lower losses is the resonant low-loss multiturn extraction. Non-linear fields (sextupole and octupole) are used to create islands of stability in phase-space. The particles are driven into the islands through tune variations. The islands are separated further in phase-space by varying the non-linear fields. The islands are then extracted in a similar manner as for the non-resonant multiturn extraction with a fast bump turn after turn.

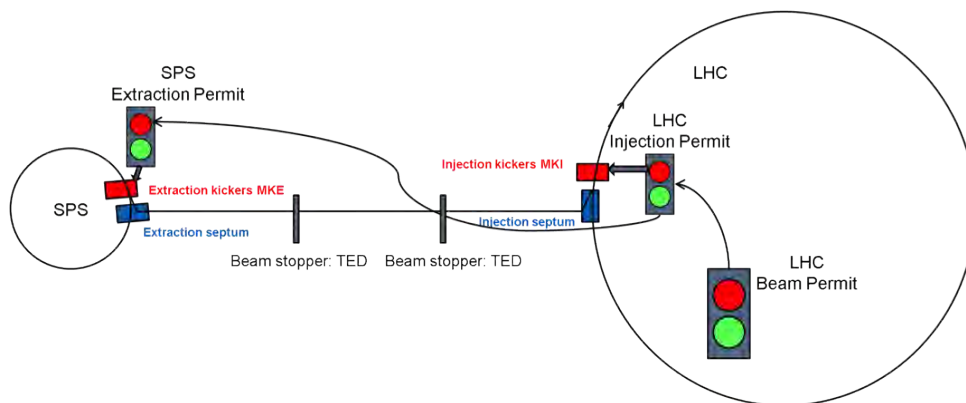
## 5 Machine protection for beam transfer

The typical concepts for machine protection systems for beam transfer will be introduced with the example of the SPS to LHC injection process and the LHC beam dumping system. The LHC is filled with 12 injections from the SPS. The extraction system in the SPS is a fast extraction using a large closed orbit bump of  $\approx 35$  mm amplitude in the horizontal plane and horizontal kickers and septa. The LHC transfer lines are 3 km long. The LHC injection system is a single-turn injection consisting of horizontal septa magnets and vertical kicker magnets that deflect the beam on the LHC closed orbit. The LHC beam dump system is a fast extraction system consisting of a horizontal kicker system and vertical septum magnets. The beam is deflected into a 900 m dump channel, at the end of which the beam dump block is located.

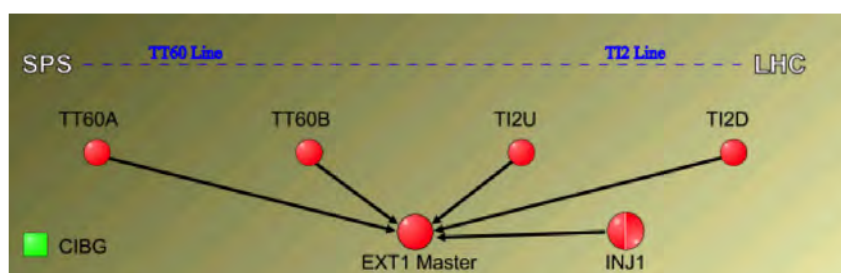
The design parameters of the LHC proton beams in the SPS at the moment of extraction and after acceleration in the LHC are summarized in Table 1. The equipment damage limit has been calculated and measured to be of the order of  $100 \text{ J/cm}^3$ , which corresponds to only  $\approx 5\%$  of the injected intensity in the LHC.

The basics of injection and extraction protection can be summarized as:

- **Ensure correct settings:** e.g., energy tracking systems, power supply surveillance;
- **Ensure kicker synchronization:** particle-free gaps (i.e., gap cleaning), locking extraction kicker triggers to RF revolution frequency, etc.;
- **Provide passive protection system:** in the form of collimators and absorbers against fast possible failure mechanisms and unavoidable kicker failures.



**Fig. 18:** Layout of the SPS to LHC transfer with the extraction kicker MKE in the SPS, the transfer line (TT60 and TI2) and injection kicker MKI in the LHC, with corresponding extraction and injection permits.



**Fig. 19:** Monitoring systems associated with equipment in the different areas involved in the SPS to LHC transfer are combined into sub-results, which are input to the extraction permit. Each circle in the drawing represents a 'beam interlock controller'. INJ1 is the injection permit of LHC beam 1.

### 5.1 Monitoring, permits, and kickers

Different interlocking systems are installed in the SPS, the LHC, the LHC injection region, and the SPS extraction region and transfer lines. A number of systems are monitored and a Boolean status from these systems is transferred to the interlock system. The result of the combination of the different inputs is called a *permit*, i.e., injection permit, extraction permit, SPS beam permit, etc. These permits are each connected to a kicker system. The SPS beam permit and LHC beam permit are connected to the dump kicker systems. If the beam permit becomes *false*, the dump kickers fire and dump the beam on the respective beam dump blocks. Without an extraction permit in the SPS, the beam is not extracted from the SPS. The kickers will not fire. Similarly, the LHC injection kickers will not fire without an injection permit.

All permits involved in the SPS to LHC transfer are connected hierarchically. The beam from the SPS must not be extracted if the LHC injection permit is not *true*. The SPS extraction permit therefore depends on the LHC injection permit. Also, the beam should not be injected into the LHC if the LHC is not ready (i.e., if the LHC beam permit is *false*). The LHC beam permit is input to the LHC injection permit. Figure 18 shows an overview of the permit hierarchy of the SPS to LHC transfer. The hardware systems in the SPS and LHC to which the different monitoring systems are connected and which provide permits are LHC-type beam interlock controllers (BICs) [5]. The interlocks of equipment of the different areas in the transfer lines are combined in local BICs (e.g., TT60A, TI2U for transfer lines TI 2 and TT60, respectively), see Fig. 19.



**Fig. 20:** Graphical user interface displays of two beam interlock controllers in TT40, the upstream part of transfer line TI 8. The vacuum system, the warm magnet interlock system (WIC), the extraction kicker MKE4 state, the septa MSE and MST states, the TT40 power supply currents, the septa currents, the orbit bump corrector magnet currents, the fast magnet current change monitors (FMCM), which survey fast changes in case of failures instead of absolute current values, the optical transition radiation (OTR) screen positions, the beam loss monitors (BLM) and the extraction orbit with BPM LSS4 are all input to these interlock controllers.

**5.1.1 Which systems are monitored?**

Almost every element in the SPS to LHC transfer regions is monitored and interlocked:

- the orbit at the SPS extraction point, measured using a number of dedicated beam position monitors;
- power supply currents, measured 2 ms before SPS extraction for all extraction and transfer line circuits, including trajectory correcting dipole magnets;
- septa currents, kicker state, and kicker charging voltage;
- vacuum valves and optical transition radiation (OTR) screens, which both have to be out for high-intensity beams;
- the settings and gaps of passive protection devices;
- the beam loss reading of the beam loss monitors – if one shot generates losses above threshold, the next shot will be inhibited;
- the trajectory via the reading of beam position monitors.

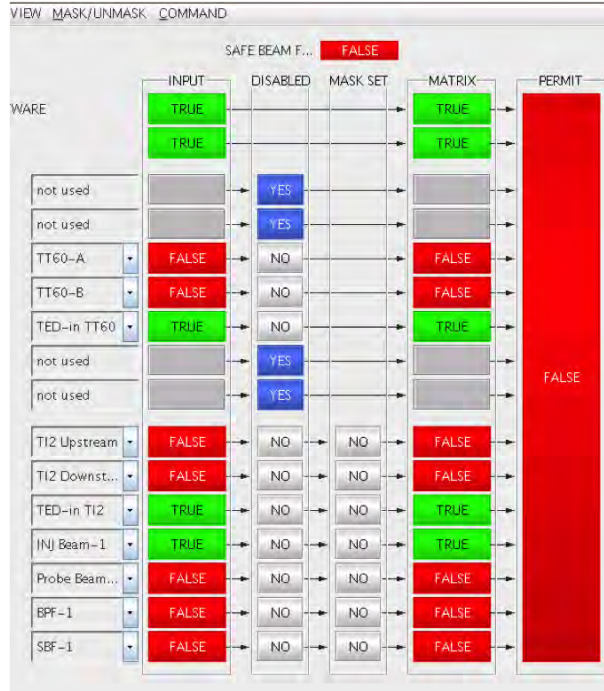
Figure 20 shows the graphical user interface display for the beam interlock controllers installed in the first part of the LHC transfer line TI 8, called TT40.



**Fig. 21:** Graphical user interface display of a beam interlock controller in TT60, which is the upstream part of LHC transfer line TI 2. Inputs 8–14 are maskable and can be disabled in case of intensity below the *set-up beam flag*. In this example, the set-up beam flag is *false* and the masks do not have an effect, i.e., in the column ‘matrix’ the masked inputs return *false* (red), despite the mask.

### 5.1.2 Masking of interlocks

Appropriate settings (e.g., of corrector magnet power converters or collimators that have to be aligned around the beam trajectory) can frequently only be established with the beam. For the first set-up, these types of interlock should, therefore, not be taken into account when generating permits. In LHC jargon, the disabling of interlocks is called ‘masking’. A number of systems can have maskable interlocks in the beam interlock controller matrix. For the LHC-type beam interlock controllers, masking is only allowed at low intensity, the so-called *set-up intensity*. The set-up intensity is derived from the SPS or LHC beam current transformers and is distributed across the machines in the form of the *set-up intensity flag*. A dedicated distribution system has been put in place – the *Safe Machine Parameter System* [6]. The interlock masks are automatically ignored if the measured intensity is above the set-up intensity limit. Figure 21 shows the graphical user interface of one of the beam interlock controllers involved in the SPS to LHC transfer. A number of inputs are maskable. Several masks have been set. As the set-up beam intensity flag is *false*, the masked interlocks are nevertheless taken into account.



**Fig. 22:** Graphical user interface of the beam interlock controller generating the SPS extraction permit for LHC beam 1. At the very top, the SPS set-up beam flag state can be seen, labelled ‘Safe Beam Flag I’. The last three inputs to the controller are the *SPS probe beam flag*, the *LHC beam presence flag* and the *LHC set-up beam flag*.

### 5.1.3 Set-up intensity and other concepts

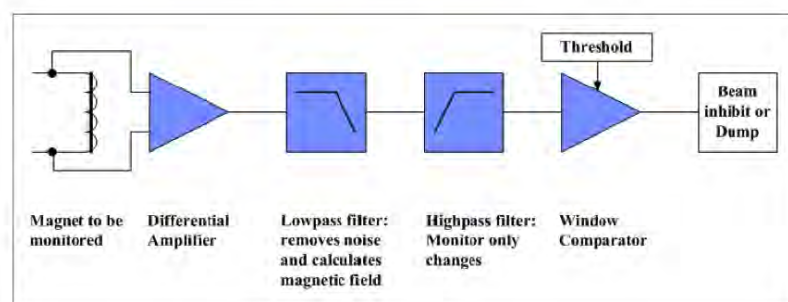
The set-up intensity has to be reasonably high, for testing under representative conditions, but must be below the equipment damage limit. For the LHC, the damage limit has been simulated and measured as  $\approx 2 \times 10^{12}$  protons at 450 GeV with a normalized emittance of  $3.5 \mu\text{m}$  [7]. Based on this limit, the set-up intensity has been defined as  $1 \times 10^{12}$  protons at injection energy. Its energy dependence has been obtained from simulation:

$$\left( \frac{E [\text{GeV}]}{450 [\text{GeV}]} \right)^{1.7} \times I [\text{p}] \leq 1 \times 10^{12}. \quad (6)$$

The LHC injection protection concept cannot rely only on monitoring systems and interlocking monitored systems. The LHC consists of thousands of sub-systems, which all have to be in the correct state to receive the beam. Only a fraction of these are directly connected to the interlock system. The SPS extraction permit for a high-intensity beam is therefore only given if there is already a beam circulating in the LHC. The so-called *probe intensity* has to be injected first, to check whether everything really is ready for the beam. The *probe intensity* is  $< 5 \times 10^9$  protons, almost a factor of 100 below the intensity of 1 nominal bunch. The *probe intensity* information is generated from the SPS beam current transformers and the *beam presence* information is generated from the sum signal of the beam position monitors in the LHC. Both are distributed in the form of flags over the *Safe Machine Parameter System*. The condition that is hard-coded in the beam interlock controller that generates the SPS extraction permit, Fig. 22, is

$$\text{SPS probe beam flag} \vee \{ \text{beam presence} \wedge [\neg(\text{LHC set-up beam flag}) \vee \text{SPS set-up beam flag}] \}. \quad (7)$$

This condition translates as: ‘high-intensity extraction from the SPS above the *probe intensity* is only allowed if there is a beam circulating in the LHC’. The set-up beam flag in the LHC has to be forced to *false* to remove potential masks if the SPS intensity is above the *set-up intensity*.



**Fig. 23:** The fast current change monitor measures the voltage of a power supply and calculates real-time current changes using a low-pass filter followed by a high-pass filter. The low-pass filter removes the noise and calculates the current; the high-pass filter removes the d.c. part and shows only current changes.

## 5.2 Protection against fast failures

Beam transfer using fast extraction systems or single-turn injection cannot be aborted once the beam transfer is triggered. All involved systems have to be checked and give the OK before the trigger arrives. For the power supplies in the SPS to LHC transfer, the last current verification happens  $\approx 2$  ms before the extraction kickers fire. Some of the circuits, however, have very low time constants for the current decay in the case of a power supply failure. The magnetic field can, therefore, still be significantly wrong at the moment of the beam passage, even if the power supply only fails at the very last moment. An example is the extraction septum circuit, with a time constant of only 23 ms and large bending angles of 12 mrad. The trajectory would be changed by roughly  $40\sigma$  within 1 ms in the case of a power supply failure. The aperture in the transfer line is only  $10\sigma$ . Instead of relying only on the classic power converter current surveillance, a faster system had to be put in place for the septa circuits and several other circuits. This system is the so-called *fast magnet current change monitor* (FMCM). Instead of measuring small currents, it measures changes in the voltage of a critical power supply. It can detect relative current changes of  $10^{-4}$  in less than 1 ms. It was successfully deployed for septa current surveillance, with a reaction time of 50  $\mu$ s for a current change of 0.1% [8]. Figure 23 shows the FMCM system.

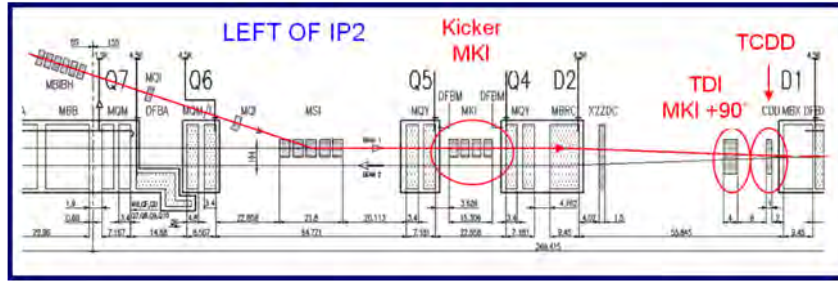
Kicker failures can develop on microsecond time-scales. No surveillance system would be fast enough to stop the beam in time. If a failure occurs during the kick pulse, other means of protecting downstream equipment from the beam impact have to be envisaged. These means are passive protection systems, in the form of absorbers and collimators.

### 5.2.1 Passive protection

The design criteria for a passive protection system can be obtained by answering a number of questions.

- Against which failure should the passive protection devices protect? This will identify where to put the devices and how many will be needed. Examples will be given later on.
- What is the damage level, quench level, or allowed radiation level of downstream equipment? This will define the required attenuation by the passive protection device.
- What is the maximum possible particle density that can impact the protection device according to the failure scenario? At which repetition rate can the failure happen? This will define the required robustness of the device and thus the maximum possible material density. Together with the required attenuation, this gives the required length of the device.
- What is the aperture that has to be protected? If it has to be set a few  $\sigma$  from the beam centre, it will have to be made movable. The required protection settings need to be evaluated.





**Fig. 24:** A top view of the LHC injection region with the horizontal MSI injection septum, the vertical inject kickers MKI, followed 90° downstream by the TDI injection protection device and the TCDD mask. The red line indicates the beam direction. Part of the transfer line can also still be seen.

The design of passive protection devices involves several different disciplines in accelerator physics and material science. Simulations for particle tracking, energy deposition in material, and thermomechanical response are required.

### 5.2.2 Examples of passive protection devices in the LHC injection protection system

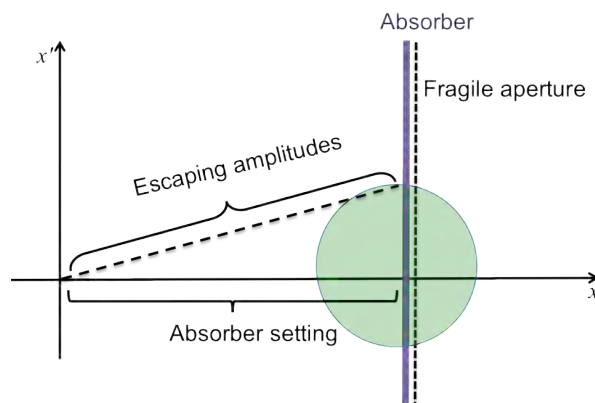
If absorbers are to protect against failures of one particular dipole error source, they are placed at 90° phase advance downstream of the error location. This is the case of the TDI absorber, which protects against LHC injection kicker failures. It has to withstand the high-intensity LHC beams and is therefore made of low-intensity materials, to be sufficiently robust. It consists of a sandwich of materials, starting with the low-density BN5000 ( $\rho = 1.92 \text{ g/cm}^3$ ) over a length of 2.85 m, then 0.6 m of Al, and finally 0.7 m of Cu–Be alloy. A large quantity of secondary showers still escapes the TDI if there is an impact. The TCDD mask protects the next magnetic element, the superconducting D1 separation dipole, from the shower particles. Figure 24 shows an overview of the injection region at LHC point 2 with the injection equipment, the TDI protection device, and the TCDD mask.

The setting of a protection device, i.e., the distance from the nominal orbit or trajectory and thus the amplitude cut in phase-space, has to be chosen so as to include a margin for orbit variations, beta beat, mechanical tolerances, set-up tolerances, etc. The extent of the particle distribution in phase-space also has an impact on the required setting. This is illustrated in Fig. 25. If the phase-space is cut at a single location, a fraction of the surviving particles will still have oscillation amplitudes larger than the absorber setting. This effect can be reduced by adding another absorber a few tens of degrees further downstream. In the LHC injection region, the TDI absorber is complemented by two shorter graphite auxiliary collimators of 1 m length at  $\pm 20^\circ$  phase advance from the TDI, with two jaws each.

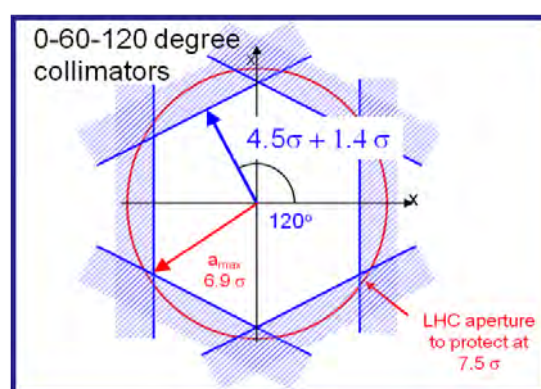
The LHC transfer lines are equipped with a generic passive protection system to protect against any failure during SPS extraction or transfer. It provides full phase-space coverage in the horizontal and vertical planes. In a single pass, this can only be achieved by installing collimators at several phase-space locations. In the case of the LHC transfer lines, three collimators per plane with 60° phase advance between two neighbouring collimators are chosen. They are located at the ends of the lines, to cover as many failures as possible. The maximum amplitude escaping such a system for 60° between the collimators can be calculated from the collimator setting  $n[\sigma]$ :

$$a_{\max} [\sigma] = \frac{n [\sigma]}{\cos(60^\circ/2)}. \quad (8)$$

An illustration of the phase-space cut for the LHC transfer line collimation system with a setting of  $4.5\sigma$  and  $1.4\sigma$  set-up accuracy is shown in Fig. 26.



**Fig. 25:** A fraction of the surviving beam will still have larger oscillation amplitudes than the absorber setting. This has to be taken into account when choosing an appropriate absorber protection setting.

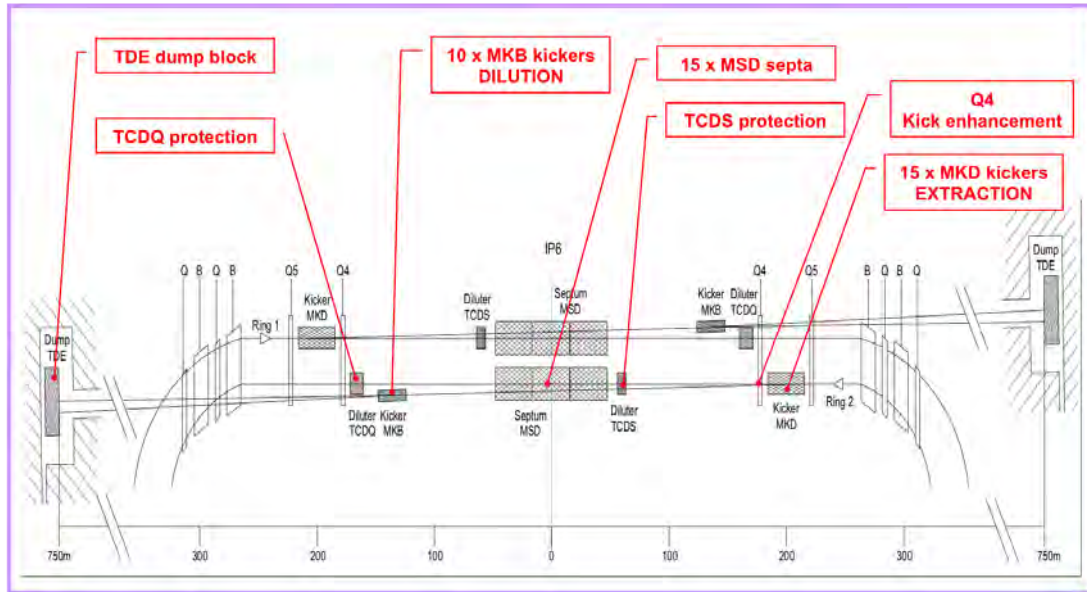


**Fig. 26:** The phase-space coverage of the LHC transfer line collimation with three collimators per plane and  $60^\circ$  phase advance between two collimators. The nominal setting is  $4.5\sigma$  and the set-up accuracy from collimator and trajectory in a single pass is  $1.4\sigma$ . The maximum escaping amplitudes for primary beam are, therefore,  $6.9\sigma$ .

### 5.3 Beam dumping

The beam dump system is one of the main components of the machine protection system. It is connected to the machine's beam permit and is triggered when the permit disappears. Fast-rising kicker magnets (and sometimes septa) are used, as for a fast extraction system, to steer the beam onto a beam dump block. The beam dump block can be installed in the circular machine, if it is an internal beam dump, or at the end of a transfer line, if it is an external beam dump. The system has to work for all energies and beam types of the accelerator. As kicker systems are involved, the beam dump system itself can cause very fast and hence very severe failures. The design has to take the failure scenarios in account and foresee sufficient redundancy and surveillance. In the LHC, the allowable failure rate is  $10^{-7}/h$  to  $10^{-8}/h$ , which corresponds to safety integrity level 3.

In contrast with a pure fast extraction system, where the pulse forming networks only have to be charged a few milliseconds before the extraction request, the beam dump pulse forming networks are always charged. The charging voltage follows the energy of the accelerator. A system to derive the energy from the synchrotron is required. In the LHC and SPS, this system is called the beam energy tracking system (BETS). The energy information is derived from the main dipole currents [9]. The LHC BETS surveys the correct energy tracking of the pulse forming networks and other involved equipment, such as the dump septa currents. If a tracking error occurs, the beam dump is triggered immediately before the error becomes too large.



**Fig. 27:** The two beam dump systems installed in long straight section 6 of the LHC for beams 1 and 2

The LHC beam dump system is installed in long straight section 6. The distance between the kicker magnets and the dump block is 975 m. An overview of long straight section 6, the dump equipment and dump lines is shown in Fig. 27.

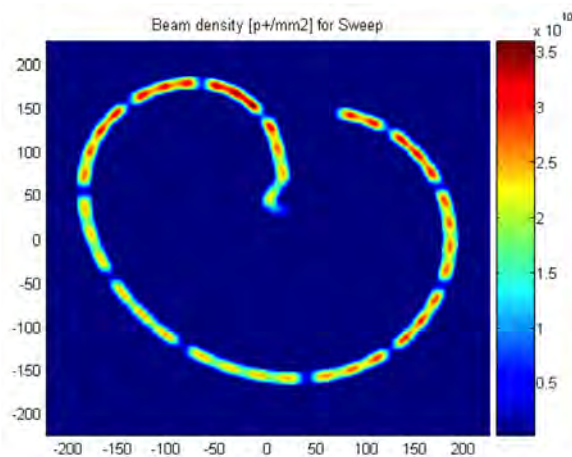
The beam dump block itself is made of 0.7 m and 3.5 m lengths of graphite with a density of  $1.73 \text{ g/cm}^3$ , interspersed with 3.5 m lengths of lower-density graphite ( $\rho = 1.1 \text{ g/cm}^3$ ). This is followed by 1 m of Al and 2 m of Fe at the end. The core is in a steel cylinder and in inert  $\text{N}_2$  gas. It is surrounded by about 900 t of radiation shielding blocks. The beam dump is designed such that no structural damage is to be expected during 20 y of operation with ultimate LHC intensities ( $1.7 \times 10^{11}$  protons per bunch in  $3.5 \text{ } \mu\text{m}$  emittance and 2808 bunches). Nevertheless, the dump block is only sufficiently robust for 7 TeV protons with beam dilution.

A set of vertical and horizontal dilution kickers is installed in the extraction channel, to sweep the beam onto the beam dump block. Not all bunches, therefore, will end up on the same spot. The resulting LHC beam dump sweep shape is shown in Fig. 28, and the waveforms of the horizontal and vertical dilution kickers are shown in Fig. 29.

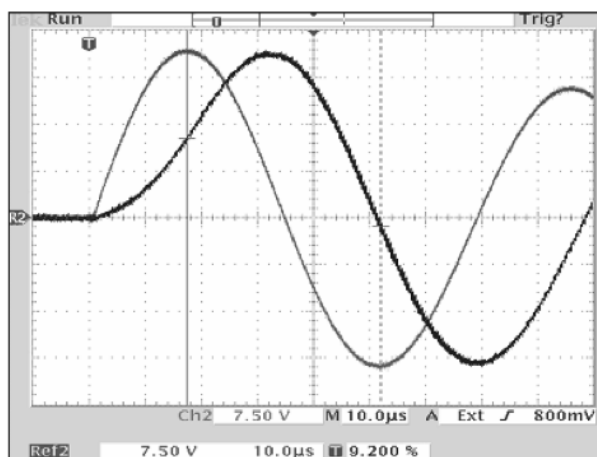
### 5.3.1 Failure scenarios of the LHC beam dump system

The LHC dump kicker magnets have a rise time of  $3 \text{ } \mu\text{s}$ . No particles are allowed in the so-called abort gap, whose length corresponds to the kicker rise time, Fig. 30. Even if the gap is kept particle-free, the kickers can still trigger spontaneously or asynchronously. Passive protection is therefore installed at the front face of the septum magnets (TCDS) and  $90^\circ$  downstream of the MKD dump kickers to protect the LHC circulating aperture (TCDQ). The TCDQ is a 9 m long single-sided movable absorber made of graphite. A fixed mask in front of the superconducting Q4 quadrupole protects the Q4 from damage from secondary showers generated in the TCDQ in the case of an impact.

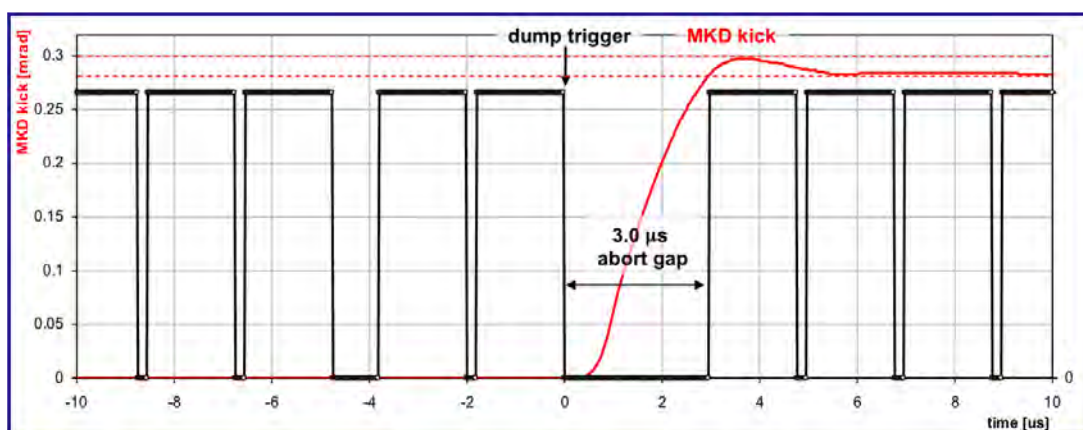
The different failure scenarios of the LHC beam dump system have been analyzed and divided into three categories.



**Fig. 28:** Simulated swept beam density on the front face of the LHC beam dump block. The sweep is the result of the excitation of horizontal and vertical dilution kickers with waveforms as illustrated in Fig. 29.



**Fig. 29:** LHC horizontal and vertical dilution kicker waveforms



**Fig. 30:** No particles are allowed in the abort gap of the dump kickers. The black line indicates the particle bunch trains. The smallest gaps between the trains correspond to the injection kicker rise time in the SPS; the 1  $\mu$ s gaps come from the LHC injection kicker rise time requirements. There is one large gap of 3  $\mu$ s for the dump kicker rise time. The beam must not be injected into the gap. Mechanisms are arranged to prevent this and any uncaptured beam is cleaned out of the gap after transverse feedback. The red line indicates the dump kicker (MKD) waveform.

### 5.3.1.1 LHC beam dump failure scenario: beyond design

This category of failures should not happen in the LHC lifetime, as they would result in severe damage to the machine. Several layers of redundancy are built into the system to cover against them:

- not receiving the trigger from the beam interlock system after a failure in the LHC;
- failure of beam energy tracking and extracting the entire beam at the wrong angle;
- fewer than 14 of the 15 dump kicker magnets firing during a dump;
- fewer than 3 of the 10 dilution kickers firing during a dump.

### 5.3.1.2 LHC beam dump scenario: active surveillance

The beam dump system and clean dump action depend on a number of conditions that must be monitored. If these conditions degrade, a dump is issued while the conditions are still sufficient for a clean dump. Examples are:

- failures of general services (electricity, vacuum, cooling, ethernet, ...);
- bad beam position in dump region;
- magnet powering failure of extraction equipment.

### 5.3.1.3 LHC beam dump scenario: passive protection

This category of failures cannot be prevented by surveillance systems and involve failures of the kicker systems. Redundancy has been built into the number of required dump kickers to ensure correct extraction. These failures can be tolerated and will not cause damage:

- one missing extraction kicker magnet;
- missing dilution kicker magnets;
- erratic behaviour of a dilution kicker magnet.

Finally there is a set of possible kicker failures where passive protection is required. They cannot be tolerated without correct settings of the passive protection element TCDQ:

- erratic behaviour of an extraction kicker magnet (spurious asynchronous trigger).

## 6 Final remarks

A number of additional concepts have been invented around and for the machine protection of LHC beam transfer systems, which have not been introduced in this lecture. Some of these will be covered in the lecture by J. Wenninger on *Machine Protection and Operation for the LHC*. A very important issue concerns the automatic *post-operational checks* and post-mortem system, which verify the correct functioning of the relevant systems after each beam transfer by launching a quick analysis of a number of recorded datasets from equipment and beam instrumentation [10, 11]. Degradation of the systems and increased likelihood of a failure are detected by these checks. This allows experts to intervene before an actual failure occurs.

## References

- [1] M. Bei, Workshop on Beam-induced Quenches, CERN, 15–16 September 2014, <http://indico.cern.ch/event/323249/>.
- [2] CERN Accelerator School, General Accelerator Physics, Introductory Level, <http://cas.web.cern.ch/cas/Granada-2012/Granada-after.html>.

- [3] M. Barnes *et al.*, Injection and extraction magnets: septa, in Proceedings of the CAS-CERN Accelerator School: Magnets, Bruges, Belgium, 16-25 June 2009, edited by D. Brandt, CERN-2010-004 (CERN, Geneva, 2010), pp. 167-184, <http://dx.doi.org/10.5170/CERN-2010-004.167>
- [4] M. Barnes *et al.*, Injection and extraction magnets: kicker magnets, CERN Accelerator School: Magnets, Bruges, Belgium, 16-25 June 2009, edited by D. Brandt, CERN-2010-004 (CERN, Geneva, 2010), pp. 141-166, <http://dx.doi.org/10.5170/CERN-2010-004.141>
- [5] B. Puccio *et al.*, The CERN beam interlock system: principle and operational experience, Proc. IPAC, 2010, <http://accelconf.web.cern.ch/AccelConf/IPAC10/index.htm>
- [6] V. Kain *et al.*, Functional specification: safe machine parameters, LHC-CI-ES-0004 rev 1.0 (2009), <https://ab-div-bdi-bl-blm.web.cern.ch/ab-div-bdi-bl-blm/Specification/LHC-CI-ES-0004-10-00.pdf>
- [7] V. Kain *et al.*, Material damage test with 450 GeV LHC-type beam, Proc. PAC, 2005, <http://accelconf.web.cern.ch/AccelConf/p05/INDEX.HTML>
- [8] M. Werner *et al.*, A fast magnet current change monitor for machine protection in HERA and the LHC, Proc. ICALEPS, 2005, <http://accelconf.web.cern.ch/AccelConf/ica05/>
- [9] E. Carlier *et al.*, The beam energy tracking system of the LHC beam dumping system, Proc. ICALEPS, 2005, <http://accelconf.web.cern.ch/AccelConf/ica05/>
- [10] N. Magnin *et al.*, External post-operational checks for the LHC beam dumping system, Proc. ICALEPS, 2011, <http://accelconf.web.cern.ch/AccelConf/icalepcs2011/index.htm>
- [11] L. Drosdal *et al.*, Automatic injection quality checks for the LHC, Proc. ICALEPS, 2011, <http://accelconf.web.cern.ch/AccelConf/icalepcs2011/index.htm>

## Beam-Induced Damage Mechanisms and their Calculation

A. Bertarelli

CERN, Geneva, Switzerland

### Abstract

The rapid interaction of highly energetic particle beams with matter induces dynamic responses in the impacted component. If the beam pulse is sufficiently intense, extreme conditions can be reached, such as very high pressures, changes of material density, phase transitions, intense stress waves, material fragmentation and explosions. Even at lower intensities and longer time-scales, significant effects may be induced, such as vibrations, large oscillations, and permanent deformation of the impacted components. These lectures provide an introduction to the mechanisms that govern the thermomechanical phenomena induced by the interaction between particle beams and solids and to the analytical and numerical methods that are available for assessing the response of impacted components. An overview of the design principles of such devices is also provided, along with descriptions of material selection guidelines and the experimental tests that are required to validate materials and components exposed to interactions with energetic particle beams.

### Keywords

Beam-induced damage; thermal shocks; beam intercepting devices; collimators; novel materials; particle beam experiments.

## 1 Introduction to beam-induced damage

When subatomic particles or ions interact with matter, they tend to transfer part of their energy to the medium they traverse [1]; this energy loss is ultimately turned into heat and leads to an increase of the temperature in the impacted target. Depending on the amount and distribution of the deposited energy and the time-scale of the phenomenon, i.e. the density of deposited power, different effects may result.

If the density of the deposited energy is relatively small, of the order of  $10 \text{ W cm}^{-3}$  or less, and extended over a relatively long period of time (of the order of seconds), the structural response can be reduced to a (quasi-) steady-state or slow transient thermomechanical problem. This class of problems can often be linearized and solved using ordinary simulation methods, either analytical or, more usually, numerical, such as standard *finite element method* tools, which allow one the change of material properties with temperature to be taken into account. Examples of these problems include, for instance, simulation of so-called slow losses of particle accelerators in the collimation regions [2, 3].

Conversely, if the deposited power density is much higher and the duration of the interaction is very short (of the order of a few milliseconds or less), dynamic responses will be induced, principally because the thermal expansion of the impacted material is partly prevented by its inertia [4, 5]. These effects, often referred to as thermal shocks, generate dynamic stresses, which propagate through the material at the velocity of sound, in much the same way as when a structure is struck by another body. Depending on the amount of the deposited energy and the melting point of the impacted material, the temperature increase induced by the impact may lead to the formation of shock waves, changes of phase, or the ejection of molten material.

The nature and intensity of the dynamic responses depend on several parameters, mainly the total amount of energy deposited, its distribution, the duration of the impact, the thermophysical and mechanical properties of the impacted material, and the form and dimensions of the device interacting with the beam.

Figure 1 provides an overview of the different types of dynamic response that might be induced in a structure as a function of the density of the deposited power and the duration of the interaction. One can easily observe that the severity of the response is broadly proportional to the deposited power density and inversely proportional to the duration of the interaction; in other words, dynamic response depends on the specific energy deposited on the material.

In spite of the large influence of material properties, some approximate general figures can be extrapolated and used with caution to predict the type of response, regardless of the actual impacted material and boundary conditions. If the deposited energy is below  $100 \text{ J cm}^{-3}$ , the dynamic response will probably remain within the *elastic dynamic regime*, meaning that the induced vibrations and stress waves will not exceed the elastic limit of the material and that the structure will return to its initial undeformed state at the end of the process. Between roughly  $100 \text{ J cm}^{-3}$  and  $10 \text{ kJ cm}^{-3}$ , we may expect that the stress waves will locally exceed the elastic limit of the material, inducing permanent plastic deformations that cannot be recovered once the dynamic response has waned out (the *plastic dynamic regime*). Above  $10 \text{ kJ cm}^{-3}$ , the stress waves will be strong enough to generate major changes of density and extensive damage to the material, such as fragmentation or explosion. If the impacted material is metal, phase transitions with the formation of liquid, gas, or plasma usually occur: this is usually referred to as the *shock wave regime*. If a significant reduction in density has occurred in the impacted material while particle bunches are still hitting it, the beam will penetrate more and more deeply, given that fewer atoms are available to interact with the incoming particles: this effect is sometimes called *hydrodynamic tunnelling* and may arise within the shock wave regime when the duration of the impact is sufficiently long to allow changes of phase to develop (several hundred nanoseconds or more).

The shock wave regime is also sometimes referred to as the *hydrodynamic regime*, implying that impacted materials start behaving as fluids, losing their mechanical strength (see Section 2.4.3). Since extensive shock-induced damage, such as fragmentation or mechanical spalling, may occur long before the material completely loses its strength, the term ‘shock wave regime’ appears more comprehensive.

These phenomena, with the possible exclusion of those belonging to the elastic regime, may severely affect the integrity and the functionality of the impacted equipment. A correct understanding and prediction of beam-induced damage is therefore extremely important in the design of any component exposed to direct interaction with intense and energetic particle beams, such as collimators, absorbers, dumps, scrapers, or windows. The same damage mechanisms apply to any device accidentally and rapidly interacting with energetic beams, such as vacuum chambers, magnet components, RF cavities, or beam instrumentation devices.

These lectures will address these topics, particularly focusing on dynamic events that have the potential to generate structural damage. Other energy release mechanisms (e.g. of stored magnetic energy or RF impedance-induced heating) are not explicitly covered here, although their effects might be dealt with in a similar way if the time-scale of the phenomena are relatively short.

Longer-term phenomena, such as radiation damage, which do also play a fundamental role in the design of devices interacting with particle beams, are treated elsewhere (see, for instance, Ref. [6]).

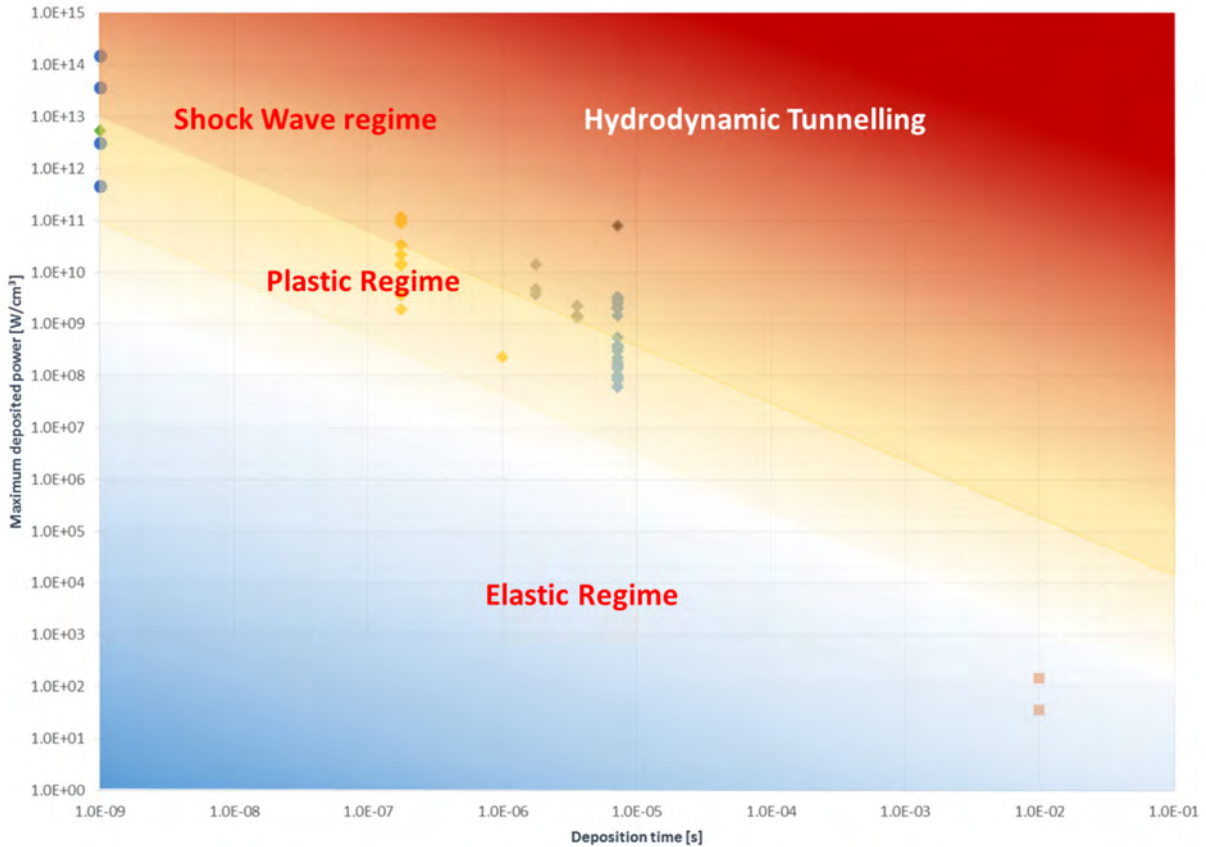
We will first provide a few cases of notable accidents and tests that have occurred in various accelerators in the world over the past few decades, to exemplify the effects of beam impacts on matter.

To explore the mechanisms of beam interaction with matter leading to the various responses and damage states previously described, we will start by introducing the concepts of linear thermo-elasticity, in particular, the case of beam impacts on circular discs and cylinders inducing responses in the elastic



domain of the material: these relatively simple cases, although not leading to permanent damage, can be treated analytically and are very useful to gain some insights and physical understanding of the mechanisms of thermally induced stresses and dynamic responses.

For the sake of simplicity, the analysis will mostly deal with isotropic, homogeneous materials. However, these principles and methods can be extended to anisotropic or non-homogeneous materials with some additional mathematical complexity.



**Fig. 1:** Plot of maximum deposited power versus duration of deposition, showing the different dynamic responses that can be induced in matter by interaction with particle beams. Points represent cases of beam impacts (real or simulated).

We will then introduce non-linear dynamic responses, which require, to be correctly treated, the use of numerical tools ranging from standard finite element methods to sophisticated highly non-linear wave propagation codes, making use of complex material constitutive models.

Next, the principles that should guide the design of equipment subjected to beam-induced accidents will be briefly set out, introducing relevant figures of merit for such cases, and finally we will describe the experimental tests that are essential to validate the numerical simulations and qualify the design of such components.

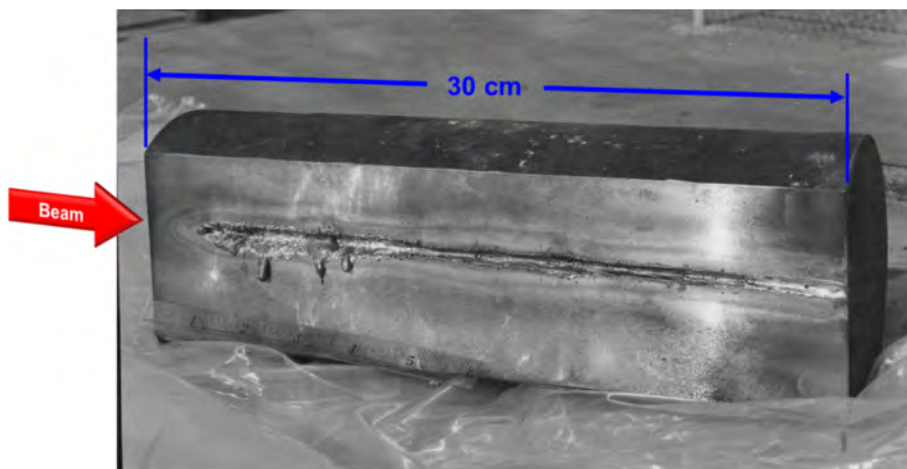
**1.1 Examples of beam-induced accidents**

Figure 2 shows thin rods that were part of the first neutrino target station installed in the Super Proton Synchrotron (SPS) at CERN [7]; they were impacted by a beam of  $1 \times 10^{13}$  protons at 300 GeV/c, with a cross-section of roughly 2 mm, with a certain offset with respect to the centre. These rods were made of beryllium and are 100 mm long, 3 mm in diameter; the pulse duration was roughly 23  $\mu$ s. The accident took place in the early 1970s.



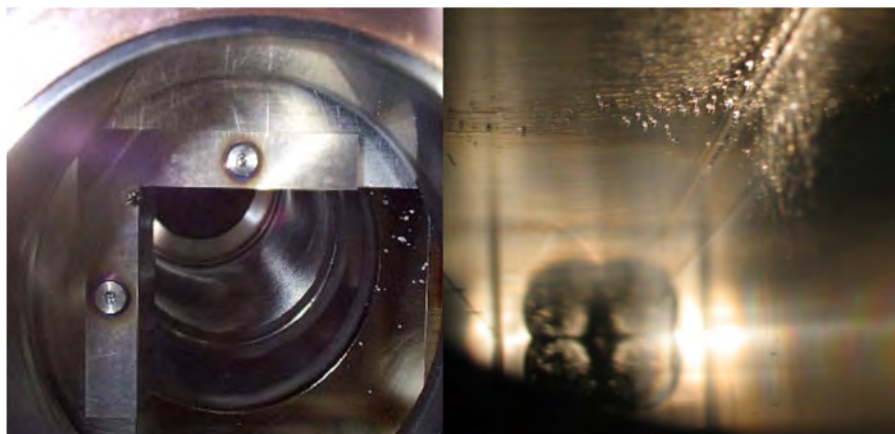
**Fig. 2:** Beryllium rods for first neutrino target installed at CERN-SPS. The unit on top was accidentally hit by an off-axis beam, leading to permanent bending and failure of the rod.

Figure 3 depicts a 30 cm long copper block onto which a beam damage test was performed in 1971 at SLAC [8]; the block was meant to simulate a collimator. An 18 GeV/c  $e^-$  beam of  $\approx 2$  mm diameter, with a power of roughly 500 kW, impinged the edge of the block. The impact lasted roughly 1.3 s and led to extensive melting in the impacted area. Although this was a relatively slow accident, the type of damage it generated can be compared to those created by faster events.



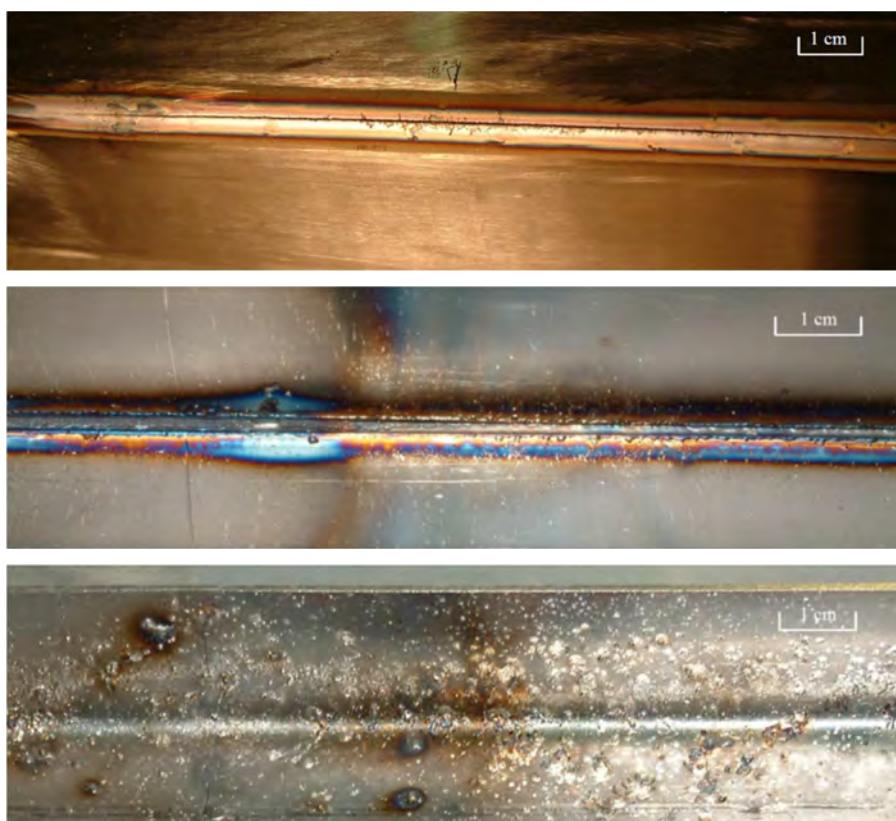
**Fig. 2:** Copper block used for damage test at SLAC in 1971

Figure 4 illustrates the effects of an accident that took place at Tevatron (FNAL). In 2003, a Roman pot accidentally moved towards the Tevatron beam, generating an intense cascade of secondary particles, which led to fast quenches in several superconducting magnets [9]. This in turn caused a drift of the 980 GeV/c proton beam, which eventually hit a primary collimator, made of tungsten alloy, and a secondary collimator, made of stainless steel, resulting in a 2.5–3 mm hole in the 5 mm thick tungsten unit and an extended groove several centimetres long on the stainless steel collimator. The peak density of deposited energy on tungsten alloy is in the range of  $1 \text{ kJ g}^{-1}$  ( $\approx 20 \text{ kJ cm}^{-3}$ ).



**Fig. 4:** Tevatron collimators accidentally hit by proton beam: left, tungsten alloy primary collimator; right, stainless steel secondary collimator.

During high-intensity extraction from SPS (CERN) in 2004, an incident occurred in which a stainless steel vacuum chamber of a magnet in TT40 transfer line was badly damaged. The beam was a 450 GeV full Large Hadron Collider (LHC) injection batch of  $3.4 \times 10^{13}$   $p^+$  in 288 bunches extracted from SPS with a wrong trajectory. The beam drift was induced by the switch-off of a septum magnet [10]. This provoked a 110 cm long groove and a cut of 25 cm on the side of the impact with projection of molten steel on the opposite side (Fig. 5). Both the vacuum chamber and the magnet had to be replaced.



**Fig. 5:** Damages on a vacuum chamber of TT40 magnet (SPS-LHC transfer line). Top: Outside of the vacuum chamber. Centre: Inside of the vacuum chamber, beam impact side. Bottom: Inside of the vacuum chamber, side opposite to the beam impact

The accident presented in Fig. 5 occurred while the beam was being prepared for a series of tests, including one which aimed at determining the damage threshold of several materials under the impact of a 450 GeV proton beam from the SPS [11]. The target consisted of a series of tightly packed plates made of metals commonly used in accelerators, such as copper, stainless steel, and Inconel™ Ni–Cr superalloy. Pulses at four intensities, ranging from  $1.32 \times 10^{12}$  p<sup>+</sup> to  $7.92 \times 10^{12}$  p<sup>+</sup> with an average root-mean-square beam size of 0.85 mm ( $\sigma_x$ , 1.1 mm,  $\sigma_y$ , 0.6 mm). Effects of the impacts on copper are presented in Fig. 6: letters correspond to different intensities. At  $1.32 \times 10^{12}$  (A), no signs of damage are visible, at  $2.64 \times 10^{12}$  (B), coloration starts to appear, at  $5.28 \times 10^{12}$  (C) and  $7.92 \times 10^{12}$  (D) melting becomes clearly visible. These results are essentially in accordance with the value of the calculated peak energy deposition (Fig. 7): for a beam size of 1 mm at 450 GeV, melting in copper is expected to begin at an intensity of  $\approx 2.5 \times 10^{12}$  protons (see Section 2.1 for calculation method).

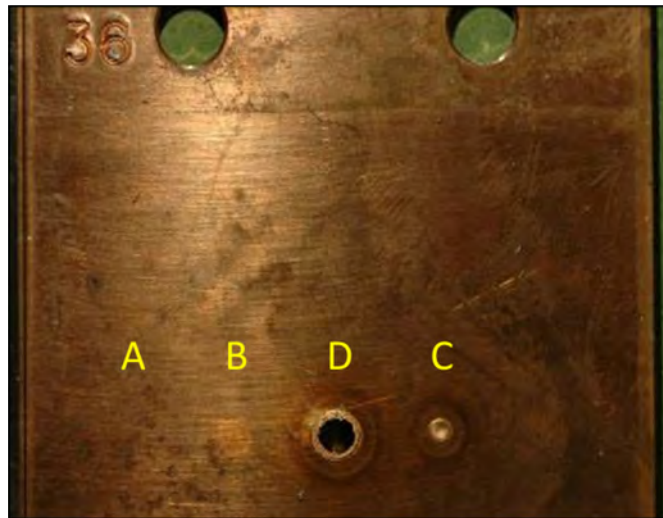


Fig. 6: Effect of beam impacts at different intensities on copper target

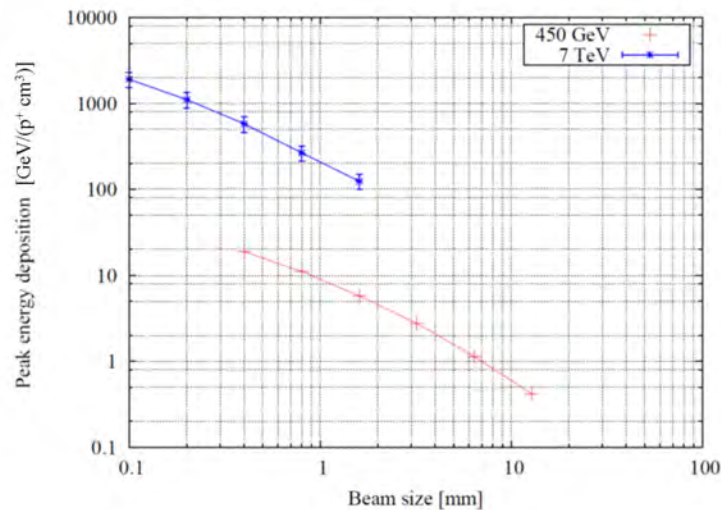
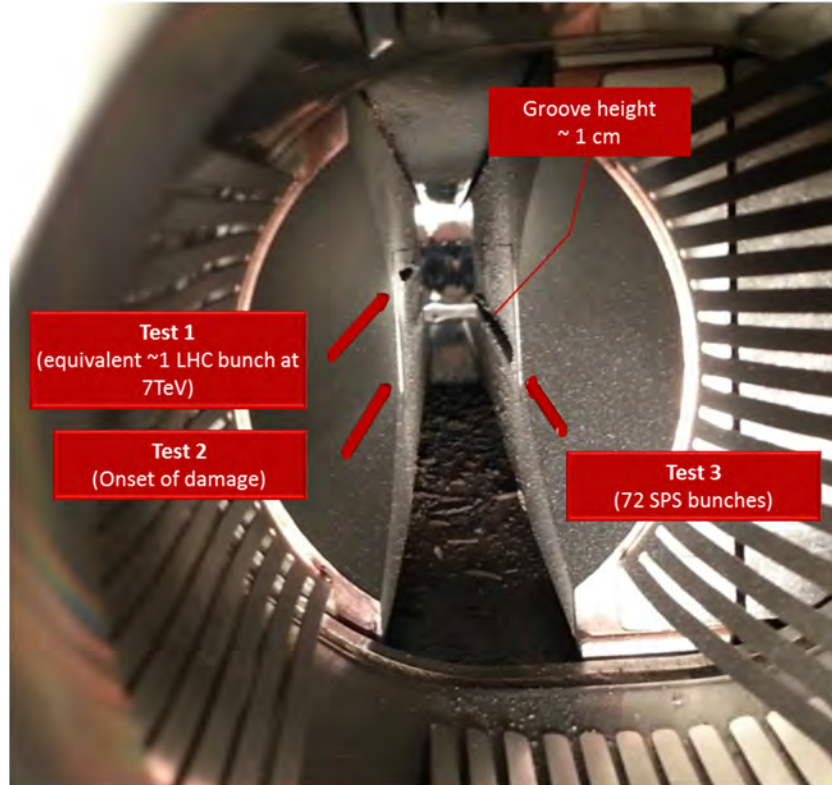


Fig. 7: Peak energy deposition in a copper target at 0.45 and 7 TeV as a function of beam size [11]

In 2012, an experiment was carried out at CERN, in the recently commissioned HiRadMat facility [12] to test the behaviour of a LHC tertiary collimator in case of direct beam impact [13]. The beam was extracted from SPS at an energy of 450 GeV/c.

Figure 8 shows a collimator jaw, whose active part is made of five blocks of tungsten heavy alloy (Inermet® IT180 from Plansee, Austria), after three distinct tests performed at various beam intensities.

In test 1, 24 SPS bunches with a total intensity of  $3.36 \times 10^{12} p^+$  hit the Inermet blocks: a groove several centimetres long and a few millimetres high was produced. This impact was intended to produce an energy distribution similar to that induced by one full LHC bunch ( $1.15 \times 10^{11} p^+$ ) at 7 TeV. In test 2, the impacting beam intensity was  $1.04 \times 10^{12} p^+$ : according to simulations this was the threshold value at which plastic deformations were first induced, without material fragmentation. In test 3, a train of 72 SPS bunches impacted the jaw with a total intensity of  $9.34 \times 10^{12} p^+$ . It can easily be inferred that a small fraction of one LHC bunch is sufficient to generate extensive damage, with changes of phase, material ejection, and fragmentation, on high-Z materials, such as tungsten alloys.



**Fig. 8:** Effects of three impact tests on a LHC tertiary collimator jaw at various beam intensities

## 2 Analysis of beam interaction with matter

As the preceding examples prove, damage phenomena induced by high-energy, high-intensity particle beams bring matter to extreme states, where practical experience and material knowledge is very limited. Hence, the accurate prediction of the structural response to such events becomes very complex. The analysis of these phenomena must rely on methodologies that integrate and couple several fields of science and engineering, numerical tools and experimental verification in a multidisciplinary approach.

From an engineering perspective, these problems can be attacked by dividing the procedure into three successive steps.

- The physical problem. The main goal of this step is to determine how much energy, and where, was deposited onto the relevant body.
- The thermal problem. The objective is to determine which temperature distribution, at which moment in time, was induced in the body by the deposited energy computed in the first step.
- The thermomechanical problem (which may be linear or non-linear). Given the temperature field, the goal is to determine which strains, stresses, deformations, dynamic responses, and phase transitions were generated in the body.

## 2.1 The physical problem

Particles interact with matter through various mechanisms, which typically depend on particle species and energies, and on the density, atomic number, and atomic mass of the impacted material.

Since, as mentioned already, we are not interested in long-term damage mechanism induced by particle irradiation and the changes they induce in material properties, what only matters, for the problems we are dealing with, is that the part of beam energy that is lost in the target during particle-matter interaction is ultimately transformed into heat. These interactions occur during extremely short time-scales, of the order of  $10^{-11}$  s [14]: this is sufficiently short to consider the heat generation by each particle as an instantaneous process.

Monte Carlo interaction and transport codes, such as FLUKA, MARS or Geant4, are typically used to simulate these phenomena and predict the distribution of energy deposited per interacting particle [1].

The energy deposition process lasts as long as the particles interact with matter. This depends on the bunch length, the number of interacting bunches and their time spacing. The total deposited heat can be simply calculated by multiplying the single particle energy distribution by the total number of particles.

It is interesting to note that the linear scaling of the deposited energy with the number of particles holds, provided that the material density does not change during the interaction. Substantial changes of density do occur if the temperature increase leads to phase transitions (melting, vaporization, plasma generation, etc.) or if severe shock waves physically displace matter in the region of the impact. In this case, particular caution must be taken, since, if the interaction with the beam is still ongoing as density varies, the energy deposition distribution will be modified by the change of density, typically reducing energy peaks in the upstream part and extending the interaction along longer portions of the target (see Section 2.4.4 for details).

### 2.1.1 Temperature distribution

Once the energy deposition distribution is available, the quasi-instantaneous temperature distribution can be easily calculated by taking into account the material specific heat,

$$q_V(x_i) = \int_{T_i}^{T_f} \rho \cdot c_p(T) \cdot dT(x_i) , \quad (1)$$

where  $q_V$  is the *deposited energy per unit volume* ( $\text{J cm}^{-3}$ ) at the location  $x_i$  ( $i$  ranges from 1 to 3),  $\rho$  is the *density* of impacted material ( $\text{g cm}^{-3}$ ),  $c_p$  is its *specific heat capacity* at constant pressure ( $\text{J g}^{-1} \text{K}^{-1}$ ) and  $T(x_i)$  is the local *temperature* (K), with  $T_i$  and  $T_f$  being the temperatures at the beginning and the end of the energy deposition process under analysis.

Attention must be paid when using Eq. (1) to determine the temperature distribution  $T(x_i)$ ; in fact, this relationship implicitly assumes that no heat diffusion occurs while heat generation occurs, i.e. the *duration of the energy deposition* process  $\tau$  is much shorter than the *thermal diffusion time* (see next sections for details).

In general,  $c_p$  is a function of *temperature*,  $T$ . When this dependence is not too strong, to a first approximation, an average value,  $\bar{c}_p$ , can be taken to obtain the quasi-instantaneous temperature increase. In such a case, the temperature increase can be derived explicitly; if, for convenience, we set the initial temperature to zero, we obtain

$$T(x_i) \cong \frac{q_V(x_i)}{\rho \cdot \bar{c}_p} . \quad (1a)$$

## 2.2 The thermal problem

To assess the stress state in a body submitted to thermal shocks, it is fundamental to determine the initial temperature distribution and its evolution over time. The temperature evolution is governed by a diffusion process, the *heat equation* (also known as *Fourier's equation*):

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot (\lambda \nabla T) + \dot{q}_V, \quad (2)$$

where  $\dot{q}_V$  is the energy deposition rate or *heat generation rate* ( $\text{W m}^{-3}$ ), and  $\lambda$  is the *thermal conductivity* ( $\text{W m}^{-1} \text{K}^{-1}$ ).

If we assume that the material is *homogeneous* and *isotropic*, so that physical properties do not change from point to point and with material orientation, we get

$$\frac{\partial T}{\partial t} = a \nabla^2 T + \frac{\dot{q}_V}{\rho c_p}, \quad (2a)$$

where  $a = \frac{k}{\rho c_p}$  is the thermal diffusivity ( $\text{m}^2 \text{s}^{-1}$ ).

It is interesting to note that Eq. (2) fails to predict heat transfer phenomena for very short time-scales, given that it implies infinite speed of heat signal propagation. This is usually not relevant in the problems we are dealing with, but can play a role in ultra-short phenomena, such as high-frequency laser pulsed heating lasting of the order of femtoseconds.

If thermophysical properties are also constant with temperature, Eq. (2a) becomes a partial differential equation with constant coefficients, which, in some cases, can be solved analytically.

Once energy deposition is completed ( $t \geq \tau$ , that is, the time  $t$  is much greater than  $\tau$ , the duration of the energy deposition), Eq. (2a) becomes a homogeneous linear partial differential equation.

$$\frac{\partial T}{\partial t} = a \nabla^2 T. \quad (2b)$$

If we assume that the initial temperature distribution is known and that, at least for the short time-scales we are interested in, the system is adiabatic regardless of the actual boundary conditions (e.g. active cooling), analytical solutions to Eq. (2b) become available for simple geometries, usually involving Fourier series, Bessel series, Laplace transforms, etc. [15].

A useful case is that of a circular cylinder or disc with an axially symmetrical energy distribution that is constant along the axis: this may well approximate impacts at the centre of thin circular windows or cylindrical targets. In this case, the solution, in cylindrical coordinates, to Eq. (2b) with an initial temperature distribution obtained from Eq. (1a) and adiabatic boundary conditions is given by

$$T(r, t) = \sum_i C_i J_0 \left( \frac{\beta_i}{R} r \right) \cdot e^{-\frac{a}{R^2} \beta_i^2 t}, \quad (3)$$

where  $R$  is the outer radius of the disc or cylinder,  $J_0 \left( \frac{\beta_i}{R} r \right)$  is a Bessel function of the first kind of order zero with  $\frac{\beta_i}{R}$  being the eigenvalues of the problem obtained by imposing the adiabatic boundary condition and  $C_i$  are numerical coefficients derived from the initial temperature distribution [16].

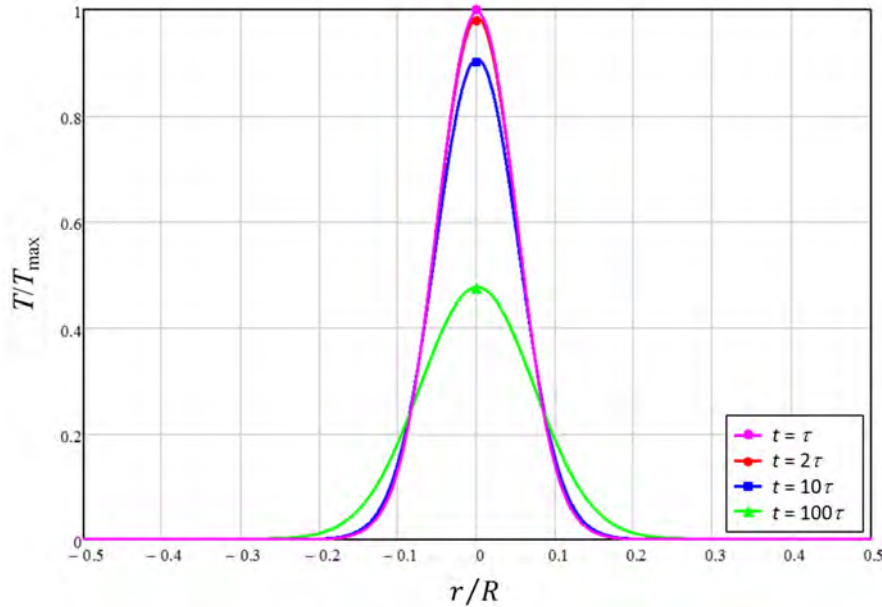
Analytical solutions can also be derived for more complicated cases, such as beams with rectangular cross-section or discs and cylinders with off-axis energy deposition [15].

Assuming that the energy deposition profiles can be approximated with an axially symmetrical normal Gaussian distribution, as is often the case, the initial temperature field in a disc or circular cylinder takes the form

$$T(r, \tau) = T_0(r) = T_{\max} e^{-\frac{r^2}{2\sigma_b^2}}, \quad (3a)$$

where  $\sigma_b$  is the standard deviation of the distribution and  $T_{\max}$  is the maximum initial temperature, obtained from Eq. (1a).

As an example, the temperature distribution obtained from solving Eq. (3) for a thin graphite disc with outer radius  $R = 5$  mm is shown in Fig. 9, assuming a Gaussian round beam with  $\sigma_b = 0.05R$ .



**Fig. 9:** Temperature distribution (normalized to maximum initial temperature  $T_{\max}$ ) in the central area of a graphite circular disc impacted at its centre by a Gaussian round beam with  $\sigma_b = 0.05R$  at different instants in time ( $\tau$  is the duration of the energy deposition).

It is very important to note that in practically all analytical solutions, regardless of their mathematical complexity, a characteristic time, called the *thermal diffusion time*,  $t_d$  can be identified:

$$t_d = \frac{B^2}{a}. \quad (3b)$$

This parameter is related to the time required to reach, by heat diffusion processes, a uniform temperature distribution in a region whose relevant dimension is  $B$  (e.g. the radius of a disc).

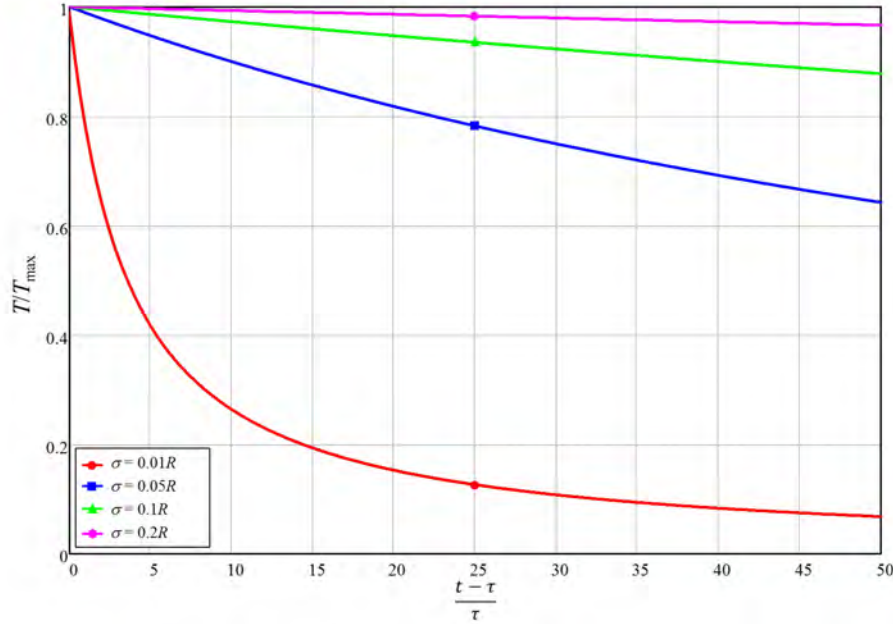
For the typical dimensions of interest (several millimetres or more), the thermal diffusion time lasts from several to many milliseconds, which is usually much longer than the duration of beam impacts we are concerned with (of the order of microseconds or less): this is why the assumption of instantaneous heat deposition is generally acceptable. Heat diffusion times for materials of interest for accelerator components exposed to interaction with the beam are given in Table 1.

Even if at the impacted component global scale the impact can be considered instantaneous when the diffusion time is much longer than the impact duration, it is important to note that very sharp and narrow temperature peaks may start to flatten even during the energy deposition process: this is because at the sub-millimetric scale the diffusion time becomes much shorter and comparable to the duration of the impact.



Observing Fig. 9, it can be seen that the temperature at the centre of the beam spot tends to decrease relatively slowly: at  $t \cong 2\tau$ , i.e. after a time equal to the duration of the energy deposition, the maximum temperature has decreased by roughly 1%. In such a case, the assumption of instantaneous energy deposition seems acceptable.

However, for smaller beam sizes, temperature, at the centre of the disc, drops at a much faster rate: as shown in Fig. 10, for  $\sigma_b = 0.01R$  (with  $R = 5$  mm) the temperature at the centre has already fallen by more than 20% when  $t = 2\tau$ . In such a case (*a fortiori* with smaller beam sizes), neglecting the heat diffusion processes occurring during the impact is not appropriate and the initial temperature obtained through Eq. (1a) would be largely overestimated.



**Fig. 10:** Temperature at the centre of the disc (normalized to  $T_{max}$ ) as a function of time (normalized to impact duration  $\tau$ ) for different beam sizes ( $R = 5$  mm).

**Table 1:** Thermal diffusion times on several length scales for materials of interest

| Material                      | Properties at room temperature |  |  |  | Thermal diffusion time [ms] |            |            |
|-------------------------------|--------------------------------|--|--|--|-----------------------------|------------|------------|
|                               | Density [kg m <sup>-3</sup> ]  | Specific heat [J kg <sup>-1</sup> ·K <sup>-1</sup> ] | Thermal conductivity [W m <sup>-1</sup> ·K <sup>-1</sup> ] | Thermal diffusivity [mm <sup>2</sup> s <sup>-1</sup> ] | $B = 0.1$ mm                | $B = 1$ mm | $B = 1$ cm |
| Copper (Glidcop)              | 8 900                          | 391  | 365  | 104.9  | 0.10                        | 9.5        | 953        |
| Tungsten alloy (Inermet180)   | 18 000                         | 150  | 90.5   | 33.5   | 0.30                        | 29.8       | 2983       |
| Molybdenum                    | 10 220                         | 251  | 138  | 53.8   | 0.19                        | 18.6       | 1859       |
| Titanium alloy (Ti6Al4V)      | 4 420                          | 560  | 7.2  | 2.9  | 3.44                        | 343.8      | 34378      |
| Aluminium alloys              | 2 700                          | 896  | 170  | 70.3   | 0.14                        | 14.2       | 1423       |
| Molybdenum-graphite (MG-3110) | 2 500                          | 740  | 770  | 416.2  | 0.02                        | 2.40       | 240        |
| Graphite                      | 1 850                          | 780  | 70   | 48.5   | 0.21                        | 20.6       | 2061       |
| Beryllium                     | 1 844                          | 1925   | 216  | 60.9   | 0.16                        | 16.4       | 1643       |

## 2.3 The linear thermomechanical problem

### 2.3.1 Stresses and strains

Any body submitted to a *mechanical stress*, defined as the limit of the ratio between a force (vector) and the surface it is acting upon, responds by deforming. The ratio of stress-induced deformation to the initial dimension is called *mechanical strain*. For a slender body loaded along its axis, the mechanical (*normal* or *axial*) strain is defined as the change in length,  $\delta$  per unit of the original length,  $L$ , of the body:  $\varepsilon_M = \delta/L$ . The normal strain is positive if the material ‘fibres’ are stretched and negative if they are compressed.

More generally, on a given plane of an infinitesimal volume of a body, the stress vector can be decomposed into a component perpendicular to the plane and two orthogonal in-plane components (Fig. 11). The component normal to the surface is called the *normal stress* and the components that act in-plane are the *shear stresses*. These three components, in combination with the three main planes ( $x$ ,  $y$  and  $z$  or 1, 2, 3), form the nine components of the *stress tensor*. Strain components of the *strain tensor* are defined in a similar way.

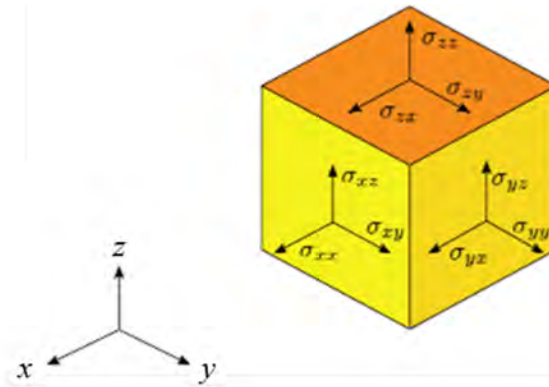


Fig. 11: Stress components acting on an infinitesimal volume

### 2.3.2 Linear elasticity

In *linear elasticity* it is postulated that a linear relationship exists between stresses and strains. Mathematically, this is expressed by *Hooke's law*, which constitutes an idealization of the behaviour of most materials submitted to low or moderate stresses.

In indicial notation, for an isotropic body, this relationship takes the expression

$$\varepsilon_{ij}^M = \frac{1}{E} \left[ (1 + \nu) \sigma_{ij} - \nu \delta_{ij} \sigma_{kk} \right] , \quad (4)$$

where  $E$  is the *Young modulus* (Pa),  $\nu$  is the *Poisson ratio*, and  $\delta_{ij}$  is the Kronecker delta.

The Poisson ratio expresses the tendency of the material to expand in the two directions perpendicular to the direction of compression. Conversely, if the material is stretched, it usually contracts in the directions transverse to the direction of stretching;  $\nu$  is the negative ratio of transverse to axial strain.

If only one component of normal stress is acting, Eq. (4) reduces to the well-known linear stress–strain relationship:

$$\varepsilon_{11} = \frac{\sigma_{11}}{E} . \quad (4a)$$

2.3.3 Linear thermo-elasticity

It is well known that an unrestrained body submitted to a change of temperature undergoes a dimensional change called *thermal deformation*.

Strains caused by thermal deformation on unrestrained bodies heated from an initial reference temperature (usually uniform and equal to ambient temperature), are called *free thermal strains*,  $\epsilon^T$ .

The rate of linear change of dimension per unit temperature variation is called the *linear coefficient of thermal expansion (CTE)*,  $\alpha$ .

$$\alpha(T) = \frac{dL}{LdT} . \tag{5}$$

The CTE has units of  $K^{-1}$  and is, in general, a function of temperature (Fig. 12); at very low temperatures (usually below 80 K),  $\alpha$  tends to zero. However, over limited temperature ranges above or around room temperature (RT), it can be averaged to a constant value.

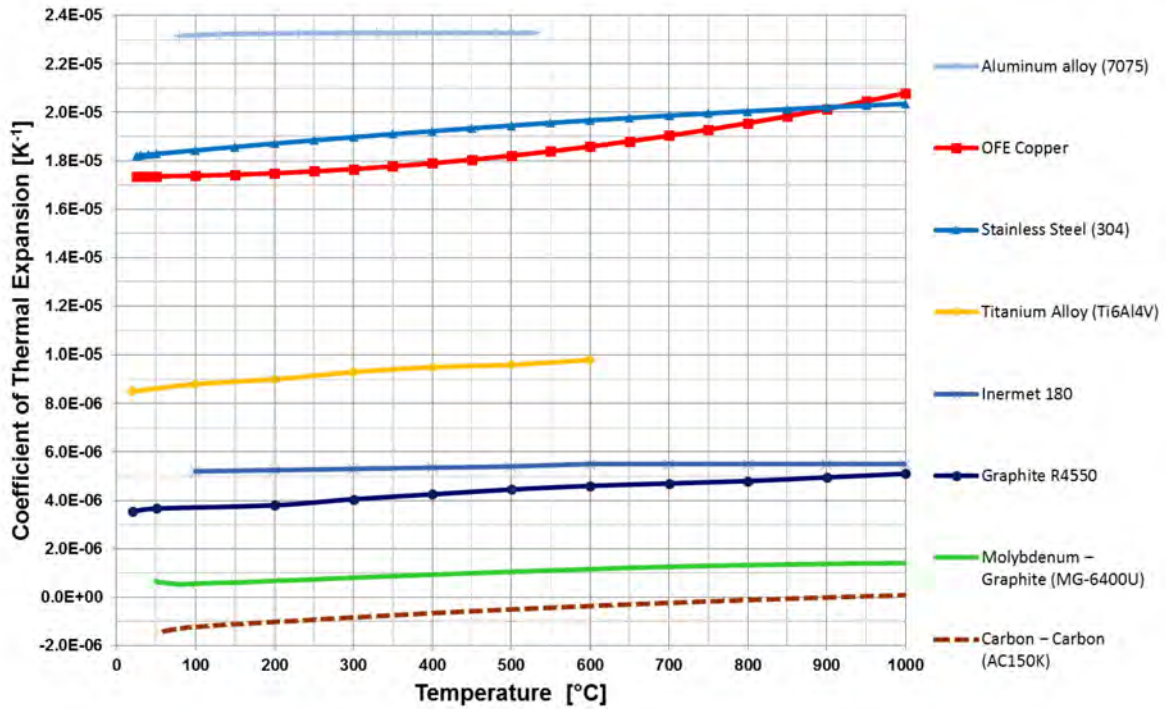


Fig. 12: Linear CTE for selected materials as a function of temperature

The linear CTE is related to the *volumetric coefficient of thermal expansion*,  $\beta$ , by the expression  $\alpha = 1/3 \beta$ .

In the nineteenth century, Hooke’s law was extended by Duhamel and Neumann to include first-order thermal effects (*linear thermo-elasticity*). They assumed that the *total strain*,  $\epsilon$ , at a point consists of two components: the *mechanical strain*,  $\epsilon^M$ , and *free thermal expansion*,  $\epsilon^T$ .

We then have

$$\epsilon_{ij} = \epsilon_{ij}^M + \epsilon_{ij}^T \tag{6a}$$

In indicial notation, the strain caused by free thermal expansion in an *isotropic* and *homogeneous* body is expressed as

$$\varepsilon_{ij}^T = \alpha \delta_{ij} T, \quad (6b)$$

with the initial reference temperature assumed uniform and taken identically equal to zero for convenience.

We note from Eq. (6b) that, for a homogeneous and isotropic body, only normal strain components are affected by a temperature change, that is, the deformation is only volumetric and, if the temperature change is uniform throughout, the shape of the body is maintained.

The *Duhamel–Neumann law* can then be expressed as

$$\varepsilon_{ij} = \frac{1}{E} [(1+\nu)\sigma_{ij} - \nu\delta_{ij}\sigma_{kk}] + \alpha\delta_{ij}T. \quad (7a)$$

It is important to note that, in an isotropic body, shear strains are never induced by free thermal expansion.

In general, stresses may be caused by mechanical loads, temperature gradients creating internal constraints to uniform expansion, or geometric restraints preventing free thermal expansion (*hyperstatic design*).

When such stresses remain well below the *yield strength*<sup>1</sup> of the material, defined as the conventional stress at which the material begins to deform plastically, the material is said to be in an *elastic regime*.

It can be observed that the smaller the CTE, the smaller the thermal strains and, hence, the total stresses: this is a fundamental concept in the design of devices directly interacting with the beam, since, for this type of equipment, mechanical loads are typically negligible and the design is usually isostatic, allowing free thermal expansion. The main (often single) source of stress is non-uniform temperature distribution (or non-homogeneity of CTE, e.g. in composite structures).

In the limiting case where CTE is zero everywhere, no thermal stresses are induced, regardless of the temperature increase!

Inverting Eq. (7a), *quasi-static total stresses* can be obtained:

$$\sigma'_{ij} = \frac{E}{(1+\nu)(1-2\nu)} [(1-2\nu)\varepsilon_{ij} + \nu\delta_{ij}\varepsilon_{kk}] - \delta_{ij} \frac{E\alpha T}{1-2\nu}. \quad (7b)$$

Although Eq. (7b) is time-dependent, since the temperature distribution obtained from Eq. (2) is a function of time, the stress distribution obtained can be considered quasi-static, given that the mass inertia effects are not yet taken into account (this is why the stress tensor components are primed).

Quasi-static stresses can be computed by combining Eq. (7b) with the equations of equilibrium and compatibility and the boundary conditions.

Some cases of special interest for isotropic bodies can be easily derived in the case of particular boundary conditions.

If no deformation is allowed, i.e. all total strain components are zero, as is the case for a fully constrained massive body, we observe that all shear stresses are zero while normal stresses in any element of the body are compressive and given by

---

<sup>1</sup> Although the terms *elastic limit*, *yield strength* and *flow stress* possess, strictly speaking, slightly different meanings, we will use them interchangeably in these lectures.

$$\sigma'_{ii} = -\frac{E\alpha T}{1-2\nu} . \quad (7c)$$

For a two-dimensional body, such as a thin, large plate, for which one can reasonably assume that in the through-thickness direction ( $z$  or 3) normal stress is zero and deformation is free, taking all other directions as constrained, it can be shown that Eq. (7b) reduces to

$$\sigma'_{11} = \sigma'_{22} = -\frac{E\alpha T}{1-\nu} . \quad (7d)$$

Finally, if only one direction is constrained, the other being free to expand, as in longitudinally clamped thin, long beams and rods, it can be shown that only the axial normal stress is non-zero; this is given by

$$\sigma'_{11} = -E\alpha T . \quad (7e)$$

In general, for more complex structures and boundary conditions, the study of the elastic thermomechanical response relies on numerical methods (typically implicit, linear *finite element analysis*); however, analytical solutions to Eq. (7) exist for simple geometries: among the most important solutions are those available for long circular cylinders and thin discs.

### 2.3.4 Thermo-elastic stresses in thin discs and long circular cylinders

For discs and cylinders, assuming an axially symmetrical,  $z$ -independent thermal distribution  $T(r, t)$  with adiabatic boundary conditions, we get, in a cylindrical reference system, for the radial and circumferential stresses

$$\sigma'_r(r, t) = \frac{E\alpha}{\zeta} \left[ \frac{1}{R^2} \int_0^R T(r, t) r \, dr - \frac{1}{r^2} \int_0^r T(r, t) r \, dr \right] , \quad (8a)$$

$$\sigma'_\theta(r, t) = \frac{E\alpha}{\zeta} \left[ \frac{1}{R^2} \int_0^R T(r, t) r \, dr + \frac{1}{r^2} \int_0^r T(r, t) r \, dr - T(r, t) \right] , \quad (8b)$$

where  $\zeta = 1$  for discs and  $\zeta = 1 - \nu$  for cylinders.

We note that at  $r = 0$ , radial and circumferential (or hoop) stresses are identical (compressive) and equal to:

$$\sigma'_r(0, t) = \sigma'_\theta(0, t) = \frac{E\alpha}{\zeta} \left[ \frac{1}{R^2} \int_0^R T(r, t) r \, dr - \frac{1}{2} T_0(t) \right] . \quad (8c)$$

This can be easily verified by observing that

$$\lim_{r \rightarrow 0} \frac{1}{r^2} \int_0^r T(r, t) r \, dr = \frac{1}{2} T_0(t) .$$

Since the body is free to expand radially, at  $r = R$ , radial stress is zero, while hoop stresses are always larger or equal to zero.

Making use of Eq. (1a), one can observe that the first term in Eq. (8a) is proportional to the *total deposited energy* (per unit length)  $Q_d$ , which, for an adiabatic problem, once the impact is concluded, remains constant over time and is therefore proportional to the uniform *final temperature*,  $T_F$ . Stresses at the centre and outer rim can then be easily computed once the maximum temperature (which is proportional to the peak energy) is known.

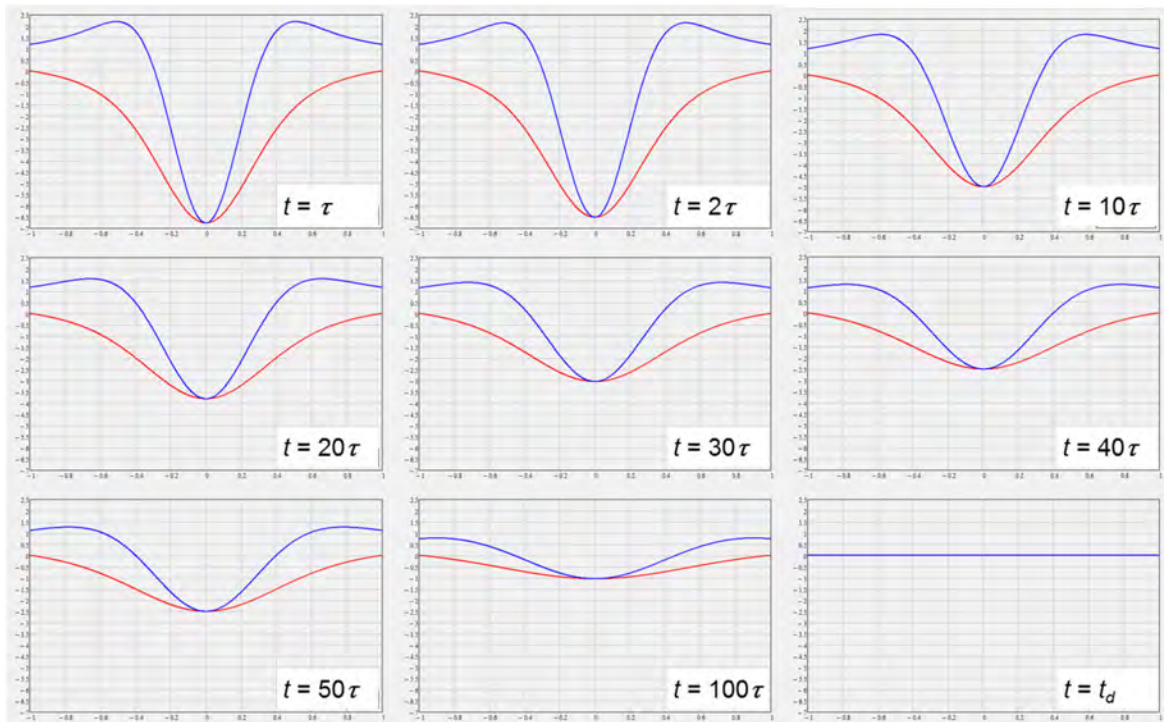
In particular,

$$\sigma'_r(0,t) = \sigma'_\theta(0,t) = \frac{E\alpha}{\zeta} \left[ \frac{Q_d}{2\pi R^2 \rho \bar{c}_p} - \frac{1}{2} T_0(t) \right] = \frac{E\alpha}{2\zeta} [T_F - T_0(t)] , \text{ for } t \geq \tau \quad (8d)$$

and

$$\sigma'_\theta(R,t) = \frac{E\alpha}{\zeta} \left[ \frac{Q_d}{\pi R^2 \rho \bar{c}_p} - T_0(t) \right] = \frac{E\alpha}{\zeta} [T_F - T_0(t)] , \text{ for } t \geq \tau . \quad (8e)$$

Figure 13 shows radial and circumferential stresses for a typical axially symmetrical energy distribution at various times. At times larger than  $t_d$ , when the temperature becomes uniform and equal to  $T_F$ , radial and hoop stresses go to zero everywhere.



**Fig. 13:** Radial and circumferential stresses in a circular cylinder as a function of radial coordinates for an axially symmetrical normal distribution at various times.

For  $t < \tau$ , assuming that no thermal diffusion has occurred yet, one can simply scale linearly with time to the stress values at  $t = \tau$ .

The axial stress component is usually disregarded for thin discs, since through-thickness stresses are negligible; conversely, in the case of long, slender structures, such as rods, bars, or beams, axial stresses become very important and are the main cause of the dynamic response.

For such structures, to compute axial stresses, it is initially assumed that the body is restrained at its ends, i.e. the axial strain is zero throughout ( $\varepsilon_z = 0$ ). In this hypothesis (known as the method of strain suppression [17]), quasi-static axial stresses can easily be derived from Eq. (7b).

For a long cylinder, using cylindrical coordinates, we get

$$\sigma'_z = \nu(\sigma'_r + \sigma'_\theta) - E\alpha T , \quad (9)$$

using radial and circumferential components obtained from Eqs. (8a) and (8b).

The distribution of axial stresses integrated over the cross-section results in a net compressive force  $R(t)$  (remember that we are blocking axial expansion, for convenience), which is a function of time and is equal to

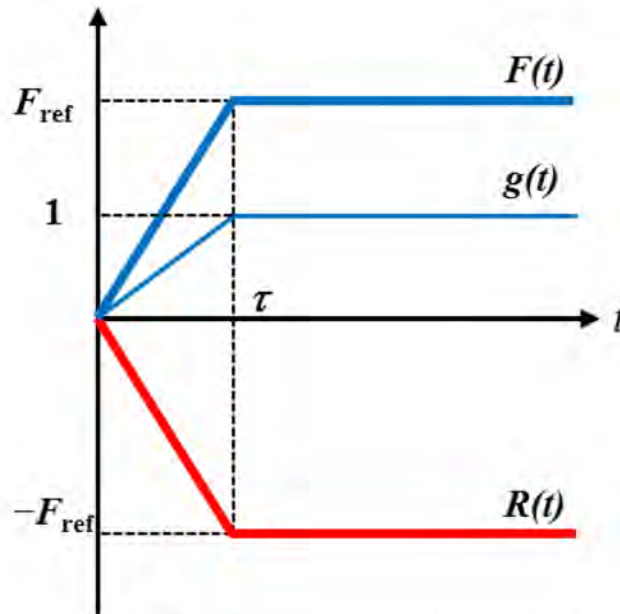
$$R(t) = 2\pi \int_0^R \sigma'_z(r, t) dr = -E\alpha \frac{Q_d}{\rho c_p} = -E\alpha T_F \pi R^2 = -F_{\text{ref}}, \text{ for } t \geq \tau. \quad (10)$$

We observe that the resultant axial force  $R(t)$  is proportional to the *total deposited energy* (per unit length)  $Q_d$ .

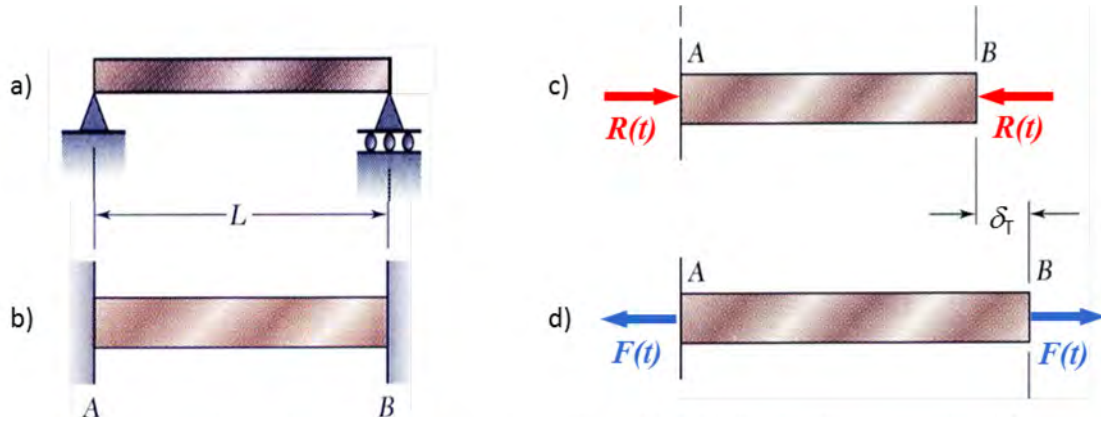
Very importantly, since  $Q_d$  is conserved after the impact (we are assuming an adiabatic problem), for  $t \geq \tau$ ,  $R(t)$  remains constant and proportional to the final uniform temperature  $T_F$ , regardless of the actual deposited energy distribution.

For  $t < \tau$ ,  $R(t)$  increases, following the trend of the deposited energy, so, disregarding the actual bunched structure of the beam, to a first approximation, it can be assumed that  $R(t)$  increases linearly from zero to a constant value, so that  $R(t) = -F_{\text{ref}}g(t)$ ,  $g(t)$  being the unit function shown in Fig. 14.

If the structure is simply supported and free to expand (as is usually the case for an isostatic structure), to restore the free-end boundary condition and allow thermal expansion  $\delta_T$ , a traction force opposed to the compressive resultant ( $F(t) = -R(t)$ ) can be superposed at the two ends of the rod (Fig. 15).



**Fig. 14:** Time history of the compressive force  $R(t)$  acting on a rod clamped at its ends and impacted by a beam (impact duration  $\tau$ ) and of the traction force  $F(t)$  to be superposed to restore the free-end boundary conditions. The unit function  $g(t)$  having the same trend of  $F(t)$  is also shown.

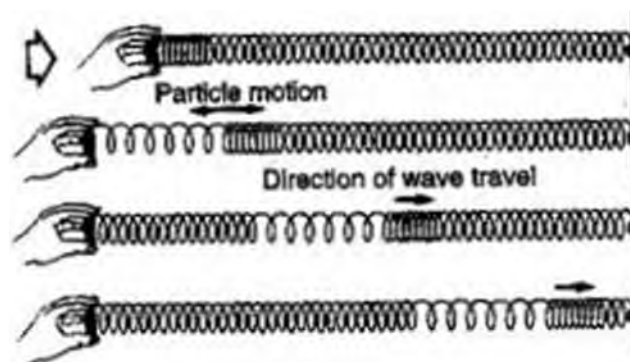


**Fig. 15:** Strain suppression and restoration approach. a) The rod is initially simply supported and free to expand. b) For convenience, the axial strain is suppressed, as if the rod were clamped. c) This leads to the generation of two compressive forces at the ends. d) The simply supported boundary condition is restored by superposing a traction force at the two ends.

Thanks to this approach, we have reduced the problem of a slender structure submitted to rapid heating to the well-known mechanical problem of the dynamic response of a beam to a pulsed axial excitation  $F(t)$  with rise time  $\tau$  (Fig. 12) applied at its ends. This response generates dynamic axial stresses  $\sigma_{z_d}$  (uniform on the rod cross-section) that can be superposed on the quasi-static stresses given by Eq. (9). The total axial stress is therefore given by

$$\sigma_z = \sigma'_z + \sigma_{z_d} . \quad (11)$$

Dynamic stresses appear because of the coupling between the rapidly applied load (thermal or mechanical) and the inertia of the material: since the heating occurs in a very short time  $\tau$ , during this process, thermal expansion in the bulk material is partly prevented by its mass inertia. However, at the two free ends, expansion is allowed to occur from the very beginning since nothing prevents particle displacement in the material. Expansion starts from the two rod ends, propagating towards the centre of the structure at the *speed of sound*,  $C_0 = \sqrt{E/\rho}$ . In this way, two elastic stress waves are generated; to some extent, this is equivalent to an axial spring that is rapidly moved outwards at its end (Fig. 16). Since these are expansion waves, dynamic tensile stresses propagate and are superposed on the compressive axial stresses, which are due to the initially ‘clamped’ state.



**Fig. 16:** Analogy of elastic wave propagation in a rod: rapid displacement at the free end generates a wave, which is propagated towards the centre.

The mechanical response of a simply supported cylindrical rod to a pulsed force with respect to time is a well-known problem in the theory of vibrations. It can be solved by resorting to such procedures as the *mode-summation method*. The *axial displacement*,  $u(z,t)$ , is expanded in terms of longitudinal *natural modes*,  $\phi_z(z)$ , and *generalized coordinates*,  $q_z(t)$ :



$$u(z, t) = \sum_i \phi_{z_i}(z) q_{z_i}(t) . \quad (12)$$

The solution can be obtained by means of *Lagrange's equation* for each independent mode:

$$\frac{d^2 q_{z_i}}{dt^2} + \omega_i^2 q_{z_i} = F_{z_i} . \quad (13)$$

The *natural modes* and the *natural (circular) frequencies* of longitudinal vibration for a simply supported beam of length  $L$  are given, respectively, by

$$\phi_{z_i}(z) = \sqrt{2} \cos\left(i\pi \frac{z}{L}\right), \quad (13a)$$

$$\omega_{z_i} = \frac{i\pi}{L} \sqrt{\frac{E}{\rho}} . \quad (13b)$$

The generalized forces  $F_{z_i}$  are given by

$$F_{z_i}(t) = \frac{F_{\text{ref}} \sqrt{2}}{m} \left[1 - (-1)^i\right] \cdot g(t) . \quad (13c)$$

Generalized coordinates  $q_z(t)$  are obtained from the response of a system with a single degree of freedom excited by a ramp function given by Eq. (13c):

$$q_{z_i}(t \geq \tau) = \frac{F_{z_i}}{(\omega_{z_i})^2} \left(1 - \frac{\sin(\omega_{z_i} t)}{\omega_{z_i} \tau} + \frac{\sin[\omega_{z_i}(t - \tau)]}{\omega_{z_i} \tau}\right), \quad (14)$$

$$q_{z_i}(t < \tau) = \frac{F_{z_i}}{(\omega_{z_i})^2} \left(\frac{t}{\tau} - \frac{\sin(\omega_{z_i} t)}{\omega_{z_i} \tau}\right).$$

Finally, the dynamic longitudinal stress component is calculated as follows:

$$\phi'_{z_i}(z) = -\sqrt{2} \cdot \frac{i\pi}{L} \cdot \sin\left(i\pi \cdot \frac{z}{L}\right), \quad (15a)$$

$$u'_z(z, t) = \sum_i \phi'_{z_i}(z) \cdot q_{z_i}(t) , \quad (15b)$$

$$\sigma_{z_d}(z, t) = E \cdot u'_z(z, t) , \quad (15c)$$

where  $u'_z(z, t)$  denotes the first derivative of  $u_z$  with respect to  $z$ , i.e. the longitudinal strain.

Figure 17 shows the evolution of dynamic axial stresses induced along the rod by the elastic wave at different times.

At the beginning of the impact, a tensile stress wave starts travelling at the speed of sound  $C_0$  from both ends towards the centre while the force  $F(t)$  (as well as axial stresses) linearly build up, generating a ramped wavefront. At the end of the impact ( $t = \tau$ ),  $F(t)$  and the axial stress stop increasing, reaching constant values of  $F_{\text{ref}}$  and  $\sigma_{\text{ref}}$ , respectively; the two wavefronts have now covered a length equal to  $\tau \cdot C_0$ . It can be shown that the value of the axial stress reached at this moment is simply given by

$$\sigma_{\text{ref}} = \frac{F_{\text{ref}}}{\pi R^2} = E\alpha T_F \quad (16)$$

The head of each elastic wave reaches the rod centre after one quarter of the *wave period*  $t_M = 2L/c$ ; after this time, the two waves start superposing, continuing to increase the axial stress at the centre during a time equal to  $\tau$ , after this time a maximum stress equal to  $2\sigma_{\text{ref}}$  is attained. The waves continue to propagate towards the other end of the rod, where the stress value always remains equal to  $\sigma_{\text{ref}}$ , as imposed by the boundary force  $F_{\text{ref}}$  acting there. At half the wave period, both wave heads reach the opposite ends of the rod and begin to become reflected as a compressive wave, so decreasing the stress value. After one full wave period, the head of the reflected wave has reached the departure end, being reflected again as an expansion wave; after an additional time equal to  $\tau$ , all the wave ramped front has been reflected and the cycle starts again.

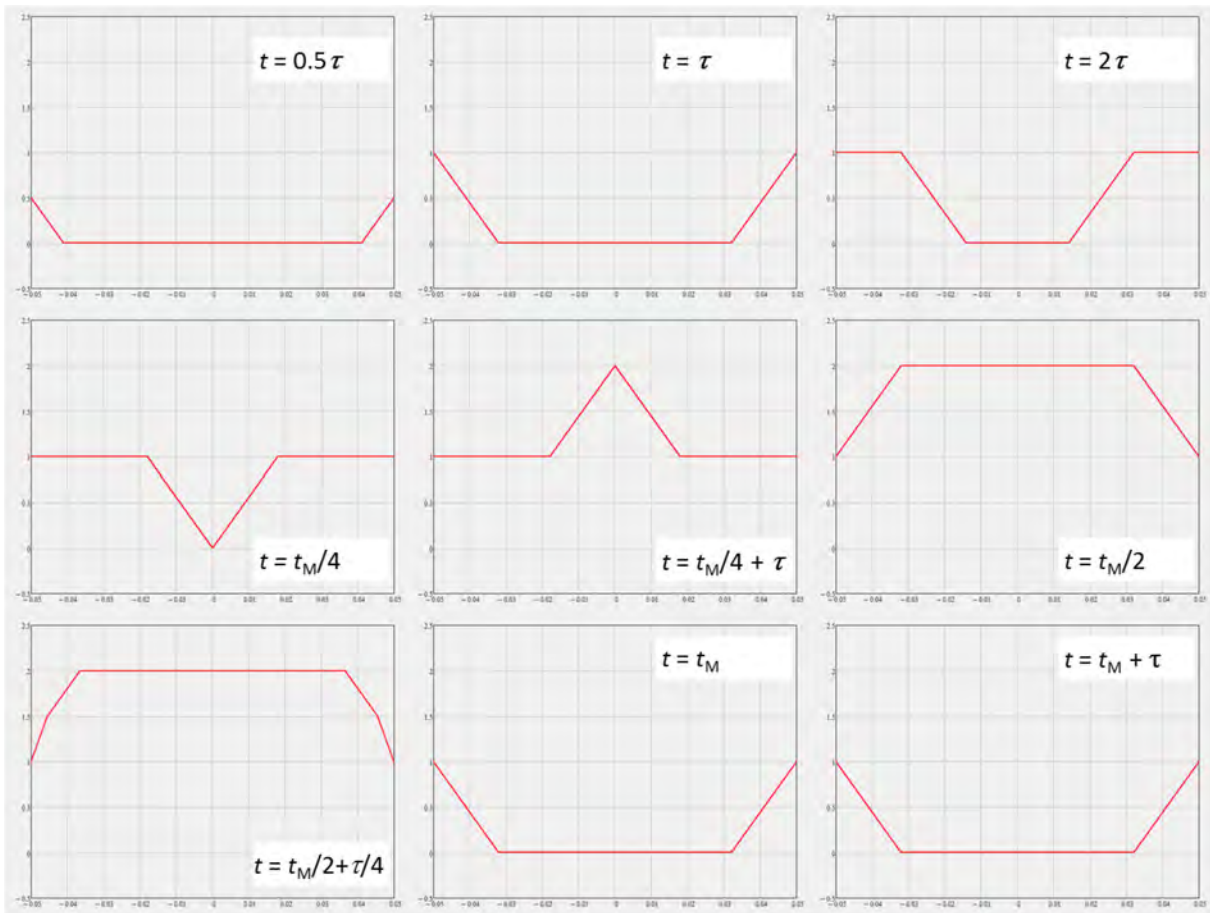
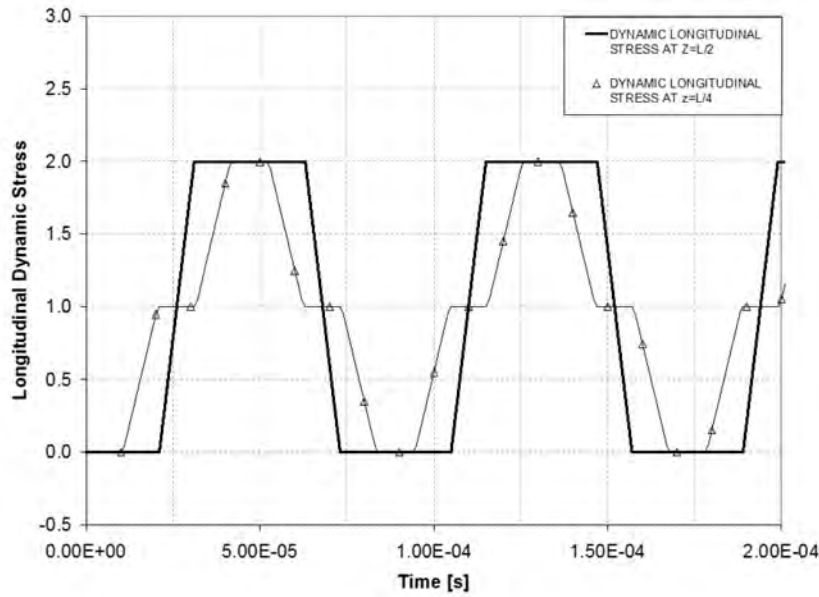


Fig. 17: Dynamic axial stress along a rod, scaled to  $\sigma_{\text{ref}}$ , at various times

Figure 18 shows the evolution over time of the axial dynamic stress.



**Fig. 18:** Dynamic axial stress scaled to  $\sigma_{ref}$  as a function of time at the centre and one quarter of the rod [15]

A similar approach, although more cumbersome, can be used for dynamic radial stresses. However, this component is small compared with quasi-static stresses in slender structures and can usually be neglected without affecting the general result [15].

If the beam hits the rod at a certain offset with respect to the centre, the energy deposition is no longer axially symmetrical and the problem becomes mathematically more complex, although it can still be solved analytically [18]. In such a case, dynamic bending stresses do appear, in addition to the dynamic stresses induced by the axial force. This can be explained by the fact that the resulting force at the two ends has an offset with respect to the centre of the beam, generating a bending moment, which varies in time. This effect is the probable cause of the permanent bending of the target rods depicted in Fig. 1: in that case, the beam, which hit the rod off its axis, probably caused dynamic bending stresses that exceeded the yield strength of the material, inducing a permanent bending of the rod.

#### 2.4 The non-linear thermomechanical problem

High-energy accelerator components are usually designed to work in the *elastic domain*; however, in the case of highly energetic beam accidents, the dynamic response of the structure can largely exceed this regime and lead, depending on the intensity of the phenomenon, to permanent deformations, very high pressures, changes of material density, phase transitions, intense stress waves, material fragmentation, and explosions [19].

When a fast transient load generates stresses with an amplitude exceeding the elastic limit of the material, the response will decompose into an elastic and a plastic wave. The plastic stress wave usually propagates at velocities lower than the elastic speed of sound ( $C_0$ ). However, if the energy is high enough to provoke stresses and rates of deformation (*strain rate*) exceeding a critical threshold of the order of  $10^4 \text{ s}^{-1}$  (Fig. 19), an energetic shock wave is formed, propagating at a velocity higher than  $C_0$ , and potentially leading to severe damage to the affected component [20].

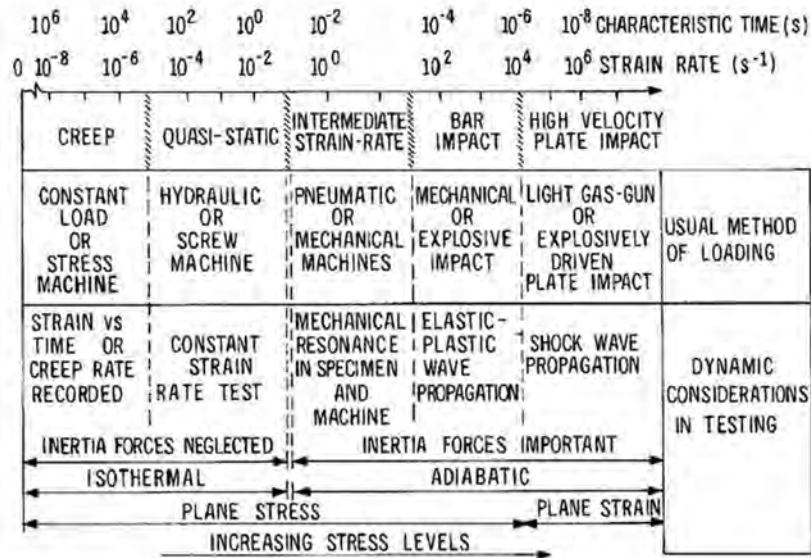


Fig. 19: Mechanical behaviour with changing strain rates and load duration [21]

Although an unambiguous identification of the critical threshold depends on the type of shock and on the geometrical conditions and is not always straightforward, as we have seen in Section 1, it is convenient to distinguish the responses according to their severity in a *plastic wave regime* or *plastic dynamic regime* when well below the threshold and in a *shock wave regime* above it. The former is usually associated with a limited permanent deformation induced by the beam impact but not with catastrophic failure, which is typical for the latter regime.

The treatment of both problem classes using pure analytical methods is virtually impossible and one must resort to numerical methods. However, the complexities of the tools required to compute the effects of these two types of regime are usually quite different, the analysis of events implying intense shock waves requiring more sophisticated numerical tools.

## 2.4.1 Numerical methods for beam-induced dynamic phenomena

### 2.4.1.1 Time integration methods

Two substantially different numerical approaches are available for the study of non-linear dynamic phenomena: when the response is relatively long and slow and the interest lies in the long-term global behaviour of a complex structure (such as oscillations and permanent deformation), rather than in capturing highly dynamic effects at a local scale, finite element tools relying on implicit time-integration schemes are preferred. On the one hand, these algorithms have the advantage of being unconditionally stable, allowing large time steps; on the other hand, they are computationally expensive (a stiffness matrix inversion is required at each step) and affected by numerical damping. All standard finite element codes, such as Ansys, Abaqus, and Nastran, belong to this category.

When large physical variations, such as large changes of density, phase transitions (melting, vaporization, plasma formation), fragmentation, or explosions, occur in a very short time, one must resort to an advanced class of numerical tool called *wave propagation codes* or *hydrocodes*. These are strongly non-linear finite element tools, using explicit time-integration schemes, which are conditionally stable, so that a short time step must be chosen according to the element dimension to ensure scheme stability (Courant–Friedrichs–Lewy condition). However, they are computationally efficient, since no stiffness matrix inversion is required. They are typically employed to study very fast and intense loading on materials and structures (high velocity impacts, explosions, crashes, etc.).

### 2.4.1.2 Mesh schemes

Several types of mesh scheme are available to describe the governing equations and their discretization in highly non-linear structural analyses.

The Lagrangian description is the most widely adopted scheme for structural analysis, both in standard finite element methods and in wave propagation codes: a Lagrangian mesh moves and distorts with the material it models as a result of forces from neighbouring elements; mesh nodes correspond to and move with ‘physical’ material points. This algorithm is usually very efficient; however, convergence problems can be met when material deformations are very severe, since elements, following the material, become highly distorted. When this occurs, mesh re-zoning is possible, but this is burdensome and introduces errors.

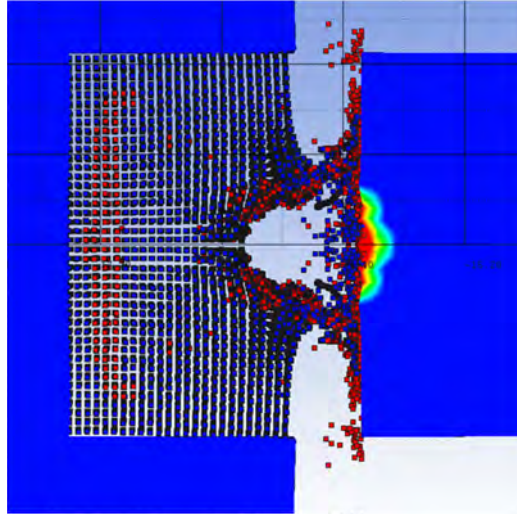
In such cases, alternatives must be found. In an Eulerian description, space is divided into fixed cells through which material flows: it is very well suited for problems involving extreme material movements (hypervelocity impacts, fluid mechanics, gas dynamics). It is computationally intensive, requiring higher element resolution and finer meshes than a Lagrangian scheme; moreover, the treatment of constitutive equations is complicated, owing to the convection of materials through the elements. This method is available in certain hydrocodes and is very extensively adopted for computational fluid dynamics calculations.

A compromise between Euler and Lagrange description is represented by the *arbitrary Lagrange Euler* (ALE) formulation: this hybrid technique tries to capture the advantages of both Lagrangian and Eulerian formulations. Typically, nodes on mesh boundaries and material interfaces move with the material (Lagrangian description), while all other interior nodes may either move with the material (Lagrange) or remain fixed in space (Euler). Most modern hydrocodes allow selection of this formulation, which is typically adopted to treat problems involving fluid/structure interaction.

An additional, relatively new technique for solving computational continuum dynamics problems is so-called *smoothed-particle hydrodynamics* (SPH). This is a mesh-free method ideally suited for certain types of problem with extensive material damage and separation. In this computational method, the material is modelled by a lattice of discrete elements (particles) with a spatial distance of interaction (smoothing length) over which their properties are weighted by a kernel function. Particles are interpolation points from which values of functions and their derivatives can be estimated at discrete points in the continuum. SPH particles can interact with Lagrange, Shell, and ALE meshes.

This method offers the possibility of studying crack propagation inside a body or the motion of expelled material fragments or liquid droplets. It is, therefore, well suited to the study of extreme beam impacts, in which explosion and *mechanical spalling* (ejection of material fragments from a surface of the impacted body) are involved (Fig. 20).

The SPH interaction points (particles) must, generally, be very small and packed to model the material accurately: a compromise must be found between accuracy and computation time.



**Fig. 20:** Simulation of the impact of a 7 TeV bunch of  $1.3 \times 10^{11}$  protons on a LHC tertiary collimator jaw. The left jaw is partly modelled by a SPH and Lagrangian mesh, while the opposite jaw is purely Lagrangian. Note the interaction between the ejected SPH particles and the Lagrangian mesh.

### 2.4.2 The plastic dynamic regime

When stresses exceed the elastic limit, materials typically undergo irreversible, non-linear plastic deformations, so that, on the removal of loads, the affected component will experience a permanent change of shape. This may occur under both quasi-static and dynamic conditions.

In dynamic conditions, the inelastic behaviour not only depends on the intensity of the applied load, as in quasi-static conditions, but also on the rate at which these loads are applied: this behaviour is called *rate-dependent* plasticity or *viscoplasticity*.

Plastic deformation, particularly for metals and alloys, basically occurs through a slip mechanism linked to the motion of dislocations, and the flow stress is essentially defined by the resistance to dislocation motion. The motion of dislocations inside the lattice is typically prevented by two types of obstacle: long- and short-range barriers.

Short-range barriers are strictly correlated to the material lattice and may include the lattice resistance itself, the resistance due to point defects, such as vacancies and self-interstitials, other dislocations intersecting the slip plane, alloying elements, and solute atoms. These dislocations can surmount short-range barriers partly by the action of shear stress due to externally applied loads and partly by increasing their thermal energy: higher temperatures tend to decrease the force required to move the dislocations. The strain rate has an opposite effect with respect to temperature: the dislocation is given less time to overcome the obstacle, attenuating the effect of the thermal energy and, consequently, increasing the force required to move dislocations. This contribution to the flow stress is called the *thermal* component and is a decreasing function of temperature and an increasing function of strain rate.

The long-range barriers may include grain boundaries, far-field forests of dislocations, and other microstructural defects with far-field influence. The resistance due to long-range barriers is often referred to as the *athermal* component of the flow stress. This type of obstacle cannot be overcome by additional thermal energy. The athermal part increases with increasing accumulated dislocations whose elastic field hinders the motion of mobile dislocations. While this elastic field does not explicitly depend on temperature, it is affected by temperature through the elastic moduli and their dependence on the temperature, and through the effect of the temperature history on the density of far-field dislocation forests. At suitably high temperatures, materials anneal, leading to a reduction in the dislocation density

and hence in the corresponding elastic field stress; this is reflected in a reduction of the athermal component.

In its simplest setting the flow stress is expressed as

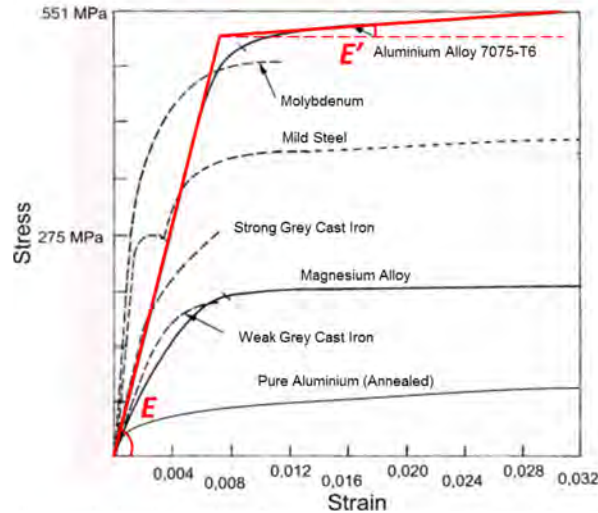
$$\sigma_y = \sigma_{th}(\varepsilon, \dot{\varepsilon}, T) + \sigma_{ath}(\varepsilon), \quad (17)$$

where  $\sigma_{th}$  and  $\sigma_{ath}$  are the thermal and athermal components of the resistance to the dislocation motion, respectively.

At moderate strains, the stress–strain curve exhibits non-linear trends like those depicted in Fig. 21; in some cases, the behaviour can be simplified, approximating the material response with a bilinear hardening law:

$$\sigma = E\varepsilon_{el} + E'\varepsilon_{pl}. \quad (18)$$

In Eq. (18),  $E'$  is the slope of the plastic linear function, sometimes called the *tangent modulus*,  $E$  is the Young (elastic) modulus, and  $\varepsilon_{el}$  and  $\varepsilon_{pl}$  are the elastic and plastic components of the strain, respectively. If  $E' = 0$ , the material is said to be *elastic–perfectly plastic*. If more accuracy is sought, the stress–strain curve can be approximated by a multilinear elastic–plastic function.



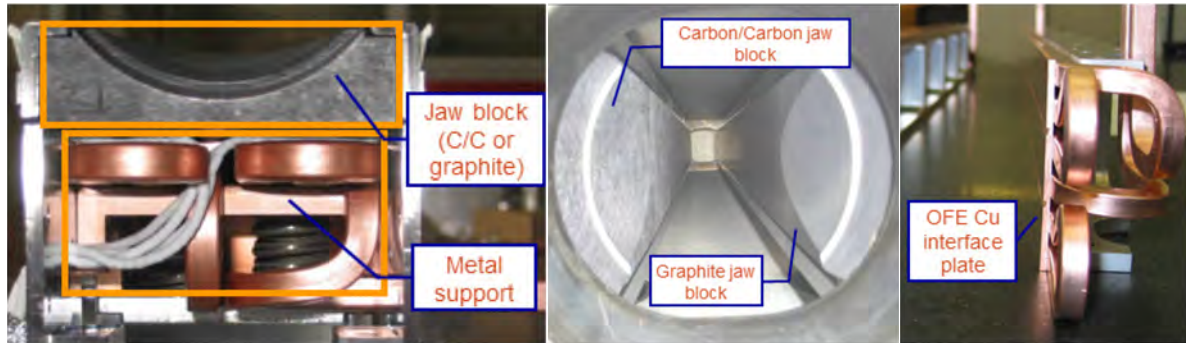
**Fig. 21:** Stress–strain curves beyond elastic region for several relevant materials. A bilinear approximation for an aluminium alloy is shown.

Along with permanent deformations, when large plastic strains occur, material density is also slightly affected: however, changes of density within plastic regime are in general small and can still be considered negligible.

Taking these assumptions into consideration, dynamic responses can be treated at an acceptable degree of approximation with non-linear, implicit *finite element method* codes, such as Ansys.

An example of this type of analysis is provided by a simulation of the effects of a LHC injection error on LHC secondary collimators: this scenario may lead to the impact on the collimator carbon/carbon (C/C) jaw of a full SPS batch of 288 bunches at 450 GeV ( $3.2 \times 10^{13}$  protons over 7.2  $\mu$ s), with transverse impact amplitudes up to 5–6  $\sigma_b$  [22]. In 2004, full-scale robustness tests were performed on collimator prototypes in the SPS extraction line to study such accidental cases. The two jaws of the prototype were submitted to a series of impacts at 450 GeV in two different conditions: (1) with increasing beam intensities at a fixed beam impact depth of 5 mm from the jaw surface and (2) with beam impact depths from 1 mm to 6 mm at a beam intensity of  $3.2 \times 10^{13}$  protons. For material

comparison, one of the jaw blocks was made of C/C (as for the series production), while the other was made of isotropic graphite (Fig. 22).



**Fig. 22:** Front view of the jaw assembly of a LHC secondary collimator (left); the two jaws in the collimator tank after completion of robustness tests in 2004 (centre); jaw metal support, showing the thin interface plate and the cooling pipes brazed on it (right).

Visual inspection of the jaw blocks after completion of the tests revealed no sign of mechanical damage; however, measurements performed on jaw assemblies revealed a permanent deformation of the metal support on both jaws of roughly 300  $\mu\text{m}$ , with a well-repeated pattern (the maximum deflection was located towards the downstream end of the support, where the highest temperatures occurred).

An explanation for the residual deformation can be inferred on the basis of the dynamic stresses developing in solids in the case of very fast heating due to material inertia partially preventing free thermal expansion, as explained in previous sections.

The maximum temperature increase expected in the 3 mm thick, oxygen-free electronic copper (OFE Cu) interface plate of the jaw metal support is of the order of 70°C. A simple analytical assessment of the elastic stress can be made by assuming that no in-plane expansion is possible; in this hypothesis, using Eq. (7b) for thin plates and using the thermo-elastic properties of copper, we find, for the in-plane stresses,

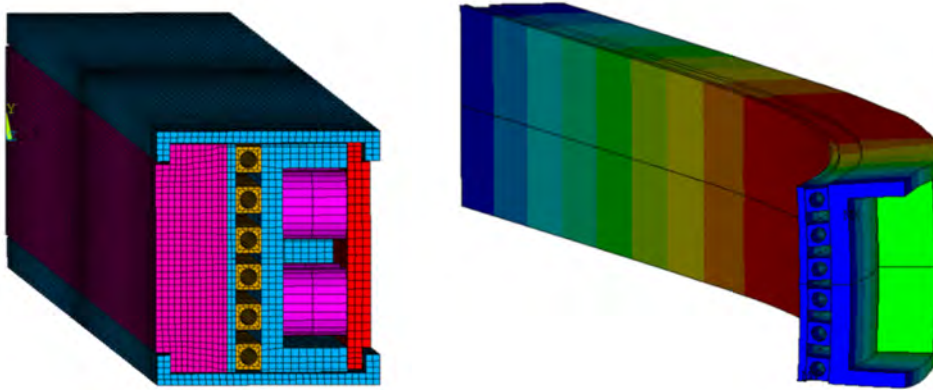
$$\sigma_{x_{\max}}^{\text{lin}} = -\frac{E\alpha\Delta T_{\max}}{1-\nu} \cong -210 \text{ MPa} .$$

This value largely exceeds the (compressive) yield strength of annealed OFE copper, which is limited to 50–70 MPa. Hence, residual strains will be present when the thermal load wanes. These compressive plastic strains are eccentric with respect to the neutral axis of the metal support and will lead to a permanent deflection of the metal support with a maximum sag towards the end experiencing higher temperatures.

This analytical assessment allows the permanent deflection and its shape to be justified qualitatively; however, it is practically impossible to estimate the magnitude of the effect quantitatively. To do so, it is necessary to resort to a non-linear, implicit, finite element analysis, including the effects of temperature, contacts, time, and plasticity (fast transient, coupled-field, elastic–plastic analysis).

Plasticity in metal components was modelled with both bilinear and multilinear kinematic hardening; results are shown in Fig. 23. As anticipated by the analytical estimate, the largest residual plastic strains are found on the thin copper plate, their magnitude (up to 0.12%) and extension being compatible with the simplified approach. The calculated permanent deflection (357  $\mu\text{m}$ ) of the metal support is close to the measured values and matches the actual deformed shape well.

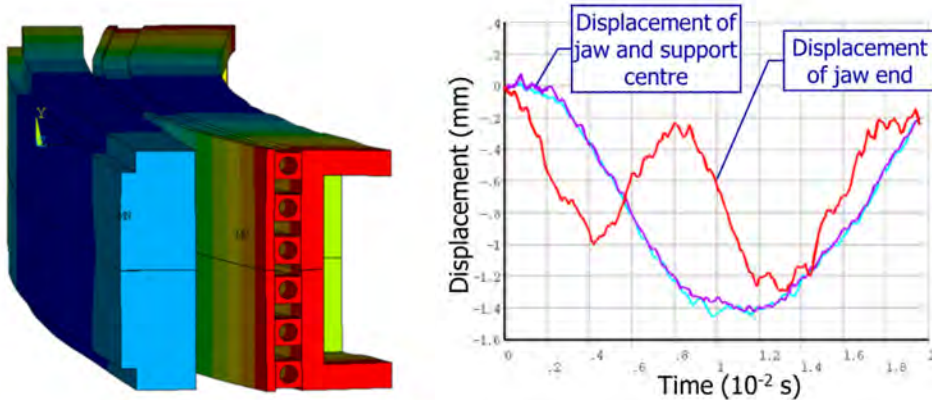




**Fig. 23:** Finite element model of the jaw assembly (left) and computed residual deflection of the jaw metal support (357  $\mu\text{m}$  max) (right).

On the basis of these results, it was decided to modify the jaw assembly series design by changing the thin plate material from OFE copper to Glidcop®, a copper alloy reinforced with a fine dispersion of alumina, which has a much higher yield strength (>200 MPa): analysis of an updated model of the series jaw gave a permanent deflection of 16  $\mu\text{m}$ . This improvement was achieved because plasticity is no longer attained on the thin plate and only occurs in a limited portion of the cooling pipes.

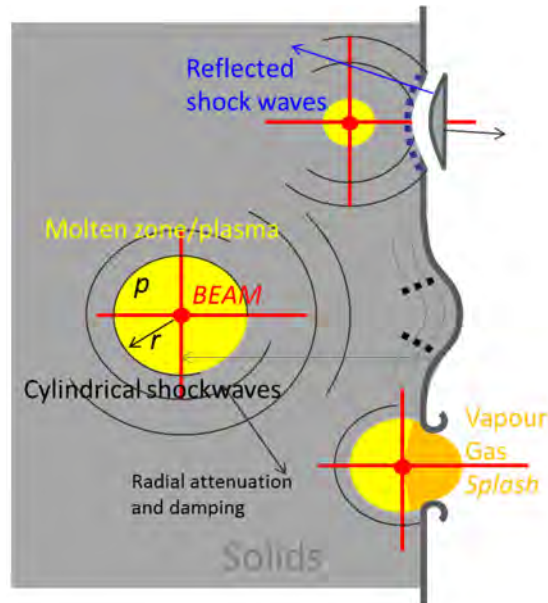
Another interesting outcome of the transient computations is the amplitude of the transverse oscillations occurring during the shock: as shown in Fig. 24, the maximum deflection at the centre of the C/C jaw reaches almost 1.5 mm after  $\approx 12$  ms; it is also worth noting that, during the transient part of the shock, the ends of the jaw may depart from the support by as much as 1.3 mm. The flexural frequency of oscillation is about 45 Hz.



**Fig. 24:** Deflection of the jaw assembly after  $\approx 12$  ms (peak value) and time history of lateral displacement for jaw and support.

### 2.4.3 The shock wave regime

As mentioned previously, when the deposited energy is high enough to provoke stresses and strain rates exceeding a critical threshold, an energetic shock wave is formed, propagating at a velocity higher than  $C_0$ . A shock wave is characterized by a sharp discontinuity in pressure, density, and temperature across its front. This event may be associated with a number of severe effects on matter, including phase transitions, explosions, and spalling (Fig. 25).



**Fig. 25:** Mechanism and effects that may be induced in solid by high-intensity particle beams at various regions of the impact [23].

In this respect, it is interesting to note that for metallic materials undergoing fast heat deposition, shock waves do not usually appear unless phase changes occur: if one assumes *uniaxial strains*, critical strains required to generate shock waves are in the range of 15% for tungsten and 7.5% for copper, whereas the total deformation at the melting point is in the range of 2% for both metals [19].

Conversely, for graphitic materials or other highly refractory materials, such as ceramics, the shock wave regime can be attained much sooner than the occurrence of extensive phase transitions.

As already mentioned, standard, implicit finite element techniques are not adapted to treat this class of problem and wave propagation codes or hydrocodes must be invoked [24].

In the usual continuum mechanics treatment, the complete stress tensor, which describes the material condition state, is divided into two components: deviatoric and hydrostatic tensors [20]. The name hydrocode stems from the original assumption of purely hydrostatic (fluid-like) behaviour of the impacted solids, which is typically acceptable when achieved stresses greatly exceed the flow strength of the material and the stress tensor can be approximately reduced to its hydrostatic component only; nowadays, the deviatoric component, responsible for material strength, is also taken into account but the original name is still widely used. Examples of codes used extensively to treat thermally induced fast dynamic phenomena are Autodyn, BIG2, and LS-Dyna. BIG2 is a two-dimensional code with a pure hydrodynamic solver which neglects the deviatoric component of the stress tensor. It was developed by Fortov *et al.* [25] for hypervelocity impacts and detonations, based on a Godunov-type numerical scheme. LS-Dyna [26] is a general-purpose transient dynamic finite element program including an implicit and explicit solver with non-linear thermomechanical capabilities. Finally, Autodyn [27] is a commercial explicit analysis tool particularly suitable for modelling the non-linear dynamics of solids, fluids, and gases, and their interactions.

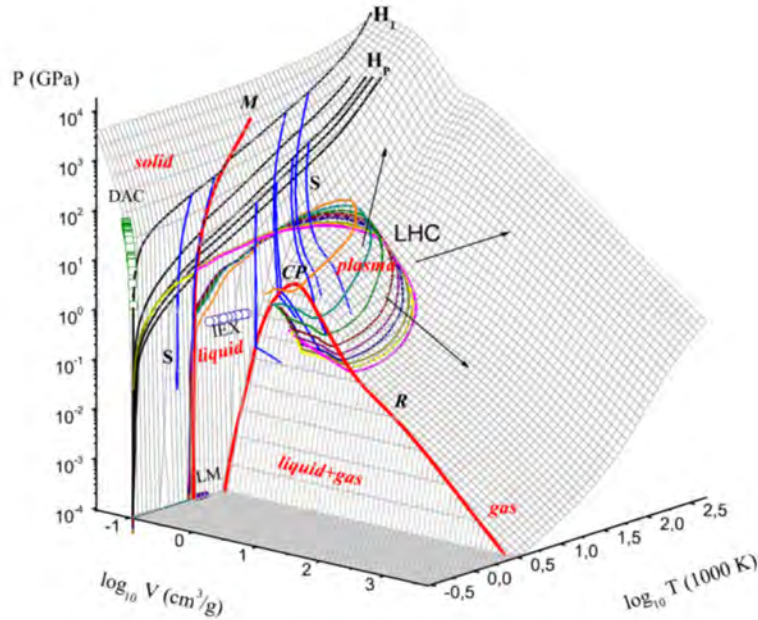
#### 2.4.3.1 Equations of state

The hydrostatic response in a hydrocode is governed by the *equation of state* (EOS), which expresses the relation between thermodynamic variables (such as pressure,  $P$ , internal energy,  $E$ , entropy,  $S$ , density,  $\rho$ , and temperature,  $T$ ). All these variables define the thermodynamic state of the matter. For a thermodynamic system that is in equilibrium, the state of the system is completely defined if two

independent and intensive variables are known. Usually, in hydrodynamics, the internal energy replaces temperature as independent variable. In this case, the EOS assumes the general form

$$P = P(\rho, E) \quad (19)$$

An EOS represents a set of surfaces, on which it is possible to define one-dimensional paths, which identify isothermal, isobaric, isochoric, isentropic processes (Fig. 26).



**Fig. 26:** Three-dimensional pressure–volume–temperature surface for copper [28]

The EOS implemented in commercial hydrocodes can be analytical or tabular. Analytical EOSs include, for example, the ideal gas law, linear EOSs (equivalent to linear thermo-elastic stress–strain relationships), polynomial EOSs, and the Mie–Grüneisen, GRAY, PUFF, and Tillotson EOSs: their use is limited, since they can usually describe only a single phase region. A tabular EOS, such as those provided by the SESAME database maintained by the Los Alamos National Laboratory, can be employed to evaluate material behaviour over different phases without loss in precision. Moreover, polynomial EOSs can be interpolated from tabular ones [20].

#### 2.4.3.2 Strength models

The deviatoric behaviour of a material is usually expressed by the *strength model*. In very fast and intense phenomena, the mechanical strength of the material is largely affected by the mechanical and thermodynamic variables that contribute to the material deformation process: the material may experience sharp discontinuities in pressure and temperature; the inner volume can melt, losing its yield strength, while the surrounding zone is still solid and subjected to heavy plasticity, generated by the shock wave propagation. The variables governing the material behaviour in the plastic regime are typically deformation (both plastic and elastic), strain rate, temperature, and pressure.

In previous decades, numerous material models in computational plasticity were proposed, for the description of deviatoric behaviour. Models are classified according to their nature in empirical, semi-empirical, and physically based models. Empirical models, such as the one proposed by Johnson and Cook [29], do not possess any physical basis and are phenomenological, obtained by interpolation of experimental data. The physically based models are obtained by starting from the transformation in the material occurring during a deformation process. An example of a semi-empirical model is the Zerilli–Armstrong model [30], which is based on the dislocation mechanics theory and presents a

different formulation for body-centred cubic and face-centred cubic materials. An alternative semi-empirical model is the Steinberg–Cochran–Guinan–Lund model [31], which was first developed for the description of high strain rate behaviour and subsequently extended to low strain rates. A completely physical-based and more complex model is the mechanical threshold stress [32]. Most of these material models are usually implemented in commercial FE codes, such as LS-Dyna and Autodyn.

When strain rate starts to play an important role, the Johnson–Cook model is one of the most popularly adopted strength models. This is an empirical multiplicative model, in which the effects of plastic strain, strain rate, and temperature are uncoupled; it is particularly suitable for metals and ductile materials. According to the Johnson–Cook model, the flow stress is defined as

$$\sigma_y = \left( A + B \varepsilon_{pl}^n \right) \left( 1 + C \ln \frac{\dot{\varepsilon}_{pl}}{\dot{\varepsilon}_0} \right) \left[ 1 - \left( \frac{T - T_r}{T_m - T_r} \right)^m \right], \quad (20)$$

where  $\varepsilon_{pl}$  is the equivalent plastic strain,  $\dot{\varepsilon}_{pl}$  is the plastic strain rate,  $A$  is the quasi-static elastic limit, and  $B$  and  $n$  are the work hardening parameters, which influence the slope and shape of the flow stress in the plastic domain. The parameter  $n$  usually assumes values between 0 (for a perfectly plastic model) and 1 (for a piecewise linear model).  $C$  expresses the sensitivity to the strain rate, while  $\dot{\varepsilon}_0$  is the effective plastic strain rate of the quasi-static test used to determine the yield and hardening parameters  $A$ ,  $B$  and  $n$  (in the original formulation, it was set equal to 1). The thermal effects are described by the thermal softening coefficient  $m$ , the actual temperature  $T$ , the reference temperature  $T_r$  used when determining  $A$ ,  $B$ , and  $n$  and the melting temperature  $T_m$  at which the material loses its shear strength and starts to behave like a fluid. The thermal parameter  $m$  determines the concavity of the temperature function: if  $m < 1$ , the function is convex, if  $m > 1$ , it is concave and if  $m = 1$ , the temperature influence is linear.

These parameters can be obtained through a set of experimental tests, which include Hopkinson bars, Taylor cylinders, and tensile and compression quasi-static tests at different temperatures [20]. Values of JC parameters for various materials are provided in Table 2.

**Table 2:** Parameters of Johnson–Cook strength model for selected materials

| Material            | Melting point<br>[K] | $A$<br>[MPa] | $B$<br>[MPa] | $n$  | $C$   | $m$  |
|---------------------|----------------------|--------------|--------------|------|-------|------|
| OFHC copper         | 1356                 | 90           | 292          | 0.31 | 0.025 | 1.09 |
| Cartridge brass     | 1189                 | 112          | 505          | 0.42 | 0.009 | 1.68 |
| Nickel 200          | 1726                 | 163          | 648          | 0.33 | 0.006 | 1.44 |
| Armco iron          | 1811                 | 175          | 380          | 0.32 | 0.060 | 0.55 |
| Electrical iron     | 1811                 | 290          | 339          | 0.40 | 0.055 | 0.55 |
| 1006 steel          | 1811                 | 350          | 275          | 0.36 | 0.022 | 1.00 |
| 2024-T351 aluminium | 775                  | 265          | 426          | 0.34 | 0.015 | 1.00 |
| 7039 aluminium      | 877                  | 337          | 343          | 0.41 | 0.010 | 1.00 |
| 4340 steel          | 1793                 | 792          | 510          | 0.26 | 0.014 | 1.03 |
| S-7 tool steel      | 1763                 | 1539         | 477          | 0.18 | 0.012 | 1.00 |
| Tungsten alloy      | 1723                 | 1506         | 177          | 0.12 | 0.016 | 1.00 |

### 2.4.3.3 Failure models

A *dynamic failure model* is typically used in association with a strength model to describe the failure mechanism of a material submitted to rapidly applied loads and determine its structural limits.

The factors that influence dynamic failure are typically the material properties and microstructure, the applied loads and the conditions they induce (stress, strain rate, and temperature), and the ambient environment.

As will be shown in more detail in Section 3.1, depending on the failure mode, the materials are classified as brittle (such as ceramics or glass) or ductile (such as metals or polymers). The former class is characterized by very limited plastic deformation before failure and nearly flat fracture surfaces originating from a single crack propagation. Ductile materials exhibit large plastic deformations, usually with necking phenomena and the typical cup-and-cone shaped failure surface, which is the result of nucleation, growth, and coalescence of voids in the material.

Dynamic failure models can be divided into two categories. In the first category, the material is supposed to fail when locally it overcomes a limit for one or more variables (such as strain to fracture, tensile hydrostatic stress, or maximum principal stress). This type of failure mechanism could be used to describe dynamic brittle failure or phenomena such as spalling. The second category includes failure models that are based on cumulative damage mechanisms: the material starts to be damaged if some property limits are exceeded; damage evolution is then controlled by a damage parameter that can increase until complete failure is achieved. This type of failure mechanism is used to describe dynamic ductile failure.

Examples of dynamic failure models are the maximum plastic strain failure criterion, the minimum hydrostatic pressure failure criterion ( $P_{\min}$ ), and the Grady spall model.

As exemplified in Fig. 25, if intense particle beams impact a solid close to a free surface, the compressive shock wave is immediately reflected and turned into a tensile wave, causing bulk failure and material ejection (spalling) if its amplitude is higher than the material hydrostatic strength.

This mechanism is usually reproduced in hydrocodes through the  $P_{\min}$  model, which broadly corresponds to the maximum normal stress criterion for slower load conditions; material models may also take into account the energy necessary for crack formation, calculated on the basis of the material fracture toughness.

One example of hydrocode computations is the simulation of accidental beam impacts of one or more full bunches on a tertiary collimator for the LHC [33]. The analysis was carried out by making use of Autodyn and simulating the whole collimator jaw assembly (Fig. 27). The jaw section directly interacting with the beam is composed of five Inermet 180 blocks, each 200 mm long, fixed with stainless steel screws to a housing made of OFE Cu. The copper housing is, in turn, brazed to cooling pipes made of copper–nickel alloy (90% Cu, 10% Ni), which are then brazed to a back stiffener made of Glidcop.

Two complementary three-dimensional models were implemented in Autodyn, based respectively on (a) a Lagrangian mesh of the full jaw assembly, to study the shock wave propagation and assess possible damage in each element of the jaw assembly, and (b) a SPH model of the most loaded Inermet block, to study the high-speed ejection of tungsten particles and their impact on the tank and on the opposite jaw.

Table 3 provides details of the constitutive models used for each of the relevant materials. It is worth noting that water in the cooling pipes was also included in the analysis.

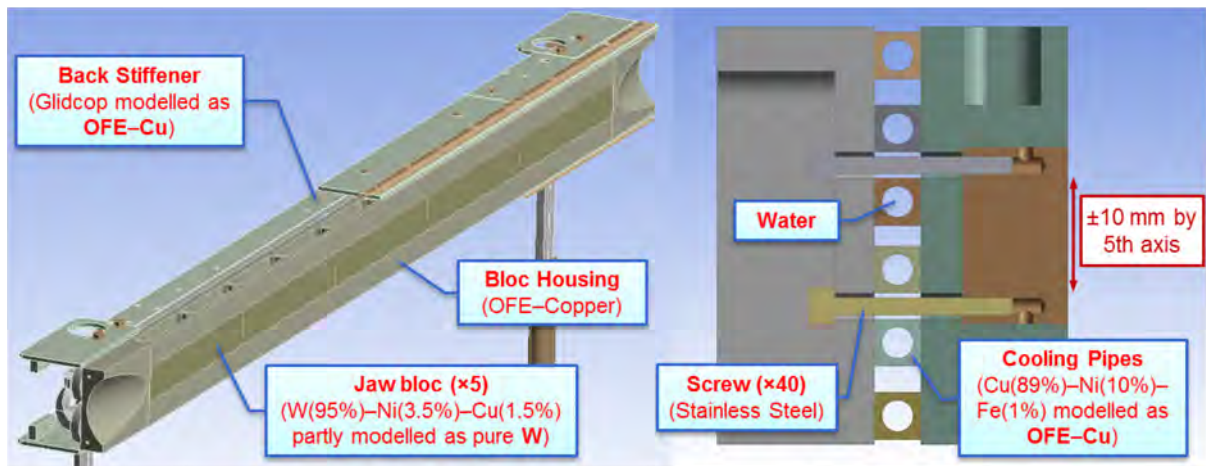


Fig. 27: Three-dimensional view and cross-section of the jaw assembly of a LHC tertiary collimator

Table 3: Material constitutive models for LHC tertiary collimator analysis

| Material                 | EOS             | Strength model | Failure model                     |
|--------------------------|-----------------|----------------|-----------------------------------|
| Inermet/tungsten         | Tabular(SESAME) | Johnson–Cook   | Max plastic strain and $P_{\min}$ |
| OFE copper               | Polynomial      | Johnson–Cook   | Johnson–Cook                      |
| Stainless steel AISI 316 | Shock           | Johnson–Cook   | Max plastic strain                |
| Water                    | Shock           | –              | $P_{\min}$                        |

Seven accident cases, with different degrees of severity and probability, were identified (Table 4). All of the cases are based on an asynchronous beam abort event, assuming that each bunch has the same impact parameter (2 mm). The beam energy corresponds to the values that were expected for run 1 of the LHC. The impinging proton pulses constitute trains of bunches of  $1.3 \times 10^{11}$  particles with energy up to 5 TeV, spaced by 25 ns.

Table 4: List of accident cases

| Case | Beam energy<br>[TeV] | Normal emittance<br>[ $\mu\text{m rad}$ ] | No of impacting<br>bunches | Energy on jaw<br>[kJ] |
|------|----------------------|---|----------------------------|-----------------------|
| 1    | 3.5                  | 3.50                                      | 1                          | 38.6                  |
| 2    | 5                    | 7   | 1                          | 56.2                  |
| 3    | 5                    | 3.5                                       | 1                          | 56.5                  |
| 4    | 5                    | 1.75                                      | 1                          | 56.6                  |
| 5    | 5                    | 1.75                                      | 2                          | 111.3                 |
| 6    | 5                    | 1.75                                      | 4                          | 216.1                 |
| 7    | 5                    | 1.75                                      | 8                          | 429.8                 |

A complete FLUKA model of the collimator was set up and full shower simulations were carried out for each case to provide the deposited energy distribution.

The impact of every bunch with matter leads to a sudden increase in temperature and pressure, which in turn generates an outbound shock wave. The expansion wave that follows may in turn lead to a substantial reduction in density. Hence, one should, in principle, update the FLUKA model for each bunch and re-perform the simulations with the new density map, since changes in densities substantially affect the deposited energy distribution (FLUKA–hydrocode coupling). However, for the simulated accident cases, the change of density during the impact of the bunch train was found to be negligible: therefore, the energy distribution map calculated in the initial state could be used throughout the whole simulation.

To determine consequences on the collimator and LHC operation, three different damage levels were defined:

- Level 1 – Collimator not to be replaced. Limited jaw damage: an intact spare surface can be found relying on the 5th axis movement, which permits a maximum vertical shift of  $\pm 10$  mm; negligible permanent jaw deformation.
- Level 2 – Collimator to be replaced. Damage to the jaw incompatible with 5th axis travel; other components (e.g. screws) may also be damaged.
- Level 3 – Long downtime of the LHC. Very severe damage to the collimator, leading to water leakage into beam vacuum.

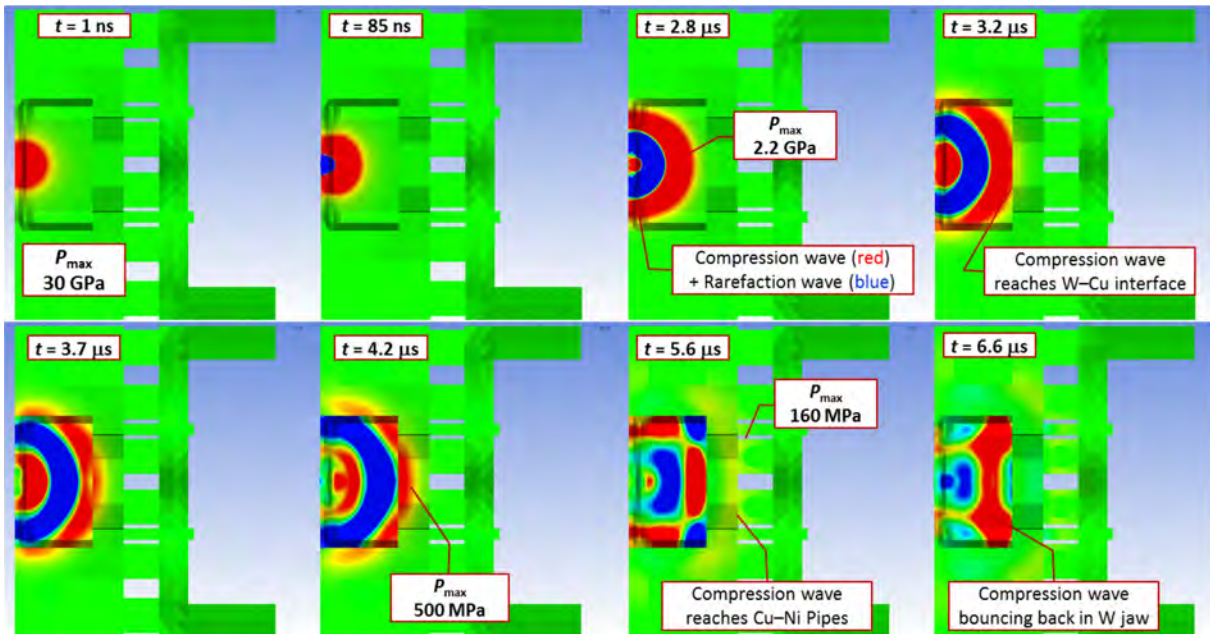
Results predict that all the single-bunch cases, both at 3.5 and 5 TeV, at all emittances, fall within damage level 1. The primary variable determining damage extent on the jaw is the total energy deposited: the size of the damaged region is already much larger than the beam size so that no sensible difference is found when varying the beam emittance. Even in the less destructive cases, a sizeable plastic deformation is found on the copper support and on the cooling circuit; a groove on the surface of impacted Inermet blocks, with an extension roughly proportional to the bunch energy, well reproducible with the SPH method, is also generated, while Inermet fragments are projected towards the opposite jaw.

It was also found that a key role in determining the damage extension induced by beam impacts on a composite structure is played by the shock impedance matching between adjoining components. Shock impedance in a given material is defined as

$$Z = A\rho_0U_s, \tag{21}$$

where  $A$  is the interface surface,  $\rho_0$  is the initial density and  $U_s$  is the shock velocity.

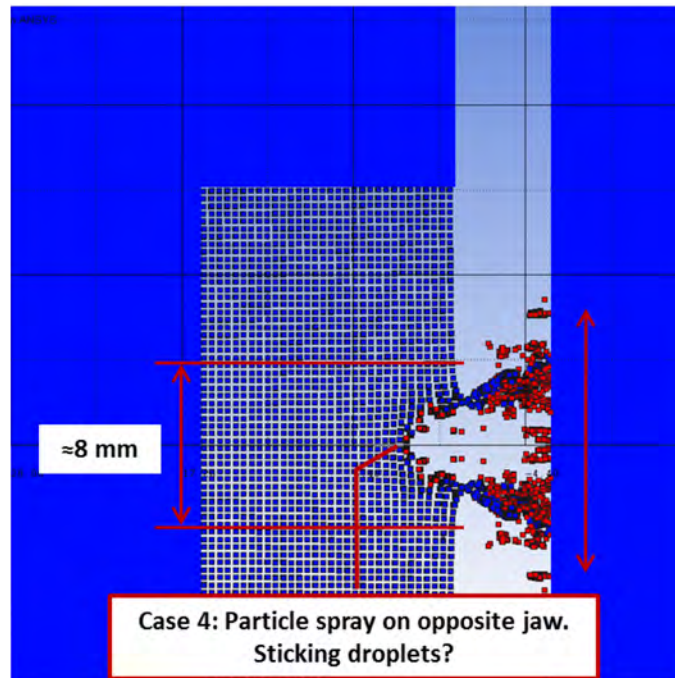
Owing to the large shock impedance mismatch between tungsten and copper (high  $Z_W$  to  $Z_{Cu}$  ratio), most of the wave energy is confined inside the Inermet®180 blocks: this limits the damage produced in other critical components, such as the cooling pipes (Fig. 28).



**Fig. 28:** Case 4: propagation of the shock wave in the jaw assembly shown at various instants. Note that the wave is mostly reflected at the W–Cu interface and only partially transmitted to the copper housing.

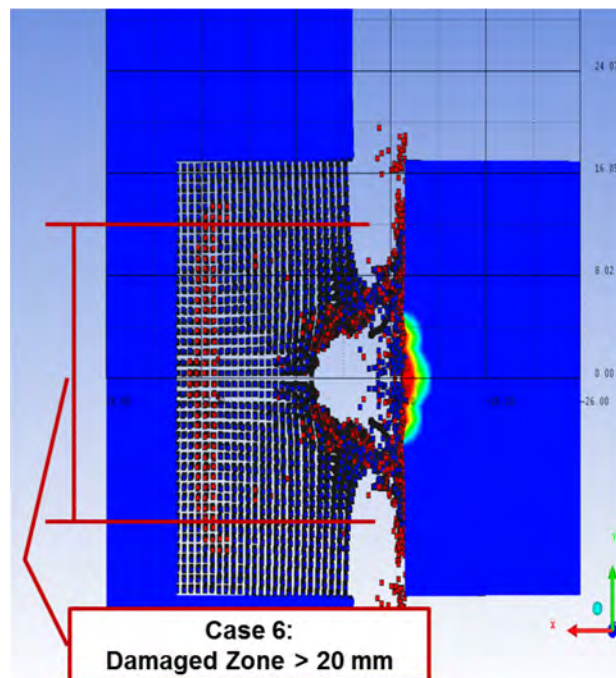
It can also be observed that the jaw damage extension at 5 TeV (case 4) is at the limit of damage level 2; plastic deformations on cooling pipes and screws remain limited, and tungsten particles are

sprayed on a larger area of the opposite jaw (Fig. 29). This jaw is not directly damaged; however, its final flatness may be affected by possible re-solidified droplets stuck on its surface.



**Fig. 29:** Case 4: damage extension on Inermet block – note particles sprayed on opposite jaw

For cases 5 and 6, the jaw damage cannot be compensated by 5<sup>th</sup>-axis travel (damage level 2). Severe plastic deformations can be observed on cooling pipes and screws, although visible failures are not detected. The SPH simulations anticipate permanent damage on the opposite jaw, provoked by tungsten particles impacting at elevated velocity (Fig. 30).



**Fig. 30:** Case 6: high-speed particle spray provoking an extended damage on the opposite jaw



The only case studied leading to damage level 3 is case 7. In this scenario, one may expect: a) high risk of water leakage due to very severe plastic deformation on the pipes (plastic strain up to  $\approx 21\%$ ); b) extended eroded and deformed zone on the tungsten jaw; c) projections of hot and fast, solid fragments ( $T \approx 2000$  K,  $V_{\max} \approx 1$  km/s) onto the opposite jaw with slower particles hitting tank covers at velocities just below the ballistic limit; d) risk of permanent bonding between the two jaws due to the projected re-solidified material (Fig. 31).

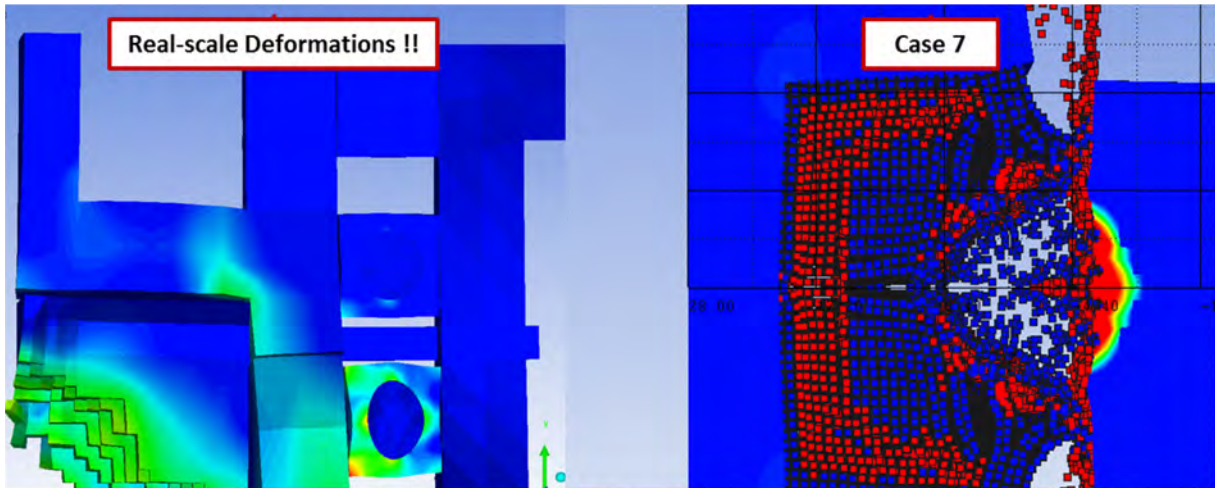


Fig. 31: Case 7: plastic strain in Cu and Cu–Ni (left) and damage extension on the two jaws (right)

#### 2.4.4 Hydrodynamic tunnelling

In the previous example, the energy deposition calculated for the first bunch on the pristine material was maintained also for subsequent bunches since the change of density induced by the impinging particles was found to be negligible for the duration of the impact. The same approach was followed for similar calculations on other structures [34].

As already discussed, however, the expansion wave, which follows the compression shock wave generated by prolonged intense impacts, when propagating radially away from the impacted region, may displace material outwards, reducing material density. Additionally, material density at the target core abruptly drops because of induced phase transitions (to liquid, gas, and plasma), along with pressure release. If subsequent bunches arrive after a lapse of time sufficiently long to allow the rarefaction wave to develop and pressure to decrease drastically, particles will experience a considerable increase of interaction length (which is density-dependent) and penetrate the matter more deeply, owing to density reduction: this extreme phenomenon is sometimes called the *hydrodynamic tunnelling* effect [28].

In such cases, a coupling between the interaction–transport code and the wave propagation code is necessary, to take into account the change of density during the interaction of the beam with matter.

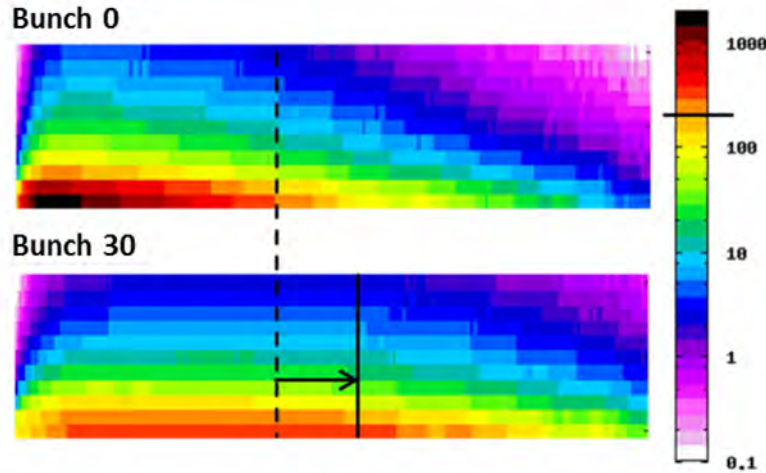
The following example is that of a tungsten cylindrical target impacted at its centre by 30 LHC full bunches at the energy of 7 TeV [35].

Results of the first FLUKA simulation, performed on pristine material, are uploaded in the LS-DYNA mechanical model. Then, for each bunch, the coupling algorithm performs the following operations:

- Immediately before the impact of bunch  $n$ , it obtains, from LS-DYNA, the density map induced by the impact of all previous bunches.
- It updates the regions of the FLUKA model that underwent significant density changes (in excess of a few percent).

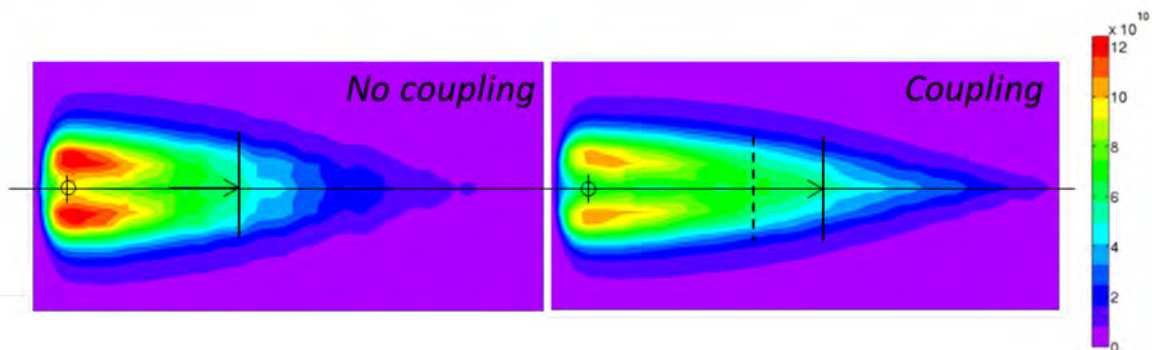
- It runs a new FLUKA calculation, to be imported in LS-DYNA, simulating the impact of bunch  $n$ .

Results show that the density variation leads to some reduction of deposited energy at each bunch: as shown in Fig. 32, the energy deposition peak penetrates more deeply at each successive bunch.

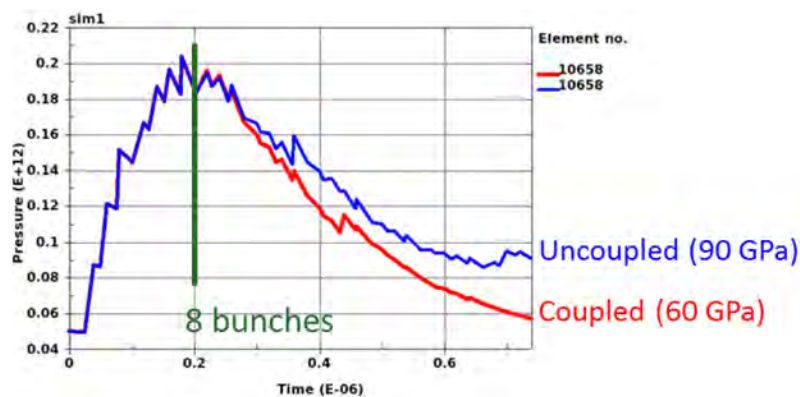


**Fig. 32:** Energy deposition ( $\text{GeV cm}^{-3}$ ) in longitudinal section for 1st and 30th bunches

Comparison with the uncoupled solution shows that pressure is also affected: its maximum value, in the beam axis direction, decreases as the shock wave penetrates the material (Fig. 33). Results also confirm that the differences between coupled and uncoupled analyses are significant only when a substantial density reduction occurs: for the studied cases more than 10 bunches are necessary (Fig. 34).



**Fig. 33:** Differences in pressure between FLUKA–hydrocode coupled and uncoupled simulations for a tungsten target impacted by 30 LHC bunches. Pressures are in Pa.



**Fig. 34:** Comparison between coupled and uncoupled solution of maximum pressure in a target element as a function of time. Differences become appreciable after roughly 10 LHC bunches.

### 3 Design principles of beam intercepting devices

#### 3.1 Introduction to failure criteria

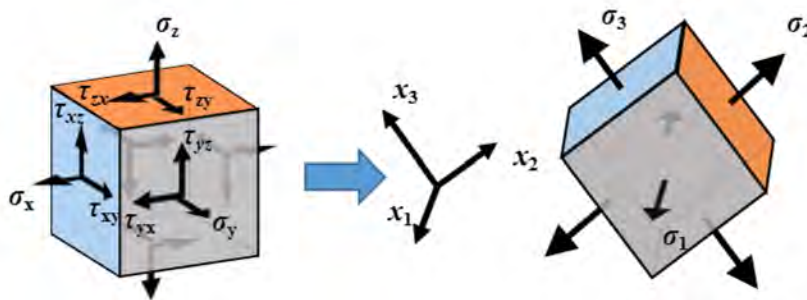
As we have seen, components directly exposed to interactions with particle beams, such as collimators, absorbers, targets, dumps, or windows, which, in short, we shall call beam intercepting devices (BIDs) are subjected, like any engineering component, to complex loadings in tension, compression, bending, torsion, or pressure, or combinations of these, so that, at a given point in the material, stresses often occur in more than one direction. If sufficiently severe, such combined stresses can act together to cause the material to yield (i.e. to exceed its elastic limit and undergo plastic, irreversible deformation) or fracture (as is more often the case for brittle materials). Predicting the safe limits for use of a component made of a given material under combined stresses in the elastic or elastic–plastic regimes requires the application of a *failure criterion*. Failure criteria in the more extreme shock wave regimes are usually replaced by the dynamic failure models introduced in Section 2.4.3, although the use of ‘standard’ failure criteria is sometimes also possible in the shock wave regime.

A number of different failure criteria are available, some of which predict the failure by yielding and others by fracture. The former are called *yield criteria* and the latter *fracture criteria*. In general, yielding is considered a form of failure in that the permanently deformed component is expected to no longer meet its design requirements, e.g. because of loss in precision, alignment, or load-carrying ability; in some cases, however, a limited plastic deformation may be tolerated, provided it does not impair component functionality.

Failure criteria are usually based on values of stress, so that their application involves in general calculation of an *equivalent stress* that condenses the complex state of stress into a single value, which is then compared with the yield or fracture strength of the material. To do so, it is always possible to identify a coordinate system in which the complete three-dimensional state of stress can be reduced to one in which only normal stresses are acting; the axes of this particular coordinate system are called *principal directions* and the corresponding normal stresses are called *principal stresses*, conventionally indicated  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_3$  (Fig. 35). Therefore, the failure criterion usually simplifies to  $f(\sigma_1, \sigma_2, \sigma_3) = 0$ .

For isotropic materials, it is often useful to rewrite the three principal stresses in terms of the so-called *invariants* of the stress tensor, because they are independent of the orientation of the coordinate system:

$$\begin{aligned} I_1 &= \sigma_1 + \sigma_2 + \sigma_3, \\ I_2 &= \sigma_1\sigma_2 + \sigma_2\sigma_3 + \sigma_3\sigma_1, \\ I_3 &= \sigma_1\sigma_2\sigma_3. \end{aligned} \tag{22}$$



**Fig. 35:** Transformation of a general three-dimensional state of stress into a state of stress with normal stresses only (principal stresses).

In some cases, however, particularly when dealing with materials for which a linear elastic behaviour cannot easily be found, it might be appropriate to base the failure criterion on strains rather than stresses.

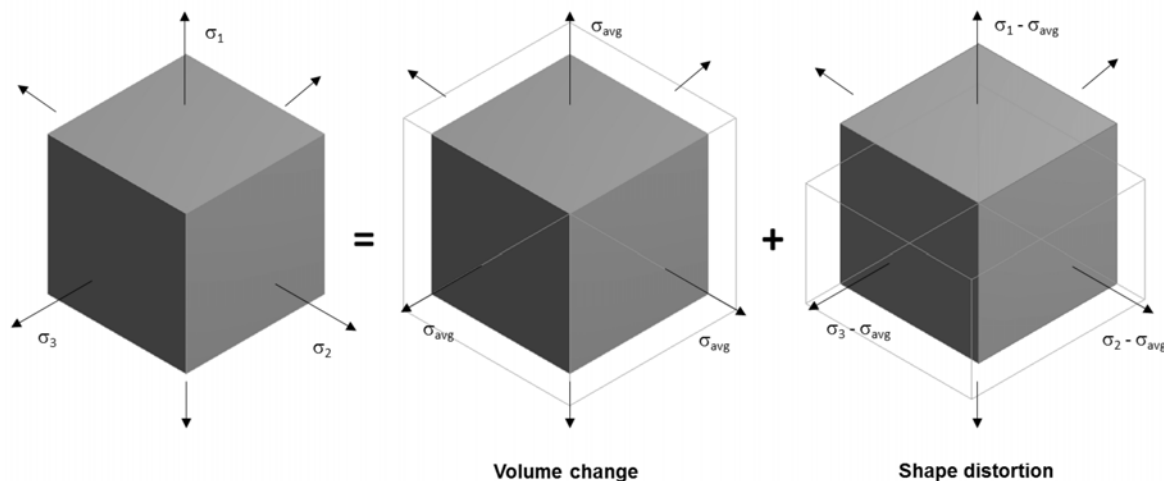
Since a given material may fail by either yielding or fracturing, depending on its properties, the state of stress, and the loading conditions (quasi-static or dynamic), no single failure criterion is suitable for every material under any state of stress and for all conditions: the choice of an appropriate failure criterion is therefore a critical step in the design of a structural component and must be carefully considered. An overview of the most adapted criteria for BIDs is given in the next sections.

Safety coefficients are adopted to protect against approximation of failure criteria and uncertainties in the knowledge of the state of stress.

### 3.1.1 Maximum distortion energy theory

The *maximum distortion energy theory*, also known as the *von Mises yield criterion*, *Huber–Hencky–von Mises yield criterion*, or *octahedral shear stress yield criterion* is a failure theory extensively used for ductile materials.

In applying stresses to a structural element, mechanical work is done; for a material within the elastic regime, all of this work is stored as potential energy. This internal *strain energy* can be partitioned into one part associated with *volume change* (caused by hydrostatic stress,  $\sigma_{\text{avg}} = (\sigma_1 + \sigma_2 + \sigma_3)/3$ ) and another part associated with *distortion* of the shape of the material element by the remaining portion of the principal stresses, corresponding to the deviatoric components of the stress tensor (Fig. 36).



**Fig. 36:** Contributions to the deformation of a material element: volume change is due to hydrostatic stress, shape distortion to the deviatoric portion of principal stresses.

Experimental observations have shown that ductile materials do not yield when subjected to uniform hydrostatic stresses: based on this empirical evidence, Huber proposed that material yielding occurs when the *distortion energy* per unit volume  $u_d$  reaches the distortion energy per unit volume of the same material when subjected to yielding in a tension test  $(u_d)_Y$ .

Mathematically, this reduces to the following relationship between an equivalent stress  $\sigma_{\text{eq}}$  and the yield strength  $\sigma_Y$ :

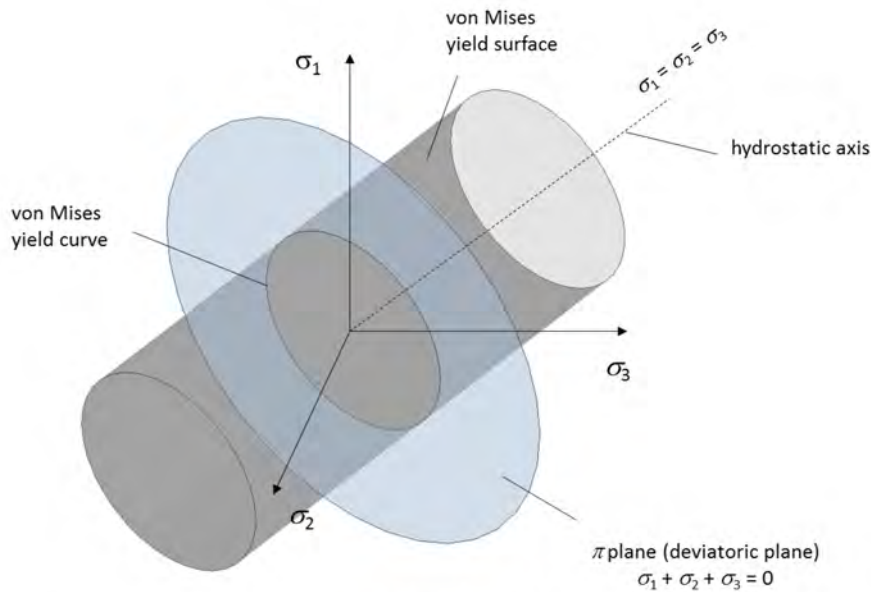
$$\sigma_{\text{eqVM}} = \frac{1}{\sqrt{2}} \sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2} = \sigma_Y \quad (23)$$

According to this criterion, yielding is supposed to occur in the material, when locally the equivalent stress (defined previously) reaches or exceeds the yield strength of that material.

It can also be shown that in Eq. (23) failure by yielding is assumed to occur when the octahedral shearing stress in the material reaches a value equal to the maximum octahedral shearing stress in a tension test at yield.

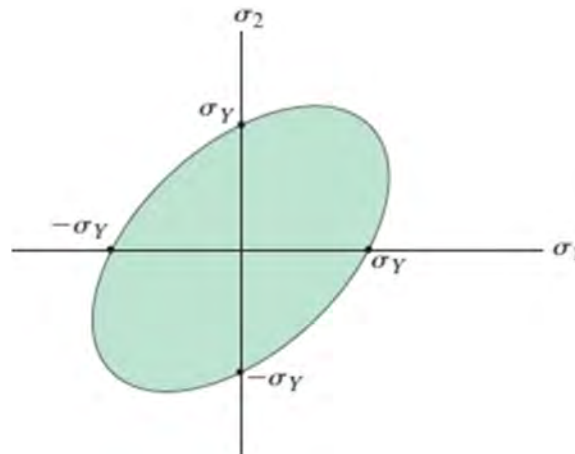
Equation (23) defines a three-dimensional surface in the principal stress space representing a circular cylinder having its axis on the line  $\sigma_1 = \sigma_2 = \sigma_3$  (Fig. 37). Any combination of principal stresses falling inside this cylindrical boundary is below the yield stress and hence safe according to the von Mises criterion, while the surface itself represent the geometrical locus of yielding.

A safety factor can be defined as the ratio between the equivalent stress and the yield strength: one can observe that in the case of purely hydrostatic stresses, the principal normal stresses are all equal and the equivalent stress is zero, so the safety factor against yielding is infinite. This state of pure hydrostatic stress is represented graphically by the axis of the circular cylinder.



**Fig. 37:** Three-dimensional yield surface of the von Mises yield criterion

If any one of the three principal stresses is zero, the intersection of the yield surface with the plane of the remaining two principal stresses gives an ellipse, as shown in Fig. 38.



**Fig. 38:** Yielding locus for the von Mises criterion for plane stress ( $\sigma_3 = 0$ )

### 3.1.2 Maximum shear stress theory

As discussed previously, yielding in ductile materials is associated with the effect of the deviatoric part of the stress tensor: on this basis, Tresca and Guest suggested that failure by yielding occurs when the maximum shear stress in any plane reaches a critical value corresponding to the maximum shear stress that causes the same material to yield when it is subjected only to axial tension.

An equivalent stress can again be identified: failure occurs when this equivalent stress equals the yield strength of the material. Mathematically, the *maximum shear stress theory* or *Tresca–Guest yield criterion* is expressed by

$$\sigma_{\text{eqTG}} = \max(|\sigma_1 - \sigma_2|, |\sigma_2 - \sigma_3|, |\sigma_3 - \sigma_1|) = \sigma_Y . \quad (24)$$

In the principal stress space, the locus of yielding corresponds to the surface of a hexagonal prism with its axis given by the line  $\sigma_1 = \sigma_2 = \sigma_3$  (which implies, like the von Mises criterion, that hydrostatic stress does not affect yielding).

For plane stress (one of the principal stresses being zero), the maximum shear stress theory is represented by a distorted hexagon, obtained by the interception of the hexagonal prism with a plane  $\sigma = 0$ .

A comparison between the failure loci provided by the von Mises and Tresca–Guest criteria for plane stress reveals that the latter is slightly more conservative, particularly for pure shear when  $\sigma_1 = -\sigma_2$  (Fig. 39).

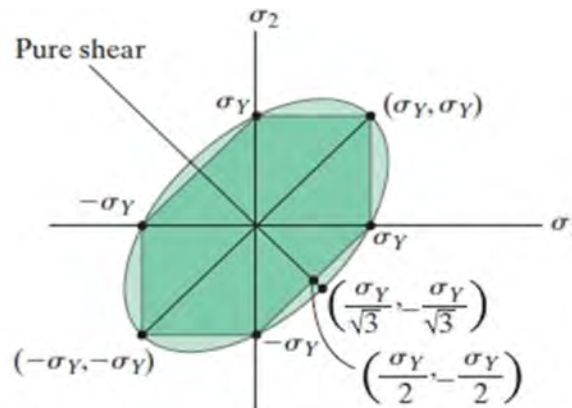


Fig. 39: Comparison between von Mises and Tresca–Guest failure loci in case of plane stress

### 3.1.3 Drucker–Prager yield criterion

In most cases, materials possess the same yield strength in tension and compression; however, in certain cases, compressive and tensile yield strength might be significantly different, the former being much larger than the second; these are said to be *uneven materials*. In such cases, an extension of the von Mises criterion, taking into account non-symmetry in stress–strain curves, can be invoked.

The so-called *Drucker–Prager yield criterion* is commonly used to model various pressure-dependent materials. It can be expressed using the principal stresses as

$$\sqrt{\frac{1}{6}[(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2]} = A + B(\sigma_1 + \sigma_2 + \sigma_3) . \quad (25)$$

Note that the left-hand side of the equation corresponds (up to a constant factor) to the von Mises equivalent stress, cf. Eq. (23).  $A$  and  $B$  are two independent material parameters that can be derived from the tensile and compressive strengths,  $\sigma_t$  and  $\sigma_c$ . The value of  $A$  defines the size of the yield locus

in principal stress space and thus relates to the overall strength of the material. The parameter  $B$  describes the dependence on hydrostatic pressure, i.e. on the first invariant of the stress tensor.

As opposed to the infinite cylinder representing the von Mises model, the Drucker–Prager yield locus forms a cone along the hydrostatic axis, which widens in the direction of compressive stress states (Fig. 40).

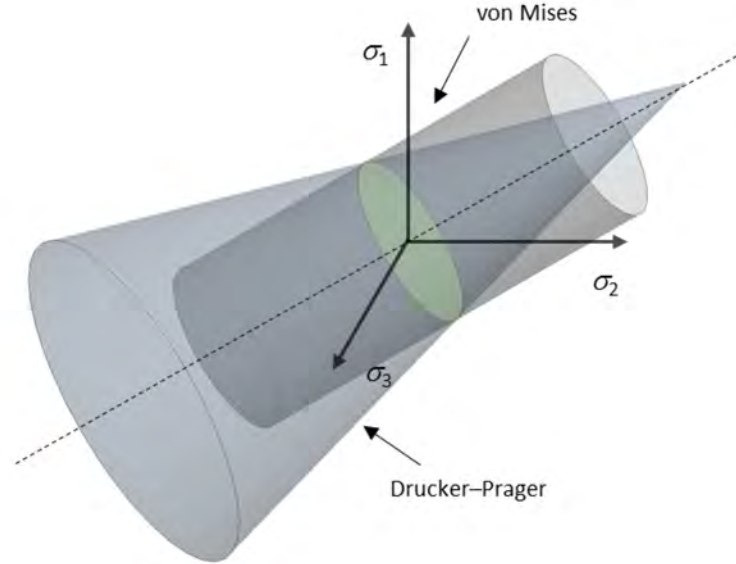


Fig. 40: Comparison between Drucker–Prager and von Mises yield surfaces

### 3.1.4 Mohr–Coulomb failure criterion

The Drucker–Prager yield criterion is closely related to the *Mohr–Coulomb model*, which is represented by a hexagonal cone inscribed inside the circular cone of the Drucker–Prager model. This is a direct analogy to the Tresca hexagonal prism inscribed inside the von Mises cylinder. According to the Coulomb–Mohr criterion, failure is supposed to occur on a given plane when a critical combination of shear and normal stress acts on this plane. At failure, the mathematical relationship between stresses is given by

$$\frac{|\tau| + \mu\sigma}{\sqrt{1 + \mu^2}} = \tau_u \quad (26)$$

where  $\tau$  and  $\sigma$  are the stresses acting on the fracture plane, while  $\tau_u$  is the pure shear failure stress and  $\mu$  is a material constant related to the angle formed by the plane of fracture with the plane of the maximum principal stress, and typically assumes values between 0.15 and 0.6.

The Mohr–Coulomb model is particularly suitable for reproducing the behaviour of brittle uneven materials in a predominantly compressive state.

### 3.1.5 Stassi–d’Alia yield criterion

A pressure dependency similar to the Drucker–Prager criterion is invoked by the *Stassi–d’Alia yield criterion*. Besides the distortion strain energy considered by von Mises, this criterion also takes into account a contribution of the hydrostatic pressure  $P = -(\sigma_1 + \sigma_2 + \sigma_3)/3$  (the negative of the hydrostatic stress  $\sigma_{\text{avg}}$ ), so that the tensile equivalent stress for which yield starts to occur is given by the root of the equation

$$k \cdot \sigma_{\text{teqSA}}^2 + 3(k-1)P \cdot \sigma_{\text{teqSA}} - \sigma_{\text{eqVM}}^2 = 0, \quad (27)$$

where  $k = -\frac{\sigma_{Yc}}{\sigma_{Yt}}$  is the ratio between the compressive and tensile yield strengths of the material.

The compressive equivalent stress is simply given by  $\sigma_{ceq} = -k\sigma_{teqSA}$ . It can easily be verified that the Stassi–d’Alia criterion reduces to the von Mises criterion when  $k = 1$ .

### 3.1.6 Maximum normal stress theory

The *maximum normal stress theory*, often referred to as the cut-off or *Rankine criterion*, is applicable to brittle materials. According to this theory, the material fails when the maximum principal stress of a given stress state reaches either the uniaxial tension strength  $\sigma_{ut}$  or the uniaxial compression strength  $\sigma_{uc}$ . This criterion can be expressed mathematically as

$$\begin{aligned}\sigma_{uc} &\leq \sigma_{\max} \leq \sigma_{ut}, \\ \sigma_{\max} &= \max(|\sigma_1|, |\sigma_2|, |\sigma_3|).\end{aligned}\tag{28}$$

Note that no interaction between the principal stresses is considered. In the principal stress space, the Rankine criterion corresponds to a cube oriented according to the three principal axes that intersect with the stress axes at the values of  $\sigma_{uc}$  and  $\sigma_{ut}$ . This model is therefore *not* independent of hydrostatic stress.

For brittle materials,  $\sigma_{uc}$  is usually much larger than  $\sigma_{ut}$ ; these materials commonly contain large numbers of randomly oriented microscopic cracks that cannot support significant tensile stresses, since these stresses tend to open these flaws and cause them to grow. If the dominant stresses are compressive, the planar flaws tend to have their opposite sides pressed together so that they have less effect on the failure behaviour. This explains the higher strengths in compression. Also, compressive failure occurs in planes aligned with planes of maximum shear.

The Rankine criterion gives reasonably accurate predictions of fracture in brittle materials as long as the normal stress having the largest absolute value is tensile, while agreement with data is much worse in compression. For such materials, the tensile ultimate strength is usually measured through flexural tests, since in a tensile test the material tends to break in correspondence with the testing machine grasps. However, flexural tests usually overestimate the stress to failure because the maximum tensile stress is only reached in one face of the bent specimen, elsewhere being smaller or even compressive. For these reasons, in recent years more advanced tests have been devised to measure the ultimate strength of brittle materials in a purely tensile state: as an example, the Hopkinson bar set-up can be configured to generate a plane tensile shock wave on the sample, also allowing the sensitivity of the material to the strain rate to be evaluated [36].

In practice, several brittle failure models consist of a combination of a Rankine criterion in tension and a more elaborate surface in compression (that allows interaction between principal stresses), such as the Mohr–Coulomb fracture criterion mentioned in Section 3.1.4: such a combination is represented for instance by the so-called *modified Mohr–Coulomb fracture criterion* (Fig. 41): the transition between tension dominated states, expressed by the Rankine criterion and the compressive dominated states, modelled by the Mohr–Coulomb criterion is usually found at ratios  $\sigma_2/\sigma_1 \approx -1$ , corresponding to the state of pure torsion.



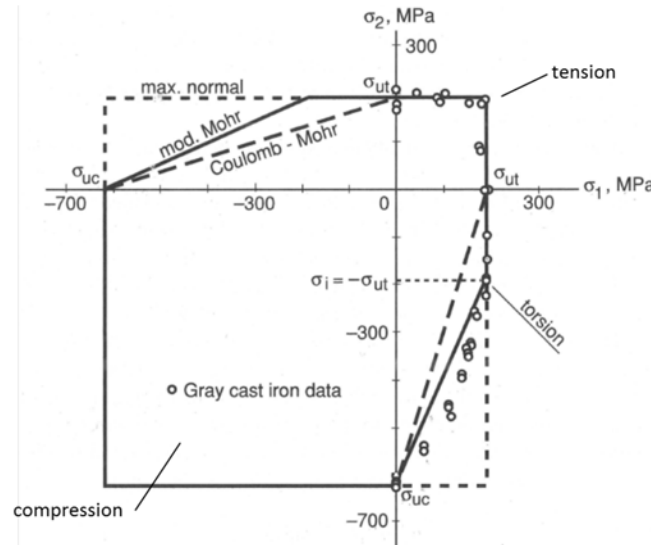


Fig. 41: Fracture data for grey cast iron compared with various failure criteria [37]

### 3.1.7 Hill criterion for orthotropic materials

All failure models presented so far are concerned with isotropic materials, thus they are represented by functions of principal stresses  $\sigma_1, \sigma_2, \sigma_3$ . This is not possible for anisotropic materials, since the orientation of the material does not permit the rotation of the stress tensor to the eigenvectors. As a result, failure criteria for orthotropic materials inevitably depend on all six components of the stress tensor.

The most commonly used orthotropic yield function is the *Hill criterion*, which is an extension of the isotropic von Mises model. It employs six independent material parameters that may be accessed experimentally from uniaxial tension and pure shear tests in the three material orientations. The criterion is of the form

$$F(\sigma_{11} - \sigma_{22})^2 + G(\sigma_{22} - \sigma_{33})^2 + H(\sigma_{33} - \sigma_{11})^2 + 2L\sigma_{12}^2 + 2M\sigma_{23}^2 + 2N\sigma_{31}^2 = 1 \quad (29)$$

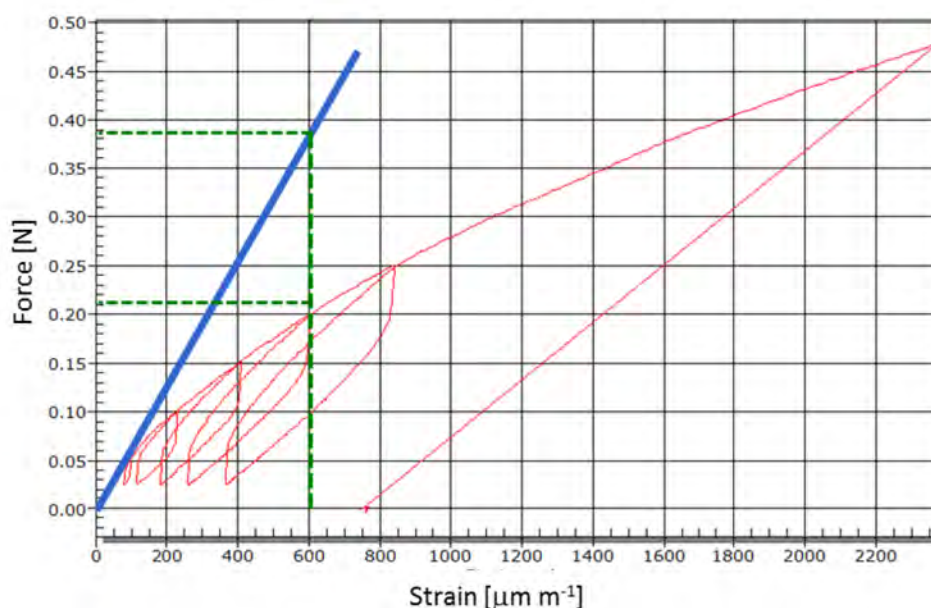
Equation (29) reduces to the von Mises criterion, albeit in terms of the stress components instead of the principal stresses, simply by setting  $F = G = H = 1/(2\sigma_Y^2)$  and  $L = M = N = 3/(2\sigma_Y^2)$ . Because orthotropic materials cannot be written as functions of the principal stresses, a graphic visualization of the yield locus is impossible.

The Hill criterion can be extended in the same fashion as the Drucker–Prager criterion to include differences in tensile and compressive strength. This results in an anisotropic Drucker–Prager model with three additional parameters accounting for the effect of hydrostatic pressure in each direction of material symmetry. Note that with increasing refinement of the models, the number of parameters increases. As a result, the applicability and accuracy of a constitutive model is often determined by the experimental accessibility of the parameters.

### 3.1.8 Criteria for non-linear materials: deformation to failure

The linear approximation is a powerful means of describing the stress–strain relationship in the elastic regime. For some materials, however, the  $\sigma$ – $\epsilon$  relationship can depart appreciably from linearity. Examples are ‘soft’ materials, such as annealed copper or aluminium and magnesium alloys, and, most interestingly for BIDs, graphitic materials.

For deformation-driven problems, such as beam-induced energy deposition, considerable overestimation can be made when considering tension as the limiting factor (Fig. 42).



**Fig. 42:** Stress–strain test for a molybdenum–graphite grade. Departure from linearity is shown, as well as overestimation of tension stress in the case where a linear stress–strain relationship is assumed.

In the case of brittle behaviour, as in the example shown in Fig. 42, it is more appropriate to replace a fracture criterion based on ultimate strength, particularly in tension, such as the Rankine theory, with one based on *deformation to fracture*: failure is reached once the maximum principal strain reaches the value of the ultimate strain obtained from a uniaxial tension or bending test.

### 3.1.9 Summary

A multitude of failure criteria exist for various applications. Some of the most prevalent isotropic and orthotropic rate-independent yield functions have been introduced (Table 5). Smooth failure surfaces are often suited to ductile materials, whereas most brittle materials exhibit a competition between different failure modes, each of which is represented by a separate patch of an overall non-smooth failure envelope.

Beam intercepting devices often make use of brittle materials for their active parts. In such cases, the use of a modified Mohr–Coulomb criterion should be considered, whereas if the material exhibits a strongly non-linear behaviour, the maximum normal strain criterion is better suited, at least in the tension region of the principal stresses space.

For ductile isotropic materials, either the von Mises or the Tresca–Guest criterion is usually chosen.

**Table 5:** Summary of relevant failure criteria, categorized according to the smoothness of the failure surfaces, dependence on hydrostatic pressure, and applicability to orthotropic materials.

|             | Pressure-independent   | Pressure-dependent   |
|-------------|--|--|
| Isotropic   | Huber–Hencky–von Mises ( <i>smooth</i> )<br>Tresca–Guest ( <i>non-smooth</i> ) | Drucker–Prager ( <i>smooth</i> )<br>Stassi–d’Alia ( <i>smooth</i> )<br>Rankine ( <i>non-smooth</i> )<br>Mohr–Coulomb ( <i>non-smooth</i> )<br>Modified Mohr–Coulomb ( <i>non-smooth</i> )<br>Maximum normal strain ( <i>non-smooth</i> ) |
| Orthotropic | Hill ( <i>smooth</i> )   | (Extended) Hill ( <i>smooth</i> )  |

### 3.2 Material selection: figures of merit

The choice of a particular material for BIDs, as much as for any other mechanical component, is driven by its performance against a large range of requirements. To general aspects, such as availability, manufacturing feasibility, costs, weight, delivery times, etc., one must add application-specific requirements, which, in the case of BIDs, typically include mechanical robustness, resistance to high temperatures, geometrical stability, cleaning efficiency, low contribution to RF impedance, and resistance to radiation.

To classify and rank potential materials against this large number of requirements, it is useful to introduce *figures of merit* which permit several material properties related to a specific requirement to be condensed into a single indicator: the higher the figure of merit, the better the material performance against that specific requirement.

A set of indices can be particularly helpful to orient material choice in the early phases of design; however, one must be aware of the fact that figures of merit rely on simplified, constant, linearized, temperature-independent material properties and on largely approximate extrapolations of certain factors, such as energy deposition. Hence, they should be used as indicative, comparative tools and not for quantitative assessment of material or component performance. Additionally, in the case of anisotropic materials, relevant properties are usually averaged over the three directions.

The most relevant figures of merit for the design of BIDs are related to:

- thermomechanical robustness;
- thermal stability;
- electrical conductivity;
- radiation resistance.

An index called the *thermomechanical robustness index* (TRI), is proposed to assess the material robustness against particle beam impacts. Given that thermal shock problems are, to a large extent, governed by the thermal deformation induced by a sudden temperature increase, it appears reasonable to base this index on the ratio between material *admissible strain* or *strain to failure*  $\varepsilon_{\text{adm}}$  and the actual strain

$$\text{TRI} = \frac{\varepsilon_{\text{adm}}}{\varepsilon_{\text{ref}}} \cdot \left( \frac{T_m}{\Delta T_q} - 1 \right)^m, \quad (30a)$$

where  $\varepsilon_{\text{ref}}$  and  $\Delta T_q$  are the strain and the temperature increase generated by a reference energy deposition given in Eqs. (30c) and (30d),  $T_m$  is the melting (or degradation) temperature and  $m$  is a coefficient related to the material loss of strength with temperature increase.

Note that TRI tends to zero when the melting point is reached.

Since *strain to failure* values are hardly available for many materials while effective *failure strength* values, usually related to fracture or yielding for brittle or ductile materials, respectively, are much easier to obtain in the literature, it is convenient to express  $\varepsilon_{\text{adm}}$  as a function of the failure strength  $R_M$ .

Incidentally, we observe that this assumption is conservative for non-linear materials, given that it involves a linear stress–strain relationship up to failure:

$$\varepsilon_{\text{adm}} = \frac{R_M}{E \cdot (1 - \nu)}. \quad (30b)$$

In Eq. (30b), we also implicitly assume that the stress distribution is the one encountered in a plane strain problem (see Eq. (7d)): this is in fact usually justified by the assumption that the beam

impact occurs relatively close to the component surface, so that thermal expansion is not constrained in the direction normal to the surface. If a deeper impact is expected, a coefficient equal to  $(1 - 2\nu)$  should be used instead.

The actual reference strain is expressed by

$$\varepsilon_{\text{ref}} = \bar{\alpha} \cdot \Delta T_q \quad (30c)$$

The temperature increase  $\Delta T_q$  can be assumed to be equal to a reference quasi-instantaneous *energy deposition*  $q_d$ , assumed to depend on the *geometric radiation length*  $X_g$  and the material density  $\rho$ , divided by the *specific heat*  $c_p$ :

$$\Delta T_q = \frac{q_d}{c_p} = \frac{C_R \rho^n}{c_p X_g} \quad (30d)$$

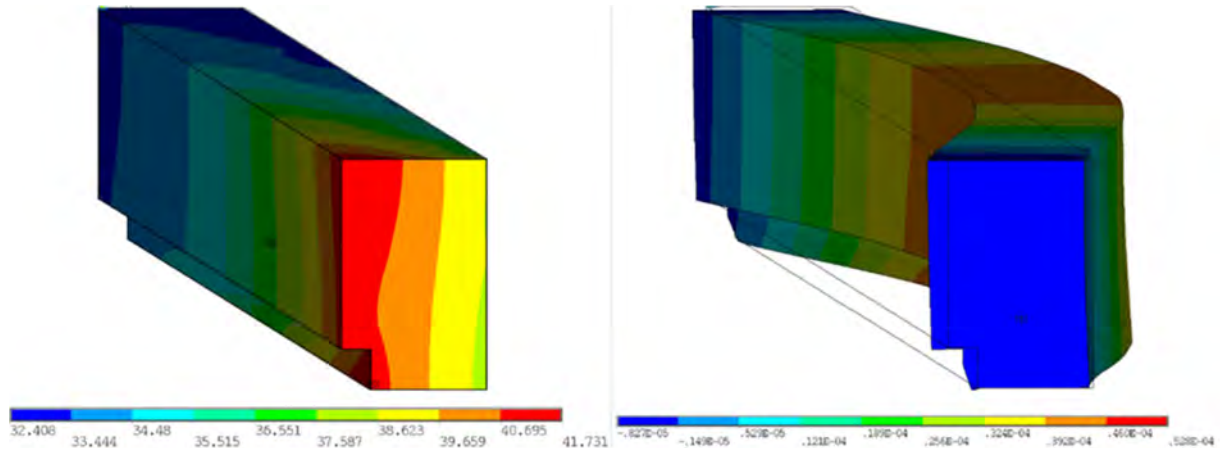
In these equations,  $\bar{E}$  is the (*averaged*) *Young modulus*,  $\nu$  the *Poisson ratio*,  $\bar{\alpha}$  the (*averaged*) CTE,  $C_R$  an arbitrary *scaling factor* and  $n$  a *coefficient* expressing the influence of density on the *energy distribution* generated by the impact.

Equation (30d) implies that the energy deposited by a given number of particles, and therefore the material temperature increase, is related to the material density and to the geometric radiation length; it has been empirically observed that the coefficient  $n$  for materials impacted by protons at several hundreds of GeV is  $\sim 0.2$ .

Combining Eqs. (30a–d), TRI can finally be written as

$$\text{TRI} = \frac{R_M c_p X_g}{\bar{E}(1-\nu)\bar{\alpha}C_R \rho^n} \cdot \left( \frac{T_m c_p X_g}{C_R \rho^n} - 1 \right)^m \quad (30e)$$

The *thermal stability index* (TSI) provides an indication of the ability of the material to maintain the geometrical stability of the component under steady-state particle losses. This is particularly important for components such as collimators and long absorbers, which are required to interact with the halo of the particle, maintaining their longitudinal straightness to a fraction of a beam transverse sigma (Fig. 43).



**Fig. 43:** Operating temperatures (in °C) (left) and thermally induced deflection (in m) (right) of a LHC secondary collimator jaw in steady-state conditions. The length of the jaw is 1 m, its deflection  $\approx 40 \mu\text{m}$ .

The TSI is proportional to the radius of curvature of an elongated structure induced by a non-uniform temperature distribution; since we are particularly interested in deformations induced by grazing beams, we can assume a steady-state energy deposition in which all the heat is flowing from the

surface exposed to the beam through the thickness of the BID. In this case, the radius of curvature of the component is given by

$$\rho_c = \frac{\bar{\lambda}}{\bar{\alpha}\dot{q}}, \quad (31a)$$

where  $\dot{q}$  is the *heat flux* in  $\text{W m}^{-2}$  and  $\bar{\lambda}$  is the (*averaged*) *thermal conductivity*. As with TRI, we can express the steady-state flux of (deposited) heat as

$$\dot{q} = \frac{C_s \rho^n}{X_g}, \quad (31b)$$

where  $C_s$  is a scaling factor.

The TSI is then given by

$$\text{TSI} = \frac{\bar{\lambda} X_g}{\bar{\alpha} C_s \rho^n}. \quad (31c)$$

Beam intercepting devices located in the accelerator ring, such as collimators and certain absorbers, are usually the machine components sitting closest to the circulating beam; therefore, their contribution to the accelerator global RF impedance is by far the highest. The part of the beam coupling impedance related to the resistive losses in the material surrounding the beam, the so-called *wall impedance*, is directly related to the material electrical resistivity. Therefore, maximizing the electrical conductivity of the materials mostly interacting with the beam can play a major role in minimizing the risk of impedance-induced beam instabilities.

A first approximation of the contribution to the total impedance at relatively high frequencies (above  $\sim 1$  MHz) by resistive objects is given by the so-called classic thick-wall regime [38]. In this regime, the transverse wall impedance of a cylindrical beam pipe is approximately given by

$$Z_t(\omega) = (1 + j) \frac{L Z_0 \mu_r}{2 \pi b^3} \sqrt{\frac{2}{\mu_0 \mu_r \gamma \omega}}, \quad (32a)$$

where  $\omega$  is the frequency,  $j$  is the imaginary unit,  $L$  is the length of the pipe,  $Z_0$  is the free-space impedance,  $b$  is the radius of the beam pipe,  $\mu_r$  and  $\mu_0$  are the relative and free-space permeability, respectively, and  $\gamma$  is the *electrical conductivity*.

We can hence define a *RF impedance index* (RFI), minimizing the material contribution to the system wall impedance, as

$$\text{RFI} = \sqrt{\frac{\gamma}{\mu_r}}. \quad (32b)$$

Irradiation of materials by energetic particles causes microstructural defects, which translate into a degradation of the thermophysical properties. Radiation resistance is defined as the ability of the material to maintain its properties under and after irradiation. An analysis of the effects induced on materials by ionizing radiation goes beyond the scope of these lectures. A review of radiation effects on materials can be found in several sources, such as Ref. [5].

### 3.3 Novel materials for beam intercepting devices

As seen, the introduction in recent years of new and extremely energetic particle accelerators, such as the LHC, brought about the need for advanced cleaning and protection systems, to safely increase the energy and intensity of particle beams to unprecedented levels. This has greatly increased the requirements for materials exposed to accidental impact from highly energetic and intense particle beam pulses; on top of outstanding thermal shock resistance, materials for halo cleaning and machine protection devices are typically required to maximize the figures of merit defined in Section 3.2, such as electrical conductivity, geometrical stability, and resistance to radiation damage. These requirements are set to become even more compelling in consideration of the High-Luminosity upgrade of the LHC (HL-LHC), expected to increase the beam intensity and its stored energy by a factor of two [39]: as an example, carbon-carbon (C-C) composites used for primary and secondary collimators in the LHC may limit the HL-LHC performance because of C-C low electrical conductivity leading to beam instability at high intensities [40], while the tungsten alloy (Inermet180) used in LHC tertiary collimators has very low robustness in case of beam impacts even at relatively low intensities.

In view of these challenges, an extensive R&D program has been launched at CERN in recent years to explore and develop a number of novel materials aiming to combine the excellent properties of graphite or diamond, specifically their low density, high thermal conductivity, and low thermal expansion, with those of metals or transition metal based ceramics, possessing high mechanical strength and good electrical conductivity. The most promising materials so far identified (early 2015) are molybdenum carbide – graphite (MoGr) and copper-diamond (CuCD).

#### 3.3.1 Molybdenum carbide – graphite

Molybdenum carbide – graphite is a novel composite developed in a collaboration between CERN and an Italian enterprise [41]: it is produced from molybdenum and graphite powders by high-temperature *fast direct hot pressing*, a pressure-assisted sintering technique in which heating is obtained by the passage of an electrical current through moulds and powders [42].

Pure molybdenum has a very high melting point and a low CTE, as well as excellent mechanical strength and electrical conductivity, while graphitic materials feature low density, extremely high service temperatures, large damping properties (particularly useful in attenuating shock waves) and, provided graphite crystallite ordering is sufficiently extended and a high graphitization degree is attained, excellent thermal conductivity and a very low CTE, at least in the direction aligned with the graphite basal plane. At high temperatures, molybdenum reacts rapidly with carbon, forming stable carbides ( $\text{MoC}_{1-x}$ ) which, in spite of their ceramic nature, retain a good electrical conductivity; in this respect, MoGr becomes a *ceramic-matrix composite*.

A broad range of compositions, powder types, and dimensions with processing temperatures ranging from 1700°C to 2600°C were developed: the best results so far were obtained for a sintering temperature of 2600°C.

The carbonaceous phase may be composed, in different grades, either of natural graphite flakes or of a mixture of natural graphite flakes and mesophase pitch-derived carbon fibres: these were selected to act as structural reinforcement and nucleation sites for enhanced graphitization, contributing to the improvement of thermal properties, thanks to their well-ordered graphitic structure (Fig. 44), and their mechanical strength.

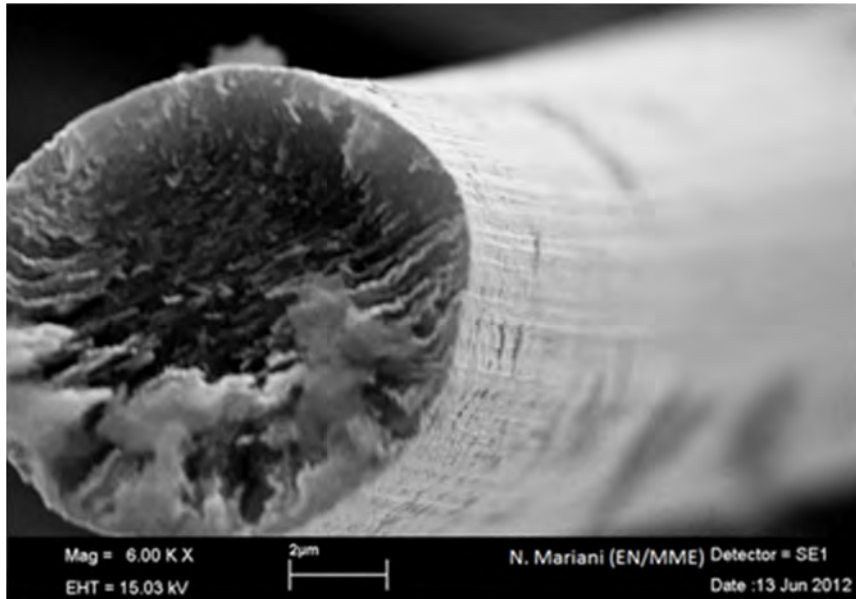


Fig. 44: Mesophase-pitch-derived carbon fibre (6000×)

High processing temperature grades are produced by *liquid-phase sintering* above the melting point of molybdenum carbide (2589°C). Scanning electron micrographs provide evidence of a very homogeneous microstructure with a regular distribution of small (5–10 µm) carbide particles and a high degree of graphitization of the carbonaceous phase (Fig. 45).

The high ordering and orientation of the graphitic phase is most likely catalysed by the presence of a carbide liquid phase at high temperatures; this lowers the required activation energy for the graphite arrangement and improves the diffusion rate of carbon atoms, with graphite crystallite growing through molten material as the graphitization process proceeds.

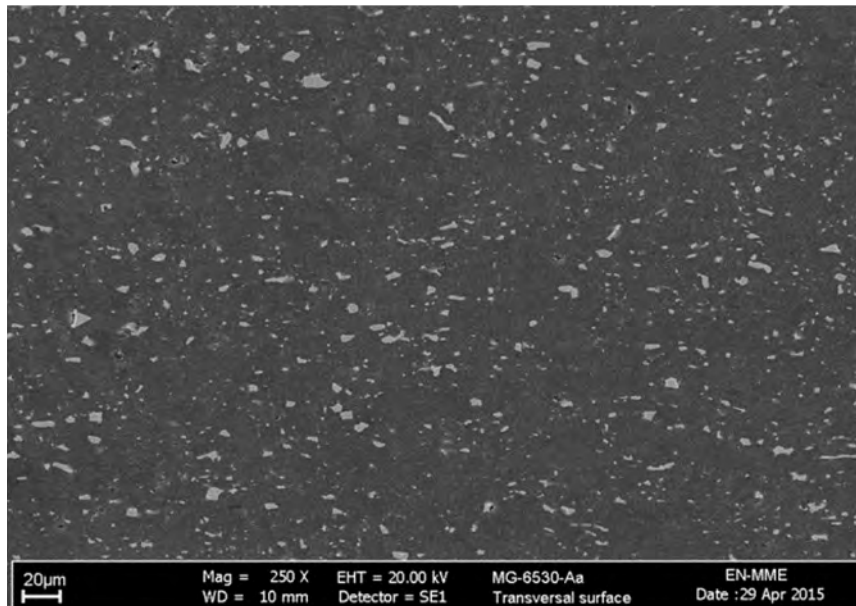
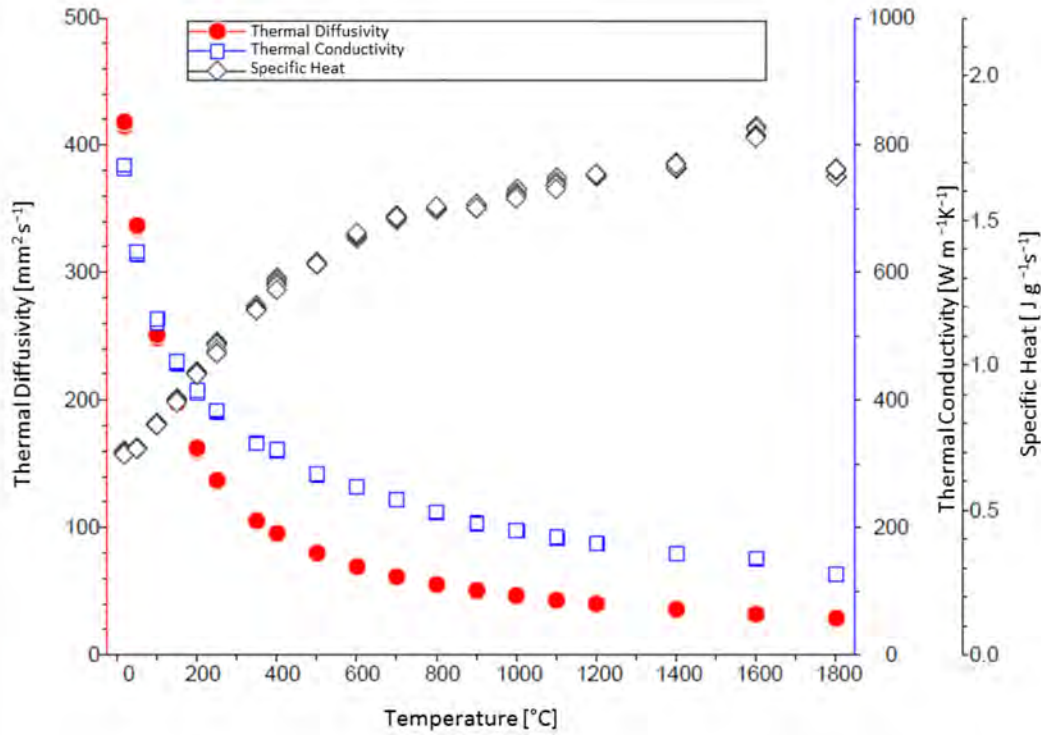


Fig. 45: Scanning electron micrograph of MoGr; note the finely dispersed carbide grains (250×)

To favour a liquid carbide infiltration and material compaction rate, a significant quantity of molten carbides is allowed to flow out of the mould during liquid-phase sintering, so that the final density of the material is reduced to  $\approx 2.5 \text{ g cm}^{-3}$ . Thanks to the extensive catalysed graphitization, MoGr has an electrical conductivity in the preferential direction (perpendicular to the pressing direction and

parallel to the basal planes of graphite crystallites) of  $\sim 1 \text{ MS m}^{-1}$ , one order of magnitude larger than that of isotropic graphite and of the C–C composites used for the LHC collimator. This property can be further increased by cladding or coating the external surface with pure molybdenum or other high electrical conductivity materials.

On top of its low density and good electrical conductivity, MoGr presents outstanding thermal properties, which are particularly useful in increasing the TSI and TRI: along the preferential direction, the thermal conductivity at RT is of the order of  $700 \text{ Wm}^{-1}\text{K}^{-1}$  (Fig. 46), almost twice that of pure copper and a factor of four greater than that of C–C, while the CTE is  $1.8 \times 10^{-6} \text{ K}^{-1}$  for temperatures spanning from RT to  $2000^\circ\text{C}$ .



**Fig. 46:** Thermal conductivity, diffusivity, and specific heat of MoGr between RT and  $1800^\circ\text{C}$

Relevant reference properties of MoGr are provided in Table 6: the subscript  $a$  indicates the preferential direction parallel to graphite basal planes, while the subscript  $c$  indicates the direction orthogonal to these planes.

**Table 6:** Selected properties of MoGr

|   |                                     |
|---|-------------------------------------|
| Density, $\rho$                                     | $2.5 \text{ g/cm}^3$                |
| CTE, $\alpha_a$ (RT to $1000^\circ\text{C}$ )       | $1.8 \times 10^{-6} \text{ K}^{-1}$ |
| CTE, $\alpha_c$ (RT to $1000^\circ\text{C}$ )       | $12 \times 10^{-6} \text{ K}^{-1}$  |
| Thermal conductivity, $\lambda_a$ (RT)              | $>700 \text{ Wm}^{-1}\text{K}^{-1}$ |
| Thermal conductivity, $\lambda_c$ (RT)              | $85 \text{ Wm}^{-1}\text{K}^{-1}$   |
| Electrical conductivity, $\gamma_a$ (RT) (uncoated) | $1 \text{ MSm}^{-1}$                |
| Electrical conductivity, $\gamma_c$ (RT)            | $0.3 \text{ MSm}^{-1}$              |
| Young modulus $E_a$ (flexural) (RT)                 | $53 \text{ GPa}$                    |
| Ultimate strength $R_m$ (flexural) (RT)             | $85 \text{ MPa}$                    |



3.3.2 Copper–diamond

Copper–diamond is produced by RHP Technology (Austria) by *solid-state sintering*; the initial volumetric composition is 60% diamond, 39% copper, and 1% boron [41, 42]. Copper is chosen for its excellent thermal and electrical conductivity, along with its good ductility, while diamond is added to reduce the density and the CTE, while contributing to the thermal conductivity.

A higher proportion of diamond would not allow a good material compaction, which is also achieved through the use of diamonds of various sizes, to optimize the filling of interstitials. Unlike MoGr, the main issue for material adhesion is the low chemical affinity between the two main elements, which leads to a lack of bonding between copper and diamond: this would jeopardize not only the material’s mechanical strength but also its thermal conductivity. Boron is added to offset such limitations, since this element promotes the formation of carbides at the diamond–copper interface, improving material internal bonding (Fig. 47).

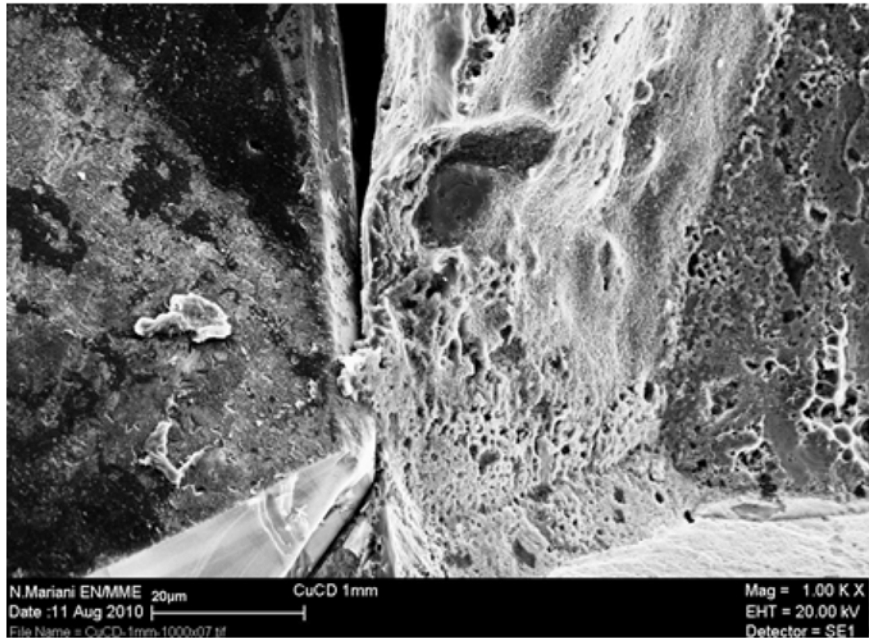


Fig. 47: High magnification scanning electron micrograph of the fracture surface of CuCD. Note the small boron carbide platelet, which is connecting the diamond grain to the detached copper matrix. 1000×.

Copper–diamond has very good thermal and electrical conductivity, and its CTE is reduced by a factor of 2–3, compared with pure copper (Table 7).

Table 7: Selected properties of CuCD

|   |   |
|---|---|
| Density, $\rho$                         | 5.4 g cm <sup>-3</sup>                  |
| CTE, $\alpha$ (RT to 900°C)             | 6–12 × 10 <sup>-6</sup> K <sup>-1</sup> |
| Thermal conductivity, $\lambda$ (RT)    | 490 W m <sup>-1</sup> K <sup>-1</sup>   |
| Electrical conductivity, $\gamma$ (RT)  | 12.6 MS m <sup>-1</sup>                 |
| Young modulus $E$ (flexural) (RT)       | 220 GPa                                 |
| Ultimate Strength $R_m$ (flexural) (RT) | 70 MPa                                  |

However, density and CTE are higher than MoGr, and the industrialization of the material is rather difficult: while thin samples of constant section can be produced via water-jet cutting, more complicated shapes with precise tolerances can only be produced by applying a pure copper cladding on the outer surfaces, which may limit material performance in the case of accidental grazing impacts (Fig. 48).



**Fig. 48:** CuCD block produced for a prototype jaw of HL-LHC secondary collimator [43]. Note the copper cladding on all functional surfaces.

### 3.4 Comparison of BID materials performances

A comparison of various materials of interest for BIDs, including the figures of merit defined in Section 3.2 is presented in Table 8. It can be seen from the table that no material perfectly meets all the requirements: material choice will therefore depend on which performance aspects must be favoured.

**Table 8:** Relevant properties and figures of merit for typical BID materials

|   | Beryllium | Carbon – carbon | Graphite | MoGr    | CuCD  | Glidcop | Molybdenum | Tungsten heavy alloy |
|---|-----------|-----------------|----------|---------|-------|---------|------------|----------------------|
| $\rho$ [g cm <sup>-3</sup> ]                        | 1.84      | 1.65            | 1.9      | 2.50    | 5.4   | 8.90    | 10.22      | 18                   |
| $Z$   | 4         | 6               | 6        | ≈6.5    | ≈11.4 | ≈29     | 42         | ≈70.8                |
| $X_g$ [cm]  | 35        | 26              | 19       | 17      | 4.8   | 1.4     | 0.96       | 0.35                 |
| $c_p$ [Jkg <sup>-1</sup> K <sup>-1</sup> ]          | 1925      | 780             | 760      | 750     | 420   | 391     | 251        | 150                  |
| $\bar{\alpha}$ [10 <sup>-6</sup> K <sup>-1</sup> ]  | 18.4      | 4.1             | 5.5      | 5.0     | 7.8   | 20.5    | 5.3        | 6.8                  |
| $\bar{\lambda}$ [Wm <sup>-1</sup> K <sup>-1</sup> ] | 216       | 167             | 70       | 547     | 490   | 365     | 138        | 90.5                 |
| $T_m$ [°C]  | 1273      | 3650            | 3650     | 2589    | ≈1083 | 1083    | 2623       | ≈1400                |
| $\bar{E}$ [GPa]                                     | 303       | 62.5            | 12       | 44      | 220   | 130     | 330        | 360                  |
| $R_M$ [MPa]   | 370       | 87              | 30       | 80      | 70    | 365     | 660        | 660                  |
| $\Delta Tq$ [K]                                     | 0.36      | 1.2             | 1.7      | 2.1     | 15.1  | 60.1    | 144        | 745                  |
| TRI   | 800       | 800–1200        | 800–1100 | 500–800 | 7     | 5       | 6.5        | 0.5                  |
| TSI   | 17        | 45              | 10       | 70      | 10    | 0.8     | 0.7        | 0.1                  |
| RFI   | 4.8       | 0.38            | 0.27     | 1       | 3.5   | 7.3     | 4.4        | 2.9                  |

Beryllium performs well in practically all aspects: unfortunately, extensive use is severely limited by its toxicity.

In general, carbon-based materials feature excellent TRI and TSI, thanks to their low atomic number and density, reduced CTE, high degradation temperature, and high thermal conductivity; however, they are penalized by low electrical conductivity when RF impedance is a critical requirement. In such a case, MoGr is the most promising compromise, particularly if coated with higher-conductivity thin films, which would further improve the electrical conductivity by a factor of 10.

The poor performance of tungsten heavy alloys as to thermomechanical robustness should be noted: this is due to a combination of factors, including, in particular, the low melting temperature of the nickel–copper matrix that is used to bind the tungsten particles and to increase material ductility.

## 4 Experimental testing and validation

### 4.1 Introduction

Advanced numerical simulation codes are powerful tools that enable the analysis of extremely complex dynamic phenomena; however, to provide reliable results, they require sufficiently accurate constitutive models for all the conditions that materials might undergo during such events.

Unfortunately, constitutive material models, in particular, at the extreme conditions generated by high-energy beam impacts, are far from being readily available and experimentally validated; many constitutive models for existing materials were obtained through military R&D and are therefore classified. The situation is even more delicate for non-conventional alloys, compounds and composite materials presently used or likely to be used in state-of-the-art BIDs for very high-energy particle accelerators, for which experimental studies have never been carried out.

Moreover, numerical simulations cannot easily predict additional, far-reaching, consequences of beam accidents on nearby equipment, ultra-high vacuum performance, electronics, etc.

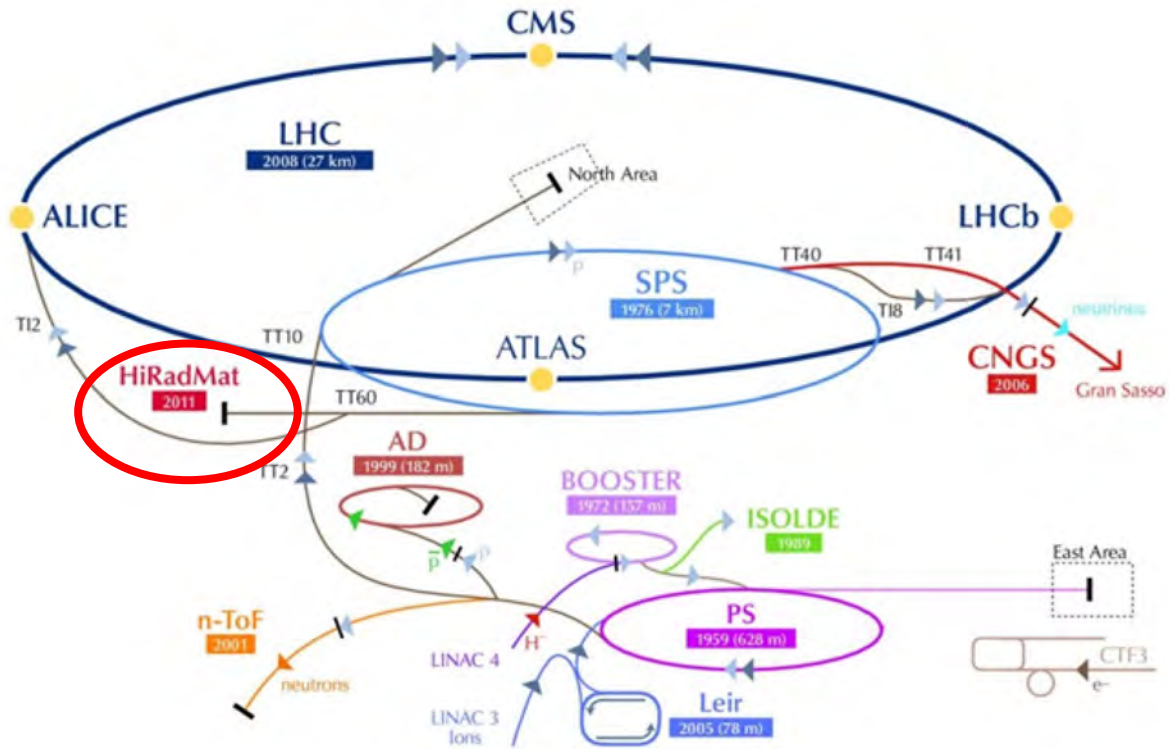
This is why only ad-hoc material tests can provide the correct inputs for numerical analyses, enabling the benchmarking and validation of simulation results on simple specimens as well as on full-scale, complex structures.

### 4.2 HiRadMat facility at CERN

A dedicated facility has been designed and commissioned at CERN to test materials and systems under high-intensity pulsed particle beams: HiRadMat (High Radiation to Materials) (Fig. 49) [12].

Given the high destructive power reached by the LHC, it has been decided that any new beam intercepting device must be tested, prior to its installation, for sufficient robustness to as realistic as possible conditions for future operation, to at least ensure that possible and unavoidable damage can be locally constrained, in order to prevent catastrophe (e.g. causing damage to nearby components, water leaks into vacuum from the cooling system, spreading of sputtered materials, or vacuum quality deterioration over long distances).

Previous tests of robustness and damage effects on BIDs and materials were performed in ad-hoc installations in the TT40 transfer beam line to LHC and CNGS in 2004 and 2006. The difficulty in performing such important tests on temporary installations and the potential impact on operating transfer lines were the main motivations for building HiRadMat, which was purposely designed to study beam shock impacts on materials and accelerator components.



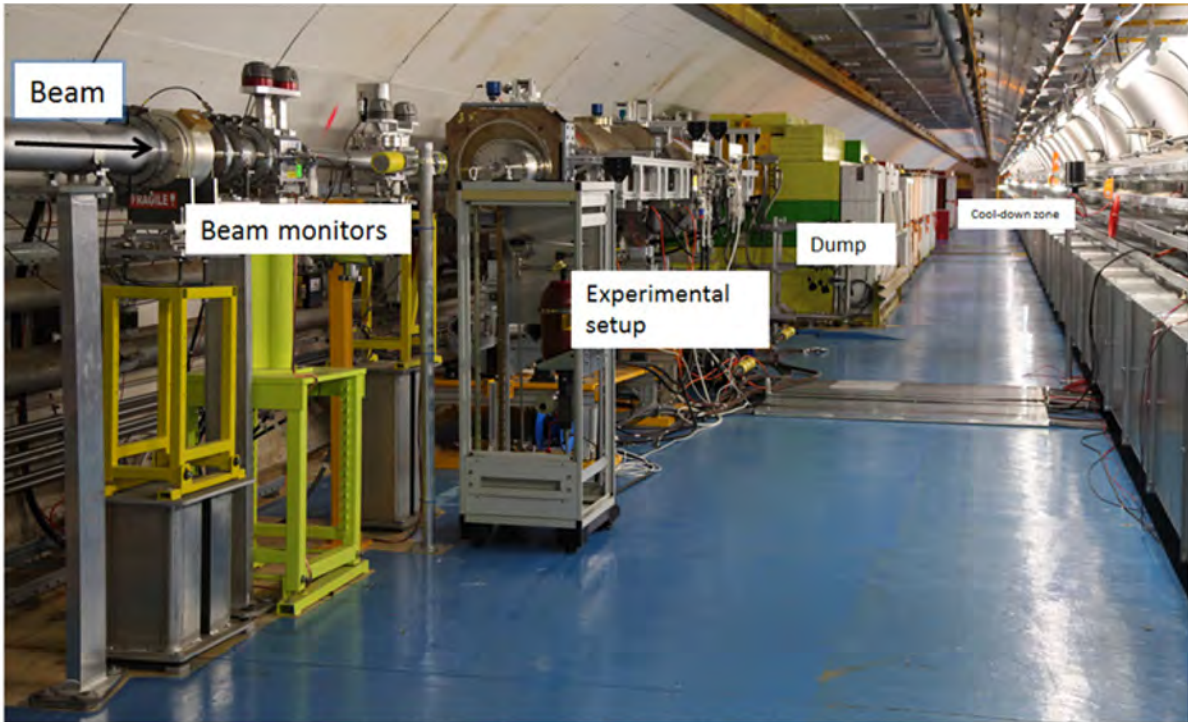
**Fig. 49:** Location of HiRadMat in CERN accelerator complex

HiRadMat uses an extracted primary proton or ion beam from the LHC SPS. The main beam parameters are listed in Table 9. The beam spot size at the focal point of the experiment can be varied from 0.5 to 2 mm<sup>2</sup>, which, together with the variable beam intensity, offers sufficient flexibility to test materials at different deposited energy densities (Fig. 50).

**Table 9:** Key parameters for the HiRadMat beam

|                       | <b>Protons</b>                      | <b>Ions (Pb<sup>82+</sup>)</b> |
|-----------------------|-------------------------------------|--------------------------------|
| Energy                | 440 GeV                             | 173.5 GeV u <sup>-1</sup>      |
| Bunch intensity (max) | $1.7 \times 10^{11}$ p <sup>+</sup> | $7 \times 10^9$ ions           |
| No of bunches         | 1 to 288                            | 52                             |
| Pulse intensity (max) | $4.9 \times 10^{13}$ p <sup>+</sup> | $3.6 \times 10^9$ ions         |
| Pulse energy (max)    | 3.4 MJ                              | 21 kJ                          |
| Bunch length          | 11.24 cm                            | 11.24 cm                       |
| Bunch spacing         | 25, 50, 75, 150 ns                  | 100 ns                         |
| Pulse length          | 7.2 $\mu$ s                         | 5.2 $\mu$ s                    |

Beyond the needs of CERN, HiRadMat is open to other users and is also included in the EUCARD FP7 European Project as transnational access to facilitate its use by European teams.



**Fig. 50:** HiRadMat facility experimental set-up

HiRadMat is not an irradiation facility, where large doses on equipment can be accumulated. It is rather a test area, designed to perform single experiments to evaluate the effect of high-intensity pulsed beams on materials or accelerator component assemblies in a controlled environment. The facility is designed for a maximum of  $10^{16}$  protons per year, distributed among 10 experiments, each having a total of  $10^{15}$  protons or about 100 high-intensity pulses. This limit allows reasonable cool-down times for the irradiated objects (from a few months to a year) before they can be analysed in specialized facilities.

Nine experiments were performed in the facility in 2012 before LHC Long Shutdown 1; three of them are described next.

### 4.3 HiRadMat 12 experiment

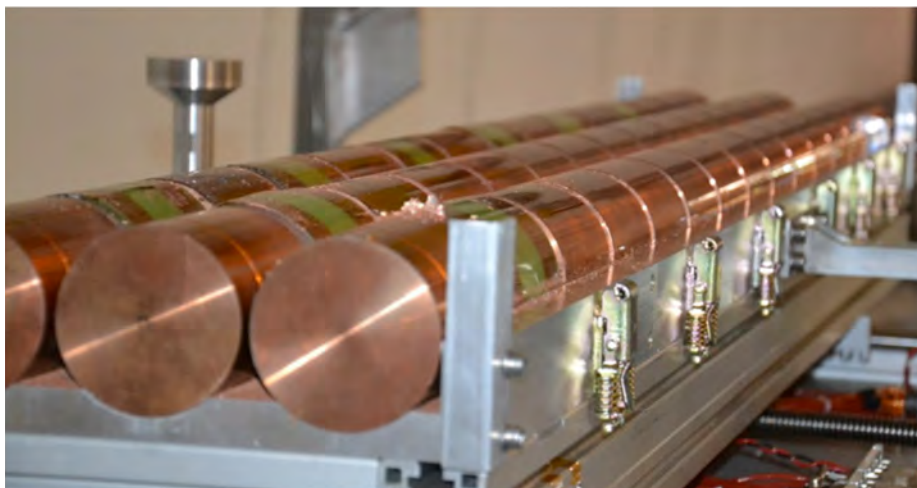
In normal operating modes, each of the 362 MJ LHC beams is safely extracted into a 700 m long transfer line, its energy density diluted and finally absorbed in large graphite blocks. However, several failure modes could abnormally deflect the beam into a graphite absorber, septum magnets, or superconducting magnets. It is therefore extremely important to predict the consequences of such events.

Extensive simulation studies of the full impact of the ultra-relativistic proton beam generated by the LHC on solid targets of different materials were carried out: they predicted that the energy deposited in the target by the protons in the first 10 bunches and their hadronic shower would lead to the phenomenon of hydrodynamic tunnelling described in Section 2.4.4. The strongly heated material would undergo phase transitions that include liquefaction, evaporation, and even conversion into weakly ionized strongly coupled plasma. The high temperature in the absorption zone would induce high pressures in the core, generating a radially outgoing shock wave, which causes substantial density depletion at the axis. As a consequence, the protons that are delivered in subsequent bunches would penetrate the target much more deeply. For example, the range of a hadronic shower of 7 TeV protons in solid carbon, which is about 3 m, would be extended to around 25 m for the full LHC beam with 2808 bunches because of hydrodynamic tunnelling [44].

This phenomenon therefore has very important implications for machine protection system design. To check the validity of these theoretical considerations, especially the existence of the

hydrodynamic tunnelling, a dedicated experiment was performed at the HiRadMat facility using the SPS proton beam [45].

The set-up consisted of three targets of 15 copper cylinders each, spaced by 1 cm to allow visual inspection after irradiation. Each cylinder had a radius of 4 cm and length of 10 cm. Figure 51 shows the targets before the installation. The targets were mounted on a movable table that could be moved to four different positions: target 1, target 2, target 3, and an off-beam position. The set-up was equipped with pCVD diamond particle detectors, PT100 temperature sensors, strain gauges and secondary electron emission particle detectors to obtain additional information during the beam interaction.



**Fig. 51:** HRMT-12 set-up; each individual target is made of 15 copper cylinders

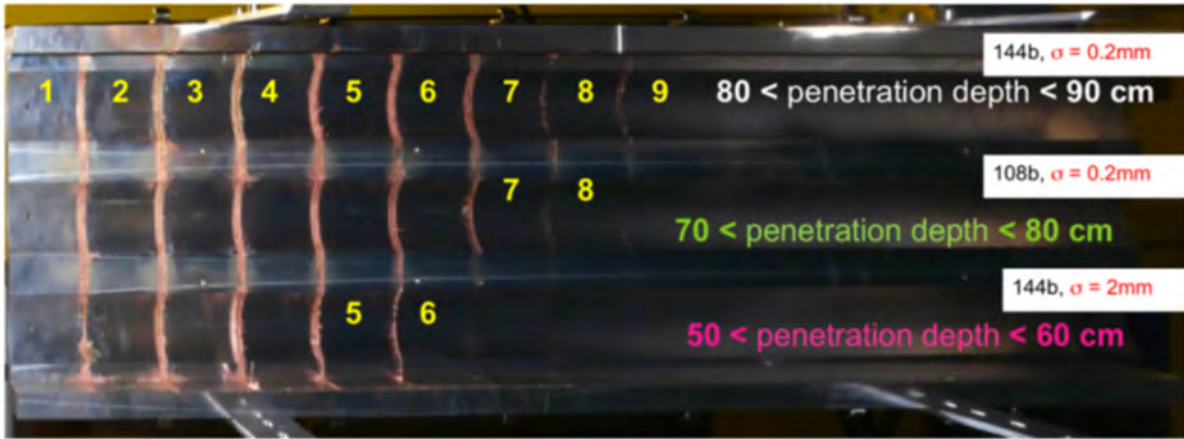
For all the experiments, the proton energy was 440 GeV, bunch intensity  $1.5 \times 10^{11}$  protons, bunch length 0.5 ns, and bunch separation 50 ns. Target 1 was irradiated with 144 bunches with a beam sigma of 2 mm. Target 2 was irradiated with 108 bunches, whereas target 3 was irradiated with 144 bunches; in both these cases, the beam had a much smaller focal spot size, characterized by  $\sigma = 0.2$  mm. The beam parameters used in these three experiments are summarized in Table 10.

The target was opened for visual inspection after 8 months of cool-down. Droplets and splashes of molten and evaporated copper were found on the copper cylinders, on the aluminium housing at the position of the gaps between cylinders, and in the front aluminium caps.

**Table 10:** Experimental beam parameters used in the three experiments

| Target | Number of bunches | Beam $\sigma$ [mm] | Beam energy [MJ] | Expectation            |
|--------|-------------------|--------------------|------------------|------------------------|
| 1      | 144               | 2.0                | 1.52             | Some tunnelling        |
| 2      | 108               | 0.2                | 1.14             | Moderate tunnelling    |
| 3      | 144               | 0.2                | 1.52             | Significant tunnelling |

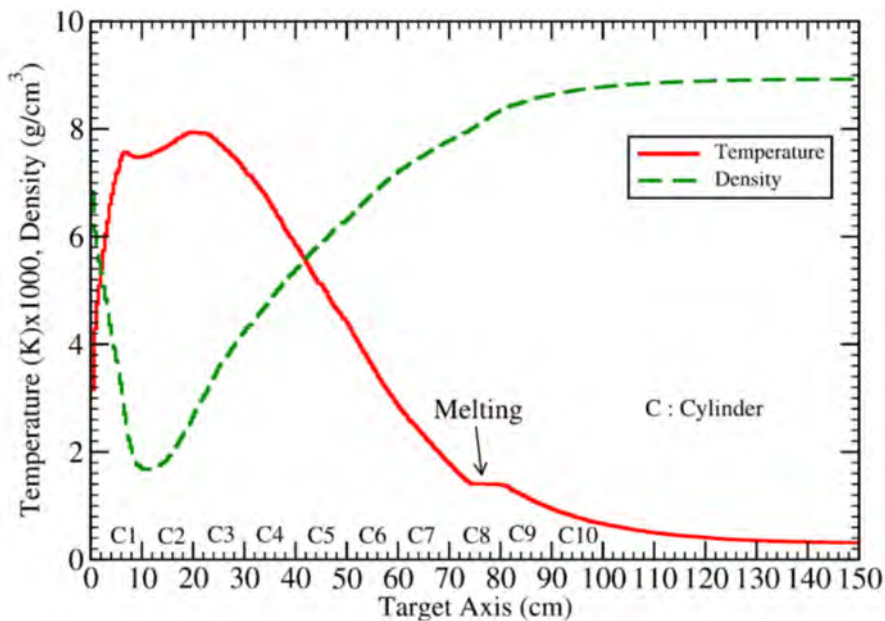
Figure 52 shows the aluminium cover that was placed on top of the target assembly. After the beam impact, molten or evaporated material is projected outwards and deposited on the cover. The traces of the projected copper between the 10 cm long cylinders are clearly visible. It can be seen that in the experiment using 144 bunches and a beam focal spot of  $\sigma = 2.0$  mm (bottom picture), the splash of molten copper occurs up to the gap between the fifth and sixth cylinders. This means that the material melted or evaporated over a length of  $55 \pm 5$  cm. In the second experiment, with 108 bunches and a beam focal spot of  $\sigma = 0.2$  mm (middle picture), the melting and evaporation zone extends to the eighth cylinder, indicating a damage length of  $75 \pm 5$  cm. In the experiment with 144 bunches and a beam focal spot of  $\sigma = 0.2$  mm (top picture), the melting and evaporation zone extends to the ninth cylinder, a length of  $85 \pm 5$  cm.



**Fig. 52:** Top cover of the experimental set-up after the irradiation. Traces of projected copper between the 10 cm long cylinders of the targets indicate the length of the melting and evaporation zone [45].

Detailed numerical simulations were carried out running the energy deposition code FLUKA and the two-dimensional hydrodynamic code BIG2 iteratively, using a time step of 700 ns, corresponding to the time during which the target density changes by about 15% at the target centre because of hydrodynamic processes. A semi-empirical multi-phase EOS was chosen to model different phases of the copper target during and after the irradiation.

In Fig. 53, density and temperature are plotted along the axis at  $t = 5800$  ns, for the 108 bunches of case 2. It can be seen that the flat part of the temperature curve that represents the melting region lies within  $L = 75$  and  $80$  cm, which is equivalent to the second half of the eighth cylinder. The temperature curve also shows that the material along the axis up to 75 cm is liquefied or even evaporated, depending on the temperature. The liquefied material escapes from the left face of cylinder number 8 and collides with the molten or gaseous material ejected from the right face of cylinder number 7. As a result of this collision, the material is splashed vertically and is deposited at the inner surface of the target cover above the gap between cylinders 7 and 8. The simulations are therefore in full agreement with experimental observations. Similar conclusions can be drawn for cases 1 and 3.



**Fig. 53:** Temperature and density on axis of target 2 after 5800 ns (case 2, 108 bunches)

Figure 54 the physical state of the target up to a radius of 0.5 cm. It can be seen that different parts of the beam-heated region lie in different phases of high energy density matter. These include gas, two-phase liquid–gas, liquid, and melting states. This suggests that the HiRadMat facility is very much suited to the important research area of high energy density physics.

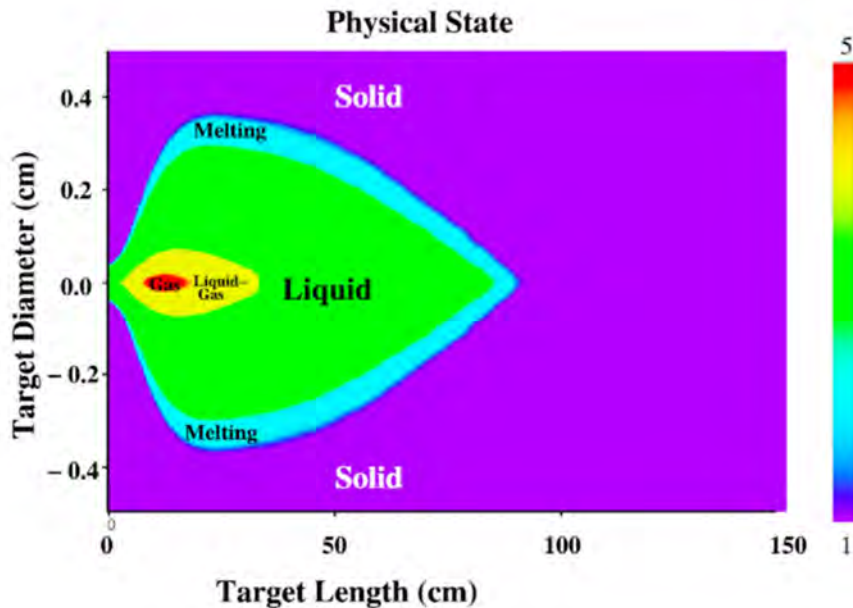


Fig. 54: Physical state of target 3 material after 7850 ns (case 3, 144 bunches) [45]

#### 4.4 HiRadMat 09 experiment

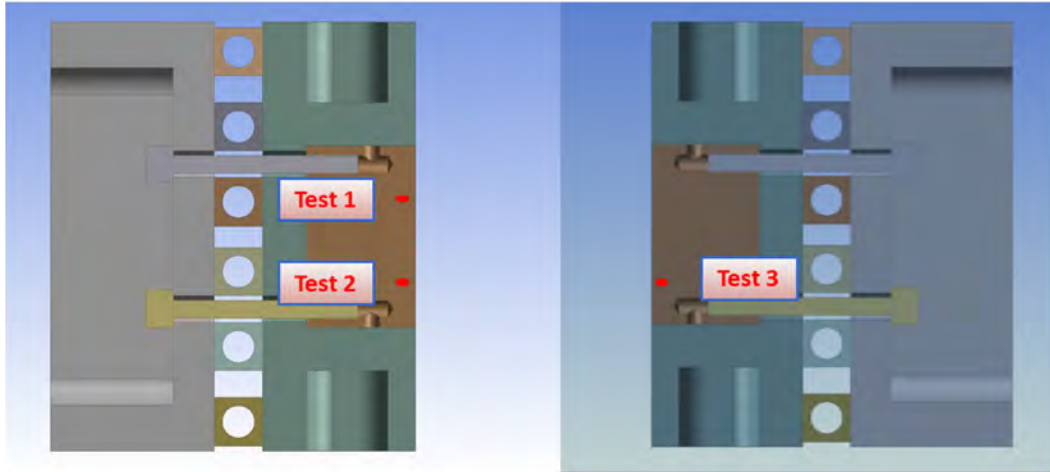
As anticipated in Section 2.4.3, a thorough numerical analysis of a tertiary collimator of the LHC was completed to simulate the effects of several asynchronous beam abort cases with different values of beam emittance, energy, and intensity. This computation relied on advanced simulations performed with the wave propagation code Autodyn, applied to a multicomponent three-dimensional model [33].

The most important issue of these simulations concerned the reliability of constitutive models of relevant materials, especially at extreme conditions of temperature, pressure, and energy induced by the beam impact. To probe and evaluate such models, two experiments were performed in the HiRadMat facility in 2012. The first experiment, known as HRMT09, entailed the destructive test of a complete tertiary collimator, to assess not only the mechanical damage provoked to the structure but also other consequences of the beam accident, such as degradation of vacuum pressure in the beam line, contamination of the inner walls of the vacuum vessel, and impacts on collimator dismounting procedure.

The aim of the experiment was to verify the robustness and performance integrity of a fully assembled tertiary collimator under direct beam impact [46]. Three different tests were performed, with different beam intensities and different goals (Fig. 55).

- Test 1: investigate the effects of asynchronous beam dump under an impact equivalent to 1 LHC bunch at 7 TeV [47].
- Test 2: identify the onset of plastic damage on blocks made of tungsten heavy alloy.
- Test 3: reproduce a destructive scenario, inducing severe damage on the collimator jaw (damage on the collimator equivalent to four bunches at 5 TeV).





**Fig. 55:** Schematic diagram of three tests performed on tertiary collimator during HRMT09 experiment

For each test, the intensity and emittance of the SPS pulse was calculated so that the mechanical damage on the jaw would be equivalent to the one induced by the LHC (Table 11). For example, a SPS pulse with  $3.36 \times 10^{12}$  protons is expected to produce a mechanical damage on the jaw equivalent to one LHC nominal bunch at 7 TeV.

A visual inspection performed a few months after the irradiation revealed the effects anticipated in Fig. 5. The damage provoked by tests 1 and 3 is clearly visible; an impressive quantity of tungsten alloy was ejected, partly stuck on the opposite jaw, partly fallen on the tank bottom or towards entrance and exit flanges.

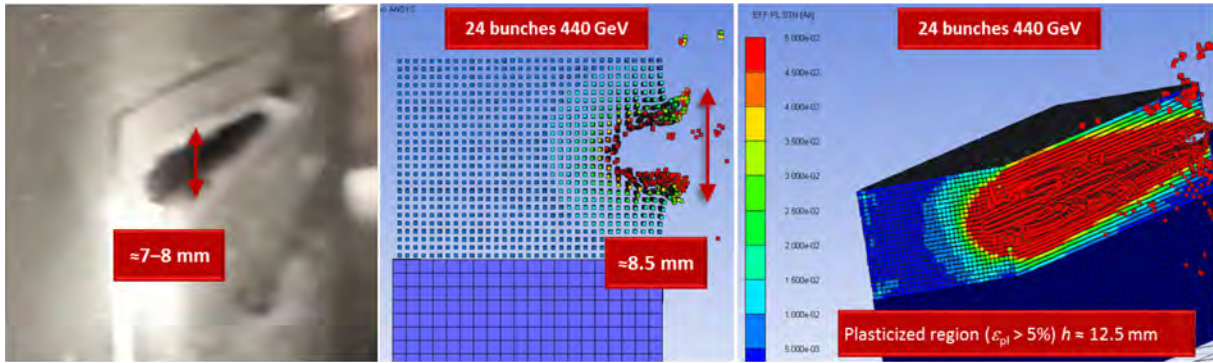
The observation also highlighted other possible issues.

- Contamination of bellows, tank, and vacuum chambers, owing to activated tungsten fragments; procedures for maintenance, intervention, and replacement must take this into account.
- Ejected particles may affect the correct functionality of movable parts (RF fingers sliding on upper and lower rails).
- Degradation of ultra-high vacuum along the beam line.

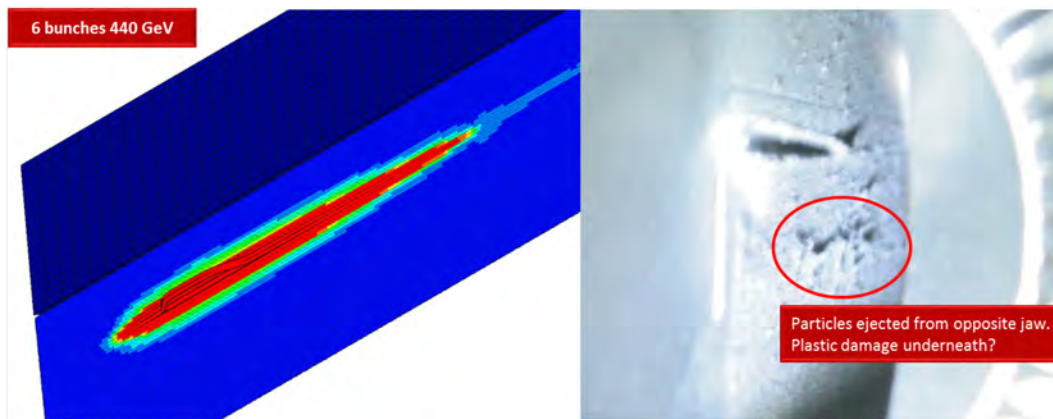
**Table 11:** Beam parameters and impact positions of tests performed during HRMT09

|  | Test 1                      | Test 2                      | Test 3                      |
|--|-----------------------------|-----------------------------|-----------------------------|
| Beam energy                              | 440 GeV                     | 440 GeV                     | 440 GeV                     |
| Pulse intensity                          | $3.36 \times 10^{12}$ p     | $1.04 \times 10^{12}$ p     | $9.34 \times 10^{12}$ p     |
| No of bunches                            | 24                          | 6                           | 72                          |
| Bunch spacing                            | 50 ns                       | 50 ns                       | 50 ns                       |
| Beam size [ $\sigma_x \times \sigma_y$ ] | 0.53 mm $\times$<br>0.36 mm | 0.53 mm $\times$<br>0.36 mm | 0.53 mm $\times$<br>0.36 mm |
| Impact location                          | Left jaw, +10 mm            | Left jaw,<br>-8.3 mm        | Right jaw,<br>-8.3 mm       |
| Impact depth                             | 2 mm                        | 2 mm                        | 2 mm                        |
| Jaw half-gap                             | 14 mm                       | 14 mm                       | 14 mm                       |

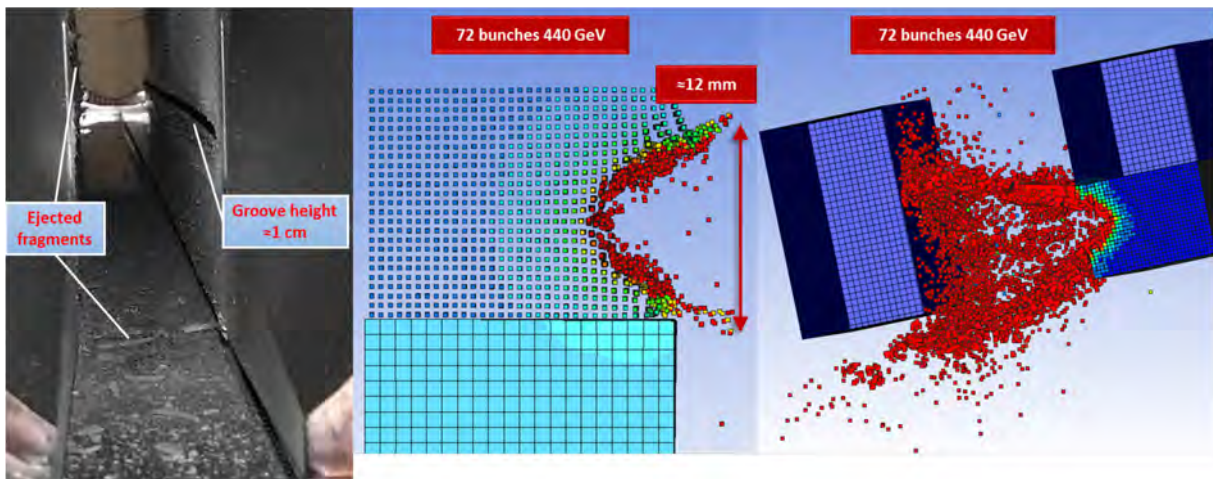
A qualitative comparison of visible damaged areas with Autodyn simulations is provided in Figs. 56 to 58.



**Fig. 56:** Comparison between actual damage and numerical simulation for test 1 beam impact



**Fig. 57:** Comparison between numerical analysis and actual damage for test 2 beam impact. Simulations predict a rather extended plasticized region, but only a tiny groove that might have been covered by the particles ejected from the opposite jaw during test 3.



**Fig. 58:** Comparison between actual damage and numerical simulation for test 3 beam impact. Note the spray of ejected particles, reaching velocities close to  $1 \text{ km s}^{-1}$ .

Simulations of test 1 and test 3 show very good agreement with visual inspections, while it is impossible to visualize the plastic deformation produced by test 2. The zone is, in fact, covered with particles ejected from the opposite jaw during test 3, which reached a velocity of about  $1 \text{ km s}^{-1}$  according to simulations; however, no signs of a significant groove are visible, in line with simulations.

#### 4.5 HiRadMat 14 experiment

The main goal of the HRMT14 experiment was to derive new material constitutive models collecting, mostly in real time, experimental data from different acquisition devices: strain gauges, a laser Doppler vibrometer, a high-speed video camera, and temperature and vacuum probes [48].

The material sample holder constituted a vacuum vessel and a specimen housing featuring 12 material sample tiers arranged in two arrays of six (Fig. 59).

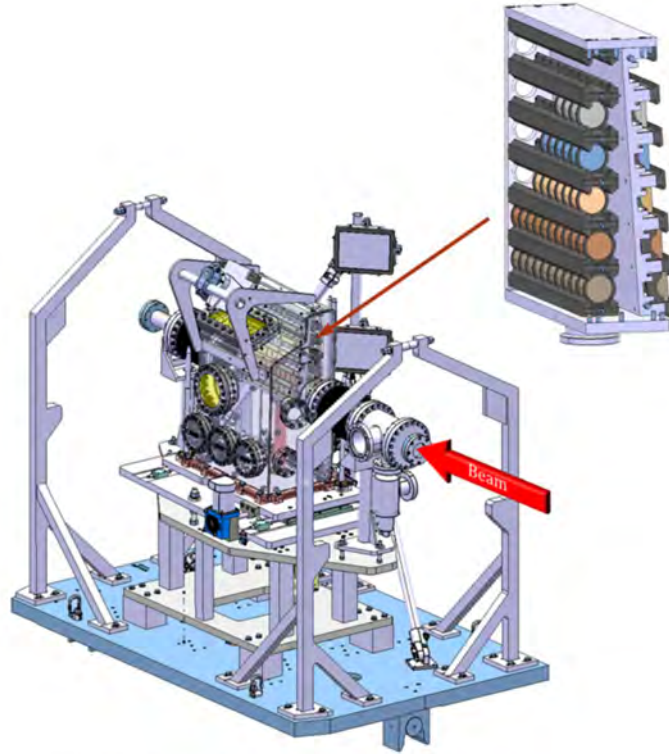
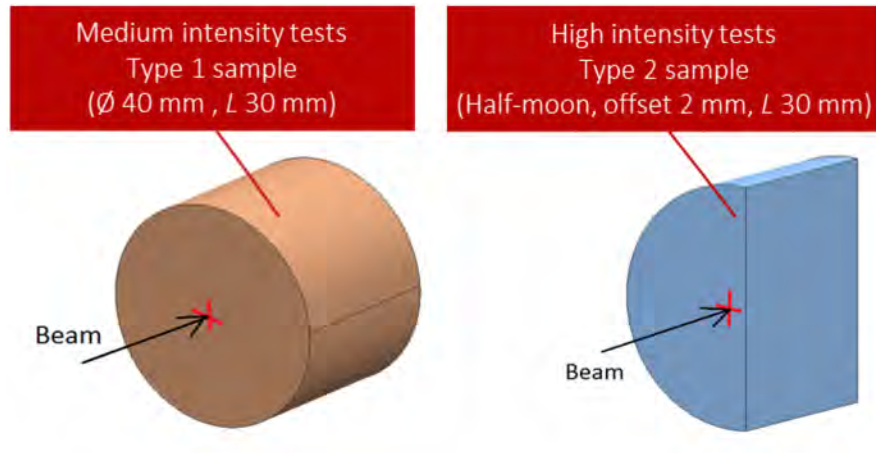


Fig. 59: General assembly of the HRMT-14 test-bench

Each tier hosted specimens made of materials currently used for collimators, such as tungsten heavy alloy (Inermet 180), Glidcop® AL-15 LOX (dispersion-strengthened copper), and molybdenum, as well as novel materials under development, i.e. molybdenum–copper–diamond (MoCuCD), copper–diamond (CuCD), and molybdenum carbide–graphite (MoGr) composites.

Two different specimen shapes were chosen for each tested material: cylindrical discs (type 1) for medium-intensity tests, to measure axially symmetrical shock waves, and semicircular prisms (type 2) for high-intensity tests, to allow extreme surface phenomena (melting, material explosion, debris projections, etc.) to be visualized and imaged (Fig. 60).

Part of the instrumentation was installed directly on the specimens; resistive strain gauges measured the strain produced on samples by shock wave propagation, to benchmark time-dependent simulations (Fig. 61). Temperature sensors, vacuum pressure gauges, and microphones were also installed inside or in the vicinity of the tank. Optical devices (a laser Doppler vibrometer and a high-speed camera) were installed remotely in a concrete bunker, to protect them from the effects of radiation. The laser Doppler vibrometer measured the radial velocity on the outer surface of one cylindrical sample per tier. The high-speed camera filmed the particle projection produced by high-energy impacts on type 2 specimens; the lighting necessary for the acquisition was provided by a battery of radiation-hard xenon flashes mounted atop the tank.



**Fig. 60:** Material specimen shapes for medium-intensity (type 1, left) and high-intensity (type 2, right) tests

Table 12 and Table 13 report the characteristic values of the most intense pulses shot on medium-intensity (type 1 specimens) and high-intensity (type 2 specimens) tiers respectively.

**Table 12:** Beam parameters for most intense pulses shot on each material on medium-intensity tiers

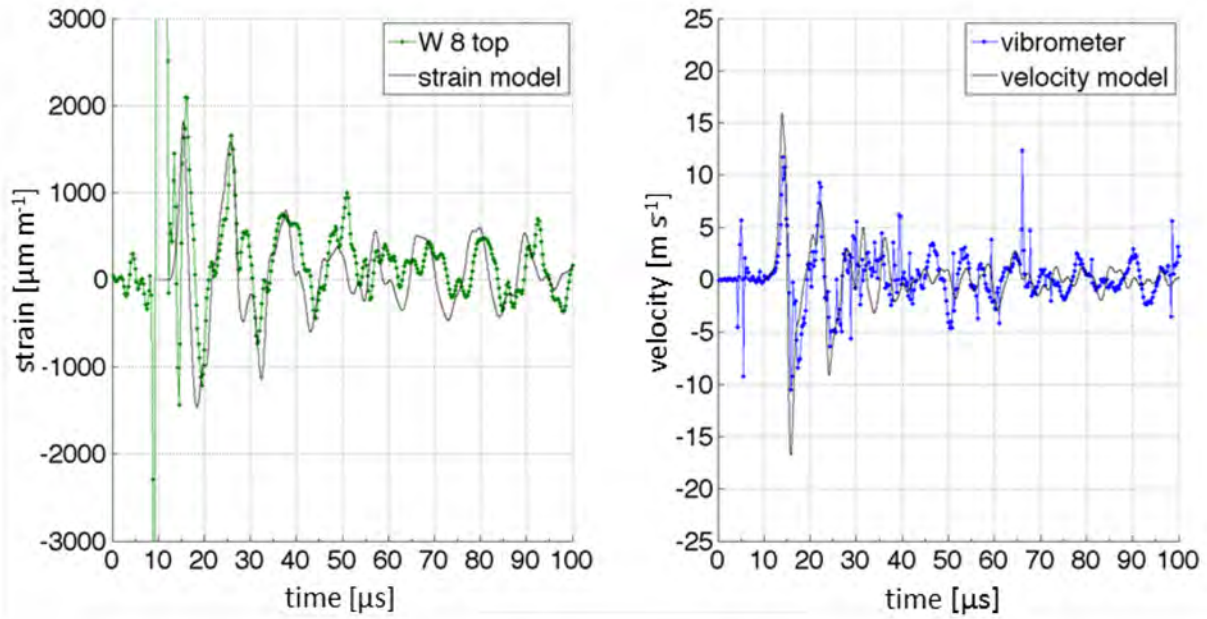
| Type 1 Specimens | Bunches (maximum) per pulse | Delivered protons     | Beam size ( $\sigma_x \times \sigma_y$ ) [mm $\times$ mm] | Pulse energy [MJ] |
|------------------|-----------------------------|-----------------------|---|-------------------|
| Inermet 180      | 24                          | $2.70 \times 10^{12}$ | $1.4 \times 2$  | 0.19              |
| Molybdenum       | 72                          | $4.75 \times 10^{12}$ | $1.35 \times 1.25$  | 0.33              |
| Glidcop          | 72                          | $4.66 \times 10^{12}$ | $1.35 \times 1.25$  | 0.33              |
| MoCuCD           | 72                          | $7.62 \times 10^{12}$ | $1.8 \times 1.8$  | 0.54              |
| CuCD             | 72                          | $7.57 \times 10^{12}$ | $1.8 \times 1.8$  | 0.53              |
| MoGr             | 72                          | $7.82 \times 10^{12}$ | $1.8 \times 1.8$  | 0.55              |

**Table 13:** Beam parameters for most intense pulses shot on each material on high-intensity tiers

| Type 2 Specimens | Bunches (maximum) per pulse | Delivered protons     | Beam size ( $\sigma_x \times \sigma_y$ ) [mm $\times$ mm] | Pulse energy [MJ] |
|------------------|-----------------------------|-----------------------|---|-------------------|
| Inermet 180      | 72                          | $9.05 \times 10^{12}$ | $2 \times 2$  | 0.64              |
| Molybdenum       | 144                         | $1.95 \times 10^{13}$ | $2 \times 2$  | 1.37              |
| Glidcop          | 72                          | $9.03 \times 10^{12}$ | $1.9 \times 1.9$  | 0.64              |
| MoCuCD           | 144                         | $1.96 \times 10^{13}$ | $2 \times 2$  | 1.38              |
| CuCD             | 144                         | $1.95 \times 10^{13}$ | $2 \times 2$  | 1.37              |
| MoGr             | 144                         | $1.95 \times 10^{13}$ | $2 \times 2$  | 1.37              |

Strain gauges measured axial and hoop strains on the external surface of type 1 samples, while a laser Doppler vibrometer measured the radial velocity. Experimental data were then compared with the results of numerical simulations (Fig. 61).

A strong electromagnetic disturbance, induced by the particle beam, perturbed the strain gauge measurements during the first few microseconds after the impact, covering the first deformation peak. However, this effect rapidly disappeared, enabling the remainder of the phenomenon to be recorded. Measured and simulated signals are in good accordance, especially during the first reflections of the shock wave.



**Fig. 61:** Comparison between measurements (dotted lines) and simulations (continuous lines) of the impact of  $2.7 \times 10^{12}$  p ( $\sigma \approx 1.4$  mm) on Inermet type 1 sample (slot no 8) at  $r = 20$  mm,  $L = 15$  mm; axial strain (left) and radial velocity (right).

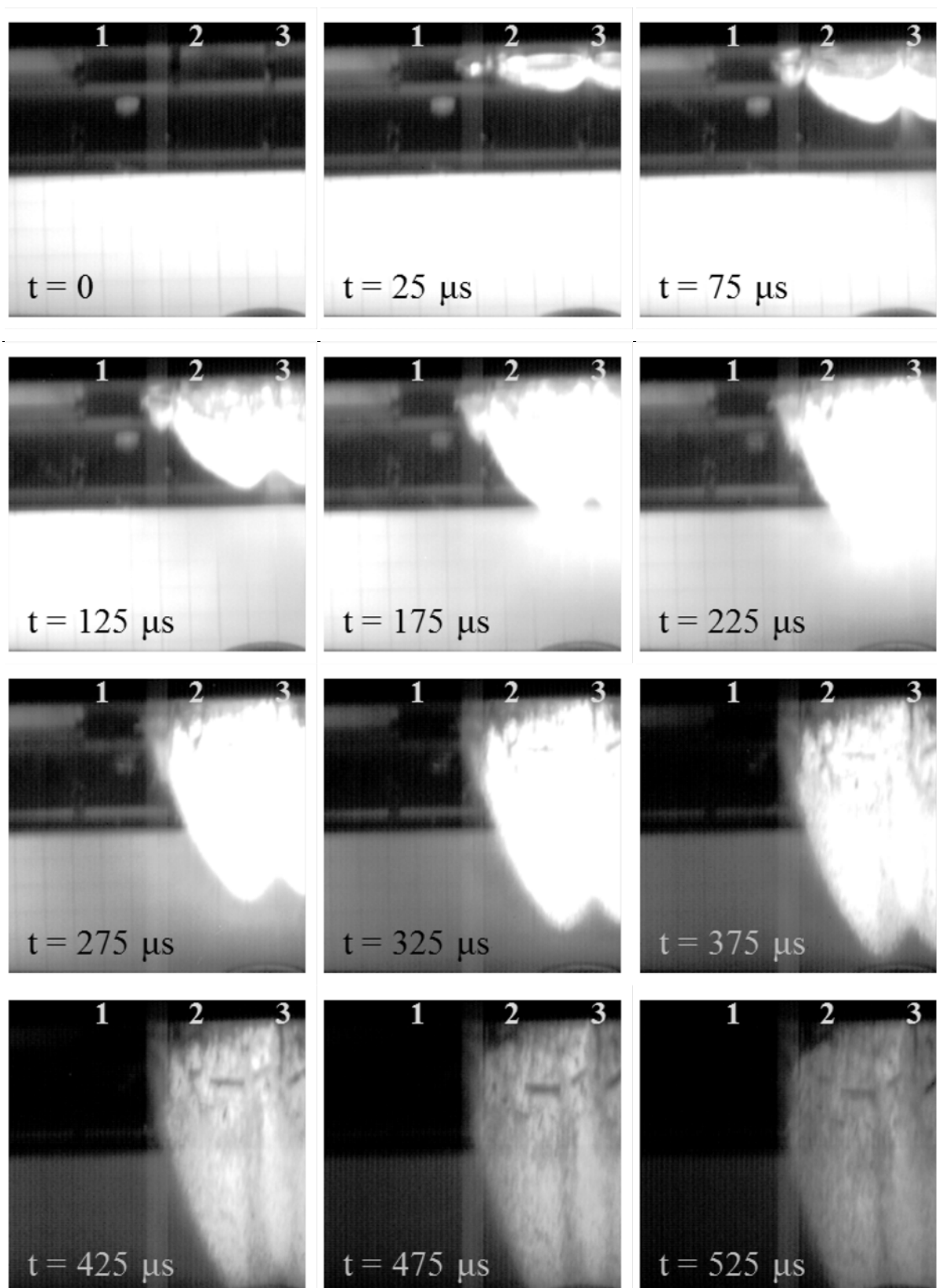
The high-speed camera and flash systems enabled images of the impact of a hadron beam on solid targets and of the effects induced to be recorded for the first time. The most remarkable phenomena occurred during beam impact on type 2 specimens made of Inermet, the material with the highest stopping power.

As shown in Fig. 62, a large quantity of hot material was ejected at high velocity from the two most loaded Inermet 180 specimens; the high temperatures reached are confirmed by the intense light emitted by the fragments over a few hundred microseconds.

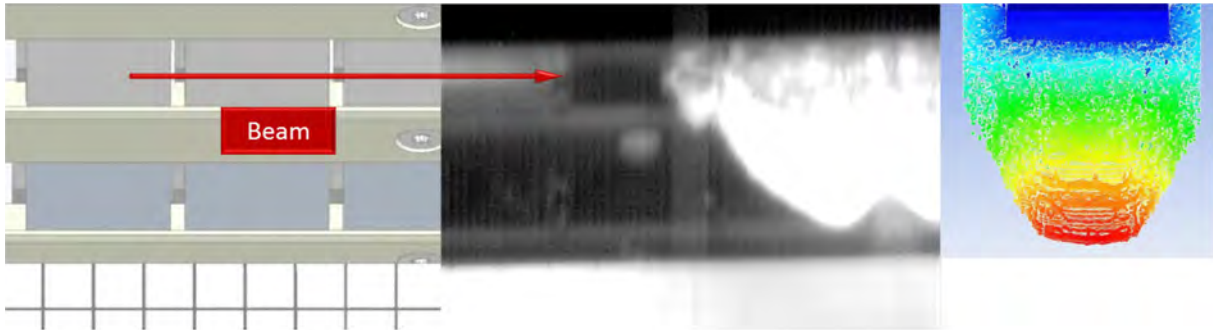
Both the front shape and velocity of the ejected particles are consistent with data acquired using the high-speed camera (Fig. 63), even considering the differences in beam size between the real ( $\sigma = 1.9$  mm) and simulated ( $\sigma = 2.5$  mm) scenarios.

Smoothed-particle hydrodynamics simulation results are consistent with the ejected particle front shape and velocity acquired by the high-speed camera (Fig. 64), even considering the differences in beam size between real and simulated scenarios. The velocity of the fragment front has been estimated by measuring the displacement between two successive frames and is  $\approx 275$  m s<sup>-1</sup>, well matching the simulated velocity of 316 m s<sup>-1</sup>.

The excellent fit between numerical results and experimental measurements confirms the reliability of the simulation techniques and provides a positive indication of the validity of the equation of state and strength model for Glidcop. The match between captured pictures of the Inermet explosion and SPH simulations is also good. Similar analyses will be performed in the near future on molybdenum, for which constitutive models exist, although they are less well established than for copper and tungsten.



**Fig. 62:** Image sequence of the impact on Inernet of a 72 bunch SPS proton pulse. The beam is coming from the left; three Inernet samples are partially visible (numbered 1 to 3).

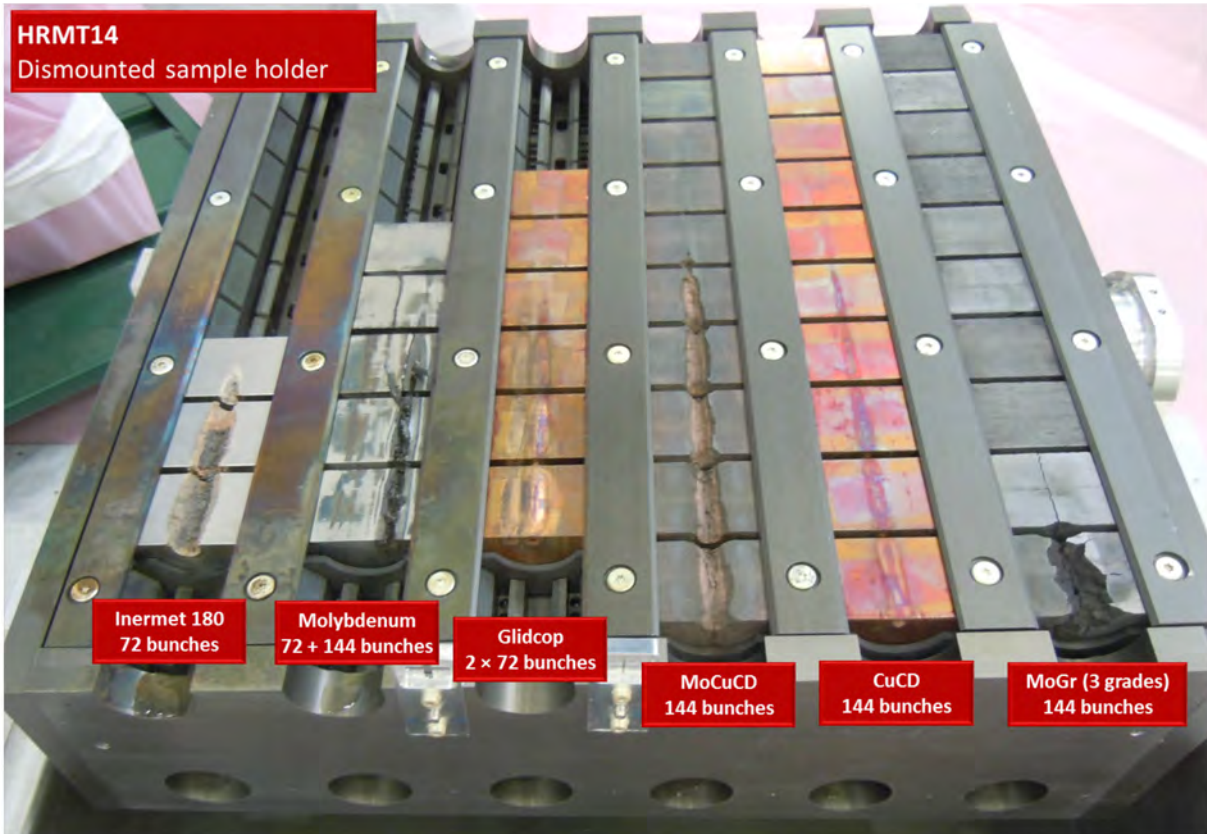


**Fig. 63:** Comparison between SPH simulation and acquired image 125  $\mu\text{s}$  after the impact. Orientation of the Inermet samples and direction of the beam are provided for convenience. Calculated maximum velocity of the fragment front is 316  $\text{m s}^{-1}$ .

Post-irradiation observations of specimens that underwent high-intensity tests (Fig. 65) confirmed that CuCD and MoGr resisted the impact of 144 bunches of the SPS, with CuCD showing coloration of the surface and a possible slight superficial deformation. Three MoGr grades were tested with densities ranging from 3.8 to 5.3  $\text{g cm}^{-3}$ : apart from an older grade of higher density, which has since been abandoned, the lighter MoGr grades showed no sign of degradation after visual inspection and non-destructive testing. It is worth noting that since 2012 newer grades of MoGr have been developed with even lower density (down to 2.5  $\text{g cm}^{-3}$ ) and better thermophysical properties.

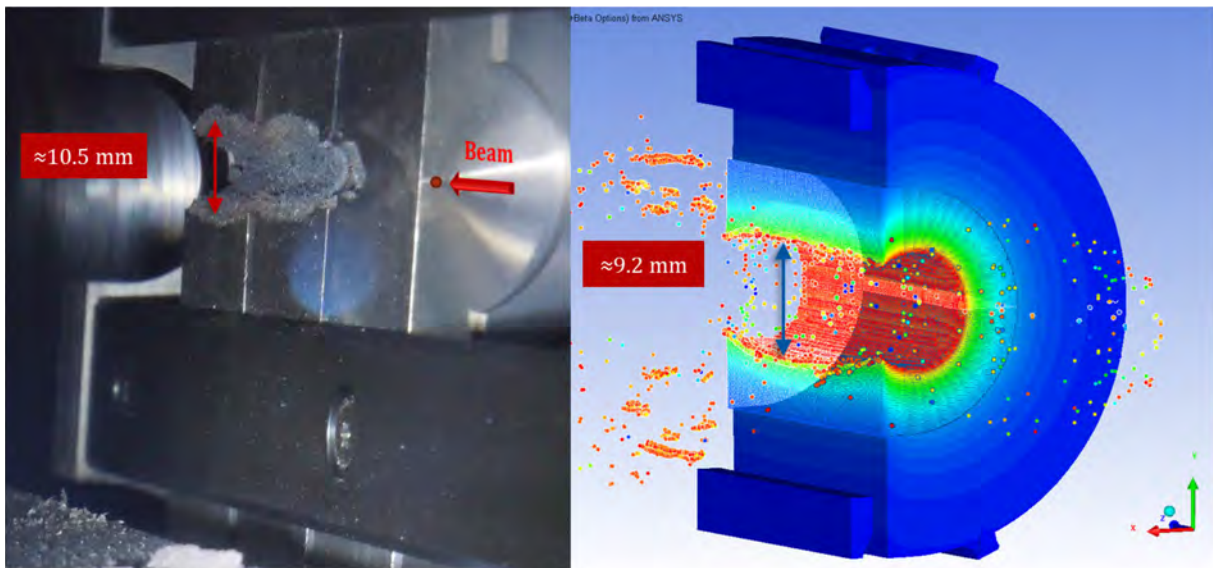
All higher-Z materials were damaged, to a variable degree of severity. MoCuCD experienced a catastrophic brittle failure and has been since abandoned; Glidcop suffered ejection of molten material at the point of impact (2 mm below the surface) of 72 bunch pulses, although, thanks to its ductility, the specimens' surfaces were largely deformed but not fractured.

Molybdenum exhibited somehow surprising behaviour: under a 72 bunch pulse (hitting the centre), the three last and most loaded specimens did not show evident signs of damage (although a later, more accurate, inspection found some small cracks), while the second sample in the series, which was less loaded, revealed a deep crack extending across most of the specimen. A second pulse at double the intensity (144 bunches) was delivered 10 mm apart from the first: in this case, a groove was produced on the most loaded specimens (although less extended than on Inermet samples at half the intensity), while cracks were induced, particularly on the third samples, several millimetres away from the point of impact; this behaviour may be explained by the temperature increase induced by the energy deposition: when the so-called ductile-to-brittle transition temperature is exceeded, materials shift from a highly brittle to a ductile behaviour. The ductile-to-brittle transition temperature in pure molybdenum is typically several tens of degrees above RT; therefore, up to a certain point, beam-induced heating may have had a beneficial effect in increasing ductility to a level that effectively countered the higher induced stresses.



**Fig. 64:** Post-irradiation observation of HRMT-14 sample holder. Beam arrived from the top. Note that the two last specimens in MoGr were from an obsolete grade.

Inermet experienced a brittle failure, with no signs of plastic deformation on the brim of the damaged area and on the flat surface. The low melting point of copper and nickel probably played an important role in determining the extent of damaged zone. The simulated damage extension is consistent with experimental observations (Fig. 65).



**Fig. 65:** Post-irradiation observation of Inermet 180 samples (left) and simulated failure (right)



## Acknowledgments

I would like to heartily thank Federico Carra, Alessandro Dallochio, and Marco Garlaschè (CERN, EN/MME) to whom I am indebted for their support in preparing and proofreading the lectures and these proceedings and, even more, for the stimulating discussions and exchanges which, through the years, allowed me to put together the material presented here.

I would also like to express my gratitude to Paolo Gradassi, Linus Mettler, and Jorge Guardia Valenzuela (CERN, EN/MME) for their important contributions to these proceedings.

## References

- [1] N. Mokhov and F. Cerutti, Beam material interaction, heating and activation, these proceedings.
- [2] S. Redaelli *et al.*, in L. Rossi *et al.* Eds., HL-LHC Preliminary Design Report, CERN-ACC-2014-0300 (2014).
- [3] A. Bertarelli *et al.*, The mechanical design for the LHC collimators, EPAC04, Lucerne (2004).
- [4] A. Bertarelli and T. Kurtyka, Dynamic thermomechanical phenomena induced in isotropic cylinders impacted by high energy particle beam, Proc. 8th Int. Conf. on Structures under Shock and Impact – SUSI VIII, Crete, 2004 (WIT Press, Southampton, 2004) p. 29.
- [5] A. Dallochio, Ph.D. thesis, Politecnico di Torino, 2008.
- [6] R.J.M. Konings *et al.*, *Comprehensive Nuclear Materials* (Elsevier, Amsterdam, 2012).
- [7] W. Kalbreier *et al.*, Target stations and beam dumps for the CERN SPS, CERN-SPS/ABT/77-3 (1977).
- [8] D. Walz *et al.*, Tests and description of beam containment devices and instrumentation – a new dimension in safety problems, SLAC-PUB-1223(A) (1973).
- [9] N. Mokhov *et al.*, Beam-induced damage to the Tevatron components and what has been done about it, 39th ICFA Adv. Beam Dynamics Workshop High Intensity High Brightness Hadron Beams, HB2006, Tsukuba (2006).
- [10] B. Goddard *et al.*, TT40 damage during 2004 high intensity SPS extraction, CERN AB-Note-2005-014 BT (2005).
- [11] V. Kain, Ph.D. thesis, Universität Wien, 2005.
- [12] I. c *et al.*, HiRadMat: a new irradiation facility for material testing at CERN, IPAC11, San Sebastián (2011). <http://accelconf.web.cern.ch/AccelConf/IPAC2011/index.htm>
- [13] M. Cauchi *et al.*, *Phys. Rev. ST Accel. Beams* **17**(2) (2014) 021004.  
<http://dx.doi.org/10.1103/PhysRevSTAB.17.021004>
- [14] G.S. Was, *Fundamentals of Radiation Material Science: Metals and Alloys* (Springer-Verlag, Berlin, 2007).
- [15] A. Bertarelli *et al.*, *J. Appl. Mech.* **75**(3) (2008) 031010. <http://dx.doi.org/10.1115/1.2839901>
- [16] A. Bertarelli, Analytical study of axisymmetric transient thermal stresses in graphite target rods for the CNGS Facility, CERN TN EST-ME 2003-005 (2003).
- [17] S.P. Timoshenko and J.N. Goodier, *Theory of Elasticity* (McGraw-Hill, New York, 1970).
- [18] A. Bertarelli, An analytical model to study transient thermal stresses in graphite target rods hit by off-axis beam for CNGS facility, CERN TN EST-ME-2003-06 (2003).
- [19] A. Bertarelli *et al.*, High energy beam impacts on beam intercepting devices: advanced numerical methods and experimental set-up, IPAC11, San Sebastián (2011).  
<http://accelconf.web.cern.ch/AccelConf/IPAC2011/index.htm>
- [20] M. Scapin, Ph.D. thesis, Politecnico di Torino, 2013.

- [21] U.S. Lindholm, in *Techniques of Metals Research*, Ed. R.F. Bunshah (John Wiley and Sons, New York, 1971), Vol. 5, Part 1, p. 199.
- [22] A. Bertarelli *et al.*, Permanent deformation of the LHC collimator jaws induced by shock beam impact: an analytical and numerical interpretation, Proc. European Particle Accelerator Conf., EPAC06, Edinburgh (2006).
- [23] M. Scapin *et al.*, “Thermo-Mechanical Modelling of High Energy Particle Beam Impacts” in *Numerical Modeling of Materials Under Extreme Conditions*, Eds. N. Bonora and E. Brown (Springer-Verlag, Berlin, 2014), Chap. 3, pp. 87–106.  
[http://dx.doi.org/10.1007/978-3-642-54258-9\\_4](http://dx.doi.org/10.1007/978-3-642-54258-9_4)
- [24] A. Bertarelli *et al.*, *J. Phys. Conf. Ser.* **451** (2013) 012005.  
<http://dx.doi.org/10.1088/1742-6596/451/1/012005>
- [25] V. Fortov *et al.*, *Nucl. Sci. Eng.* **123** (1996) 169.
- [26] B. Gladman *et al.*, LS-DYNA® Keyword User’s Manual, (Livermore, Livermore Software Technology Corporation, 2007) Vol. I, Ver. 971.
- [27] ANSYS Autodyn User Manual, (Canonsburg, ANSYS Inc., 2010), Release 13.0.
- [28] N.A. Tahir *et al.*, *Phys. Rev. E* **79**(4) (2009) 046410.  
<http://dx.doi.org/10.1103/PhysRevE.79.046410>
- [29] G.R. Johnson and W.H. Cook, A constitutive model and data for metals subjected to large strains, high strain rates and high temperatures, Proc. 7th Int. Symp. Ballistics, the Hague (1983).
- [30] F.J. Zerilli and R.W. Armstrong, *Appl. Phys.* **61**(5) (1987) 1816.  
<http://dx.doi.org/10.1063/1.338024>
- [31] D. J. Steinberg *et al.*, *J. Appl. Phys.* **51**(3) (1980) 1498. <http://dx.doi.org/10.1063/1.327799>
- [32] P.S. Follansbee and U.F. Kocks, *Acta Metall.* **36**(1) (1988) 81–93.  
[http://dx.doi.org/10.1016/0001-6160\(88\)90030-2](http://dx.doi.org/10.1016/0001-6160(88)90030-2)
- [33] A. Bertarelli *et al.*, Proc. Chamonix 2011: Workshop on LHC Performance, CERN-ATS-2011-005 (2011).
- [34] M. Scapin *et al.*, *J. Nucl. Mat.* **420**(1–3) (2012) 463–472.  
<http://dx.doi.org/10.1016/j.jnucmat.2011.10.036>
- [35] M. Scapin *et al.*, *Comput Struct.* **141** (2014) 74–83.  
<http://dx.doi.org/10.1016/j.compstruc.2014.05.008>
- [36] L. Peroni *et al.*, *Key Eng. Mat.* **569–570** (2013) 103–110.  
<http://dx.doi.org/10.4028/www.scientific.net/KEM.569-570.103>
- [37] N.E. Dowling, *Mechanical Behavior of Materials* (Pearson, Boston, 2012).
- [38] E. Métral, Transverse resistive-wall impedance from very low to very high frequencies, CERN-AB-2005-084 (2005).
- [39] L. Rossi *et al.*, HL-LHC preliminary design report, CERN-ACC-2014-0300 (2014).
- [40] N. Mounet *et al.*, Collimator impedance measurements in the LHC, IPAC13, Shanghai (2013).  
<http://accelconf.web.cern.ch/AccelConf/IPAC2013/>
- [41] A. Bertarelli *et al.*, Novel materials for collimators at LHC and its upgrades, HB2014, East Lansing (2014).
- [42] N. Mariani, Ph.D. thesis, Politecnico di Milano, 2014.
- [43] F. Carra *et al.*, Mechanical Engineering and Design of Novel Collimators for HL-LHC, IPAC14, Dresden (2014).
- [44] N. A. Tahir *et al.*, *Phys. Rev. ST Accel. Beams* **15**(5) (2012) 051003.  
<http://dx.doi.org/10.1103/PhysRevSTAB.15.051003>

- [45] R. Schmidt *et al.*, *Phys. Plasmas* **21**(8) (2014) 080701. <http://dx.doi.org/10.1063/1.4892960>
- [46] M. Cauchi *et al.*, *Phys. Rev. ST Accel. Beams* **17**(2) (2014) 021004.  
<http://dx.doi.org/10.1103/PhysRevSTAB.17.021004>
- [47] E. Quaranta *et al.*, Updated simulation studies of damage limit of LHC tertiary collimators, IPAC15, Richmond (2015).
- [48] A. Bertarelli *et al.*, *Nucl. Instr. Meth. B* **308** (2013) 88–99.  
<http://dx.doi.org/10.1016/j.nimb.2013.05.007>

### **Bibliography**

- S.P. Timoshenko and J.N. Goodier, *Theory of Elasticity* (McGraw-Hill, New York, 1970).
- B.A. Boley and J.H. Weiner, *Theory of Thermal Stresses* (Dover Publications, Mineola, 1997).
- H.S. Carslaw and J.C. Jaeger, *Conduction of Heat in Solids* (Oxford University Press, Oxford, 1959).
- W.T. Thomson, *Theory of Vibration with Applications* (CRC Press, London, 1993).
- N.E. Dowling, *Mechanical Behavior of Materials* (Pearson, Boston, 2012).
- M.A. Meyers, *Dynamic Behavior of Materials* (Wiley-Interscience, New York, 1994).  
<http://dx.doi.org/10.1002/9780470172278>
- J.A. Zukas, *Introduction to Hydrocodes* (Elsevier, Amsterdam, 2004).



## Protection Related to High-power Targets

*M.A. Plum*

Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA

### Abstract

Target protection is an important part of machine protection. The beam power in high-intensity accelerators is high enough that a single wayward pulse can cause serious damage. Today's high-power targets operate at the limit of available technology, and are designed for a very narrow range of beam parameters. If the beam pulse is too far off centre, or if the beam size is not correct, or if the beam density is too high, the target can be seriously damaged. We will start with a brief introduction to high-power targets and then move to a discussion of what can go wrong, and what are the risks. Next we will discuss how to control the beam-related risk, followed by examples from a few different accelerator facilities. We will finish with a detailed example of the Oak Ridge Spallation Neutron Source target tune up and target protection.

### Keywords

Target protection; beam; proton; high-intensity.

## 1 Introduction to high-power targets

In general, high-power targets are needed to create either lots of secondary particles for applications where the interaction cross-sections are low, or to make as many secondary particles as possible when the creation cross-sections are low. Example applications include:

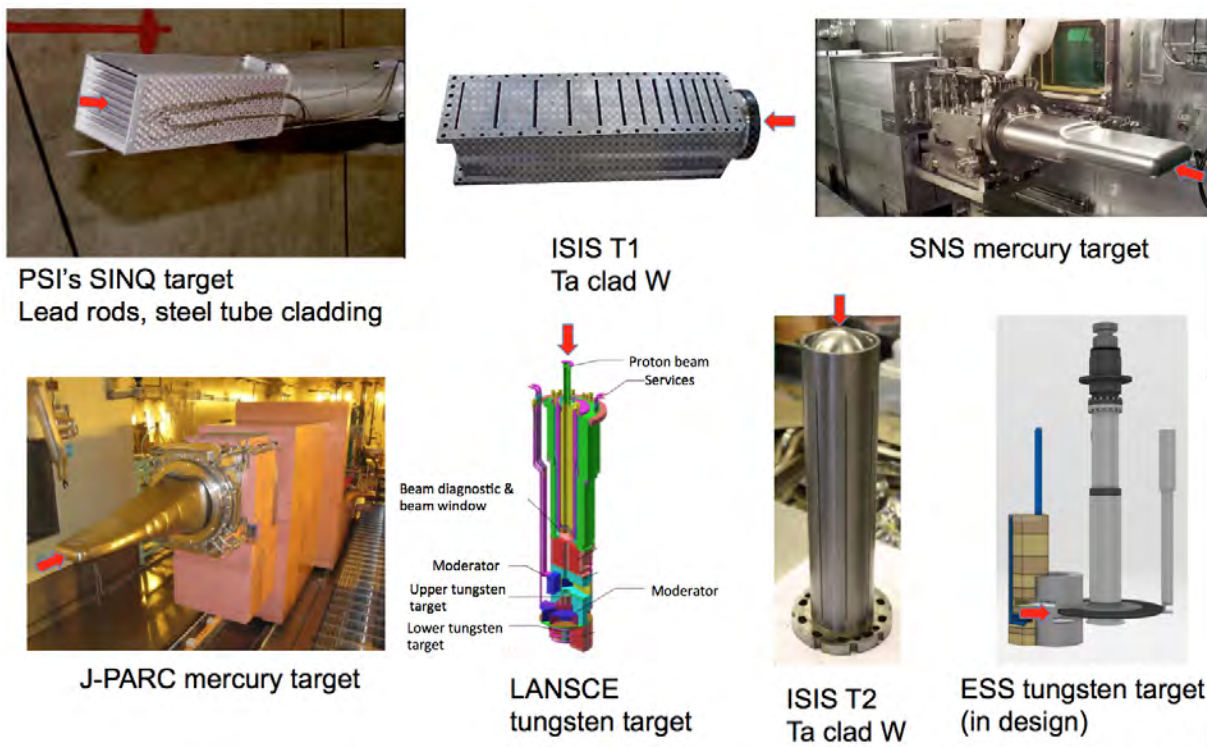
- neutron spallation targets (LANSCE, ISIS, SNS, J-PARC, PSI);
- muon production targets (J-PARC, PSI, TRIUMF, ISIS);
- Isotope Separation On-Line (ISOL) facilities (CERN, TRIUMF, IRIS);
- material irradiation studies (IFMIF);
- antiproton production (FNAL);
- neutrino production (FNAL, J-PARC, CERN).

High-power targets come in many shapes and sizes, and have many uses. Some example high-power targets are shown in Figs. 1–3, and parameters [1] from some example high-power facilities are shown in Table 1. High-power targetry is a highly developed and complex field, and there are many technological challenges. Some of the biggest beam-related challenges are:

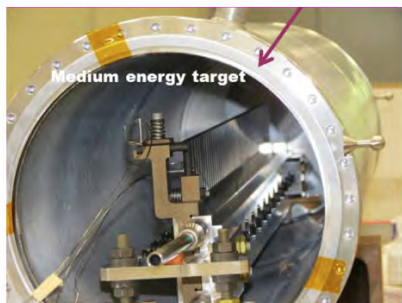
- removing the heat generated mainly by the beam, but also by nuclear decay;
- mechanical shock due to thermal stress from pulsed beams;
- radiation damage, including swelling and embrittlement;
- target handling (including installation and maintenance in a high-radiation environment);
- corrosive environment;
- beam parameters—matching target requirements to what the accelerator can deliver.

## 2 Target environment

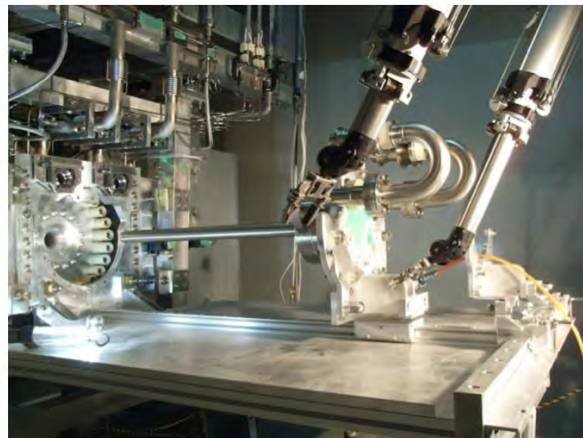
The target environment is often rough vacuum or helium. Vacuum is good because it does not interfere with the beams (primary or secondary). Helium atmosphere is good because it helps cool components



**Fig. 1:** Some example neutron spallation targets. Images reproduced from Refs. [2–6]



FNAL's NuMI carbon target



J-PARC's T2K graphite target  
26 mm dia x 910 mm long



FNAL's pbar production target

**Fig. 2:** Some example neutrino and pbar targets. Images reproduced from Refs. [7, 8]



Rutherford RIST ISOL target  
Stacks of Ta discs and washers



TRIUMF ISOL target  
Diffusion bonded Mo foils

Fig. 3: Some example isotope production targets. Images reproduced from Ref. [9]

Table 1: Some example high-power target facilities

| Facility            | Status    | Target material | Beam pulse duration ( $\mu\text{s}$ ) | Rep rate (Hz) | Proton energy (GeV)   | Time avg beam power (MW) | Peak time ave power density ( $\text{GW/m}^3$ ) | Peak energy density ( $\text{MJ/m}^3$ ) |
|---------------------|-----------|-----------------|---------------------------------------|---------------|-----------------------|--------------------------|---|---|
| ISIS                | Operating | W               | 0.4                                   | 50            | 0.8                   | 0.16                     | 0.25  | 5                                       |
| LANSCE-Lujan        | Operating | W               | 0.3                                   | 20            | 0.8                   | 0.16                     | 0.5   | 25                                      |
| NuMI                | Operating | C               | 8.6                                   | 0.53          | 120                   | 0.4                      | 0.32  | 600                                     |
| SINQ/Solid target   | Operating | Pb-SS clad      | CW                                    |               | 0.57                  | 1                        | 1   | NA                                      |
| SINQ/MEGAPIE        | Completed | Pb-Bi           | CW                                    |               | 0.57                  | 1                        | 1   | NA                                      |
| JSNS                | Operating | Hg              |                                       | 25            | 3                     | 1                        | 0.63  | 25                                      |
| SNS                 | Operating | Hg              | 0.7                                   | 60            | 1                     | 2                        | 0.8   | 13                                      |
| ESS—long pulse      | Proposed  | Hg              | 2000                                  | 16.7          | 1.3                   | 5                        | 2.5   | 150                                     |
| ESS—short pulse     | Proposed  | Hg              | 1.2                                   | 50            | 1.3                   | 5                        | 2.5   | 50                                      |
| EURISOL             | Proposed  | Hg              | 3                                     | 50            | 2.2                   | 4                        | 100   | 2000                                    |
| IFMIF               | Proposed  | Li              | CW                                    |               | 0.04 ( $\text{D}_2$ ) | 10                       | 100   | NA                                      |
| LANSCE-MTS          | Proposed  | Pb-Bi/W         | 1000                                  | 120           | 0.8                   | 0.8                      | 2.4   | 20                                      |
| US Neutrino Factory | Proposed  | Hg or C         | 0.003                                 | 15            | 24                    | 1                        | 3.8   | 1080                                    |
| AUSTRON             | Proposed  | W               | 1                                     | 10            | 1.6                   | 0.5                      |   |   |

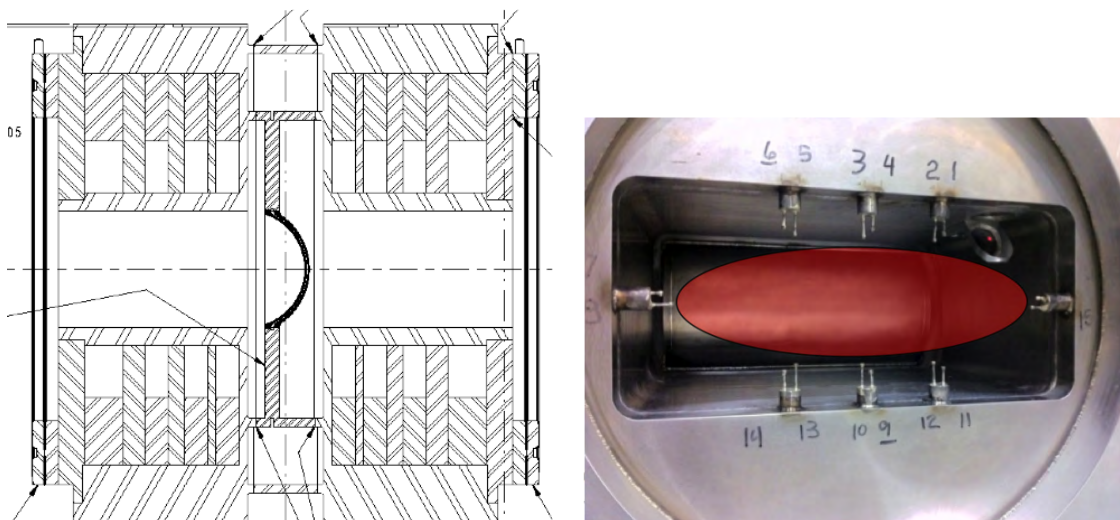
while minimizing beam scattering; however, impurities in helium can lead to corrosion of components. Beam transport lines leading up to the target need good vacuum to minimize beam loss due to scattering, so high-power targets often involve windows to separate the beam line vacuum from the target environment. Safety separation must also be considered (usually required for a liquid metal target). A partial pressure of air in the target environment can result in the formation of nitric acid, which can cause stress corrosion cracking in high-strength steel. Nitric acid can also cause vacuum leaks in thin bellows exposed to stray beam with air outside. For example, IPNS had a target bolt fail from stress corrosion cracking in a nominal helium atmosphere with air impurities when high-strength steel was substituted for stainless steel. Also, ISIS observed corrosion around the target/reflector assemblies and consequently limited impurities to below a couple of percent in helium. And, as shown in Fig. 4, Fermilab had a broken chain due to acidic vapour/condensate from air ionization.



**Fig. 4:** High-strength broken steel chain from hydrogen embrittlement caused by acidic vapour/condensate from air ionization (MiniBooNE 25 m absorber). Figure reproduced from Ref. [10].

### 3 Beam windows

Just like the target, beam windows are also challenged by high-power beams, due to heating, thermal stress, radiation damage, etc. An example beam window is shown in Fig. 5. This window is used to separate the SNS beam line vacuum of approximately  $1 \times 10^{-7}$  Torr from the target environment, which is 1 atm He. The window is made of water-cooled Inconel.



**Fig. 5:** Side view and end view of the SNS proton beam window, located about 2 m upstream of the neutron spallation target. The end view also shows the eight thermocouple halo monitors; the red ellipse indicates the approximate beam size. Images reproduced from Refs. [4, 11].

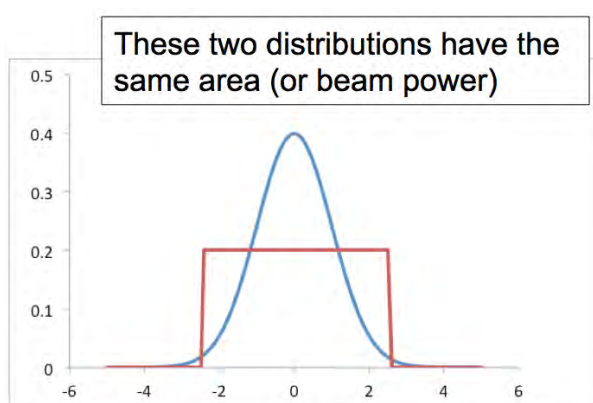
### 4 Beam parameters

The highest-power targets push the edge of achievable technology, and to live on the edge requires a very narrow range of beam parameters to avoid overpowering the target. High-power targets require the nominal beam distribution to be in the correct location, without exceeding the maximum beam current or the maximum beam density. The most important beam parameters are:



- beam position;
- beam size and shape (i.e. distribution);
- beam energy;
- beam current;
- beam pulse length, repetition rate, energy per pulse (short pulses cause a pressure pulse).

High-power targets often prefer flat or uniform beam distributions because they minimize the beam density and thereby minimize the density of energy deposited in the target. However, unless special measures are taken in the accelerator/beam delivery systems, the beam distribution will usually be nearly Gaussian. A possible target protection requirement could then be to monitor the beam distribution and to shut off the beam if the distribution exceeds requirements. For example, Fig. 6 shows two possible beam distributions. One is rectangular and the other is Gaussian. They both have the same area (or beam power), but the Gaussian distribution has twice the peak beam density.



**Fig. 6:** Two possible beam distributions. One is rectangular and the other is Gaussian. They both have the same area (or beam power), but the Gaussian distribution has twice the beam density.

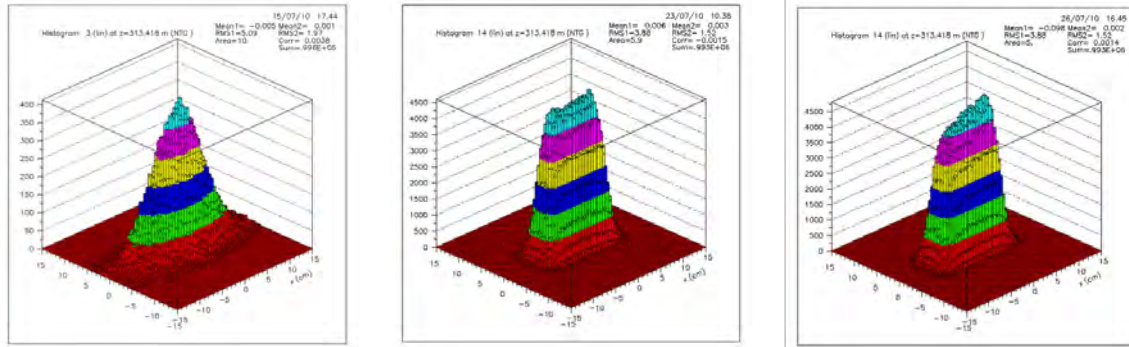
#### 4.1 Methods to flatten the beam distribution

Depending on the application, there are several ways to flatten the beam distribution, to make it more favourable for the target by decreasing the peak density. For storage rings and synchrotrons that employ multiturn injection (e.g. LANSCE, SNS, J-PARC), injection painting can be an easy way to control the distribution. Another method is to employ multipole magnets (e.g. octupoles) in the beam transport line to target. This method was recently implemented at J-PARC, as shown in Fig. 7. Note that the beam must be well centred in the octupoles for proper flattening, since off-centre beams will produce skewed distributions. Another method is a rastering system, that ‘draws’ the desired profile by quickly moving a small beam spot on the surface of the target. This is the method planned for the ESS, as shown in Fig. 8. All these methods rely on magnets or pulsed power supplies to achieve the desired beam profile. This highlights possible equipment that would be important to include as inputs to the machine protection system (MPS), and also possible beam monitoring systems, such as beam distribution monitors, that could also be part of the MPS interlock system.

### 5 What can go wrong?

In this section, we briefly describe some examples that highlight the importance of MPSs for high-power target protection.

In May 2013, a gold target (shown in Fig. 9) melted in the Hadron Experimental Facility at J-PARC and radioactive material was released from the building [15]. The root cause was traced to a



**Fig. 7:** Simulations of beam distributions on the J-PARC neutron production target. Left: no octupoles. Centre: with octupoles. Right: with octupoles and beam off centre at octupoles. Figures reproduced from Refs. [12, 13].

quadrupole magnet power supply in the main ring, which caused the beam to be spilled in 5 ms instead of the nominal 2 s. Due to this accident all beam operations were terminated for about 8 months and beam operations to the hadron hall were stopped for more than one year.

In January 2010, the neutron production target (shown in Fig. 10) failed at the ISIS second target station [16]. The root cause was traced to a combination of the high-density beam profile and a water leak. The water disassociated in the high-temperature environment and then the oxygen attacked the grain boundaries on the tantalum cladding. The cladding essentially fell apart, leaving a hole in the target. In 2004 the beam density on the PSI neutron production target increased to 3.5 times the nominal value [17, 18] because the transport line (shown in Fig. 11) was accidentally set up for the case of the muon target being inserted, but actually the muon target was not inserted. The over-density condition was caught by a newly installed beam distribution monitor (VIMOS) before any serious damage could occur.

We also need to be concerned with beam windows. They are often part of the target system, and can also be damaged by the beam. High-power beam windows must be cooled, usually with water. Windows can fail due to over-focused or off-centre beams. Radioactive water can be spilled into the beam line vacuum and/or the target environment. In 1996 the beam window at the LAMPF linac beam dump failed [19], probably due to a combination of an unusually small beam size (set up on purpose for an experiment) and a weld joint where a thermocouple was attached to the air-side surface of the water-cooled window.

## 6 Calculating the beam density

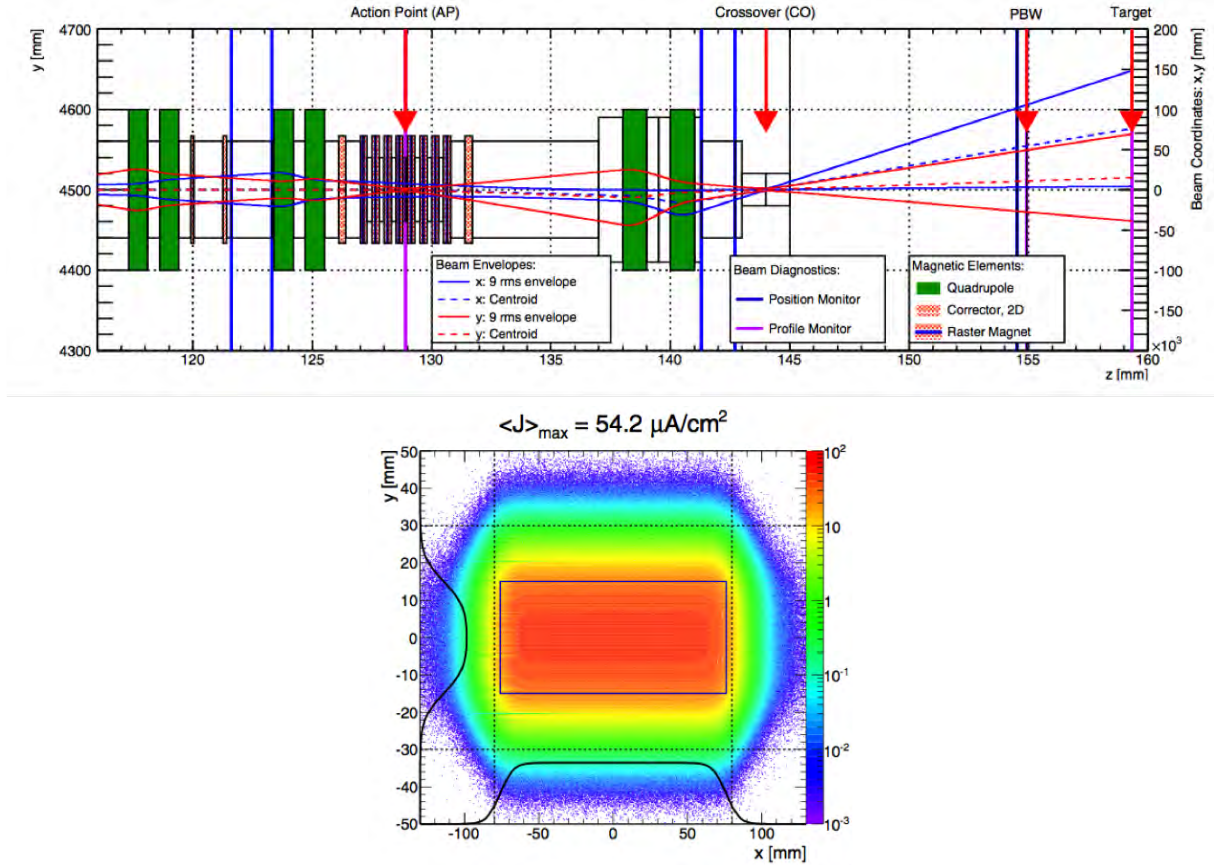
To demonstrate how the beam density depends on various beam parameters, we will consider several examples.

### 6.1 Example 1: rectangular DC beam

Assume a 2 MW DC beam of 1 GeV protons with rectangular cross-section 2 cm  $\times$  2 cm. Assume that it stops 0.75 m into the target. Also assume uniform energy deposition and no scattering. The energy deposition will be as depicted in Fig. 12. The 1-s power density is then

$$(2 \text{ MW}) / (2 \text{ cm}) / (2 \text{ cm}) / (0.75 \text{ m}) = 6.67 \text{ GW/m}^3.$$

Now assume that if the beam is too far off centre it must be shut off before it can deposit 100 J. What is the acceptable time delay (or MPS response time)? The beam must be shut off within  $(100 \text{ J}) / (2 \text{ MJ/s}) = 50 \mu\text{s}$ .



**Fig. 8:** The ESS design to achieve the desired beam distribution on the target by rastering the beam. Top: the beam transport line just upstream of the target. Bottom: simulation of the achieved beam distribution. Figures reproduced from Ref. [14].

### 6.2 Example 2: rectangular pulsed beam

Consider a pulsed version of the DC beam in Example 1. Assume that the beam is divided into 10 pulses per second, and the duration of each pulse is 1.5 μs.

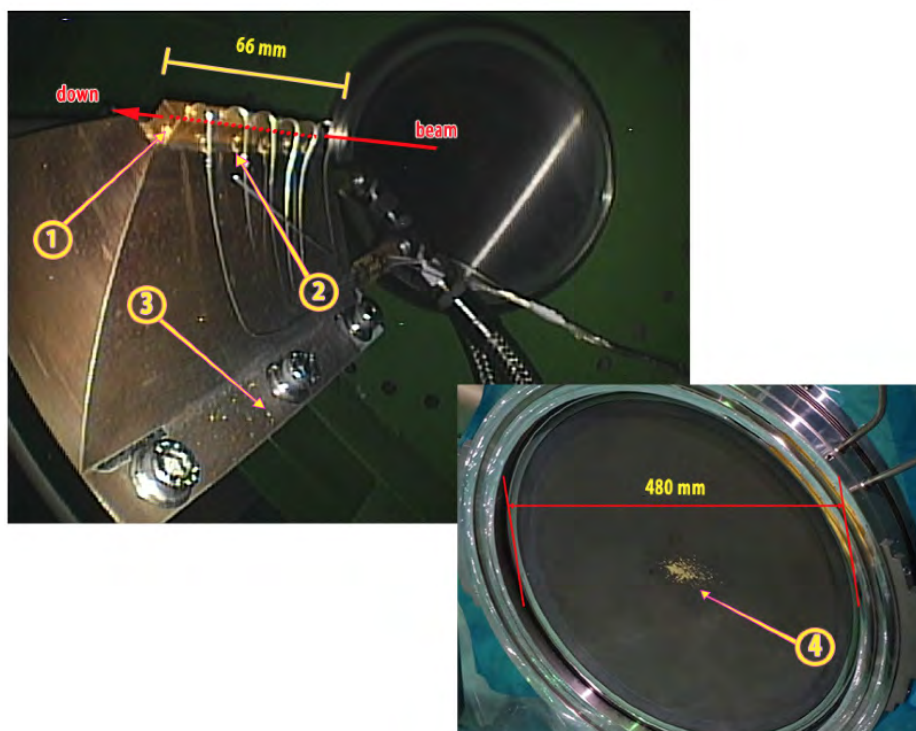
- The 1-s power density is still (2 MW)/(2 cm)/(2 cm)/(0.75 m) = 6.67 GW/m<sup>3</sup>.
- The energy per pulse = (2 MW)/(10 Hz) = 0.2 MJ.
- The energy density per pulse in target = (0.2 MJ)/(2 cm)/(2 cm)/(0.75 m) = 667 MJ/m<sup>3</sup>.
- The peak power density = (667 MJ/m<sup>3</sup>)/(1.5 μs) = 4.45 × 10<sup>14</sup> W/m<sup>3</sup>.

Each beam pulse contains more than 100 J of energy, so even one off-centre pulse will exceed the 100 J limit in Example 1. With 1.5 μs beam pulses, it is not practical to shut off the beam mid-pulse. We can only prevent the next pulse from occurring, and there is a fair amount of time to do that—about 0.1 s.

### 6.3 Example 3: pulsed Gaussian beam

Consider a Gaussian version of the pulsed beam in Example 2. Assume that the Gaussian shape is characterized by  $\sigma_x = \sigma_y = 0.5$  cm, as shown in Fig. 13. The functional form of the Gaussian is

$$f(x, y) = A \exp \left( - \left( \frac{(x - x_0)^2}{2\sigma_x^2} + \frac{(y - y_0)^2}{2\sigma_y^2} \right) \right).$$



**Fig. 9:** Photographs of the gold target in the J-PARC hadron hall, melted by a proton beam that was accidentally extracted over 5 ms. Figure reproduced from Ref. [15].

The volume under this two-dimensional Gaussian is

$$V = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy f(x, y) = 2\pi A \sigma_x \sigma_y$$

and the 1-s peak power density is

$$(2 \text{ MW}) / (2\pi) / (0.5 \text{ cm}) / (0.5 \text{ cm}) / (0.75 \text{ m}) = 17 \text{ GW/m}^3,$$

which is 2.25 times greater than the rectangular distribution case. The per-pulse peak energy density and peak power density also increase by a factor of 2.5.

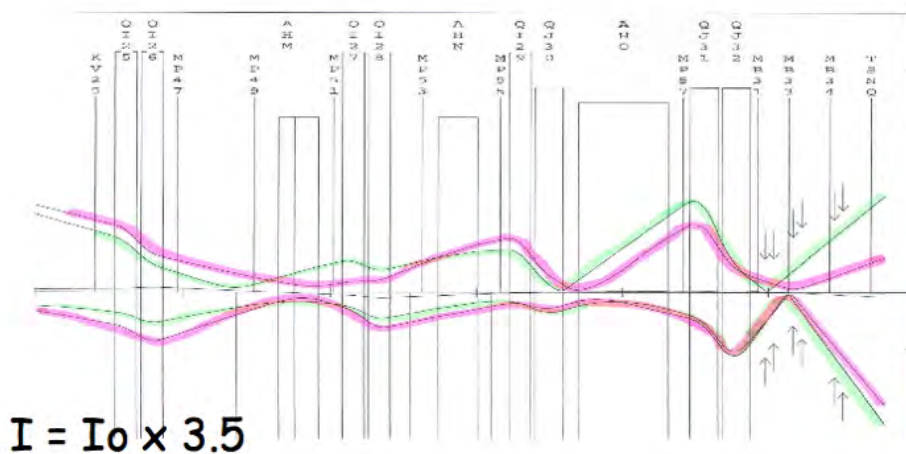
## 7 Target protection

We will consider only the beam-related protection. The targets themselves will have their own separate non-beam-related protection systems to monitor coolant flow, over temperature, etc. The main beam parameters of concern are density (sometimes derived from a beam profile measurement), beam size, beam position, and beam current (from which beam power can be calculated). Beam-related target protection involves control over these parameters, monitoring these parameters, and turning off the beam if these parameters move outside allowable limits. Types of beam-related target protection include:

- rapid and automatic beam turn off for off-normal beams;
- locking down equipment and operating control parameters to prevent accidental changes;
- alarms that alert operators when certain parameters move outside of pre-defined limits;
- collimators to partially intercept off-normal beams;



**Fig. 10:** A photograph of a failed target from the ISIS second target station. The tantalum cladding on the tungsten target failed. Figure reproduced from Ref. [16].



**Fig. 11:** The beam transport line leading to the PSI neutron production target, showing the beam sizes for the two different sets of parameters (one for the muon target inserted, the other for the muon target retracted). Figure reproduced from Ref. [17].

- beam transport designs that avoid high sensitivity—e.g. do not want to live on the edge, do not want to be very sensitive to small magnet changes;
- target designs that can tolerate off-normal beam for a short period of time (e.g. most SNS equipment is designed such that it can withstand two full-power off-normal beam pulses).

### 7.1 Protection by rapid beam turn off

The time it takes to automatically turn off the beam varies depending on the method used. Figure 14 shows some typical turn-off times for the SNS case.

The fastest method to turn off the beam is through the MPS. After all, that is the purpose of the MPS. A typical time to turn off the beam at SNS is 20 to 30  $\mu\text{s}$  for trips that originate in the linac, for example if a beam loss monitor exceeds a pre-determined threshold. Trips that originate further downstream, in the ring or in the ring-to-target beam transport, can take a few  $\mu\text{s}$  longer. An interesting

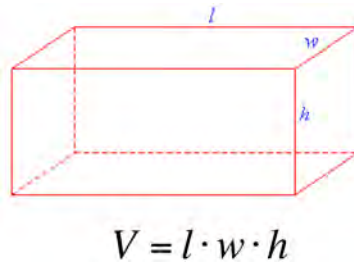


Fig. 12: A beam pulse with a rectangular cross-section

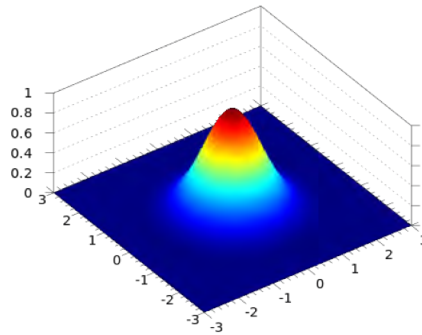


Fig. 13: A two-dimensional Gaussian beam profile

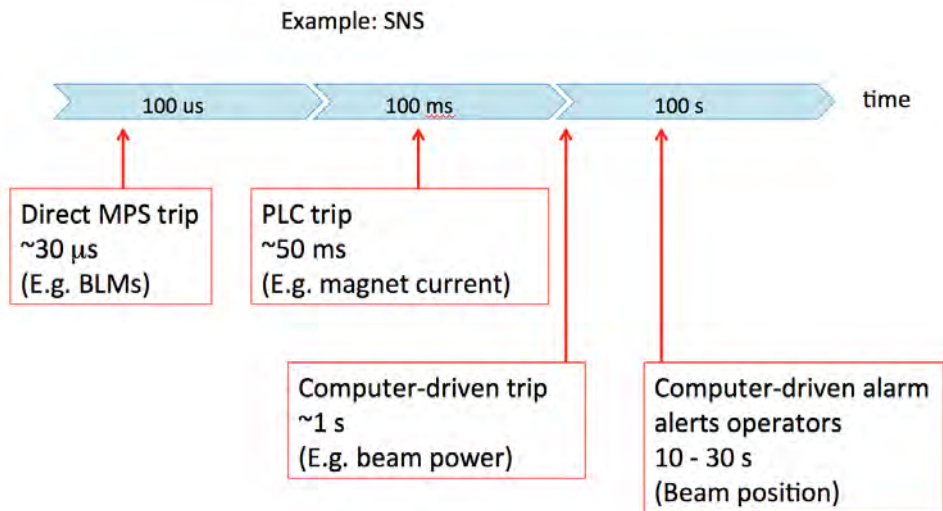


Fig. 14: Some typical turn-off times for the SNS case

PROTECTION RELATED TO HIGH-POWER TARGETS

fine point is that if the trip occurs upstream of or within the accumulator ring, the MPS can stop beam injection, but the beam that has already been injected into the ring will be sent to the target. The beam intensity can be anywhere from small to full intensity. If the trip occurs downstream of the ring, it can only prevent the next pulse from beginning the accumulation process.

For example, at SNS, PLCs monitor magnet current transformers on the power supplies for the last few magnets in the transport line to the target. If the magnet current strays outside pre-determined limits, the beam will be automatically turned off. The time for a PLC trip to turn off the beam is about 50 ms. As an additional precaution, PLC trips also turn off the high voltage to the ion source and turn off the high voltage to the first few klystrons.

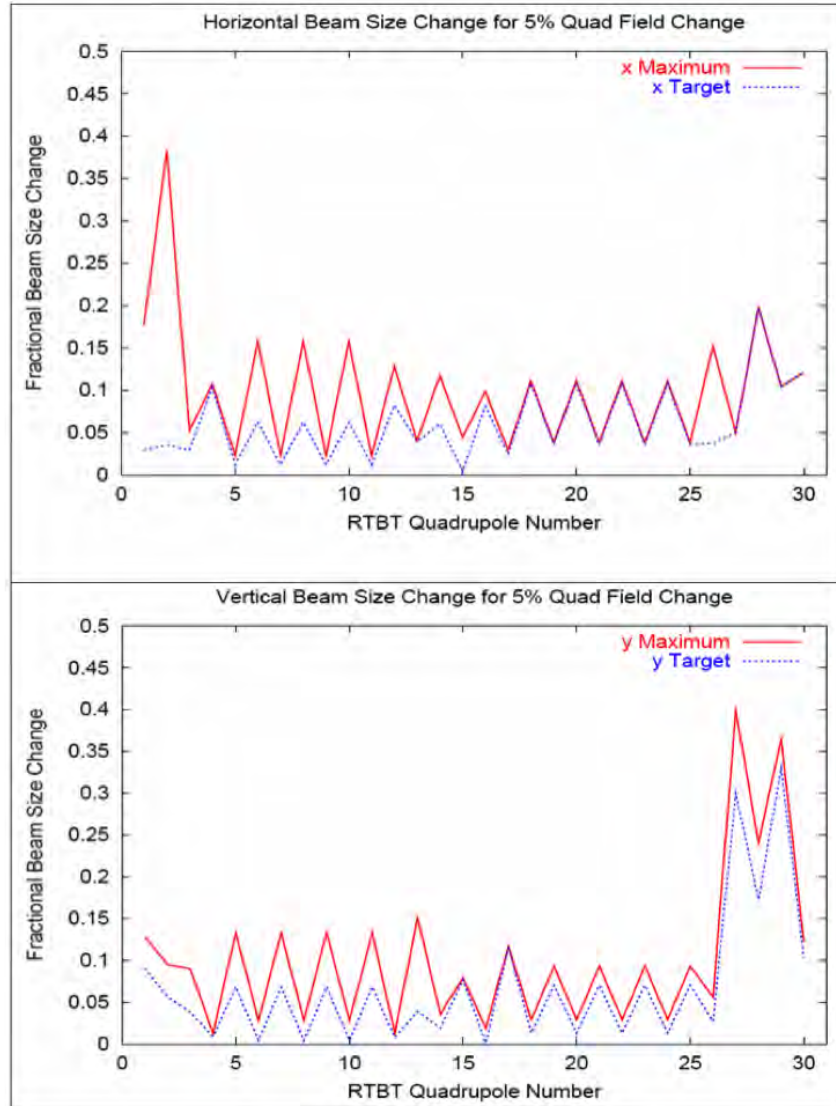
The third-fastest method is faults detected by the control computers. The computers interface to the MPS to turn off the beam. For example, at SNS, computers monitor all the magnet power supply currents in the beam transport to the target (as read by the power supply controller—not by separate current transformers). The estimated time to turn off the beam is about 1 s. For example, Fig. 15 shows a partial list of these magnets and their high- and low-trip limits.

| POWER SUPPLY STATUS |                |                |                |           | SETPOINTS |          | READBACKS           |         | Interlock Limits |          |             | PLC         |      |
|---------------------|----------------|----------------|----------------|-----------|-----------|----------|---------------------|---------|------------------|----------|-------------|-------------|------|
| PS Name             | Mode Select    | Channel Status | Control Status | PS Status | Set I     | Set B    | Trip Window<br>s, % | B       | LowTripLmt       | B Book   | HighTripLmt | Int1<br>Sta | RSet |
| RTBT_Mag_PS_ExtOptm | ON OFF STANDBY | Enabled        | Remote         | OK        | 1050.530  | 0.673359 | 5.000               | 0.673 T | 0.639691         | 0.673359 | 0.707027    | OK          | RSet |
| RTBT_Mag_PS_QV01    | ON OFF STANDBY | Enabled        | Remote         | OK        | 714.249   | 3.2719   | 5.000               | 3.280 T | 3.1003           | 3.2719   | 3.4355      | OK          | RSet |
| RTBT_Mag_PS_QH02    | ON OFF STANDBY | Enabled        | Remote         | OK        | 711.548   | 3.240    | 5.000               | 3.247 T | 3.0700           | 3.2400   | 3.4020      | OK          | RSet |
| RTBT_Mag_PS_QV03    | ON OFF STANDBY | Enabled        | Remote         | OK        | 498.442   | 3.106    | 5.000               | 3.106 T | 2.9504           | 3.1057   | 3.2610      | OK          | RSet |
| RTBT_Mag_PS_QH04    | ON OFF STANDBY | Enabled        | Remote         | OK        | 550.474   | 3.430    | 5.000               | 3.431 T | 3.2584           | 3.4293   | 3.6014      | OK          | RSet |
| RTBT_Mag_PS_QV0511a | ON OFF STANDBY | Enabled        | Remote         | OK        | 728.939   | 4.529    | 5.000               | 4.536 T | 4.0759           | 4.5288   | 4.9016      | OK          | RSet |
| RTBT_Mag_PS_QH0511b | ON OFF STANDBY | Enabled        | Remote         | OK        | 763.429   | 4.742    | 5.000               | 4.740 T | 4.2677           | 4.7409   | 5.2161      | OK          | RSet |
| RTBT_Mag_PS_QH12    | ON OFF STANDBY | Enabled        | Remote         | OK        | 553.173   | 3.444    | 5.000               | 3.445 T | 3.2718           | 3.4440   | 3.6162      | OK          | RSet |
| RTBT_Mag_PS_QV13    | ON OFF STANDBY | Enabled        | Remote         | OK        | 482.570   | 3.008    | 5.000               | 3.008 T | 2.8573           | 3.0077   | 3.1561      | OK          | RSet |

Fig. 15: Some of the power supplies in the SNS beam transport from the ring to the target, and their trip limits that are monitored by the control computer.

The fourth-fastest method is the alarm system, which alerts the operators to off-normal conditions. The alarms are generated by the control system, and require operator action to turn off the beam or to correct the off-normal condition. A strength of this method is that the control system can monitor a huge number of parameters, and also values derived from those parameters. For example, at SNS, the control system alarms if the beam position on target strays outside of pre-determined limits. The estimated time to correct the alarm state is 10 to 30 s, or possibly longer, depending on the operator reaction time and the complexity of the alarm.

Beam loss monitors (BLMs) are a standard part of machine protection for any beam transport system, and their purpose is primarily to protect beam line components. But BLMs can also be used to protect a target. The idea is that a quadrupole magnet change that can cause the beam to be too small or too large at the target will sometimes also cause beam loss upstream of the target, because the mis-focused beam could be large enough somewhere along the beam line that the beam tails will strike the beam pipe walls. BLMs can also protect against deviations from the nominal beam trajectory that could result in a bad position on the target, if the deviation is large enough to cause beam loss. At SNS, BLMs are located, and thresholds are set, such that some of the quadrupole magnet changes that can result in off-normal beams at the target will trip the BLMs. Figure 16 shows, for each quadrupole magnet in the beam line from the ring to the target, how much the beam size on the target changes for a 5% change in the quadrupole magnet gradient. It also shows the greatest amount of change anywhere in the beam line. Figure 17 shows the six of 19 quadrupole magnets that will cause a BLM trip (beam loss greater than 0.1%) before causing dangerous beam parameters at the target.



**Fig. 16:** These plots show, for each quadrupole magnet in the beam line from the ring to the target, how much the beam size on the target changes for a 5% change in the quadrupole magnet gradient (blue line). Also shown is the maximum beam size fractional change along the beam line (red line). Figure reproduced from Ref. [20].

## 7.2 Protection by equipment lock down

Another method to protect the target is to lock down certain control parameters. The idea is to administratively control critical hardware set points to prevent inadvertent changes. An example of this method is that at SNS the operators set a gate generator at the ion source that limits the possible beam pulse lengths, which limits the beam power on the target.

## 7.3 Protection by collimation

High beam power facilities often have collimators in the last part of the beam transport leading to the target, to protect against large beam position variations and overly large beam sizes, and to ensure that the beam hits the central region of the target. For example, SNS has a collimator immediately upstream of the target. The  $27.9 \times 12.7 \text{ cm}^2$  aperture is smaller than the face of the target, but slightly larger than the nominal beam size (90% of beam must fit within a  $20 \times 7 \text{ cm}^2$  rectangle). SNS also has collimators



| Magnet           | Constraint % | Reason   |
|------------------|--------------|--|
| QV1              |              | Upstream beam loss at -30%, +15% field strengths |
| QH2              |              | Upstream beam loss at -10%, +15% field strengths |
| QV3              |              | Upstream beam loss at -50%, +40% field strengths |
| QH4              | 7            | Peak current too high at -7% field strength      |
| QV5, 7, 9, 11    | 12           | Peak current too high at -12% field strength     |
| QH6, 8, 10       |              | Upstream beam loss at -20%, +10% field strengths |
| QH12             |              | Upstream beam loss at -20%, +20% field strengths |
| QV13             |              | Upstream beam loss at -40%, +30% field strengths |
| QH14             | 12           | Peak current too high at +12% field strength     |
| QV15             | 16           | Peak current too high at -16% field strength     |
| QH16             | 8            | Peak current too high at -8% field strength      |
| QV17             | 8            | Peak current too high at +8% field strength      |
| QH18, 20, 22, 24 | 24           | Peak current too high at +24% field strength     |
| QV19, 21, 23, 25 | 32           | Beam on target too low at +32% field strength    |
| QH26             | 40           | Beam on target too low at -40% field strength    |
| QV27             | 8            | Beam on target too low at -8% field strength     |
| QH28             | 7            | Peak current too high at +7% field strength      |
| QV29             | 8            | Beam on target too low at -8% field strength     |
| QH30             | 9            | Peak current too high at +9% field strength      |

**Fig. 17:** The six of 19 quadrupole magnets that will cause a BLM trip (beam loss greater than 0.1%) before causing dangerous beam parameters at the target. Figure reproduced from Ref. [20].

**Table 2:** Some example target interlocks

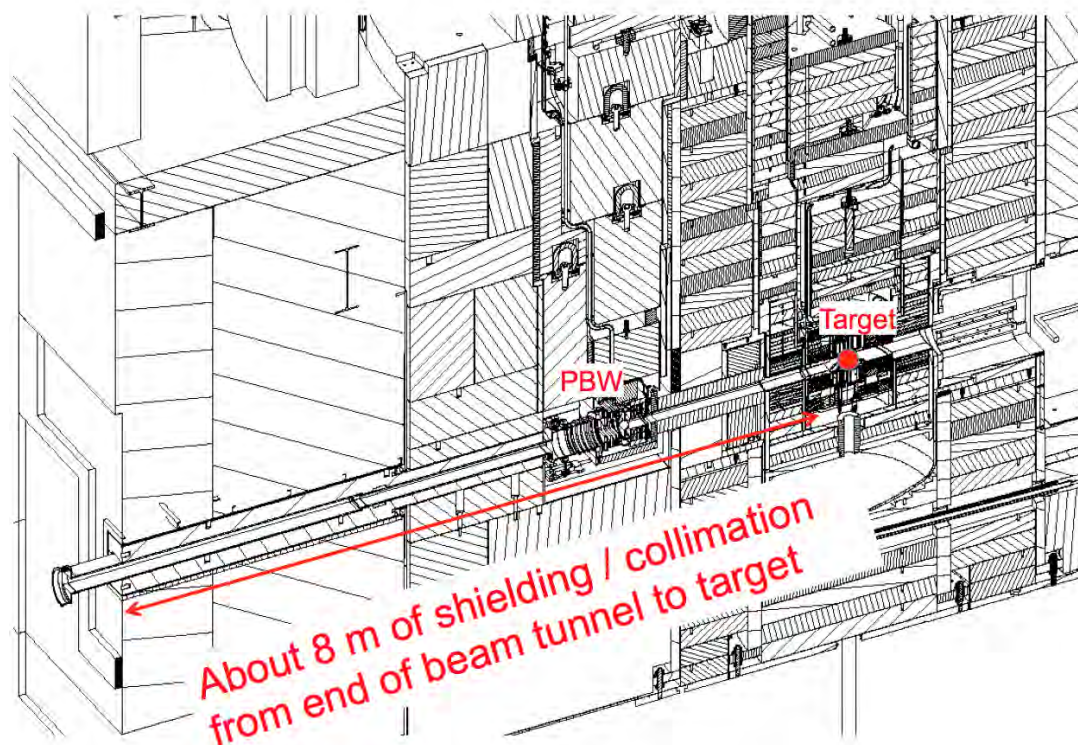
| Interlock  | Some accelerators that use this as a target interlock                            |
|--|--|
| Beam loss monitors                                       | Everybody  |
| Beam current monitors                                    | Everybody  |
| Beam position monitors                                   | PSI and FNAL have automatic beam centring based on beam position monitor signals |
| Harp profile monitors                                    | ISIS <sup>a</sup> , LANSCE <sup>a</sup> , J-PARC <sup>a</sup> , SNS <sup>a</sup> |
| Ionization profile monitors                              | PSI  |
| Glowing screen just upstream of target (profile monitor) | PSI  |
| Halo thermocouples                                       | SNS  |
| Amount of beam intercepted by collimator                 | PSI  |

<sup>a</sup> No interlock at present but this is under consideration

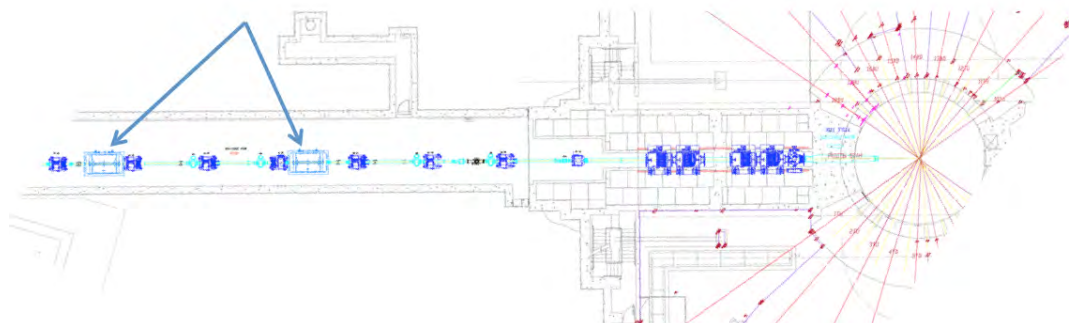
in the beam transport line that can intercept a portion of the beam that does not receive the nominal kick angle from the extraction kickers, thus protecting the target from some of the mis-kicked beam that would consequently arrive off centre at the target. Figure 18 shows the collimator that is embedded in the bulk shielding surrounding the SNS target, Fig. 19 shows the positions of the beam line collimators, and Fig. 20 shows simulations of the beam trajectory deviations due to extraction kickers that do not fire properly.

#### 7.4 Protection by interlocking on off-normal measured beam parameters

This protection category involves continuously monitoring critical beam parameters, and then automatically turning off the beam if the parameters exceed pre-determined thresholds. Table 2 shows some of the more common beam instrumentation used for this purpose.



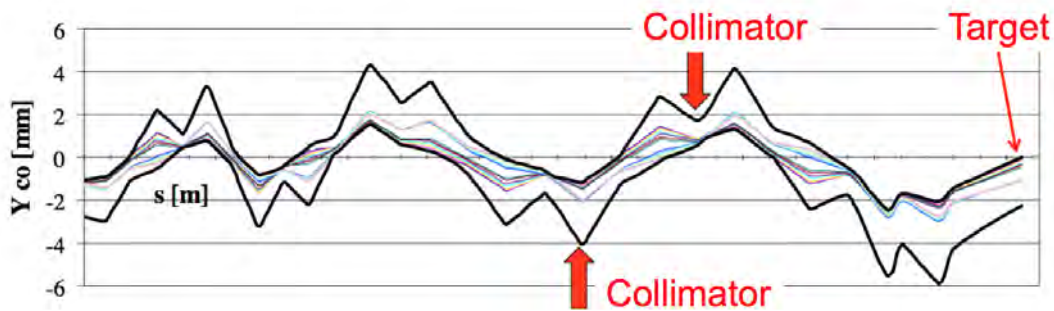
**Fig. 18:** The collimator and shielding just upstream of the SNS neutron production target. The PBW label refers to the Proton Beam Window.



**Fig. 19:** The downstream portion of the SNS RTBT beam line. The blue arrows show the positions of the RTBT collimators.

### 7.5 Protection by beam line design

It is usually desirable to design the beam transport line to avoid high sensitivity to beam parameters at the target. For example, small adjustments to dipole magnets should not cause large position changes on the target, and small adjustments to quadrupole magnets should not cause large beam size changes on the target. At SNS, one way this design practice was implemented was to require the phase advance from the extraction kickers to the target to be approximately a multiple of  $\pi$ . In this way, a missed kick from an extraction kicker will not cause a large position change on the target. This effect is illustrated in Fig. 20.



**Fig. 20:** Simulations of beam trajectory deviations caused by missed kicks in the ring extraction kicker system. The thin lines indicate trajectory deviations for individual missed kicks, and the thick lines indicate the maximum and minimum trajectory deviations for the case of two simultaneous missed kicks. Figure reproduced from Ref. [21].

## 8 Some example target protection implementations

### 8.1 J-PARC target protection

The J-PARC design parameters call for a 1 MW, 3 GeV, 25 Hz proton beam to be delivered to a liquid mercury target. To protect the target [22], there is a waveform monitor on each of the eight extraction kickers. A PLC monitors some of the quadrupole and dipole magnet power supply currents. A PLC also monitors the beam profile measurement device upstream of the target, and it is possible to connect this to the machine protection system, but presently this is not the case. A recent addition to the MPS interlock system is a fast beam current monitor to shut off the beam if the beam current exceeds a pre-defined limit.

### 8.2 FNAL NuMI target protection

The NuMI neutrino production target at FNAL now operates with 120 GeV, approximately 400 kW protons, <1 Hz, with an upgrade to 700 kW planned in 2015. A later upgrade to >1 MW is planned. Prior to beam extraction from the main injector (MI), more than 250 different inputs to the beam permit system are checked [23]. The beam is sent to the abort dump if inputs are not correct. The inputs checked include the following: beam position and angle for the MI extraction channel, possible excessive residual MI beam in the kicker beam gap, the extraction kicker status, proper NuMI power supply flat-top values, the beam readiness of the target station and absorbers, all beam loss monitor readings from the previous extraction, and the previous pulse position and trajectory at the target. The beam delivery system also has an automatic beam steering system to keep the beam centred on the target. If a beam pulse is >1.5 mm off centre, the beam is automatically turned off.

### 8.3 ISIS target protection

The ISIS facility at RAL operates two targets, TS1 and TS2. TS1 operates with 800 MeV, 160 kW, 40 Hz protons, and TS2 operates with 800 MeV, 40 kW, 10 Hz protons. A beam halo monitor (comprising eight equally spaced thermocouples in the penumbra of the beam  $\sim 100$  mm upstream of the target) indicates mis-focused and mis-steered beams. If the thermocouple signals exceed pre-determined trip thresholds, the protection system turns off the beam. The protection system [24] is capable of turning off the machine within 2 ms and hence inhibits the next beam pulse from the ion source in the 50 Hz synchrotron cycle. There is also a steering servo system to accommodate variations in the upstream beam. The servo and trip system is slow, running at  $\sim 2$  s. On TS2 there is also a harp profile monitor which sits permanently in the beam, but this is not interlocked. A similar harp is planned for TS1 as part of an upgrade project.

## 9 SNS target tune up

To provide a detailed example of target protection, we will now discuss the SNS case. We will be referring to the beam instrumentation shown in Fig. 21. The initial beam tuning to the target is performed at low intensity, low power, 1 Hz (<1 kW) beam. Once the beam position has been adjusted to centre the beam on the target, the beam power is slowly increased to full intensity, but still at 1 Hz ( $\sim 23$  kW). The beam intensity is now high enough that an image can be seen on the target imaging system (TIS), which is based on a light-emitting coating on the surface of the target, and hence gives a direct measurement of beam position and beam distribution at the target. The TIS is used to check the beam centring, and then the beam size is measured not with the TIS (due to unresolved discrepancies with the measurement technique about to be discussed), but with the four wire scanners and one harp located at various positions in the ring to target beam transport (RTBT) beam line, as shown in Fig. 21. The beam size at the target is determined by fitting the measured rms beam sizes with an online model, and varying the simulated beam parameters until the best fit to the measured rms beam sizes is obtained. The model is then used to extrapolate the beam size at the target. An example fit is shown in Fig. 22, and Fig. 23 shows the corresponding output table with the extrapolated beam size at the target.

The same online model is then used to extrapolate the peak beam density, measured with the harp, to the target. The critical beam parameters for the target have now been measured. To meet requirements:

- the beam must be centred on the target to within 6 mm horizontal and 4 mm vertical;
- the rms beam size must be <49 mm horizontal and <17 mm vertical;
- the peak density on target must be less than  $2 \times 10^{16}$  protons/m<sup>2</sup>;
- the peak density on the proton beam window must be less than  $2.9 \times 10^{16}$  protons/m<sup>2</sup>.

Once the measured beam parameters have been determined to meet requirements, the beam delivery system is locked down. This is required before increasing the beam power above 100 kW. The lock down is accomplished by engaging the following interlock thresholds.

- The last five quadrupole magnet power supplies are monitored by a PLC with current limits set to  $\pm 7\%$ .
- Both large dipoles are monitored by a PLC with current limits set to  $\pm 2$  A.
- The last two horizontal dipole correctors and last two vertical dipole correctors are monitored by a PLC with current limits set to  $\pm 5$  A.
- The injection kicker waveform monitor system is engaged. If waveforms stray outside of pre-defined windows, the MPS trips the beam.
- The extraction kicker waveform monitor system is engaged. If waveforms stray outside of pre-defined windows, the MPS trips the beam.
- All RTBT magnets are monitored by the control computer with current limits set to  $\pm 5\%$  on quadrupole magnets,  $\pm 5\%$  on large dipole magnets, and  $\pm 0.5$  A on dipole corrector magnets.
- The beam power limit is set into the control system, and monitored by the computer.

The beam pulse length and beam duty factor are also locked down, but in a way that does not cause a beam trip. It simply prevents the timing system hardware from being set in a way that could exceed a pre-determined beam power. A screen shot of this system is shown in Fig. 24.

Figure 25 shows a screen shot of the magnet currents that are monitored by a PLC. The PLC monitors those magnets that have the biggest impact on beam size and position at the target. The control system computers also monitor these magnets, as well as many more, as shown in Fig. 26, but the PLC is a more robust and reliable system compared to the control computer.

There are four horizontal and four vertical injection kickers to paint the beam distribution in the ring. An example set of waveforms is shown in Fig. 27. If for some reason the waveforms were accidentally set to paint a smaller beam, it would result in an excessively high beam density on the target. Or, the kicker system may fail in a way that could also cause the beam size to be too small. To protect against these possibilities an injection kicker waveform monitor system, based on commercially available oscilloscopes, watches the readback waveforms and automatically trips off the beam if the waveforms stray outside pre-determined envelopes.

Some interlocks are always in effect, not just when the beam power is greater than 100 kW. Examples include the beam loss monitors (direct MPS interlock), the proton beam window halo thermocouples (shown in Fig. 5, and interlocked through a PLC connected to the MPS), and the target protection system, which monitors parameters such as mercury flow, water cooling, etc (PLC interlock).

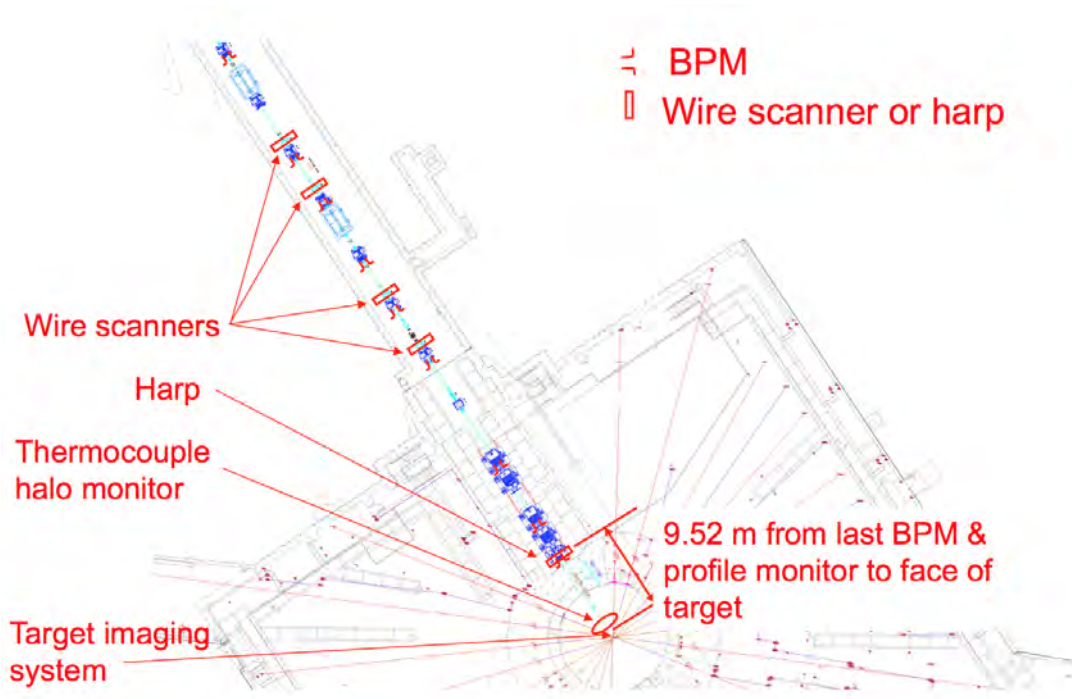
The control system computers monitor the following parameters and alert the operators if they stray outside of pre-set bounds:

- beam position on target calculated by the target imaging system;
- the beam density at the harp 9.52 m upstream of the target. The beam density at the target should be proportional to the beam density at the harp;
- the beam size at the harp 9.52 m upstream of the target. The beam size at the target should be proportional to the beam size at the harp;
- beam power;
- beam centring estimated from the proton beam window halo thermocouples (top–bottom, left–right);
- some magnet currents.

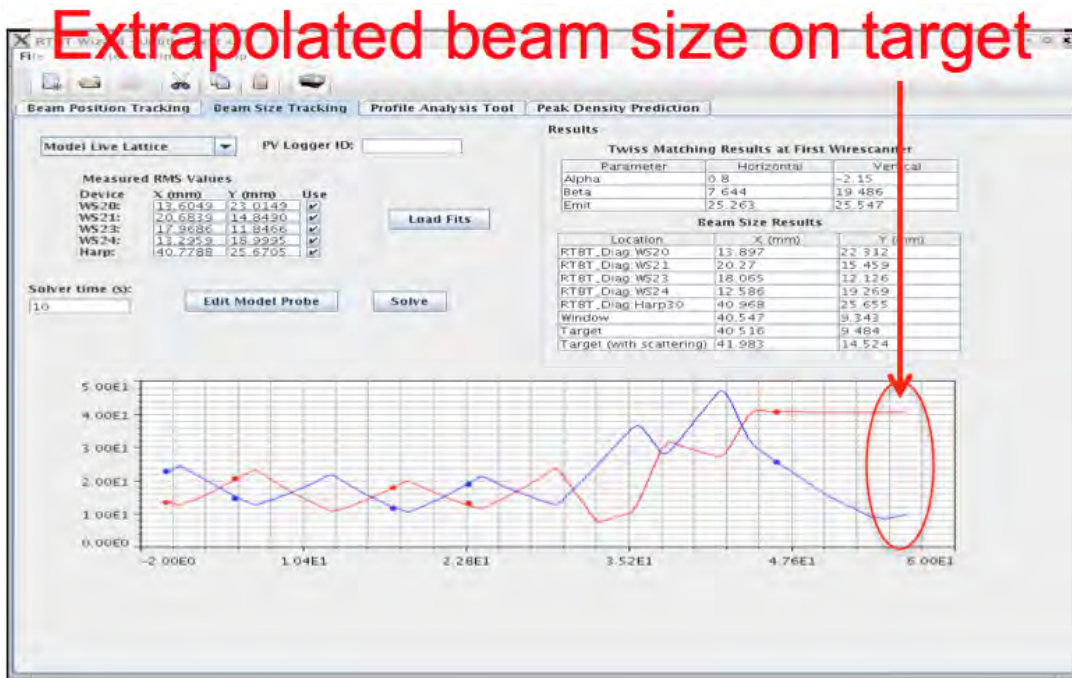
The last three items on the list are also interlocked by other methods as discussed above, but in this case the thresholds are more loosely set to give the operators a chance to correct the problem before it becomes severe enough to trip off the beam. The alarm summary screen is shown in Fig. 28.

## 10 Summary

In summary, high-power targetry is an active field that is continuously advancing. The highest power targets require tight control over the beam position, density, and distribution, and even then the target lifetimes can be short (e.g. 6 months at SNS). Machine protection systems must monitor these beam parameters and quickly activate interlocks if they stray outside of limits. Protection can include monitoring equipment set points, equipment readback parameters, beam parameters, locking down certain controls, beam line design, and collimator systems.



**Fig. 21:** A drawing of the SNS beam line upstream of the target, showing the locations of the profile and position monitors. Image reproduced from Ref. [25].



**Fig. 22:** Screen shot of the application that fits a model to measured beam sizes to extrapolate the beam size at the target. The points show the measured beam sizes and the lines show the model beam sizes.

PROTECTION RELATED TO HIGH-POWER TARGETS

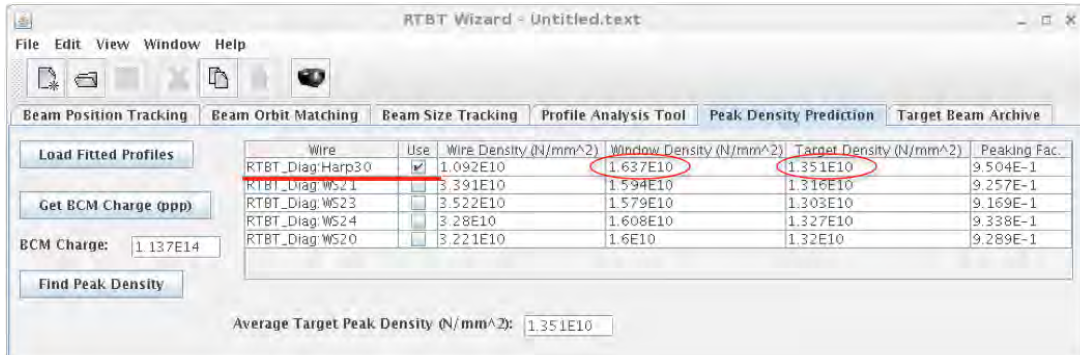


Fig. 23: Another screen shot of the application that extrapolates the beam size on the target. The resultant size predictions at the beam window and target are circled.

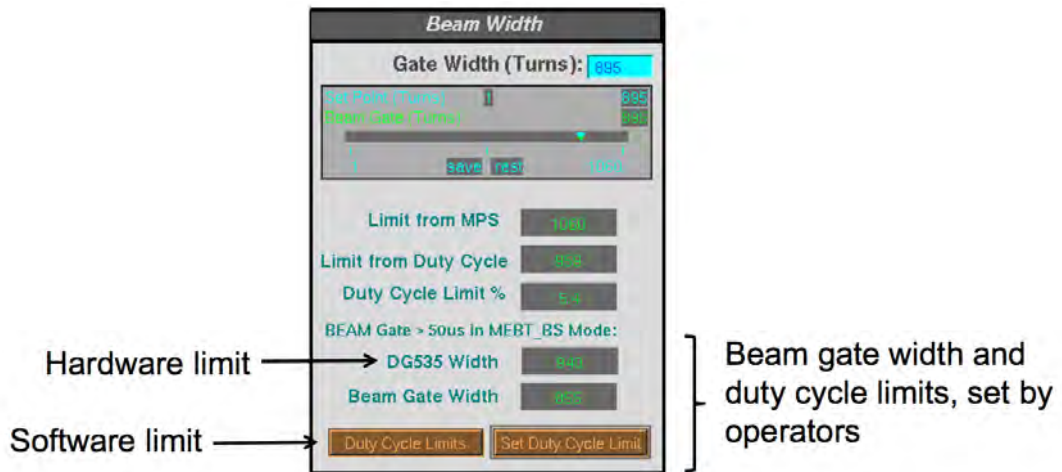


Fig. 24: A screen shot of the application showing the DG535 pulse generator settings that lock down the maximum possible beam pulse width.



Fig. 25: A screen shot showing the PLC limits on the last magnets upstream of the target



Fig. 26: A screen shot showing the magnets monitored by the control system to lock down the beam transport to the target.



PROTECTION RELATED TO HIGH-POWER TARGETS

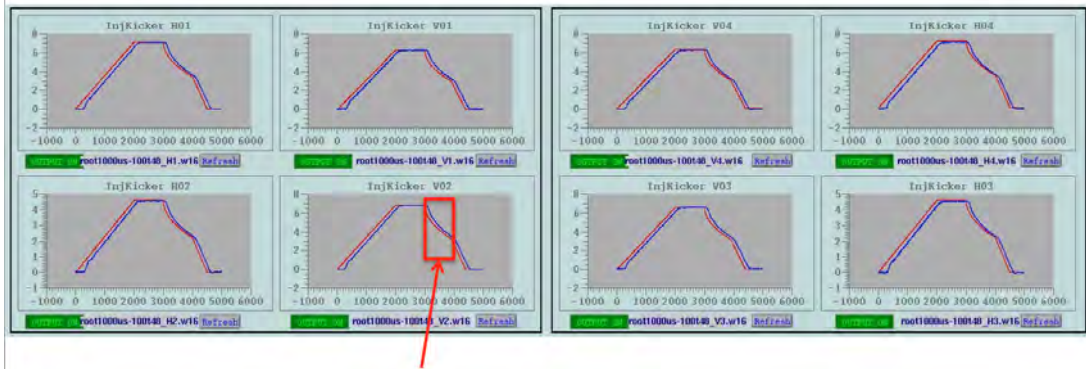


Fig. 27: (Colour) A screen shot showing the injection kicker waveforms. The region marked by the rectangle and arrow shows the portion of the waveform that is monitored for one of the kickers (note that all eight are monitored).

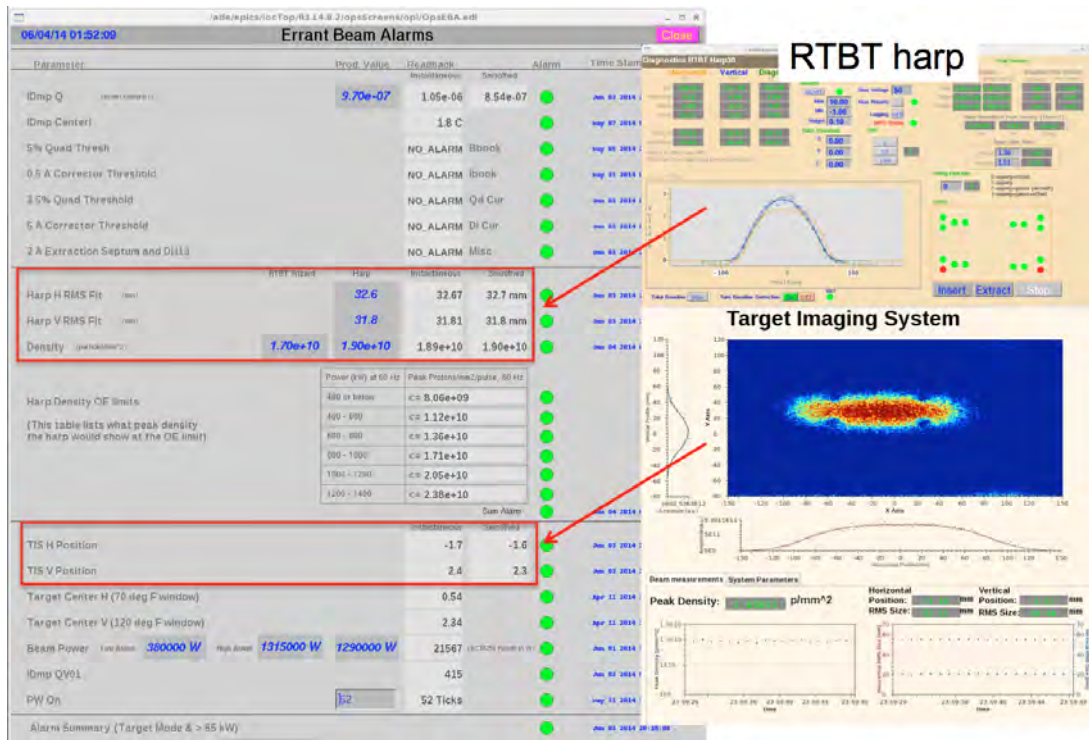


Fig. 28: A screen shot of the alarm summary screen and example screen shots of the beam profile measured by the harp and the beam distribution measured by the TIS. The arrows show how the measured beam parameters are monitored by the alarm system.

## Acknowledgements

ORNL is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy.

Notice: This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide licence to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

## References

- [1] J. Haines, U.S. Particle Accelerator School, High Power Beam Targets Class, Vanderbilt University, Nashville, Tennessee, 20 January 2009.
- [2] B. Reimer, "Spallation Source Facilities," 5th High Power Targetry Workshop, Fermilab, 20–23 May 2014. <https://indico.fnal.gov/conferenceDisplay.py?confId=7870>.
- [3] D. Haynes, "Introduction and Overview of the ISIS Target Systems," ISIS and SNS Bilateral Workshop, Oak Ridge, Tennessee, 25–26 June 2013.
- [4] T. McManamy, U.S. Particle Accelerator School, High Power Beam Targets Class, Vanderbilt University, Nashville, Tennessee, 20 January 2009.
- [5] R. Werbeck, "LANSCE Short-Pulse Target Operation," Workshop on High-power Targetry for Future Accelerators, Long Island, New York, 8–12 September 2003.
- [6] W. Wagner et al., "The SINQ solid spallation target ? operational experience and recent improvements ," High Power Targetry Workshop, Malmo, Sweden, 2–6 May 2011. Also [http://www.hep.princeton.edu/mumu/target/Wagner/Wagner\\_050311.pdf](http://www.hep.princeton.edu/mumu/target/Wagner/Wagner_050311.pdf).
- [7] N. Mokhov, "Beam-Material Interactions," Joint US–CERN Accelerator School on Machine Protection, Newport Beach, CA, USA, 5–14 November 2014.
- [8] J. Hylen, "Survey of Target Facility Landscape: Neutrino Beam Facilities," 5th High Power Targetry Workshop, Fermilab, 20–23 May 2014.
- [9] P. Bricault, "Radioactive Ion Beam Facilities ? High Power Target, Current status and Future Directions," 5th High Power Targetry Workshop, Fermilab, 20–23 May 2014. Also <https://indico.fnal.gov/getFile.py/access?contribId=91&sessionId=0&resId=0&materialId=slides&confId=7870>.
- [10] P. Hurh, Third High Power Target Workshop, Bad Zurach, Switzerland, September 2007.
- [11] T. McManamy, "Operational Experience with the SNS Target Systems and Upgrade Plans," Workshop on Applications of High Intensity Proton Accelerators, Fermi National Accelerator Laboratory, Batavia, IL, USA, 19–21 October 2009.
- [12] S. Meigo, "Development of beam flattening system using non-linear beam optics at J-PARC," Eleventh International Topical Meeting on Nuclear Applications of Accelerators, Bruges, Belgium, 5–8 August 2014.
- [13] S. Meigo et al., Beam flattening system based on non-linear optics for high power spallation neutron target at J-PARC, Proc. IPAC2014, Dresden, Germany, 15–20 June 2014.
- [14] H.D. Thomsen, A.I.S. Holm, and S.P. Møller, "A linear beam raster system for the European Spallation Source," Proc. IPAC2013, Shanghai, China, May 2013, p. 70.
- [15] [http://j-parc.jp/en/topics/20130812Accident\\_Report.html](http://j-parc.jp/en/topics/20130812Accident_Report.html); <http://j-parc.jp/en/topics/HDAccident20131217.pdf>
- [16] L. Jones, PASI Working Group Report by D.M. Jenkins, April 2013; [http://pasi.org.uk/images/2/2c/DavidJenkins\\_PASI\\_meeting\\_28\\_Feb\\_13.pdf](http://pasi.org.uk/images/2/2c/DavidJenkins_PASI_meeting_28_Feb_13.pdf).

- [17] K. Thomsen and P.A. Schmelzbach, "A dedicated beam interrupt system for the safe operation of the Megapie Liquid Metal Target," Utilisation and Reliability of High Power Proton Accelerators (HPPA5), Workshop Proc., Mol, Belgium, 6–9 May 2007.
- [18] K. Thomsen, "VIMOS, near-target beam diagnostics for MEGAPIE," *NIM A* **575**(3) (2007) 347. <http://dx.doi.org/10.1016/j.nima.2007.03.011>
- [19] W. F. Sommer et al., "Failure analysis of a radio-activated accelerator component," *Practical Failure Anal.* **3**(1) (2003) 71. <http://dx.doi.org/10.1007/BF02717412>
- [20] J. Holmes, "Quadrupole Strength Limits in the RTBT," SNS Tech. Note 162, 2005.
- [21] N. Catalan-Lasheras and D. Raparia, "The Collimation System of the SNS Transfer Lines," Proc. of the 2001 Particle Accelerator Conference, Chicago, p. 3263. <http://accelconf.web.cern.ch/AccelConf/p01/INDEX.HTM>
- [22] M. Kinsho, Private communication, 2014.
- [23] S. Childress, "NUMI Proton Beam Diagnostics and Control: Achieving 2 Megawatt Capability," Proc. HB2008 Workshop, Nashville, TN, USA, 25–29 August 2008, p. 475.
- [24] D. Adams, Private communication, 2014.
- [25] M. Plum, "SNS Injection and Extraction Systems - Issues And Solutions," Proc. HB2008 Workshop, Nashville, TN, USA, 25–29 August 2008, p. 268.



## Detection of Equipment Faults Before Beam Loss

*J. Galambos*

ORNL, Oak Ridge, TN, USA

### Abstract

High-power hadron accelerators have strict limits on fractional beam loss. In principle, once a high-quality beam is set up in an acceptable state, beam loss should remain steady. However, in practice, there are many trips in operational machines, owing to excessive beam loss. This paper deals with monitoring equipment health to identify precursor signals that indicate an issue with equipment that will lead to unacceptable beam loss. To this end, a variety of equipment and beam signal measurements are described. In particular, several operational examples from the Spallation Neutron Source (SNS) of deteriorating equipment functionality leading to beam loss are reported.

### Keywords

Beam-loss; high-power; high-intensity; proton; accelerator; machine-protection; equipment-failure.

## 1 Introduction

Prevention of beam loss is a primary concern for high-power, high-intensity proton machines, to avoid instantaneous damage and longer-term residual activation build-up. The typical rule of thumb for avoiding residual activation build-up is to maintain beam loss below 1 W/m (for beam energies above  $\sim 100$  MeV) [1]. For megawatt-level beams, this corresponds to a small fractional loss ( $10^{-6}$ ), which can be a challenge to measure, much less anticipate. Indeed, direct beam loss measurements are often the first sign of a developing equipment issue. Sudden and catastrophic equipment failures are easy to detect and diagnose, and result in direct shut-down of the beam. The more challenging task is to detect the slow gradual loss of equipment performance leading to very small impacts on beam transport, yet significant enough to increase the beam loss above the 1 W/m level. Detecting these slow and exceedingly slight equipment degradations is the subject of this paper.

Direct monitoring of the equipment directly related to beam transport is a straightforward method of anticipating issues that can lead to beam loss. This monitoring involves system measurements of magnets, power supplies, RF systems, vacuum, sources, and rotating equipment (pumps). Quantities that are monitored include temperature, voltage, and current. Examples of changes in these quantities that affect beam loss will be outlined in Section 3.

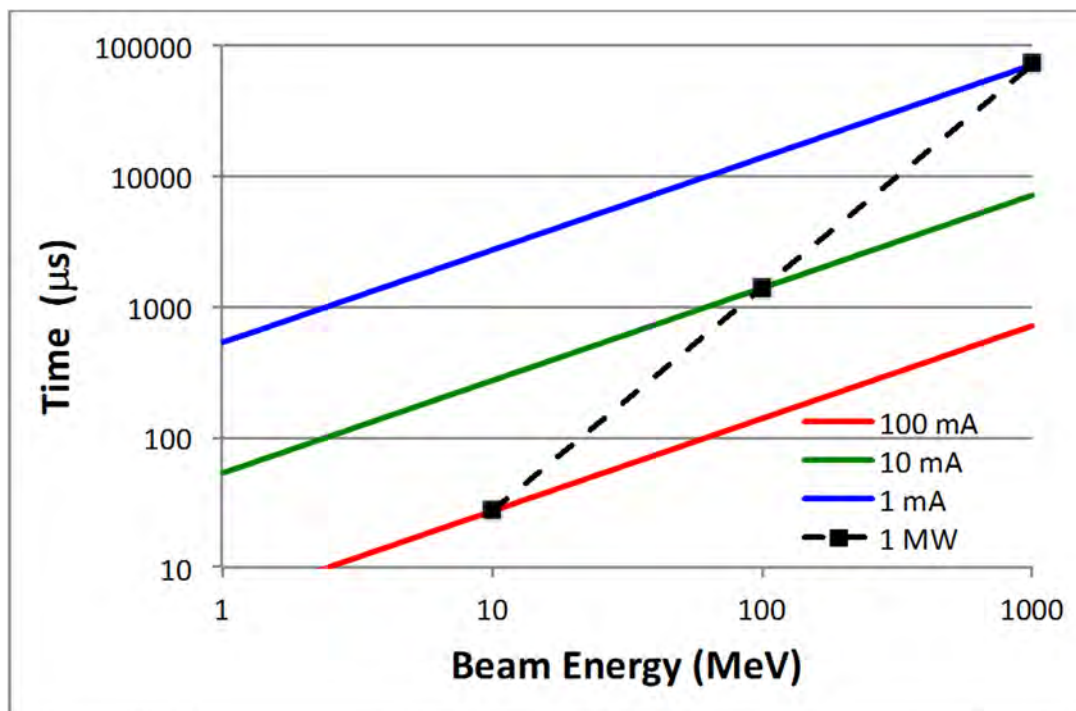
We note that the most sensitive measure of changes in the transport of high-power hadron beams is often beam loss measurement. It is often possible to continue running, even with a modest increase in beam loss (say from 0.1 to 0.2 W/m). Careful attention to changes in beam loss levels, even if they are at acceptable levels, is a valuable method of detecting incipient equipment degradation. The beam itself is a quite useful probe of equipment health. Finally, other beam measurements can be useful indicators of equipment health, as discussed in Section 4.

Most of the examples for beam loss and equipment issues are taken from the Spallation Neutron Source (SNS), which is a neutron scattering facility. It includes a high-power 1.4 MW proton accelerator [2].

## 2 Preparing for the beam

### 2.1 Reaction time-scales

Before discussing equipment failure precursors, it is useful to understand the reaction time-scales needed to protect equipment from gross beam loss. Figure 1 shows the time required for copper to increase in temperature by  $100^{\circ}\text{C}$  when subject to a  $1\text{ mm}^2$  cross-section proton beam of varying average current and energy. This can be thought of as a characteristic response time to protect the machine from catastrophic damage from the beam (or approaching the ‘melting metal’ stage). The reaction time is shorter at low energies, owing to the shorter beam penetration length. For average currents above  $10\text{ mA}$ , the reaction time is generally less than  $1\text{ ms}$ .



**Fig 1:** Time required for temperature of Cu to increase by  $100^{\circ}\text{C}$  when subjected to a  $1\text{ mm}^2$  proton beam of varying average current and energy. The dashed line indicates the  $1\text{ MW}$  beam contour.

Beyond catastrophic melting of the equipment, a major concern for high-power hadron accelerators is keeping beam loss below the level at which hands-on maintenance becomes problematic. As mentioned previously, while there is no absolute cut-off for this beam loss level, a general rule of thumb is  $1\text{ W/m}$  beam loss. To put this level of beam loss in perspective, the  $1\text{ MW}$  beam power contour is shown as a dashed line in Fig. 1. At this beam power, fractional loss should be maintained below  $10^{-6}/\text{m}$ . However, it takes a considerable period (e.g. days) of beam loss at this level to create long-lasting residual activation that inhibits maintenance. Detecting modest increases in beam loss (e.g. tens of percent) over long periods (e.g. days) can indicate an emerging equipment issue, without seriously jeopardizing machine health.

Because the build-up time for residual activation is so long, this provides some relief in the beam loss monitoring systems that protect against this hazard. This is a good circumstance, as beam loss detection for fractional loss at the  $10^{-6}$  level is often noisy. For example, with loss monitors near accelerating structures (e.g., a linac), there is often background X-ray generation that obscures the actual beam loss signal. A common practice for ‘slow’ protection against low levels of beam loss is to time average the loss signals to enhance the signal-to-noise ratio. Careful monitoring of the slow beam loss signals and correlating these with other equipment signals can shed light on emerging equipment issues. Some examples of this are shown later.

## 2.2 Slow protection

Machine protection systems will inhibit the beam if an unsafe situation exists, even before one attempts to start high-power operation. A primary example of this sort of protection is checking that all the vacuum system valves where the beam will be transported are in the open position. Other examples include ensuring that interceptive beam diagnostic devices (for example, wire scanners) are not inserted in any area in which high-power beams may be directed. These are straightforward equipment protection methods against catastrophic equipment damage, and are covered in other lectures in this series.

## 3 Measuring equipment

### 3.1 Electrical measurements

Particle accelerators use magnets as a primary means to guide and focus the particles. Typically, minimum stability control requirements on the magnetic fields are  $\sim 10^{-3}$  for single-pass systems (linacs) and  $\sim 10^{-4}$  for multipass systems (rings). Proper field levels are typically set up using beam methods. It is important to maintain the field levels at the desired values for long periods, after the set-up has been completed. Most magnets in accelerators are electromagnetic, and consist of multiturn windings to create the magnetic field. Ignoring hysteresis effects, field stability is controlled by maintaining a constant current through the circuit.

It is relatively straightforward to measure the current in a power supply, and appropriate control capability is specified when the power supplies are ordered. It is a simple matter to adopt software that monitors the power supply output to ensure that the current is at the appropriate set-point. Figure 2 shows an example of a normal quadrupole power supply current fluctuating with time, for a d.c. application.

Large accelerator facilities have large numbers of magnets to monitor, and it is standard practice to have automated software applications to ‘snapshot’ magnet parameters (e.g. current set-points and read-back values) when the machine is set up and operating well. These applications can also monitor these levels and report variations in live values relative to the ‘golden’ snapshot.

It is possible for the current read-back to be acceptable, but for there still to be a problem in maintaining the desired magnetic field. This can happen if some of the current is accidentally shunted around the desired current path, for example, if there is a turn-to-turn short in a multiturn magnet, or if there is a partial short to ground, as illustrated schematically in Fig. 3(a) (Fig. 3(b) shows an example cause of a ground fault interrupt). The latter example is referred to as a ground fault, and power supplies typically have a protective component called a ground fault interrupt to prevent damage of the cable or power supplies. However, it may be possible for a small amount of current to short-circuit the desired path through the magnet, and still be within the acceptable range of the ground fault interrupt comparator.

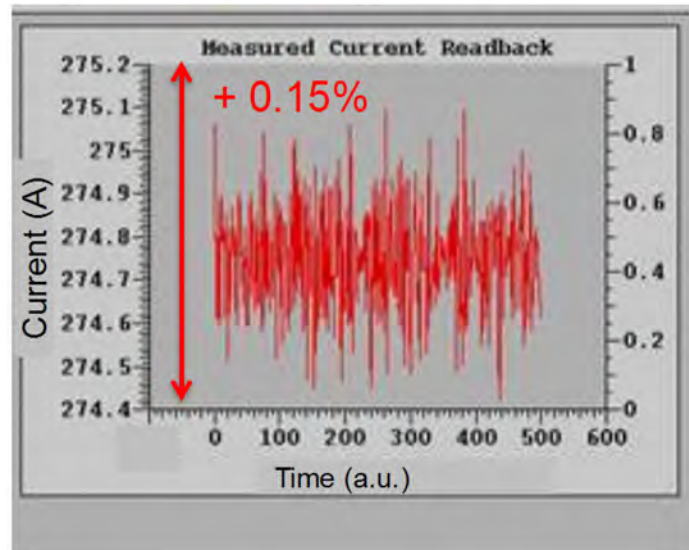
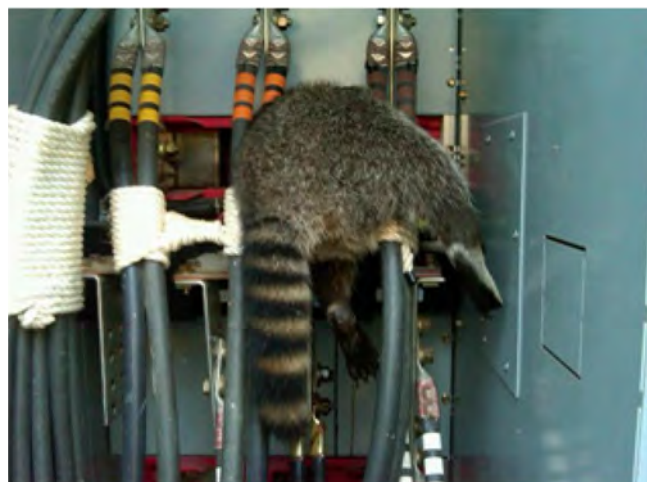
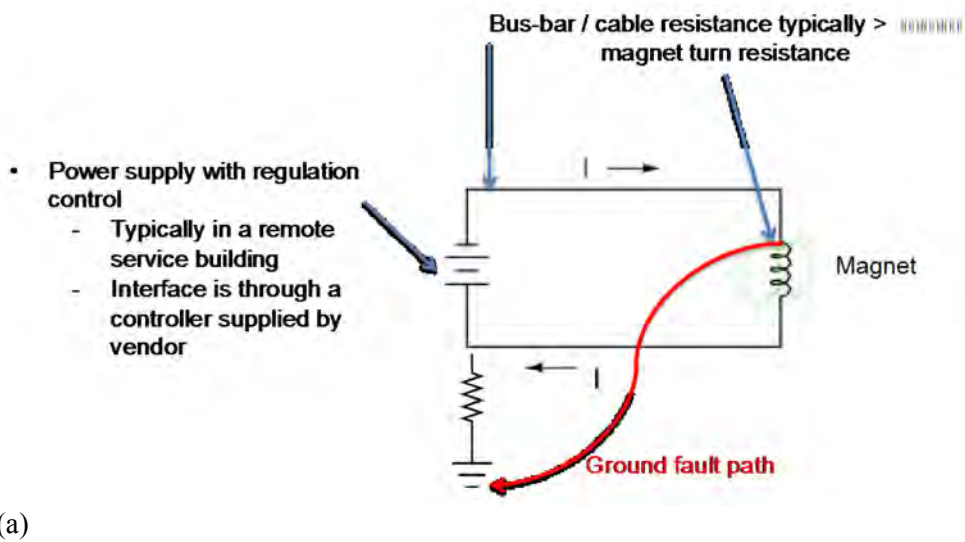


Fig. 2: Sample current fluctuation for a typical quadrupole power supply in a linac

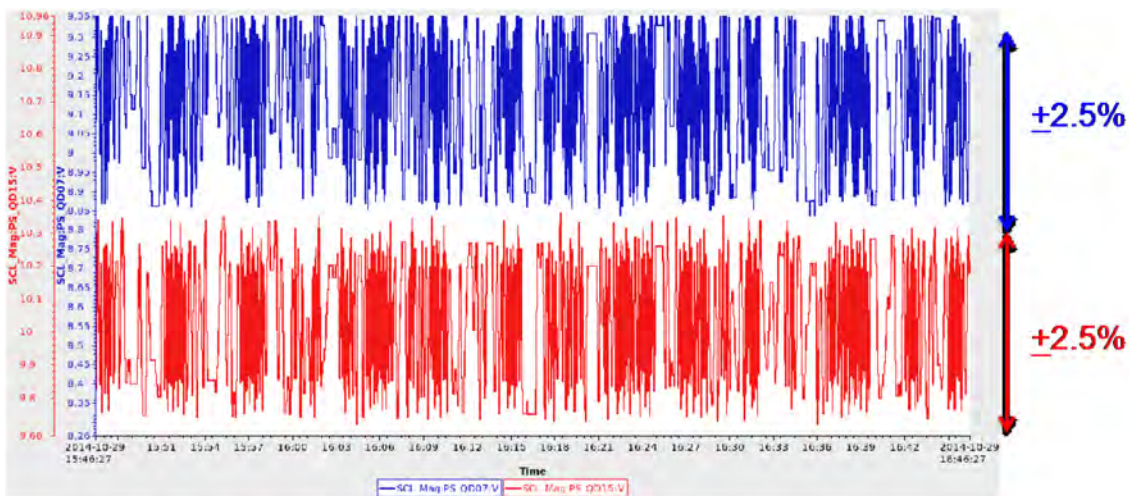


(b)

Fig 3: (a) Schematic of a ground fault that shunts part of the current around the desired path through a magnet. (b) Example cause of a ground fault interrupt.



For particle accelerators, the magnet power supplies are typically of the current-controlled type. A particular current is specified (corresponding to the desired magnetic field level), and the voltage is adjusted to produce the specified current. It is also possible to measure and monitor the voltage required to maintain the specified current. For d.c. magnets, the voltage is ideally constant; for cycled accelerators, the voltage can be quite complicated, but it should still follow the same pattern each cycle. A partial turn-to-turn short will slightly change the resistance across the magnet and affect the voltage required for the power supply to maintain the same current. However, typically in large accelerators, the resistance change is small compared with the overall circuit resistance (or impedance), which includes the effects from cables to and from the tunnel, and multiple magnets driven by the same power supply. The voltage change caused by a turn-to-turn short may not be detectable. Also, magnet power supplies are often not required to have tight tolerance on voltage read-back. Figure 4 shows a typical voltage read-back for two quadrupole power supplies in the SNS superconducting linac section. There is a large level of noise ( $>2\%$ ), which obviates the possibility of detecting small changes in the magnet coil resistance.

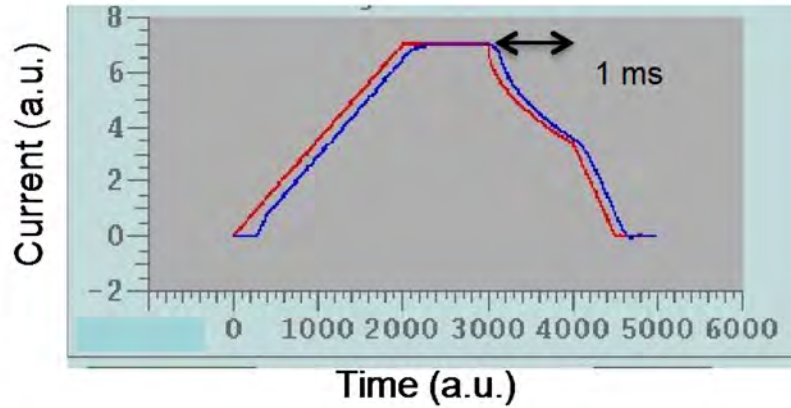


**Fig 4:** Time chart of the voltage read-back on two quadrupole power supplies (top and bottom plots) over  $\approx 1$  hour, indicating rather large noise fluctuations.

Although it may be difficult to detect equipment issues by voltage monitoring, as will be discussed in Section 4.1, it is possible to measure small changes in the applied magnetic field by monitoring changes in the beam trajectory.

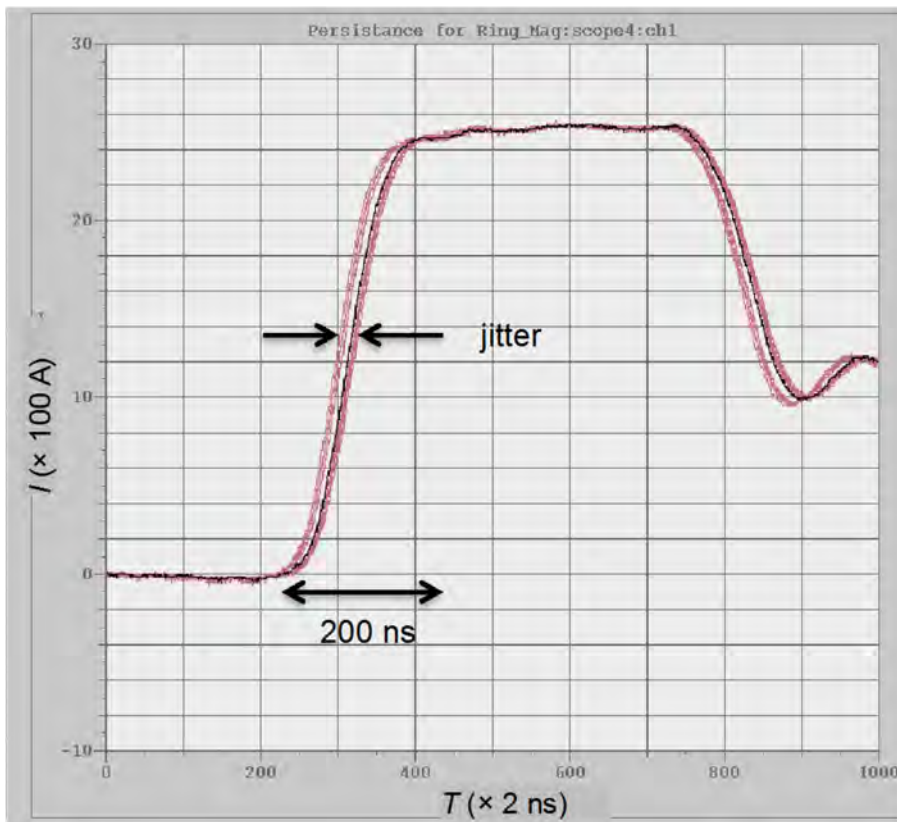
### 3.2 Pulsed magnets

As mentioned already, cycled accelerators have repeatable patterns that magnets follow each cycle. While these may be complicated, there is typically a pattern goal for the magnet current over each cycle. Figure 5 indicates a dipole magnet current over one cycle. In this case, it is for an injection kicker in the SNS ring, and occurs over a period of a few milliseconds. Note that there are two curves: a target waveform (red), and the actual measured current waveform (blue). It is possible to compare the two waveforms electronically and report an error or alarm if the difference exceeds a pre-set value. This is the typical procedure for pulsed systems.



**Fig. 5:** Current waveforms for a pulsed dipole magnet, with a pre-set target waveform (red, initially leading curve) and actual read-back waveform (blue, initially trailing curve).

Another example of a pulsed magnet is a fast kicker system. In this case, magnets reach full field in a fraction of a turn within a ring ( $\approx 200\text{--}300$  ns). The fields rise during a gap in the beam, so the precise waveform during the field rise is not important. The more critical issue in this case is the timing of the kicker firing, as premature or late firing results in the kicker affecting the beam, which is not in the gap. Figure 6 indicates a waveform display of a kicker in the SNS ring. There are actually  $\sim 6000$  waveforms displayed on the plot, but they fall into two families, and appear as only two waveforms. The appearance of multiple traces indicates the initiation of some drift in the kicker firing, which can be a precursor of an emergent kicker problem. Persistent displays of this type are useful for illustrating a drift in time of a pulsed quantity.



**Fig. 6:** A persistent display of a kicker waveform over many cycles, indicating a drift of the kicker firing

### 3.3 RF system monitoring

High-power RF systems are a major component of high-power accelerators. There are several linked subsystems within a pulsed high-power RF installation, as illustrated in Fig. 7. Power from the grid is converted from a.c. to high-voltage d.c. in a rectifier. For pulsed systems, the d.c. power is converted by some sort of a pulsed forming network to pulsed high-voltage d.c. waveforms, which are used to power an RF source. The RF power generated in the source is finally transported to the accelerating structure. A crucial part of the overall system is the low-level RF (LLRF) system, which coordinates the timing and amplitude control of the delivered RF power very precisely with the beam in the structure. Any variation or drift in the equipment of these components can cause beam loss. Some examples of identifying causes of beam loss due to RF equipment issues are shown here.

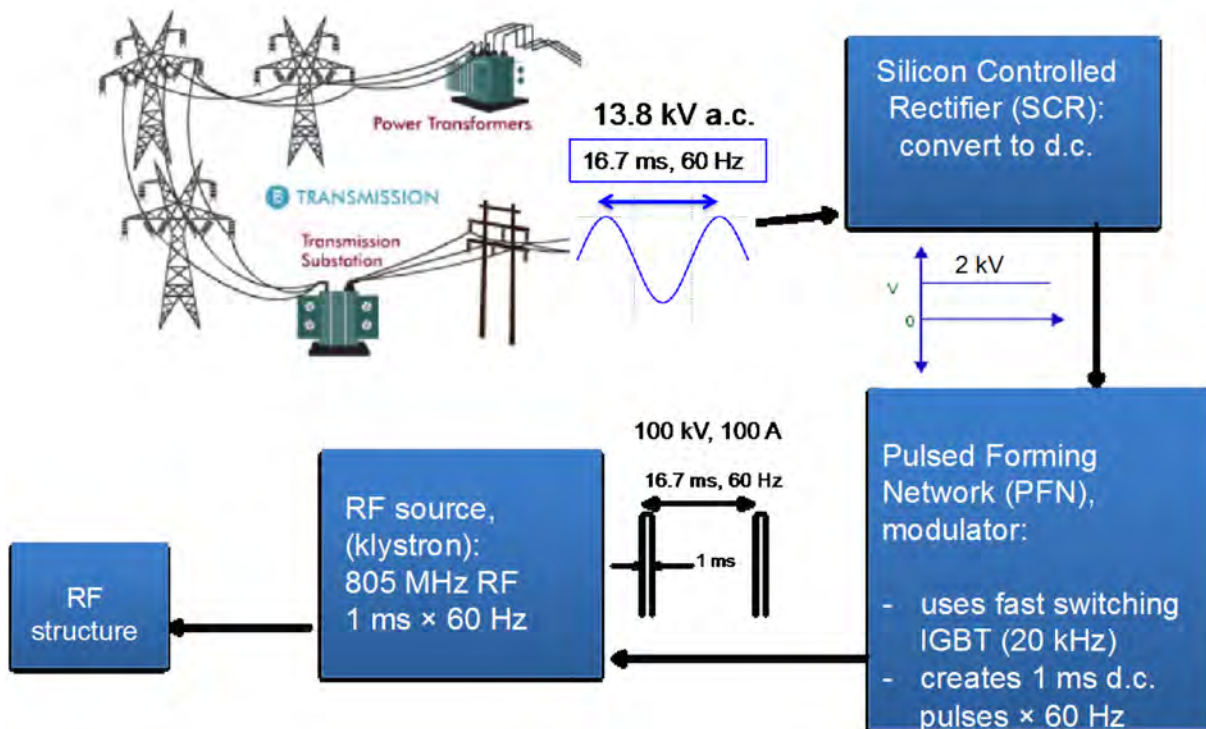
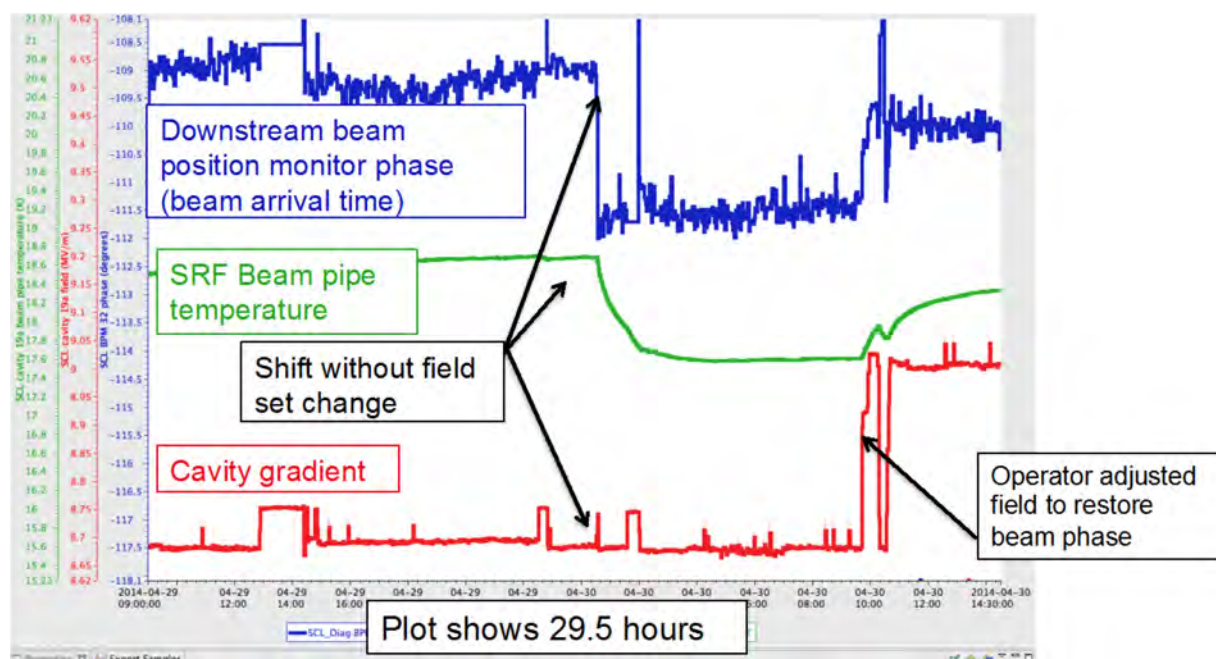


Fig. 7: Schematic of the components of a high-power RF installation

#### 3.3.1 LLRF issue

Figure 8 indicates a time history over  $\approx 30$  hours of: (1) the field gradient in a superconducting linac cavity (red), (2) the temperature of a downstream cavity beam-pipe (green), and (3) the beam arrival time in a downstream beam position monitor (blue). Near the middle of the time period, there is a sudden change in the beam-pipe temperature and arrival time of the beam, indicating a change in the beam acceleration. The beam loss monitors along the linac did not report a meaningful increase in beam loss, so the beam kept running. About 4 hours later, an operator noticed the change in the downstream beam arrival time and artificially increased the cavity gradient to restore the beam arrival time to the previous value. Subsequently, the beam-pipe temperature also returned to its previous level. In this case, the root cause of the observed changes was an electronic component failure in the LLRF system, resulting in a change in the field regulation, which the operator compensated for. The cavity gradient change was small (a few percent) so it was possible to continue running, but this was an indication of an equipment issue, which needed addressing. The elevated beam-pipe temperature was due to elevated beam loss, even though it was not detectable on loss monitors in this case.

This issue was diagnosed by correlating different signal changes along a timeline. Control systems have tools to perform this task, with both live streaming and archived data. This is a common technique for diagnosing previously unseen issues.

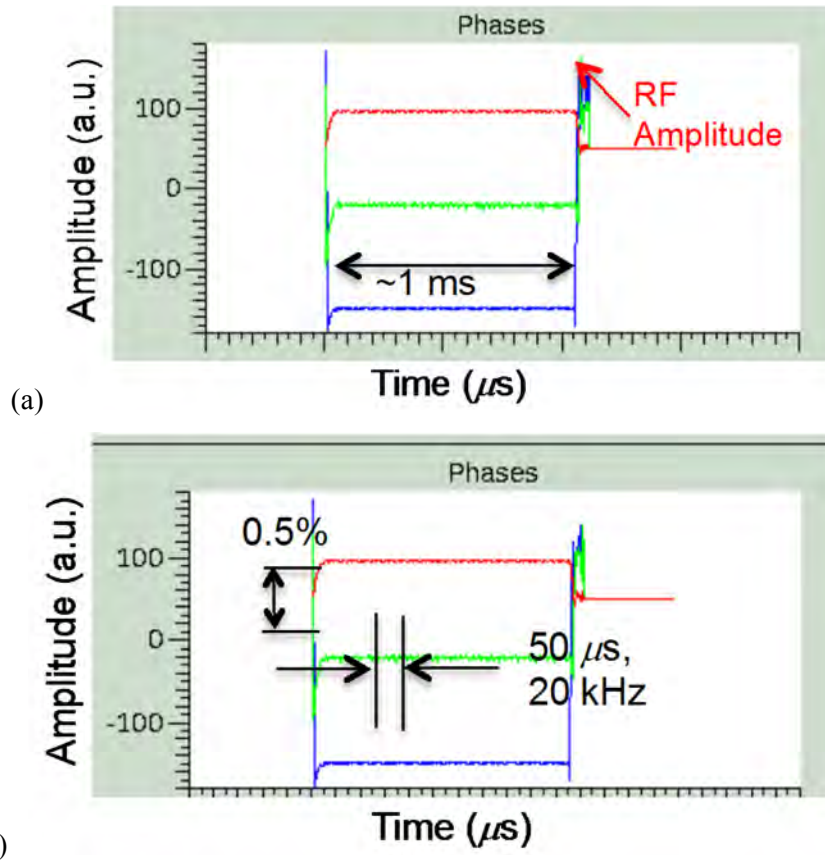


**Fig. 8:** Timeline indicating a shift in LLRF performance. The bottom curve is the cavity gradient, the middle curve is the beam pipe temperature, and the top curve is a downstream loss monitor signal. The entire time span shown is 29.5 hours.

### 3.3.2 Fast time-scale monitoring

Another technique for diagnosing RF system equipment issues is that of monitoring fast time-scale waveforms of the RF system. Figure 9(a) shows the LLRF amplitude waveform output for a 1 ms pulse in a copper cavity structure. Figure 9(b) shows a zoomed-in view of the amplitude axis. While the zoomed-out view indicates a fairly nice looking waveform, the zoomed-in image shows some amplitude noise at about the 0.5% level. This jitter is within the acceptable control margin, and is not, in itself, a concern. Analysis of the structure of the amplitude noise reveals a 20 kHz frequency component. It turns out that the pulse-forming network (PFN) uses 20 kHz solid-state switching technology to provide the high-voltage drive for the klystrons powering this cavity.

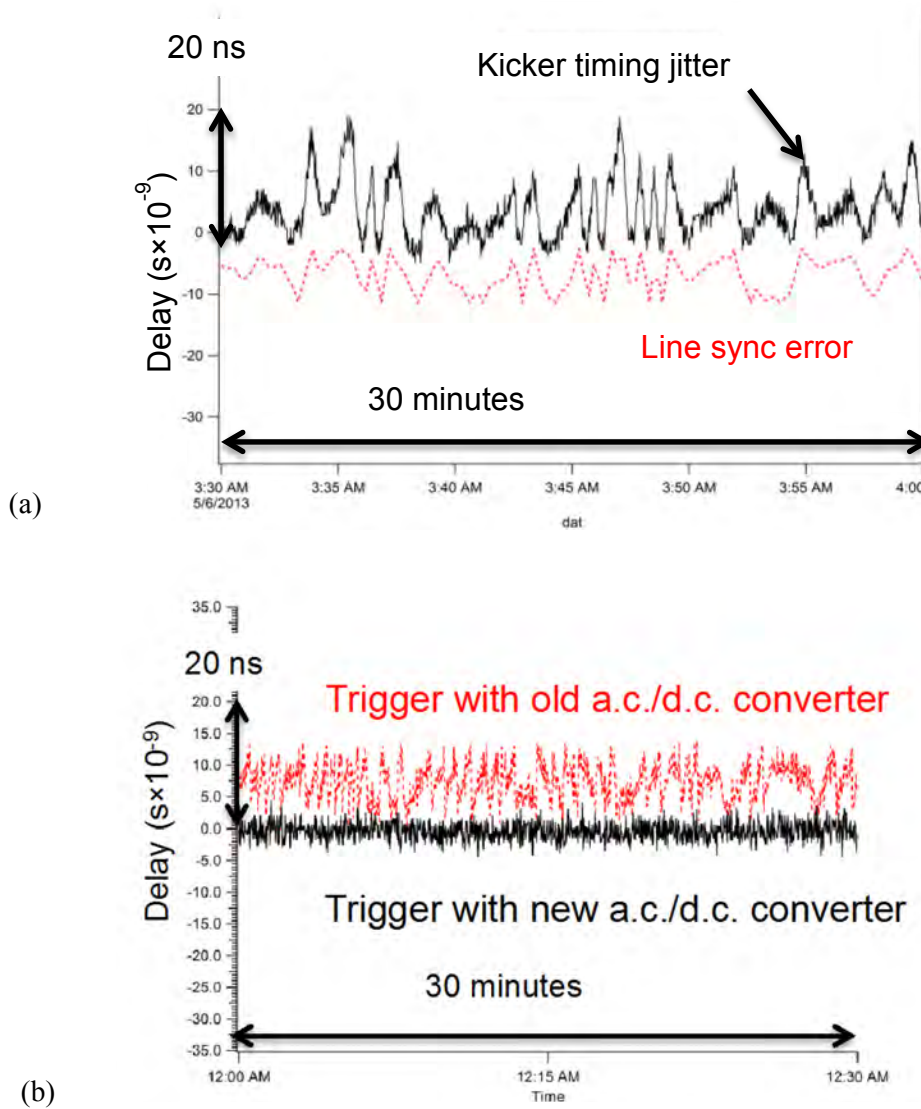
This amplitude fluctuation is an indicator of the PFN health, and can be monitored. If this ripple increases, it could become a source of beam loss, so maintenance or adjustment of the PFN unit can be identified from the downstream LLRF measurement. It is useful to have automated systems to monitor the quality of the RF waveforms and report cavities that exceed a permissible threshold of amplitude or phase variation.



**Fig. 9:** (a) A waveform display of the RF amplitude (top curve in (a)) of a copper structure cavity. (b) Enlargement of the same trace, revealing 20 kHz noise.

### 3.3.3 Line-sync issues

High-power accelerators are large electrical power consumers. Historically, pulsed accelerators were designed to run at harmonics of the electrical grid frequency, to be able to ‘ride along’ at a constant phase of the electrical grid a.c. power cycle. Although electrical grids are referred to as operating at 60 Hz, or 50 Hz, there are constant slight variations in the electrical power generation frequency to match the demand load. Previous-generation pulsed accelerators would adjust the beam pulse timing to follow the grid frequency, so as to maintain a constant phase offset from the peak of the voltage cycle. With the advent of solid-state fast switching technology to provide PFN capabilities, modern electrical systems are generally not sensitive to when the beam pulse occurs relative to the grid power cycle. This enables accelerators to run at constant frequency, and even at a frequency that is not a harmonic of the grid. However, there can be issues if the grid a.c. frequency effects ‘leak-through’ to the beam in unanticipated ways. An example is shown in Fig. 10(a), which shows jitter in a kicker timing signal (black trace) and the variation of the beam trigger with grid line cycle peak (red trace). There is a weak correlation in these traces, which led to the investigation of the kicker timing unit. Indeed, a low-cost a.c.–d.c. transformer was found to be performing below its specification (with an unacceptable a.c. component leading through to the d.c. signal). Although the beam loss was acceptable, if the issue progressed, it would become an issue. Figure 10(b) shows the improved-stability timing signal after replacement of the faulty a.c.–d.c. conversion unit.



**Fig. 10:** (a) Jitter in the timing signal of a kicker magnet (black) with a correlation in the line-synch phase variation (red). (b) Jitter in the timing signal with the faulty a.c.–d.c. convertor (red) and with a replaced unit (black).

### 3.4 Vacuum

Most high-power accelerators require high vacuum to maintain low beam loss. Beam scattering and charge state changes (e.g. e-stripping) are examples of the loss mechanisms that cause beam loss [3]. Figure 11 indicates a clear response of downstream losses in a superconducting linac section, from a purposeful change in an upstream copper linac section vacuum level (caused by turning off vacuum pumps). While there is a clear dependence of loss on increasing vacuum levels, there is also another loss component. Monitoring the health of the vacuum systems is clearly an important part of preventing beam loss. Figure 12 shows the time history of the vacuum in a transport line of the SNS accelerator. There are fairly regular vacuum ‘spikes’ in the first half of the display, but these can be ignored. They happen regularly (e.g. related to beam trips and related loss). However, in the second half of the display, a sustained increase in vacuum baseline is observed, and is cause for implementing vacuum pump repair work. In this case, there was not a serious beam loss increase, but this is a precursor to more serious issues.

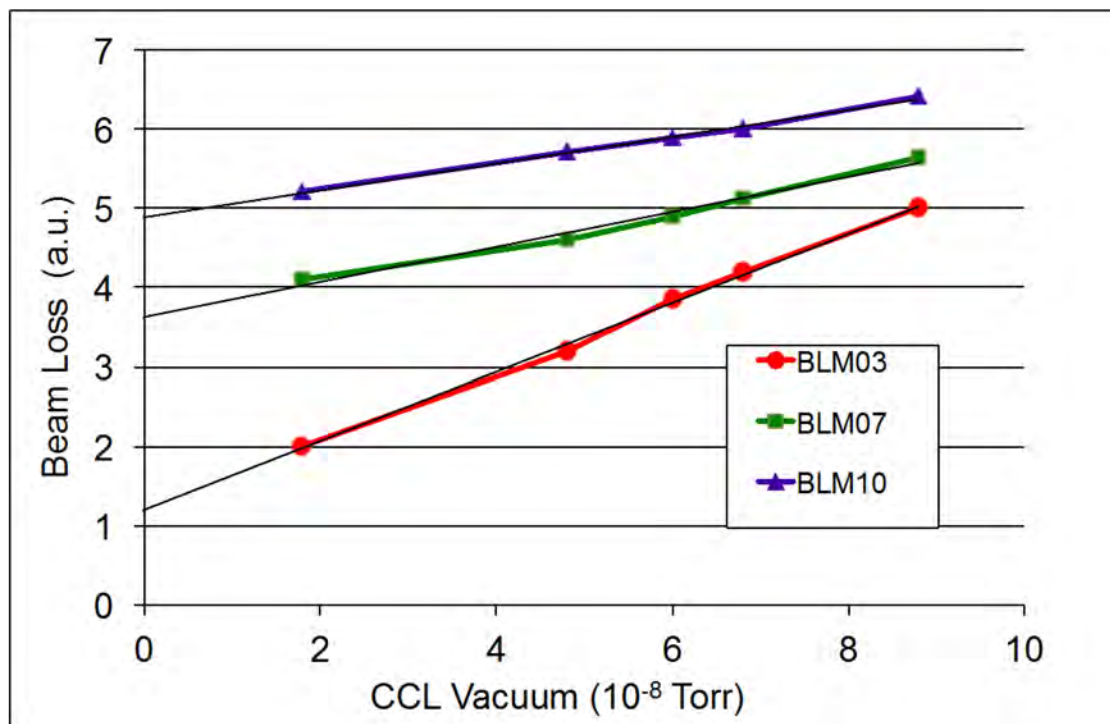


Fig. 11: Variation of downstream beam loss measurements with a purposeful variation of vacuum in an upstream region.

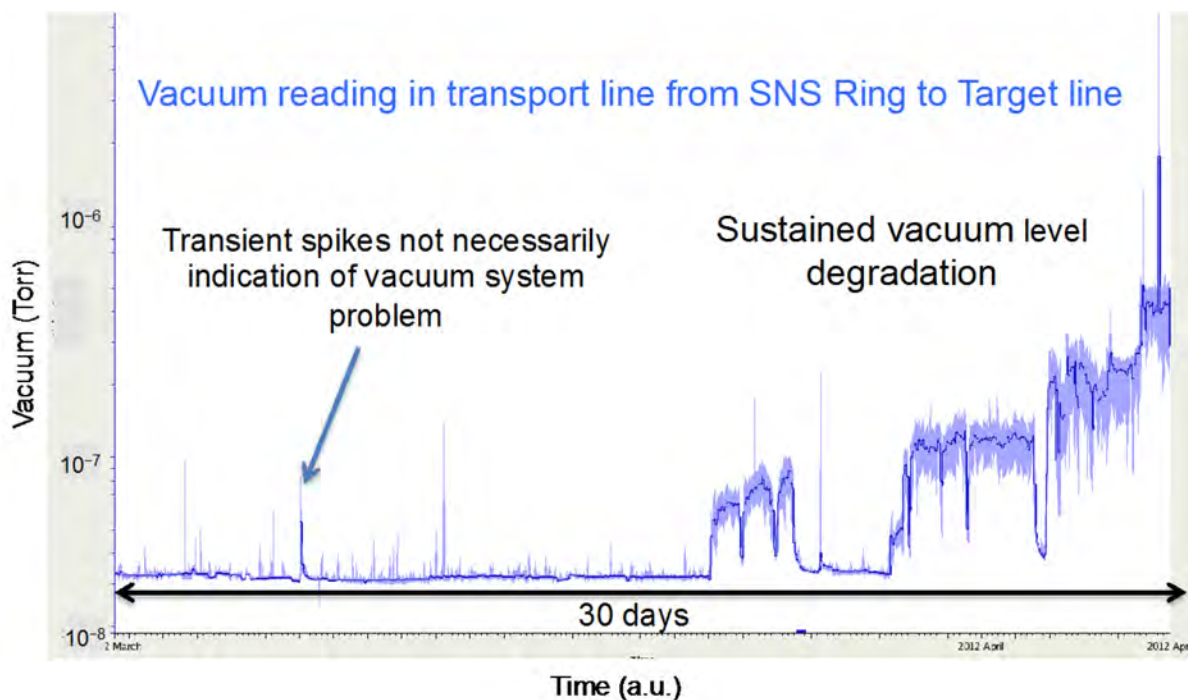


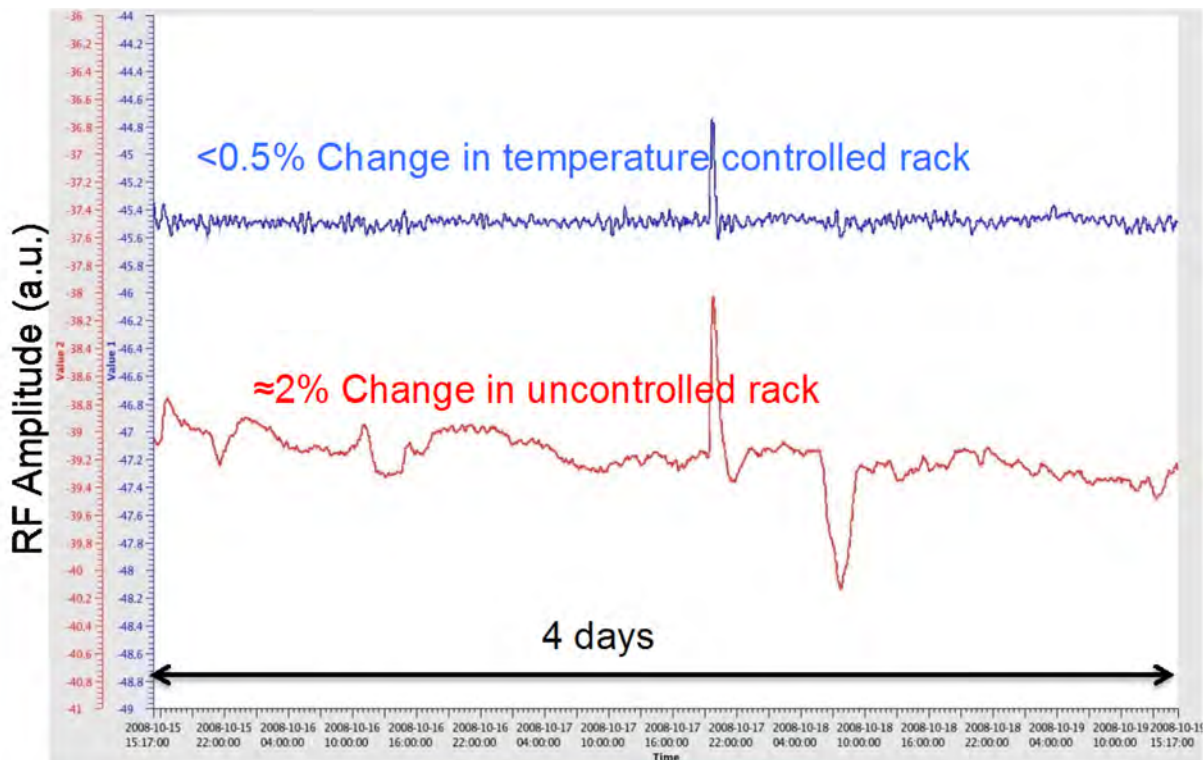
Fig. 12: One-month time history of the vacuum level in a transport line section, indicating a developing issue

### 3.5 Temperature monitoring

Measuring ambient air and equipment temperatures can be a useful tool in identifying equipment issues that cause beam loss. Much modern accelerator equipment is controlled by sensitive electronics, which may include temperature-sensitive components. For example, Ref. [4] describes how order of magnitude

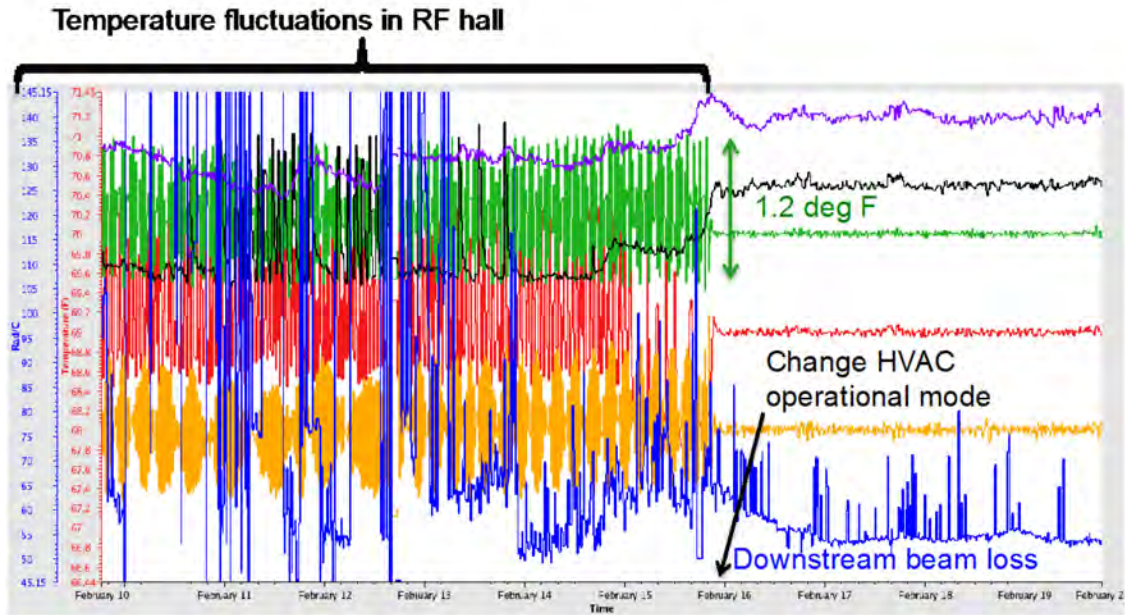
variations in the thermal stability of the LLRF analogue front end board stability were measured, dependent on the particular electrical components used on the board. Figure 13 shows the RF field amplitude stability over several hours for two cavities with similar LLRF electronics. The unit in a rack with temperature control stabilization shows an acceptable drift in the field level, whereas the unit without temperature control shows an unacceptable drift. Temperature control stabilization in the racks was instituted to alleviate the drift, as indicated in Fig. 13. The better solution is to use electronic components that are not temperature sensitive, but this is not always possible.

Another temperature sensitivity of concern is the change in cable lengths with temperature. For some critical applications, such as RF reference lines, steps are taken to minimize this effect (temperature control of the reference line). But it is prohibitively expensive to control the temperature of all cables, and some beam instrumentation cables and LLRF cables can be many tens of metres long, and subject to slight length changes, which can cause changes of a few degrees Fahrenheit in the RF phase control for high frequency ( $\sim$ GHz) systems. As an example of this effect, Fig. 14 shows the SNS klystron gallery building temperature (measured at several locations) and a downstream linac beam loss, measured over about 10 days. For the first week of the display, a new heating ventilation and air conditioning (HVAC) operational mode was attempted, which resulted in continuous temperature fluctuations of  $\sim 1^\circ\text{F}$  throughout the building (this is the building that houses the RF equipment). Finally, the old control mode was re-established, and both building air temperature fluctuations and unexplained beam loss fluctuations disappeared.



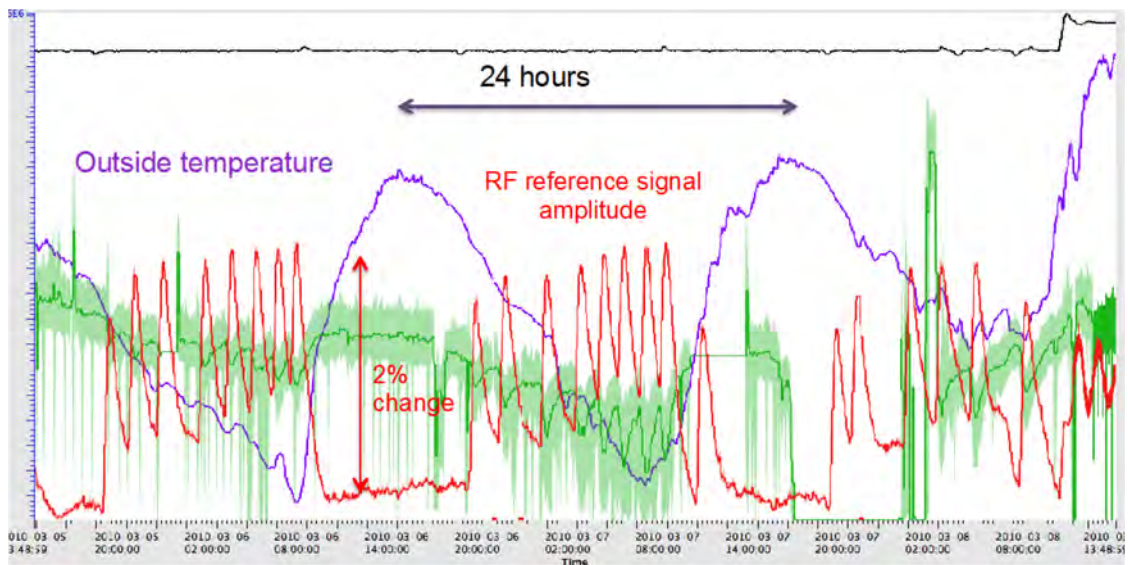
**Fig. 13:** Drifts in the RF field amplitude in a cavity with temperature-controlled LLRF electronics (blue) and in a cavity with similar LLRF electronics in a rack without temperature control (red).





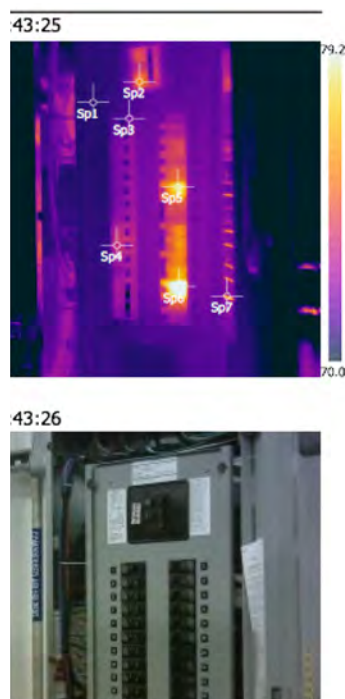
**Fig. 14:** Ten day period during which a new HVAC control mode was attempted, resulting in  $\sim 1^\circ\text{F}$  fluctuations within the linac RF gallery [green (third curve from top), red (third curve from bottom), and orange (second curve from bottom) traces], causing unstable beam loss in the linac (bottom blue) trace.

Another example of a subtle temperature change affecting beam parameters is shown in Fig. 15, which shows parameter variations over  $\approx 3$  days. In this case, unexpected erratic behaviour was observed in the beam injection area of the SNS accumulator ring (green curve labelled ‘beam missing foil’). An operator noticed a correlation of this parameter with the variation in amplitude of the linac RF reference signal (which should be constant). The erratic behaviour only occurred at night, when the building, which houses the reference line signal generator, was cooler. Replacing the reference line signal generator solved the issue, as it had developed a temperature sensitivity. This is another example of how searching for correlations helped identify the source of the equipment problem. The associated beam issues were never severe enough to stop beam operation in this case.



**Fig. 15:** Diurnal temperature variation leading to unexpected behaviour in RF reference line control and a related change in beam parameters.

Finally, we should note that it is often useful to monitor equipment temperature directly, even if direct temperature sensors are not available. Infrared imaging is a useful technique to provide highly localized temperature information. An example is shown in Fig. 16. This technique is useful for periodic monitoring of circuit breakers and magnet cable connections, as loose connections generate heat, which leads to equipment degradation or failure.



**Fig. 16:** Thermal (top) and visual (bottom) images of the same breaker panel. The thermal image shows hot spots (e.g. Sp2, Sp6, and Sp7) not seen in the visible spectrum image.

## 4 Beam measurements

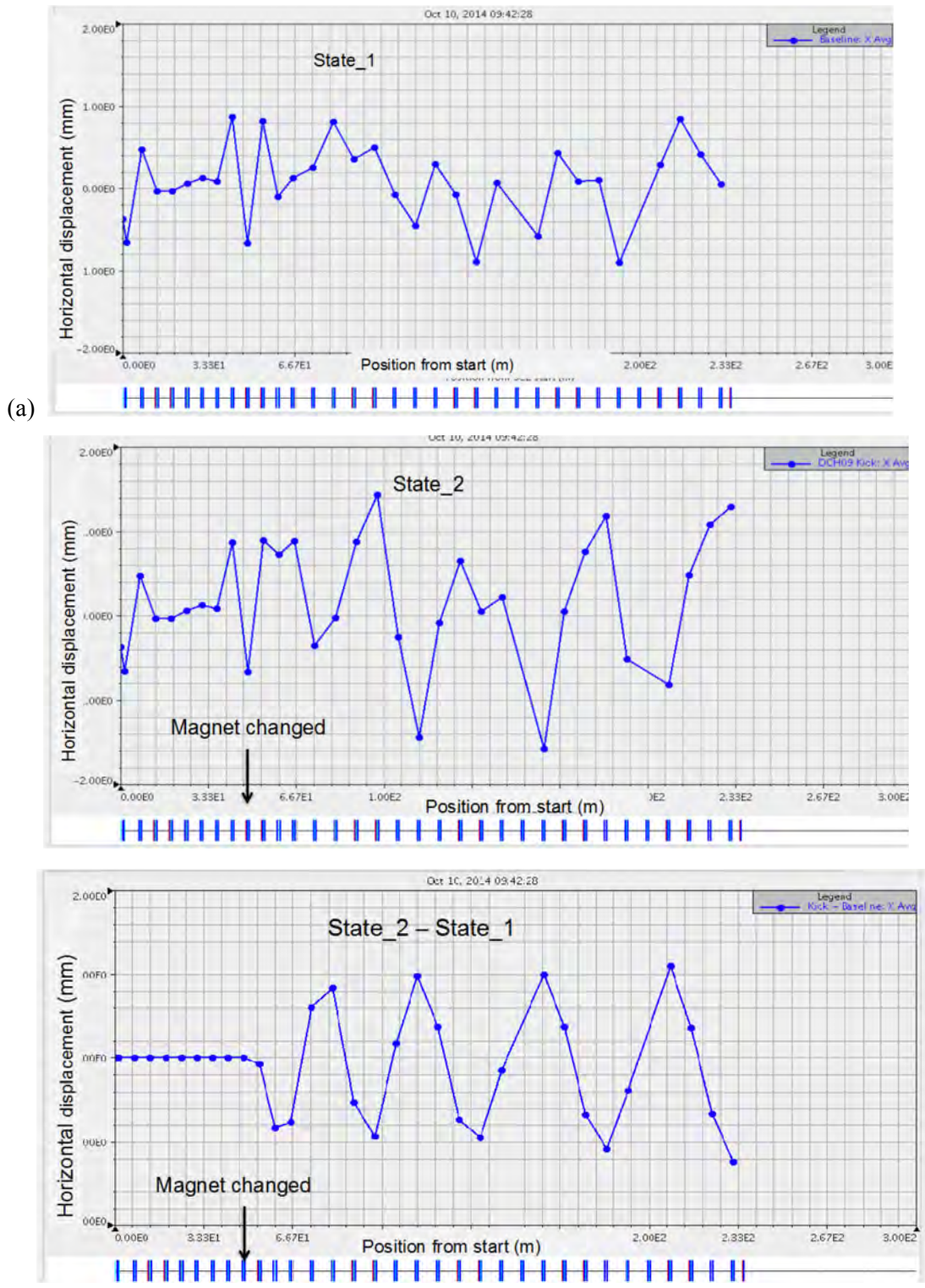
The measurements described so far have focused on monitoring signals from equipment or buildings to understand emergent issues that may lead to beam loss. In addition, beam signals can be useful for identifying equipment issues. The beam has perhaps the most sensitive response to small changes in the equipment or external environment.

### 4.1 Transverse beam measurements

Measuring changes in the transverse beam position is the most direct method to detect issues with equipment that affect the transverse beam position (e.g. magnets and RF devices for hadron beams). Figure 17(a) shows a typical beam trajectory along a linac. Ideally the trajectory lies perfectly along the axis; however, there are typically slight imperfections. This trajectory is somewhat chaotic, but acceptable. Figure 17(b) indicates the same beam trajectory, with a slight change to a steerer at the location indicated. It is not obvious from this image alone where the trajectory change originated. However, a plot of the difference between the two trajectories indicates a wave in the beam motion beginning where the steering changed. Orbit (trajectory) difference techniques are powerful methods of identifying changes in steering magnets and RF or quadrupole elements if the beam is even slightly off axis. Sometimes, the settings of the magnet or RF may not have changed. For example, if a slight turn-to-turn short is developing, the applied magnetic field can change slightly, even though the magnet current settings have not been changed. Also, if the LLRF control is not working properly (e.g. as shown in Fig. 8), a trajectory change may be observable.

# DETECTION OF EQUIPMENT FAULTS BEFORE BEAM LOSS

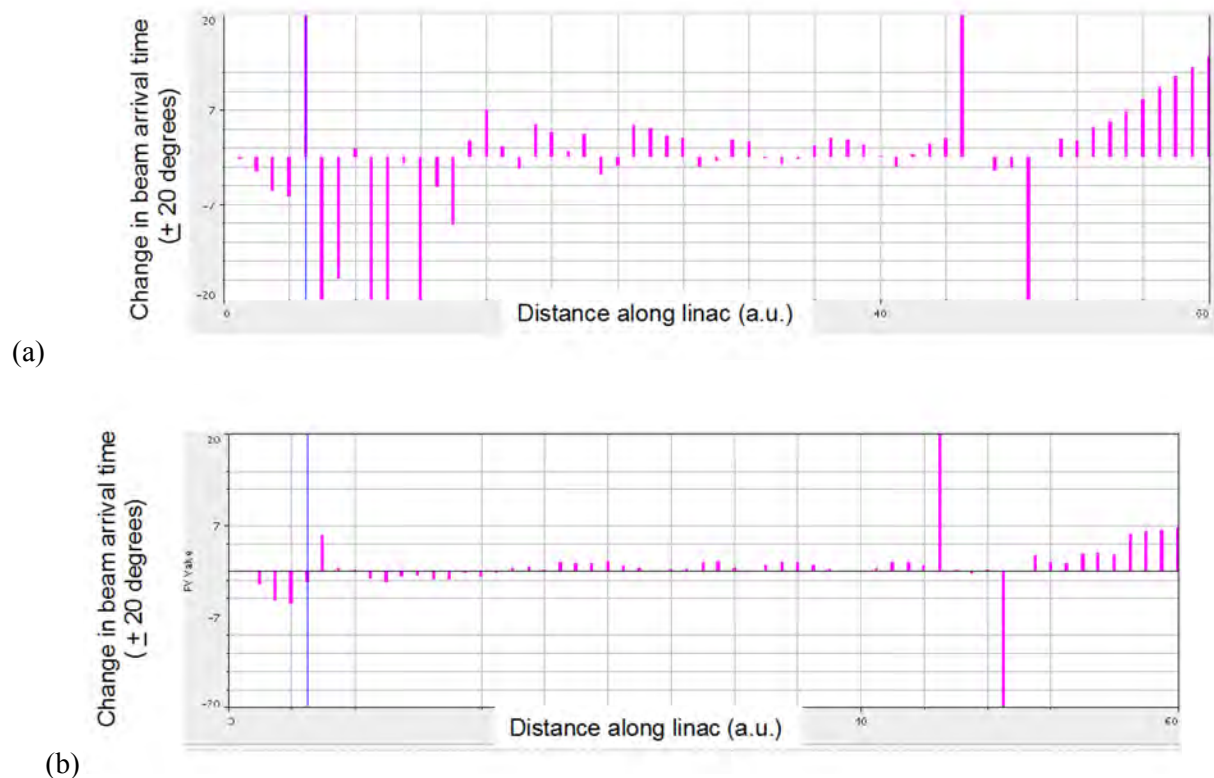
The key to applying this method is to take a snapshot of the beam transverse trajectory along the accelerator during the set-up period, when the quality is known to be in a good state. Software can be provided to highlight changes in the saved ('golden') and live beam trajectories along the accelerator.



**Fig. 17:** (a) Original beam trajectory along the horizontal axis. (b) Trajectory after a change in the steering at the indicated location. (c) Difference between the trajectories of (b) and (a). (Magnets are displayed synoptically below the x-axis).

## 4.2 Longitudinal beam measurements

The method described in Section 4.1 can also be applied in the longitudinal direction. In the longitudinal plane, the longitudinal ‘position’ is typically recorded as the beam arrival time relative to a reference RF signal, often in units of ‘degrees’ of the reference RF signal. The measuring device may be a beam position monitor (if properly designed for this capability), or an RF resonator of some sort. Figure 18(a) displays the change in beam arrival time along the SNS linac from the saved beam set-up values. A wave is evident, starting from the beginning (the magnitude of the wave changes along the linac in this case, because the units of the arrival time change, and are not corrected here). This provided information as to where the problem originated, and an operator adjusted the phase of the first cavity in the linac, resulting in the difference plot shown in Fig. 18(b), which shows that the beam trajectory is much closer to the set-up conditions. This enabled continued beam running until a maintenance day, when a poor LLRF cable contact was discovered at the first cavity.



**Fig. 18:** (a) Change in beam arrival time along the SNS linac from beam set-up conditions. (b) Same plot as (a), after adjusting the phase of an RF structure at the start of the linac.

## 5 Summary

Identifying emergent equipment issues before a problem results in unacceptable beam loss is a challenging task. For high-power accelerators, beam loss becomes intolerable at quite small fractional levels, but it is possible to operate with some small level of beam loss increase. The challenge lies in identifying the equipment that causes the change in beam loss. Different equipment properties have been identified for monitoring, and example techniques for monitoring these quantities are shown. A key element in these techniques involves identifying correlations between changes in beam parameters and equipment parameters to identify the culprit driver for the onset of beam loss increase. Finally, using direct beam measurements to localize the source of equipment issues was described.

## Acknowledgements

ORNL is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the US Department of Energy.

Notice: This manuscript has been authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the US Department of Energy. The US Government retains, and the publisher, by accepting the article for publication, acknowledges, that the US Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US Government purposes.

## References

- [1] J. Alonso, Beam loss working group report, 7th ICFA Mini-workshop on High Intensity High Brightness Hadron Beams, Sept. 1999, Lake Como, Wisconsin, p. 51.  
<https://inspirehep.net/record/537420>
- [2] S. Henderson *et al.*, *Nucl. Instrum. Methods Phys. Res. A*, **763** (2014) 610–673.  
<http://dx.doi.org/10.1016/j.nima.2014.03.067>
- [3] M. Plum, Challenges facing high power proton accelerators, Proc. IPAC 2013, Shanghai, China, 2013. <http://accelconf.web.cern.ch/AccelConf/IPAC2013/papers/moxbb101.pdf>
- [4] M. Crofford *et al.*, SNS LLRF temperature dependence and solution, Proc. Linear Accel. Conf. LINAC2010, Tsukuba, Japan, 2010.  
<http://accelconf.web.cern.ch/AccelConf/LINAC2010/papers/mop088.pdf>



## Controls and Machine Protection Systems

*E. Carrone*

SLAC National Accelerator Laboratory, Menlo Park, CA, USA

### Abstract

Machine protection, as part of accelerator control systems, can be managed with a ‘functional safety’ approach, which takes into account product life cycle, processes, quality, industrial standards and cybersafety. This paper will discuss strategies to manage such complexity and the related risks, with particular attention to fail-safe design and safety integrity levels, software and hardware standards, testing, and verification philosophy. It will also discuss an implementation of a machine protection system at the SLAC National Accelerator Laboratory’s Linac Coherent Light Source (LCLS).

### Keywords

MPS; Functional Safety; PLC; SIL; Control Systems; Cyber Security.

## 1 A software problem

On 4 June 1996, the maiden flight of the Ariane 5 launcher ended in a failure. Only 39 s after initiation of the flight sequence, at an altitude of about 3700 m, the launcher veered off its flight path, broke up, and exploded.

During those first 39 s, the software generated a number too large for the system to handle: the computer shut down and passed control to its redundant twin, which, being identical to the first, came to the same conclusion and shut down a few milliseconds later. The rocket, now without guidance, changed direction to compensate for an imagined error and collapsed in its own turbulence.

In general terms, the flight control system of the Ariane 5 is of a standard design. The attitude of the launcher and its movements in space are measured by an inertial reference system. It has its own internal computer, in which angles and velocities are calculated on the basis of information from an inertial platform, with laser gyroscopes and accelerometers. The data from the inertial reference system are transmitted through the databus to the onboard computer, which executes the flight program and controls the nozzles of the solid boosters and the Vulcain cryogenic engine, via servo valves and hydraulic actuators.

To improve the reliability of such a system, there is considerable redundancy at the equipment level: two inertial reference systems operate in parallel, with identical hardware and software. One inertial reference system is active and one is in ‘hot’ standby; if the onboard computer detects that the active inertial reference system has failed, it immediately switches to the other one, provided that this unit is functioning properly. Likewise, there are two onboard computers, and a number of other units in the flight control system are also duplicated.

The launcher started to disintegrate at about 39 s into operation because of high aerodynamic loads due to an angle of attack of more than 20° that led to separation of the boosters from the main stage, in turn triggering the self-destruct system of the launcher. This angle of attack was caused by full nozzle deflections of the solid boosters and the main engine.

These nozzle deflections were commanded by the onboard computer software on the basis of data transmitted by the active inertial reference system. Part of these data at that time did not contain proper

flight data, but showed a diagnostic bit pattern of the computer of the inertial reference system 2, which was interpreted as flight data. The reason that the active inertial reference system 2 did not send correct attitude data was that the unit had declared a failure due to a software exception.

The onboard computer could not switch to the back-up inertial reference system 1 because that unit had already ceased to function during the previous data cycle (72 ms) for the same reason as inertial reference system 2.

The internal inertial reference system software exception was caused during execution of a data conversion from 64-bit floating point to 16-bit signed integer value. The floating point number that was converted had a value greater than could be represented by a 16-bit signed integer. This resulted in an operand error.

Among the causes:

- software reused from the Ariane 4 series (a rocket with different requirements);
- an error while converting a 64-bit floating point number to a 16-bit integer caused an overflow, a custom floating point format for which the processor could have generated an exception error;
- some operations (in Ada code) on the computers are protected from bad conversions, but one was disabled;
- the primary inertial sub-computer and its back-up both shut down because of this, and the primary sub-computer started a memory dump;
- the main computer looked at the data dump and interpreted it as flight data. The nozzles swivelled to their extreme position to try to ‘right’ the rocket, causing it to break apart.

The investigation committee issued many recommendations.

- No software function should run during flight unless it is needed.
- Prepare a test facility including as much real equipment as technically feasible, inject realistic input data, and perform complete, closed-loop, system testing. Complete simulations must take place before any mission.
- Organize, for each item of equipment incorporating software, a specific software qualification review. Make all critical software a configuration controlled item.
- Review all flight software (including embedded software) and, in particular, identify all implicit assumptions made by the code and its justification documents on the values of quantities provided by the equipment. Check these assumptions against the restrictions on use of the equipment.
- Include participants external to the project when reviewing specifications, code, and justification documents. Make sure that these reviews consider the substance of arguments, rather than checking that verifications have been made.
- Give justification documents the same attention as code.

Many of these recommendations are applicable to accelerators. In this paper, we will discuss procedures, systems, and techniques to handle and mitigate the risks related to designing, deploying and operating machine protection systems.

### 1.1 Accelerator controls are complex systems

Control systems comprise many parts, and opportunities for malfunctioning are everywhere, e.g.:

- software fails unsafe;
- hardware fails unsafe;



- changes made on the wrong version of a program;
- wrong data received from sensors (but interpreted as true);
- a system was changed and cannot be brought back to a previous state;
- a system needs to be upgraded or changed, but there is not enough documentation to do it;
- system compromised by a malicious piece of code, which may go unnoticed for a long time;
- system hacked into.

Given this scenario, some risk mitigation strategies are:

- redundancy;
- life cycle management;
- fail-safe design;
- configuration control;
- quality assurance and quality control;
- standards;
- tests;
- documentation;
- cybersafety.

The concept of ‘functional safety’ is the corpus of concepts, processes, and guidelines that will enable us to mitigate those risks.

## **2 Functional safety**

### **2.1 Introduction**

For classic electrical and electronics based systems, there are three ways to improve safety: reduce component failure rate, increase diagnostics, and employ redundancy. Modern electronics, such as programmable logic controllers, microcontrollers, field programmable gate arrays and application-specific integrated circuits are powerful enough to be used to implement complex diagnostic schemes and control strategy reconfiguration on fault detection; in many cases, redundancy can be implemented with little cost increase. However, these features come at the cost of increased hardware complexity and introduction of software, which is more difficult to verify and validate for safety applications.

### **2.2 Life cycle management**

A good life cycle management strategy is imperative for traceability and control. The standard V-model, shown in Fig. 1, represents a verification and validation model. Just like the waterfall model, the V-shaped life cycle is a sequential path of execution of processes: each phase must be completed before the next phase begins. Product testing is planned in parallel with a corresponding phase of development.

The model is advantageous in that it is simple to use, and such activities as planning and test design happen well before coding, saving time and increasing the chance of success (also, since defects are found at an early stage, they do not propagate quickly). Conversely, this approach might prove rigid, and requires software and hardware to be developed during the implementation phase. Moreover, if any changes happen mid-testing, then the test documents, along with requirement documents, must be updated.

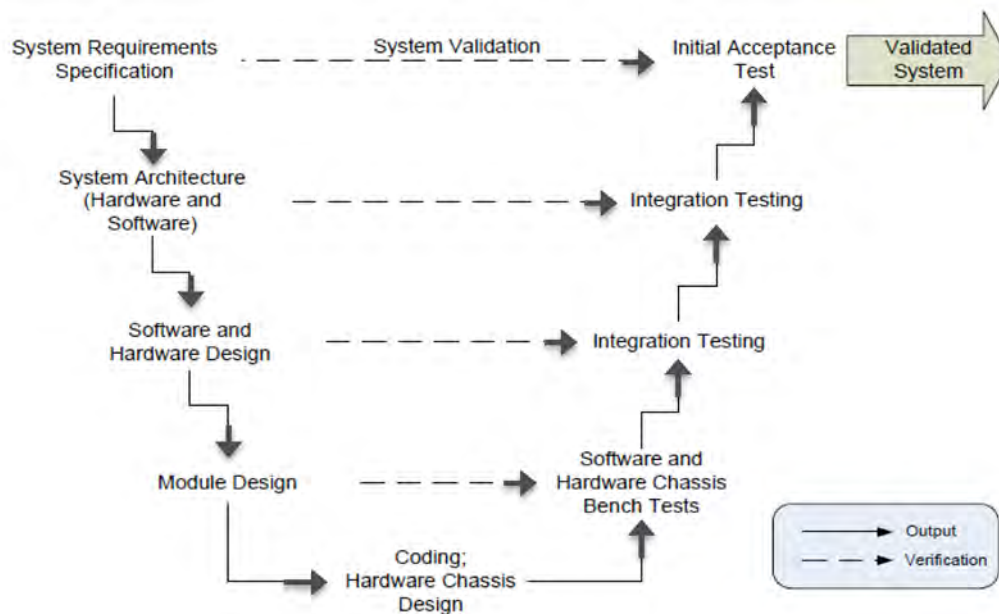


Fig. 1: V-model

### 2.3 Redundancy

Redundancy has different types and can be implemented at different levels. Sometimes it is two capacitors on a circuit-board, in case one fails; other times it is the duplication of a whole system, such as in some programmes of the 1950s, where redundancy was built into each and every component of an entire missile. The most common redundancy employed is parallel redundancy, where redundant parts, channels, or systems are active all the time. With a properly designed sensing and switching scheme, standby redundancy can also be employed.

The standards ISO 14118 *Safety of Machinery—Prevention of Unexpected Start-Up* [1] and IEC 60204-1 *Safety of Machinery, Electrical Equipment of Machines* [2] both state that reliance on a single-channel programmable electronic system is not recommended for safety. The IEC 60204-1 recommendation in particular is interpreted by many as an absolute ban on safety functions being implemented by programmable electronic systems in the sector.

IEC 61508 *Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems* [3] has been published in recognition of the increasing use of this technology throughout a wide range of industrial uses.

Failures can be divided into two categories: common cause and common mode. Common cause failure is defined as one or more event causing concurrent failures of two or more separate channels in a multiple channel system, leading to system failure. Common mode failures are failures of two or more channels in the same way, causing the same erroneous result. Hardware redundancy is very effective in improving reliability: in systems employing redundancy, common cause or common mode failures usually dominate system-level failures. Compared with systems with identical duplicating components and circuits, systems with diversity redundancy (non-identical components) are less vulnerable to common cause and common mode failures.

To get the most out of redundancy, a managerial system is also required to determine, indicate, mediate, and isolate failures such that both safety and availability can be achieved (e.g. four engines on aircrafts).

## 2.4 Choosing components

Components from manufacturers with good quality-control systems and better manufacturing quality are preferred. It is expected that components with manufacturing defects, which contribute to early failure, have been identified and blocked by the quality-control system. Moreover, a better manufacturing quality will make the device less likely to fail during the normal working life, i.e., it will have a lower failure rate.

There are other approaches to improving component's reliability. Control of operating environment is the most important. The working environment, temperature, humidity, vibration, etc., should be controlled so that it matches the required working conditions for the electronics component. Thermal stress is another important factor that affects electronics reliability. Reducing electrical stress by lowering voltage or current also helps. When the hardware design is complete and ready for reliability predication, these factors will be required as inputs for evaluation.

## 2.5 Diagnostics and fail-safe design principle

If the detection of certain failures is feasible, the fail-safe design principle should be applied. This is particular useful if there is no way to tolerate fault consequence: 'fail-safe' design makes provisions for loss of energy source or control signal. Therefore, a 'de-energize-to-trip' philosophy is adopted in safety system design, so that system safety will not be jeopardized during power loss or absence of circuit integrity.

For complex systems where multiple failure modes exist, implementation of the fail-safe principle involves diagnostics: the system's integrity will depend on the information provided by the diagnostics to determine the nature of the failure and take corresponding action. With the aid of diagnostics, the failure of a component or a system can be classified as 'detected' or 'undetected'. For 'detected' failures with mild or safe implication, the user should be alerted, while for a 'detected dangerous' failure mode, the system should be brought to a safe state to ensure safety performance.

A fail-safe device or system is expected to fail eventually but, when it does, it will be in a safe way (for example, a ratchet mechanisms is used in lifts and elevators so that they cannot drop if the cable breaks). A fail-safe physical device may also define what occurs when a user error causes the system to behave in an undesired manner. In the case of software, there is no physical strain on systems, so the concept of mean time between failures is arguably inapplicable. However, software systems can and do fail all the time; for example, the following may happen:

- underlying hardware failure (e.g., networks and servers);
- external system failure (e.g., timing system failure);
- user error.

It is tempting to try to correct a failure situation and keep on running but this can lead to a system moving into an unknown state and creating more issues, as in these examples.

- The network is not responding but the system keeps on processing inputs and queuing outputs, expecting the network to respond later. Caches and disks fill up, affecting other systems, so, even when the network functionalities are restored, the system has to process hours' worth of data.
- A sensor seems to be showing the wrong data, but the system keeps running.

The solution is to institute limits on actions for recovery situations, e.g., by retrying only three times, setting a time limit on caches, etc. It is equally dangerous to make generic assumptions about correcting data across a system. If an input seems wrong, it is better to fail it, since one has no idea why the data do not seem to make sense, and the error is being hidden.

It is important not to simply put the system into a safe state, but also to inform those who can resolve the situation: error reporting and monitoring services should be designed upfront, and should define how operators should be kept informed.

## 2.6 Functional safety and safety integrity level

With the adoption of complex electronics into safety system applications, software must coordinate with hardware, such as microcontrollers. Although field programmable gate arrays can be purely hardware-only, with no run time software, the development process is very software-intensive, using complex software to design and verify the application. Therefore, verifying software is becoming a new challenge in safety system design, and the approach is that of ‘functional safety’.

Another factor that contributes to the wide acceptance of functional safety comes with the adoption of a risk-based approach. Traditional descriptive standards and regulations list requirements for bottom-line protection; with a wide spectrum of applications, such an effort is increasingly difficult and requirements may be too conservative for some cases. A risk-based approach allows each application to carry out a risk assessment to determine the safety function and associated safety integrity level, such that there is no over-design or under-design.

The safety integrity level is the probability of a safety-related system performing the required safety function under all the stated conditions within a stated period of time (or put differently, the probability of failure on demand). ‘Functional safety’ standards originated from the IEC 61508 standard, and have spread to multiple applications, including process, machinery safeguarding, nuclear, and radiological industries.

The standard is sector independent in seven parts, the first four of which have been assigned basic safety publication status. This is the first international standard to quantify the safety performance of an electrical control system that can be expected by conforming to specified requirements, not only for the design concept but also for the management of the design process, operation, and maintenance of the system throughout its life cycle, from concept to decommissioning. These requirements, therefore, safely control failure to function resulting from both random hardware failure and systematic faults. Consequently, the standard represents a bold step, as a proactive approach to quantified, objective safety by design.

To categorize the safety integrity of a safety function, the probability of failure is considered—in effect, the inverse of the safety integrity level definition, looking at failure to perform rather than success. This is because it is easier to identify and quantify possible conditions and causes leading to failure of a safety function than it is to guarantee the desired action of a safety function when called upon.

The safety integrity level concept has emerged from the considerable effort invested in the safety of systems over the past two decades. Two factors have stood out as principal influences.

- 1) A move from the belief that a system can be either safe or unsafe, i.e., that safety is a binary attribute, to the acceptance that there is a continuum between absolute safety and certain catastrophe, and that this continuum is a scale of risk. This has led to an emphasis on risk analysis as an essential feature in the development of safety-related systems.
- 2) A huge increase in the use of software (and complex hardware, such as microprocessors) in the field of safety. This has led to a change in the balance between random and systematic faults. Previously, it was normal to assume (often implicitly) that safety could be achieved through reliability, and to deduce a value for the reliability of a system by aggregating, often through a fault tree, the random failure rates of its components. In some cases, failure rates were derived from historic use of the components and in others they were estimated, so the accuracy of the result was never beyond question. In fact, the greatest accuracy that could be achieved was that

derivable from considering only random failures, for probabilistic methods are not valid for the analysis of systematic faults (those introduced, for example, through specification and design errors). With software, which does not wear out and in which all faults are systematic, there is no possibility of deducing system reliability by a method that is restricted to the consideration of random failures.

Another feature of software is its inherent complexity. Not only is it impossible to prove the absence of faults, but it would require an impracticably long time to derive high confidence in reliability from testing. So a number of problems arise for the developer, who needs not only to achieve but also to demonstrate safety.

The first consideration is that safety requirements may result from a risk analysis that may be quantitative or qualitative. However, as software failures result from systematic and not random faults, direct measurement of the probability of failure, or the probability of a dangerous failure, is not feasible, so qualitative risk analysis must be employed. While the reduction of a given risk may be defined as the specification of a software safety function, the tolerable failure rate of that function may be defined in terms of a safety integrity level. Depending on the standard in use, the safety integrity level may or may not be equated to numerical ranges of failure rates. Once risk analysis has led to a safety integrity level, this is used to define the rigour of the development process. The higher the safety integrity level, the greater the rigour, and tables are used in the standards to identify the methods, techniques, and management processes appropriate to the various safety integrity levels.

When a safety integrity level has been used to define the level of safety to be achieved, it follows that that safety integrity level should be the criterion against which a claim for the achieved safety is made (and judged). But if numerical values for the expected failure rate of software cannot be derived with confidence, it may not be possible to adduce proof of such a claim.

The IEC standard is based on a model relying on two entities: the equipment under control, which is used to provide some form of benefit or utility, and a complementary control system.

The standard recommends that the hazards posed by the equipment under control and its control system be identified and analysed and that a risk assessment be carried out. Each risk is then tested against tolerability criteria to determine whether it should be reduced. If risks are reduced by redesign of the equipment under control, we return to the starting point and hazard identification and analysis and risk assessment should again be carried out.

When it is decided that risk-reduction facilities should be provided in addition to the equipment under control and its control system, and that these should take the form of one or more electrical, electronic, or programmable electronic systems, then the terms of the standard apply to it or them.

The risks posed by the equipment under control and its control system may be contributed to by many hazards, and each must be mitigated until its risk is considered tolerable. The reduction of the risk associated with each hazard is specified as a 'safety requirement' and, according to the standard, each safety requirement must have two components: the functional requirement and the safety integrity requirement. The latter takes the form of a safety integrity level.

In Part 4 of IEC 61508, safety integrity is defined as "the likelihood of a safety-related system satisfactorily performing the required safety functions under all the stated conditions, within a stated period of time" and a safety integrity level as "a discrete level (one of 4) for specifying the safety integrity requirements of safety functions". Thus, a safety integrity level is a target probability of dangerous failure of a defined safety function.

The totality of the safety requirements for all hazards forms the safety requirements specification. Safety requirements are satisfied by the provision of safety functions, and in design these are implemented in 'safety-related systems'. The safety integrity levels of the safety requirements become those of the safety functions that will provide them, and then of the safety-related systems on which the

safety functions are to be implemented. The separation of safety-related systems from the equipment under control and its control system (as by the provision of a protection system) is preferred. However, safety functions may also be incorporated into the control system and, when this is done, certain rules apply, to ensure that higher safety integrity level functions are not affected by the failures of lower safety integrity level functions.

Two classes of safety integrity level are identified, depending on the service provided by the safety function (Table 1):

- for safety functions that are activated when required (on demand mode), the probability of failure to perform correctly is given;
- for safety functions that are in place continuously (continuous mode), the probability of a dangerous failure is expressed in terms of a given period of time (per hour).

**Table 1:** Probability of failure

| <b>Safety integrity level</b> | <b>Mode of operation: on demand<br/>(average probability of failure to perform design function on demand)</b> | <b>Mode of operation: continuous:<br/>(probability of dangerous failure per hour)</b> |
|-------------------------------|---|---|
| 4                             | $\geq 10^{-5}$ to $< 10^{-4}$   | $\geq 10^{-9}$ to $< 10^{-8}$   |
| 3                             | $\geq 10^{-4}$ to $< 10^{-3}$   | $\geq 10^{-8}$ to $< 10^{-7}$   |
| 2                             | $\geq 10^{-3}$ to $< 10^{-2}$   | $\geq 10^{-7}$ to $< 10^{-6}$   |
| 1                             | $\geq 10^{-2}$ to $< 10^{-1}$   | $\geq 10^{-6}$ to $< 10^{-5}$   |

The standard defines a low-demand mode of operation as ‘no greater than one [demand] per year’. Since, in approximate terms, a year is taken to consist of  $10^4$  hours, assuming a failure rate of once per year, the safety integrity level 4 requirement for the low-demand mode of operation is no more than one failure in 10 000 years. If there is to be no more than one demand per year made on a protection system, the equipment under control and its control system must have a dangerous failure rate of no more than once per year, or  $10^{-4}$ . However, arriving at this conclusion can be problematic because doing so is at the very limit of practical testability.

The failure rates attached to safety integrity levels for continuous operation are even more demanding (by a factor of  $10^4$ ) and are intended to provide targets for developers. Because a system—certainly not a software-based system—cannot be shown to have met them, they are intended to define the rigour to be used in the development processes. Safety integrity level 1 demands basic sound engineering practices, such as adherence to a standard quality system, repeatable and systematically documented development processes, thorough verification and validation, documentation of all decisions, activities and results, and independent assessment. Higher safety integrity levels, in turn, demand this foundation plus further rigour.

The value of the safety integrity level lies in providing a target failure rate for the safety function or safety-related system. It places constraints on the processes used in system development, such that the higher the safety integrity level, the greater the rigour that must be applied. The processes defined as being appropriate to the various safety integrity levels are the result of value judgements regarding what needs to be done in support of a reasonable claim to have met a particular safety integrity level. However, the development processes used, however good, appropriate, and carefully adhered to, do not necessarily lead to the achievement of the defined safety integrity level. Even if, in a particular case, they did, the achievement could not be proved. But, even if evidence is insufficient to show that the safety integrity level requirement has been met, it does increase confidence in the system and its software.

Although the safety performance is the primary design objective, availability should also be considered. Large physics facilities are expensive investments; their productivity is critical financially and matters for the sake of science. Hence, there are always system-availability requirements for the

project or the facility, and the availability of the indispensable safety system sets an upper bound for the whole facility's availability.

## 2.7 Standards and guidance

Standards are documents that establish uniform engineering and technical requirements for processes, procedures, practices, and methods. As this definition implies, some standards contain industry best practices, some provide description of interfaces such that interoperability can be achieved, while other standards simply describe methods for development and testing.

There are other, less formal, guidance documents that provide equally important information. They contain standard procedural, technical, engineering, or design information about the material, processes, practices, and methods covered or required by standards.

The definition of the term 'standard' includes the following:

- common and repeated use of rules, conditions, guidelines, or characteristics for products or related processes and production methods, and related management systems practices;
- definitions of terms; classifications of components; delineations of procedures; specification of dimensions, materials, performance, designs, or operations; measurements of quality and quantity in describing materials, processes, products, systems, services, or practices; test methods and sampling procedures; or descriptions of fit and measurements of size or strength.

We need standards to build our systems efficiently:

- deliverable products must be designed and built—they make use of procured items and must themselves be procured;
- each of these phases—procurement, especially—requires specification;
- effective specification requires standards.

An additional differentiation can be based on purpose:

- a basic standard has a wide-ranging effect in a particular field, such as a standard for metal, which affects a range of products from cars down to screws;
- terminology standards (or standardized nomenclature) define words, permitting representatives of an industry or parties to a transaction to use a common, clearly understood, language;
- test and measurement standards define the methods to be used to assess the performance or other characteristics of a product or process;
- product standards establish qualities or requirements for a product (or related group of products), to assure that it will serve its purpose effectively;
- process standards specify requirements to be met by a process, such as an assembly line operation, to function effectively;
- service standards, such as for repairing a car, establish requirements to be met in order to achieve the designated purpose effectively;
- interface standards, such as the point of connection between a telephone and a computer terminal, are concerned with the compatibility of products;
- standards on data to be provided contain lists of characteristics for which values or other data are to be stated for specifying the product, process or service.

International standards have been developed through a process that is open to participation by representatives of all interested countries, and that is transparent, consensus-based, and subject to due process.

Standards may also be classified by the intended user group, for example:

- organization standards are meant for use by a single industrial organization and are usually developed internally;
- industry standards are developed and promulgated by an industry for materials and products related to that industry;
- government standards are developed and promulgated by federal, state, and local agencies to address needs or applications peculiar to their missions and functions;
- international standards are developed and promulgated by international governmental and non-governmental organizations, such as the International Organization for Standardization (ISO);
- harmonized standards can be either an attempt by a country to make its standard compatible with an international, regional, or other standard, or it can be an agreement by two or more nations on the content and application of a standard, the latter of which tends to be mandatory.

### **2.7.1 Software standards**

The ISO/IEC 12207 standard provides a common framework for developing and managing software. The IEEE/EIA 12207.0 standard consists of the clarifications, additions, and changes accepted by the Institute of Electrical and Electronics Engineers (IEEE) and the Electronic Industries Alliance (EIA), as formulated by a joint project of the two organizations. The IEEE/EIA 12207.0 standard outlines concepts and guidelines to foster better understanding and application of the standard. Thus, this standard provides industry with a basis for software practices that would be useable for both national and international business.

#### *2.7.1.1 IEEE 12207—Software Life Cycle Processes*

This standard establishes a common framework for software life cycle processes, with well-defined terminology, that can be referenced by the software industry. It contains processes, activities, and tasks that are to be applied during the acquisition of a system that contains software, a stand-alone software product, or software service, as well as during the supply, development, operation, and maintenance of software products. Software includes the software portion of firmware. This standard also provides a process that can be employed for defining, controlling, and improving software life cycle processes.

The standard applies to the acquisition of systems and software products and services, to the supply, development, operation, and maintenance of software products, and to the software portion of firmware, whether performed internally or externally to an organization.

The standard groups the activities that may be performed during the life cycle of software into five primary processes, eight supporting processes, and four organizational processes. Each life cycle process is divided into a set of activities; each activity is further divided into a set of tasks.

In addition to IEEE12207, standard IEC61508, parts 3 and 7, focuses on safety functions with the following recommendations:

- use of structured and modular design;
- restricted use of asynchronous constructs;
- design for testability;
- restrictive use of ambiguous constructs;



- transparent and easy to use code;
- defensive code and range checking (to pick up faults or anomalies and respond in a pre-determined way);
- use of comments and annotations;
- limits on module sizes and number of ports to increase readability;
- avoidance of multi-dimensional arrays and go-to type commands;
- avoidance of redundant logic and feedback loops;
- avoidance of latches, asynchronous reset.

### 2.7.2 *Hardware standards*

#### 2.7.2.1 *Computer Automated Measurement and Control (CAMAC)*

This is a standard bus and modular crate electronics standard for data acquisition and controls, defined in 1972:

- solved the low-channel density problem of nuclear instrumentation methods;
- up to 24 modules in a crate, interfaced to a personal computer;
- not hot-swappable because of backplane design;
- data way management: module power, address bus, control bus, and data bus;
- 24-bit communication between controller and selected module.

#### 2.7.2.2 *Versa Module Europa (VME)*

This standard backplane bus was defined in 1981:

- architecture not scalable for high speeds (single-ended parallel bus, not for gigabits per second);
- electromagnetic shielding not specified;
- developed for Motorola 68000 line of CPUs (the bus is equivalent to the pin of 68000 run out onto a backplane);
- faster bus (from 16 to 64 bit), up to 40 MHz (VME64).

#### 2.7.2.3 *xTCA (Telecommunications Computing Architecture)*

ATCA (Advanced Telecommunications Computing Architecture) and  $\mu$ TCA are platforms that provide:

- all-serial communications (multigigabits per second backplane);
- both complex experiment controls and large, high bandwidth and throughput data acquisition systems;
- the highest possible system performance, availability, and interoperability.

To achieve high availability in a complex physics system, three main features are required:

1. modular architecture;
2.  $N + 1$  or  $N + M$  redundancy of single-point-of-failure modules (whose malfunction could stop operation of the machine or experiment);
3. intelligent platform management interface for quick isolation of faults and hot-swap.

In addition, physics modules need a few extended features:

- intelligent platform manager interface for cooling and thermal management, control and monitor;
- built-in hot-swap;
- designed for high-reliability;
- intelligent platform manager interface for cooling and thermal management, control, and monitor;
- built-in crate and component status monitoring and remote management and diagnostics;
- independent monitoring channel within the crate.

Specifically,  $\mu$ TCA is a modular, open standard for building high-performance switched-fabric computer systems in a small form-factor. At its core are standard advanced mezzanine cards, which provide processing and input–output functions. The  $\mu$ TCA standard was originally intended for smaller telecom systems at the edge of the network but has moved into many non-telecom applications, with standardized rugged versions becoming popular in mobile, military, telemetry, data acquisition, and avionics applications. The core specification, MTCA.0, defines the basic system, including backplane, card cage, cooling, power, and management. A variety of differently sized advanced mezzanine card modules are supported, allowing the system designer to use as much or as little computing and input–output as necessary. Subsidiary specifications (MTCA.1 to MTCA.4) define more rugged versions, specifically suited for military, aeronautic, and other demanding physical environments.

Modules (for a  $\mu$ TCA):

- cooling units;
- power modules;
- advanced mezzanine card for electronics, CPU, hard drives;
- rear transition module;
- $\mu$ TCA central hub.

## 2.8 Tests

Testing is a process rather than a single activity, and starts as early as the system requirements specification. The choice of testing frequency definitely affects system reliability; the system design should accommodate such requirements, including setting up the test mode to facilitate testing. It is easy to see from the V-model that testing activities are a necessary step in completing every activity. Activities within the fundamental test process fall into the following basic steps (we will focus more on software tests, but the same principles apply to hardware tests):

1. planning and control;
2. analysis and design;
3. implementation and execution;
4. evaluating exit criteria and reporting;
5. test closure activities.

### 2.8.1 *Planning and control*

Test planning is intended to:

- determine the scope and risks and identify the objectives of testing;
- determine the test approach;

- implement the test policy or test strategy;
- determine the required test personnel and resources; test environments, hardware, etc.;
- schedule test analysis and design tasks, test implementation, execution, and evaluation;
- determine exit criteria.

A test strategy is created to inform project managers, testers, and developers of key issues of the testing process. This includes the testing objectives, method of testing, total time, and resources required for the project and the testing environments.

Test control is intended to:

- measure and analyse the results of reviews and testing;
- monitor and document progress, test coverage, and exit criteria;
- provide information on testing;
- initiate corrective actions;
- enable decision making.

### **2.8.2 Analysis and design**

Test analysis and design should:

- review the test basis;
- identify test conditions;
- design the tests;
- evaluate testability of the requirements and system;
- design the test environment set-up and identify and required infrastructure and tools.

The test basis is the information needed to start the test analysis and create test cases. It is a documentation on which test cases are based, such as requirements, design specifications, product risk analysis, architecture, and interfaces. Test basis documents help understand what the system should do once built.

### **2.8.3 Implementation and execution**

During test implementation and execution, test conditions are translated into test cases and procedures and scripts for automation, the test environment, and any other test infrastructure. (Test cases are a set of conditions under which a tester will determine whether an application is working correctly or not.)

Test implementation should:

- develop and prioritize test cases and create test data for those tests (to test a software application, for example, the tester needs to enter some data for testing most of the features: any such specifically identified data used in tests are known as test data);
- create test suites (a collection of test cases that are used to test a software program to show that it has some specified set of behaviours) from the test cases for efficient test execution;
- implement and verify the environment.

Test execution should:

- execute test suites and individual test cases, according to test procedures;
- re-execute tests that previously failed, to confirm a fix;

- log the outcome of test execution and record the identities and versions of the software under tests;
- compare actual results with expected results;
- report discrepancies between actual and expected results.

The test log is used for the audit trail. A test log records the test cases that were executed, in what order, who executed that test cases and the status of the test case (pass or fail).

#### **2.8.4 Evaluating exit criteria and reporting**

Based on the risk assessment of the project, criteria are set or each test level against which one can determine that ‘enough testing’ has been done. These criteria vary from project to project and are known as exit criteria.

Exit criteria are satisfied when:

- a maximum number of test cases are executed with a certain pass percentage;
- the software bug rate falls below a certain level.

#### **2.8.5 Test closure activities**

Test closure activities are performed when hardware or software is delivered, and include the following major tasks:

- check which planned deliverables are actually delivered and ensure that all incident reports have been resolved;
- finalize and archive test procedures, such as scripts or test environments, for later reuse;
- deliver test procedures to the maintenance organization;
- evaluate the testing process, to provide lessons for future releases and projects.

#### **2.8.6 Proof tests**

Safety system standards also require *proof tests*, to include verification of the following conditions:

- operation logic sequence given by cause and effect diagrams;
- operation of all input devices, including field sensors and single-instance storage input modules;
- logic associated with each input device;
- logic associated with combined inputs;
- trip set-point of all inputs;
- alarm functions;
- response time of the system (when applicable);
- functioning of manual actions bringing the process to its safe state, e.g., emergency stop;
- functioning of user-initiated diagnostics;
- safety system is operational after testing;
- all paths through redundant architectures should be tested.

Proof tests can identify ‘hidden’ device failures, although they cannot prevent failures from happening. The proof test interval should be large enough to catch failures, but too frequent testing also increases the likelihood of human errors in the system.

General considerations for proof testing are as follows:

- failure modes of the device and their effects on functionality; if a device failure is either self-revealed or can be detected by diagnostics, there is no need to include this device into proof testing;
- if a device has dominant age-related failure modes, preventive maintenance should be applied, in accordance with a reliability centred maintenance analysis;
- only safety-critical functions should be tested; non-safety-related functions should be included in another maintenance test or should simply be tested during an initial acceptance test and then again after a much longer interval;
- instruments and field devices that have no direct impact on safety, and are run under ‘continuous mode’ (continuous comparison among redundant devices), can either ‘run to fail’ or be tested with a much longer period (for example, signage);
- logic solvers that lack integrated diagnostic functions should still be tested annually, since they have no diagnostics and their functionalities can easily be changed;
- safety programmable logic controller-based logic solvers with strict management of change procedures have no need for a strict programmable logic controller-dedicated assurance test, however, the whole system should go through a full functional testing every 8–12 years (according to Shell standard DEP 32.80.10.10-Gen, July 2008);
- ease of testing should be considered during the system design stage, e.g., the process industry uses a ‘maintenance override service’ to facilitate online testing without tripping the process.

### **2.8.7 Test examples**

As an example, these test methodologies can be followed for an accelerator safety system.

#### *2.8.7.1 Programmable logic controller or field programmable gate array bench test*

This is part of the programmable logic controller or field programmable gate array software quality assurance activity; it demonstrates that the programmable logic controller field programmable gate array logic satisfies the specification.

#### *2.8.7.2 Interlock checks*

These checks test field components subject to accidental damage or harsh environmental conditions, especially those of an electromechanical nature or which have moving parts; they are conducted at least every 6 months.

#### *2.8.7.3 Initial acceptance test*

This is intended to test physical hardware, installations, all functions of a new safety installation (the installation includes hardware, programmable logic controller or field programmable gate array logic), including unintended functions and common mode failures, which could arise from design or implementation errors, or component malfunction. This test is carried out for a new installation or after major modifications. New safety software code downloads are currently considered a major modification.

#### *2.8.7.4 System features to be tested by initial acceptance test*

These include:

- interaction between system and human–machine interfaces;

- each safety function, either loop-oriented or a complex functionality as a whole;
- degraded mode of operation if there are any requirements defined in operations requirements;
- recovery from failure;
- redundancy;
- different operation mode of the system;
- reasonable foreseeable abnormal conditions and misuse of the system.

#### 2.8.7.5 *Safety assurance test*

This is intended to perform a maintenance function, to verify continuing operation of safety features; this test is carried out annually.

#### 2.8.7.6 *Other tests*

Different industries and industrial standards use different terms for these test activities. For Safety Instrumented Systems, ANSI/ISA 84 (IEC 61511 Mod) contains requirements for factory acceptance testing, site acceptance tests, and proof testing. For complex process automation projects, IEC 62381 defines the scope and activities for factory acceptance testing, site acceptance tests, and site integration tests. Broadly speaking, the programmable logic controller bench test falls under the scope of factory acceptance testing; and the initial acceptance test is equivalent to a site acceptance test.

## 2.9 Configuration control

Configuration management is the unique identification, controlled storage, change control, and status reporting of selected intermediate components during the life of a system. Configuration control is the activity of managing the system and related document throughout the product's life cycle.

Configuration control ensures that:

- the latest approved version of the system and its components are used at all times;
- no change is made to the product baselines without authorization;
- there is a clear audit trail of all proposed, approved, or implemented changes.

When applied to software, there are additional challenges: on the one hand, individual developers need the flexibility to do creative work, to modify code to try out what-if scenarios, and to make mistakes, learn from them, and evolve better software solutions; on the other hand, teams need stability to allow code to be shared with confidence, to create builds and perform testing in a consistent environment, and to ship high-quality products with confidence. This requires an intricate balance to be maintained. Too much flexibility can result in problems, including unauthorized or unwanted changes, the inability to integrate software components, uncertainty about what needs to be tested and working programs that suddenly stop working. Conversely, enforcing too much stability can result in costly bureaucratic overhead and delays in delivery, and may even require developers to ignore the process in order to get their work done.

How is it possible to maintain the necessary balance between flexibility and stability, as software moves through the life cycle?

Some techniques include:

- selecting the appropriate type and level of control for each software artefact;
- selecting the right acquisition point for each configuration item;
- utilizing multiple-levels of formal control authority.

## 2.10 Quality assurance and quality control

Quality assurance is process oriented; quality control is product oriented: this might be one's starting point when considering how to assure quality to products and systems, as quality assurance makes sure that one is doing the right things, the right way, while quality control makes sure the results of what one has done are what one expected.

Assuring quality means more than making sure that quality exists also means stepping in wherever there are opportunities to add or ensure quality; for instance, clarifying requirements, documenting new requirements, facilitating communication among teams and, of course, testing. Testing should not be limited to hardware or software, but should extend to requirements, understanding of requirements, etc.

Typical quality assurance activities are: quality audit, defining process, selection of tools, and training.

Typical quality-control activities are: testing, walkthrough, inspection, and checkpoint review.

Any project should begin with a clear definition of requirements and deliverables. Performance requirements are defined in a 'black box' manner: the 'how' is not defined; the size, location, number of entry points, number and location of devices, entry requirements, desired access states, and operational modes must all be defined. Any interfaces to other systems should be highlighted and described in their own sections, for clarity.

The requirements document is carefully reviewed, since many future quality assurance tests reference these requirements; ideally, requirements documents should be maintained as 'living' documents.

A formal specification document is developed to define, specifically, how the system is to be built and operate. Specific parts are identified, system architectures of technology may be selected, input and output signal lists are defined. The systems specification should describe how to meet the requirements set forth in the requirements document. Any interfaces to other systems shall be highlighted and described in their own sections for clarity.

The specification document is carefully reviewed, since engineering and design work, as well as many future quality assurance tests, is performed against it.

### 2.10.1 Technical reviews

Technical reviews are conducted to evaluate design and engineering work for accuracy and performance against the requirements, specification, and other best-practice standards.

Informal peer reviews should be utilized periodically, to assess engineering work or testing procedures.

Formal reviews are conducted to evaluate project design and engineering work for accuracy and performance against the requirements, specification, and other best-practice standards. Formal reviews are typically specified in a project quality assurance plan. Large projects will typically have an early preliminary or system architecture review, followed by a detailed or final design review. At a minimum, there will always be at least one final technical review for a project. The membership of formal reviews follows a graded approach (Table 2). The number of external reviewers, the overall number of reviewers, and the organizational distance of reviewers from the overseeing organization is dependent on the scope, complexity, and technological familiarity of the proposed design compared to common practice.

**Table 2:** Minimum recommended reviewer complement

|                                      |  |
|--------------------------------------|--|
| Minor modification, familiar methods | 1 external reviewer  |
| Minor modification, new methods      | 2 external reviewers   |
| Medium change, familiar methods      | 2 external reviewers   |
| Medium change, new methods           | 2 or more external reviewers,<br>1 external to control department                                      |
| Large change, familiar methods       | 3 external reviewers,<br>1 external to control department  |
| Large change, new methods            | 3 or more external reviewers,<br>1 or more external to control department,<br>1 external to laboratory |

## 2.11 Documentation

Document management is the process of applying policies and rules to how documents are created, maintained, and archived within an organization. Document collaboration is merely the process of checking out, checking in, and versioning a document before it is published. Records management encompasses all of the functions of document management, but applies them to a broader set of content elements—not just documents.

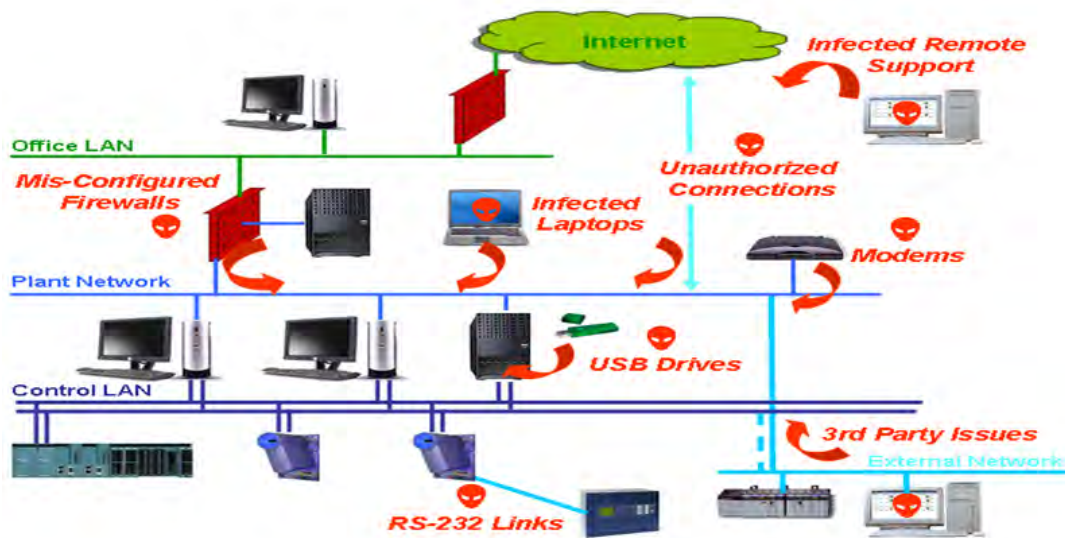
The main aspects of managing a document through its life cycle include the following.

- Creation: Methods for envisioning, initiating, and collaborating on a new document's development.
- Location: There must be a physical location where documents will be stored and accessed. Usually, most documentation management systems require single-instance storage of a document so that there is only one version of the truth.
- Authentication and approval: Methods of ensuring that a document is fully vetted and approved before it is considered to be official compliant communication from the organization.
- Workflow: This describes the series of steps needed to pass documents from one person to another for various purposes, such as to gain approval to publish the document or to collect signatures on a document.
- Filing: For electronic systems, a document is filed by placing it in a physical location and then attaching metadata to the document. The metadata files the document logically by allowing the document to be found based on the metadata values assigned to the document.
- Distribution: Methods of getting the document into the hands of the intended readers.
- Retrieval: Methods used to find the documents, such as querying the index for keywords or using search alerts to find new content that meets the query keywords.
- Security: Methods used to ensure the document's integrity and security during its life cycle.
- Retention: Organization's policies and practices that inform everyone how long different document types are retained by the organization.
- Archiving: Similar in concept to retention, the differing characteristic is that archiving is a subset of retention policies. Archiving focuses on the long-term retention of documents in a readable format after the document's active life has ended. Subsumed in this category is the expiration of documents after they no longer need to be retained.



**2.12 Cybersafety**

Traditional network security risk management techniques are often inadequate to meet the specialized needs of control systems, whose security represents a unique challenge. Generally speaking, control systems are designed for accuracy, extreme environmental conditions, and real-time response in ways that are often incompatible with the latest cybersecurity technologies, inconsistent with consumer-grade hardware and software, and in conflict with common network protocols. As a result of these performance factors and limitations, engineers (rather than IT managers) have traditionally been responsible for the design, operation, and maintenance of control systems. Yet, despite their uniqueness, control systems are increasingly reliant on common network protocols, and connectivity often exists between control systems and enterprise networks, to include the Internet (Fig. 2).



**Fig. 2:** Pathways into control systems

How does an organization ensure that its supervisory control and data acquisition system is secure? One of the answers is in standard ISA-99.02.01 (*Security for Industrial Automation and Control Systems: Establishing an Industrial Automation and Control Systems Security Program*), approved and published by the American National Standards Institute (ANSI). This readable standard lays out seven key steps for creating a cybersecurity management system for use with supervisory control and data acquisition and control systems.

The steps in ISA-99.02.01 are divided into three fundamental categories: risk analysis, addressing risk with the cybersecurity management system, and monitoring and improving the cybersecurity management system.

- 1 The first category lays out the stages an organization needs to follow to assess its current security situation and determine the security goals it wants to achieve.
- 2 The second category outlines processes to define security policy, security organization, and security awareness in the organization and provides recommendations for security countermeasures to improve supervisory control and data acquisition system security. The core idea in this section is a concept known as ‘defence in depth’, where security solutions are carefully layered to provide multiple hurdles to attackers and viruses.
- 3 The third category describes methods to make sure a supervisory control and data acquisition system not only stays in compliance with the cybersecurity management system but follows a continuous improvement programme.

### 2.12.1 *Defence in depth*

Sound strategy, regardless of whether it is for military security, physical security, or cybersecurity, relies on the concept of ‘defence in depth’. Effective security is created by layering a number of security solutions so that if one is bypassed another will provide the defence. This means, for instance, not overrelying on any single technology, such as a firewall.

Defence in depth begins by creating a proper electronic perimeter around the supervisory control and data acquisition or control system and then hardening the devices within. The security perimeter for the control system is defined by both policy and technology. First, policy sets out what truly belongs in the control-system network and what is outside; next, a primary control-system firewall acts as the choke point for all traffic between the outside world and the control-system devices.

Once the electronic perimeter of the control system is secured, it is necessary to build the secondary layers of defence in the control system itself. Control components, such as human–machine interfaces and data historians based on traditional IT-operating systems, e.g., Windows and Linux, should take advantage of the proven IT strategies of patch and anti-virus management. However, this requires prior testing and care.

For such devices as programmable logic controllers and supervisory control and data acquisition controllers—where patching or anti-virus solutions are not readily available—industrial security appliances should be used. This solution deploys low-cost security modules directly in front of each group of control devices needing protection. The security modules then provide tailored security services, e.g., ‘personal firewalling’ and message encryption, to the otherwise unprotected control devices.

Table 3 compares the cybersafety requirements of a machine protection system with those of a personnel protection system (access control)—the latter being, generally, stricter.

## 2.13 Evolution of cybersafety landscape (a US perspective)

### 2.13.1 *Framework for Improving Critical Infrastructure Cybersecurity (US National Institute of Standard, NIST, February 2014): a system of regulations and the means used to enforce them*

The framework is based on:

- core functions (activities and references);
- implementation tiers (guidance);
- a framework profile (how to integrate cybersecurity functions within a cybersecurity plan).

The framework consists of four implementation tiers, each defined for three categories—risk management process, integrated risk management programme, and external participation. Any organization will follow into one of these three categories.

#### 2.13.1.1 *Tier 1: Partial*

- Risk management process: Organizational cybersecurity risk management practices are not formalized, and risk is managed in an ad-hoc and sometimes reactive manner.
- Integrated risk management programme: There is limited awareness of cybersecurity risk at the organizational level.
- External participation: An organization may not have processes in place to participate in coordination or collaboration with other entities.

**Table 3:** Cybersafety requirement comparisons

| <b>Requirements</b>  | <b>Personnel protection</b>  | <b>Machine protection</b>   |
|--|--|---|
| Use of configuration versioning system for software  | Yes  | Yes   |
| Manage check-in and out of configuration versioning system with procedures   | Yes  | Yes   |
| Track and check checksum   | Yes; additional ‘safety signature’ available for safety-rated programmable logic controllers   | Yes   |
| Software download is password protected  | Yes  | Yes   |
| Download over network?   | No, not allowed; only local PROFIBUS (process field bus) connection allowed  | Yes   |
| Download to wrong CPU across network?  | No; isolated networks and different CPU names and Internet protocol addresses even if on same network  | No; isolated networks and different CPU names and Internet protocol addresses even if on same network |
| Protection against wrong safety program load   | Hardware configuration is loaded; safety modules have hardware dual in-line package switches; hardware configuration error causes fail-safe shutdown | No  |
| Physical isolation from controls network   | No   | No  |
| Possible accidental (or act of sabotage) download of safety-critical code from controls network  | No; local download only  | Yes   |
| Possible accidental changes (or act of sabotage) to supervisory control and data acquisition human-machine interface from controls network | Yes  | Yes   |

*2.13.1.2 Tier 2: Risk informed*

- Risk management process: Risk management practices are approved by management but may not be established as organizational-wide policy.
- Integrated risk management programme: There is an awareness of cybersecurity risk at the organizational level but an organization-wide approach to managing cybersecurity risk has not been established.
- External participation: The organization knows its role in the larger ecosystem, but has not formalized its capabilities to interact and share information externally.

*2.13.1.3 Tier 3: Repeatable*

- Risk management process: The organization’s risk management practices are formally approved and expressed as policy.
- Integrated risk management programme: There is an organization-wide approach to managing cybersecurity risk.

- External participation: The organization understands its dependencies and partners and receives information from these partners that enables collaboration and risk-based management decisions within the organization in response to events.

#### 2.13.1.4 Tier 4: Adaptive

- Risk management process: The organization adapts its cybersecurity practices based on lessons learned and predictive indicators derived from previous and current cybersecurity activities.
- Integrated risk management programme: There is an organization-wide approach to managing cybersecurity risk that uses risk-informed policies, processes, and procedures to address potential cybersecurity events.
- External participation: The organization manages risk and actively shares information with partners to ensure that accurate, current information is being distributed.

### 2.13.2 NIST Special Publication (SP) 800-53 (Computer Security Guide)—Revision 4, April 2013

This standard is based on an information security programme: it covers risk assessment; policies and procedures; subordinate plans; training; periodic testing; incident response; and continuity of operations.

The standard is mission-oriented. It is based on FIPS 199 (Federal Information Processing Standard) for Security Categorization of Federal Information and Information Systems, and it includes definitions of security control categories for information systems (based on the key aims of confidentiality, integrity, availability).

The standard is also based on the impact on an organization's capability to accomplish its mission. (There is a full catalogue, including access control, awareness and training, audit and accountability, authentication, maintenance, media protection and access.)

### 2.13.3 Other standards

#### 2.13.3.1 IEC 17799: Information Technology—Code of Practice for Information Security Management

This standard is of a high level; being broad in scope and conceptual in nature, it forms a basis to develop customized security standard and security management practices.

#### 2.13.3.2 ISA-TR99: Integrating Electronic Security into the Manufacturing and Control System Environment

This standard is a guide to user and manufacturers. It can be used to analyse technologies and determine their applicability in securing manufacturing and controls.

#### 2.13.3.3 IEC 15408 (3.1): Information Security Management Systems (ISMS)

This standard provides a framework to specify security functional and assurance requirements through the use of protection profiles. Vendors can implement security attributes and testing laboratories can evaluate products.

#### 2.13.3.4 IEC 27001:2005: Common Criteria (CC) for Information Technology Security Evaluation

This is a system to bring information security under explicit management control through policies and governance; asset management; human resources security; access control; incident management; business continuity; etc.

### 2.13.3.5 NIST SP 800-82

This standard formalizes the defence-in-depth strategy: layering security mechanisms to minimize the impact to one mechanism as a result of failure.

The standard covers:

- Internet connection sharing policies based on Department of Homeland Security threat level;
- implementation of a multi-layer network topology;
- provision of logical separation between corporate and Internet connection sharing networks;
- use of a demilitarized zone (i.e., no direct communication between Internet connection sharing and corporate use);
- fault-tolerant design;
- redundancy for critical components;
- privilege management;
- encryption.

For laboratories, the risk tolerance for ‘generic’ Internet connection sharing is different than that for personnel protection or medical technology (the protection of lives, information, assets, etc.). Boundaries and interfaces have to be identified; moreover, in highly regulated environments, once a standard is chosen and committed, the organization can be audited against it.

### 3 An example: the Linac Coherent Light Source machine protection system

The machine protection system at the Linac Coherent Light Source (LCLS-I) at the SLAC National Accelerator Laboratory is an interlock system responsible for turning off or reducing the rate of the beam in response to fault conditions that might damage or cause unwanted activation of machine parts.

The system is required to:

- turn off or limit the rate of the electron beam when faults are detected, to prevent damage to sensitive machine components;
- protect undulator permanent magnets from the electron beam, limiting the radiation dosage to below a specified amount;
- protect beamline components from excessive beam exposure, to prevent damage to the vacuum system and unnecessary activation;
- shut off the beam (detect and mitigate) within one pulse at 120 Hz (i.e., 8.33 ms for LCLS-I);
- protect the laser heater system from the injector laser;
- allow fault conditions to set different maximum rates for each mitigation device;
- allow automatic beam rate recovery (after a fault is corrected, the beam rate is raised to its before-fault value);
- bypass faults securely;
- provide a user interface that quickly identifies system trips, allows ‘post-mortem’ analysis and shows history;
- provide the ability to change the configuration of the logic and beam rate by adding and removing input signals, bypassing device fault inputs, and setting and changing fault thresholds.

The machine protection system (see Fig. 3) is able to reduce the beam rate only to below the operators' requested beam rate, and cannot raise the beam rate above operators' requested beam rate. Separate systems support the machine protection system, to protect other energized devices such as power supplies, magnets, and klystrons. A separate beam containment system ensures that no beam or radiation reaches potentially occupied areas. To perform its functions, the machine protection system relies on a set of inputs and output signals (see Fig. 4).

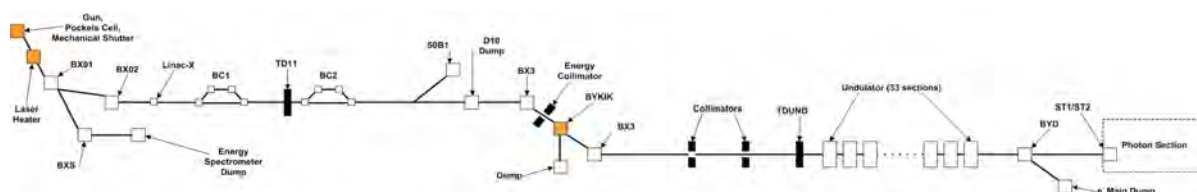


Fig. 3: Linac Coherent Light Source machine protection system

### 3.1 Inputs

#### 3.1.1 Inputs from obstructions

Obstructions cover numerous devices, including vacuum valves, tune-up dumps, beam finder wires, and profile monitor screens. The beam is turned off whenever an obstruction reads a 'not-out' status. Beam finder wires, the tune-up dump, and profile monitor screens allow a maximum 10 Hz repetition rate once they are fully inserted. All obstruction devices require two limit switches to be incorporated in the design, to indicate fully in and fully out positions. An inconsistent status between the two switches is to be treated as a machine protection system fault. Obstructions are regarded as a pre-emptive fault, where the beam is turned off before it can cause any damage, e.g. to:

- profile monitor screens;
- collimator jaws;
- dechirper plates;
- beam stoppers.

#### 3.1.2 Beam loss monitors

Beam loss is regarded as an actual fault with a requirement that the beam be shut off before the next pulse can be delivered; this means that the overall system must detect and mitigate the fault in less than 8.3 ms for 120 Hz operation.

Loss monitors with different types of sensitivity (e.g., toroids; protection ion chambers; optical fibre beam loss monitors) are deployed in different locations. The signal is gated to coincide with the beam arrival time and compared with a programmable threshold; it will indicate a fault if the threshold is exceeded. Two programmable threshold settings are required:

- exceeding the lower threshold can allow the beam rate to be lowered by the machine protection system;
- exceeding the higher threshold should cause the beam to be shut off and require the machine protection system to be reset manually.

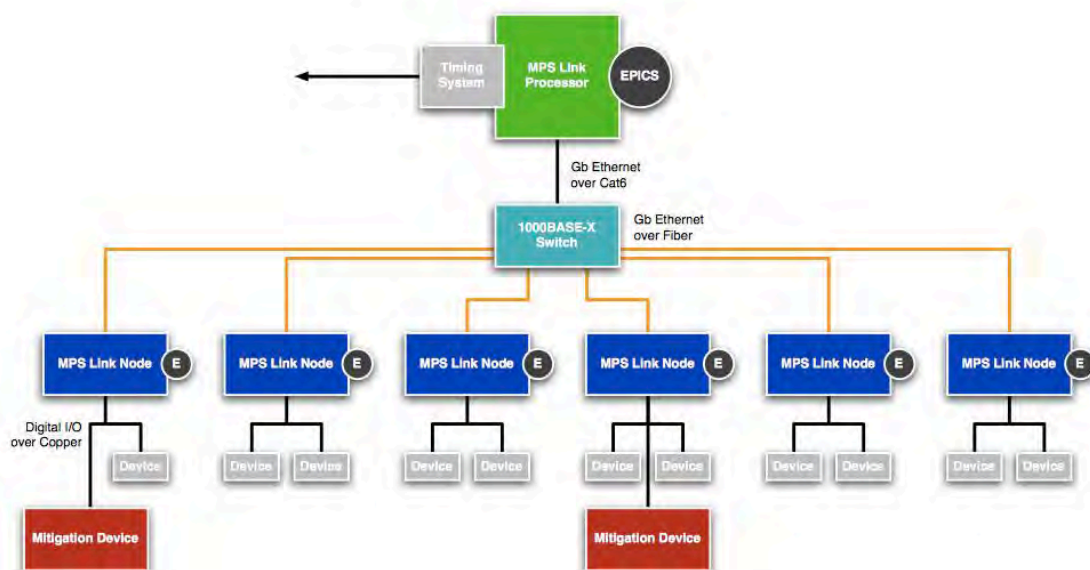


Fig. 4: Board diagram. EPICS, Experimental Physics & Industrial Control System; I/O, input–output; MPS, machine protection system.

### 3.1.3 Other inputs

These include:

- watchdog;
- vacuum valves;
- temperature readouts.

### 3.1.4 Sensors

There are a number of sensors. These include sensors for:

- vacuum valve position;
- water flow status;
- magnet power supply status;
- temperature;
- in-beam diagnostics status;
- beam position;
- beam charge;
- RF system status;
- beam containment status;
- beam loss.

## 3.2 Outputs (mitigation devices)

### 3.2.1 At the gun

The following devices interrupt the beam at the gun:

- laser heater mechanical shutter;
- photocathode laser mechanical shutter;
- gun trigger permit.

The mitigation scheme is based on shifting the timing of the gun RF from beam-time to standby time, while leaving the average rate of the laser and RF systems constant. The beam will not be extracted from the gun if the timing is shifted to standby time, since the RF is not present at the same time that the laser impinges on the cathode. The mechanical shutter is still deployed to block the light when the rate goes to zero. The shutter inhibits the injector UV laser before it hits the gun's cathode; its control is verified with optical position sensors; it faults the beam containment system mechanical shutter when the control does not match the position status.

### 3.2.2 *Pre-undulator fast kicker (BYKIK)*

The BYKIK (Fig. 5) is a pulsed dipole in LCLS-I, and is located in the middle of the DL-2 bend system in the linac-to-undulator beam line. The BYKIK is pulsed at a constant average frequency of 120 Hz and receives two input triggers, one at beam-time and the other at standby time. When the standby trigger is applied, the beam is transported unperturbed to the undulator. When the machine protection system shifts the trigger to beam-time, the beam is deflected by BYKIK onto a dump (collimator) and is not transported to the undulator. The switching of the BYKIK triggers between standby and beam-time can be done on a pulse-by-pulse basis so that either the beam is fully suppressed or bunches can be selectively allowed through to the undulator at a reduced rate. This feature is further exploited in special cases to send single shots and burst modes to the undulator on demand.

The secondary mitigation device requirements for BYKIK are derived from the need to suppress the beam to the undulator while the beam at the front end remains on at the full rate so that the machine is held stable by the beam-based feedback systems. For this to be reliable, the system must verify that BYKIK is operating correctly and dumping the beam before the undulator can be damaged. The verification is achieved at two levels. First, the BYKIK magnet control module signals a pre-emptive machine protection system fault if the magnet is out of tolerance within the specified time window of the pulse. The final verification must come from the beam itself at the time that BYKIK actually fires to kick the beam. For example, the beam position monitor immediately downstream of BYKIK should see the beam deflected by  $>1$  mm, otherwise it should also register a machine protection system fault. In the event of either of these faults occurring, the beam is shut off at the gun.

## 3.3 Architecture

The system (Fig. 6) is based on a (dedicated, private) star network (Fig. 7) consisting of two entities: link processor and link nodes (interconnected over a private Gb ethernet network).

The machine protection system determines the maximum allowed beam rate by processing device fault input signals (from link nodes and input multiplexers) with a rate-limiting algorithm (executed on the link processor).

- The link node is the collection point of all sensor signals; it integrates sensor subsystems and drives mitigation devices.
- The link processor, in turn, runs the machine protection system control algorithm and makes decisions based on sensor states and interfaces to the timing system.



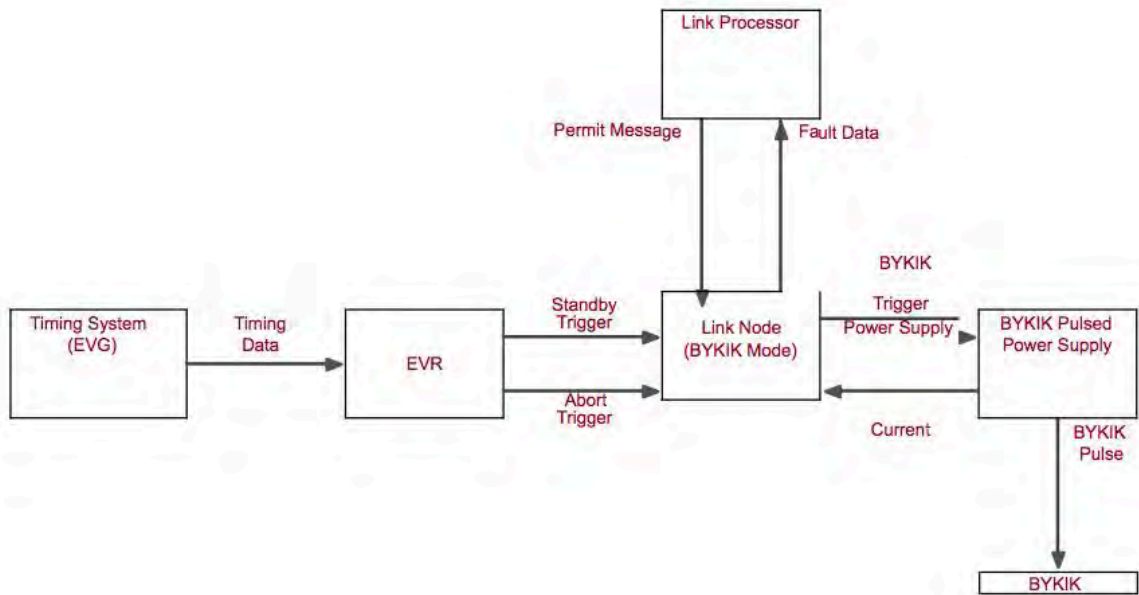


Fig. 5: BYKIK architecture: EVG, event generator; EVR, event receiver

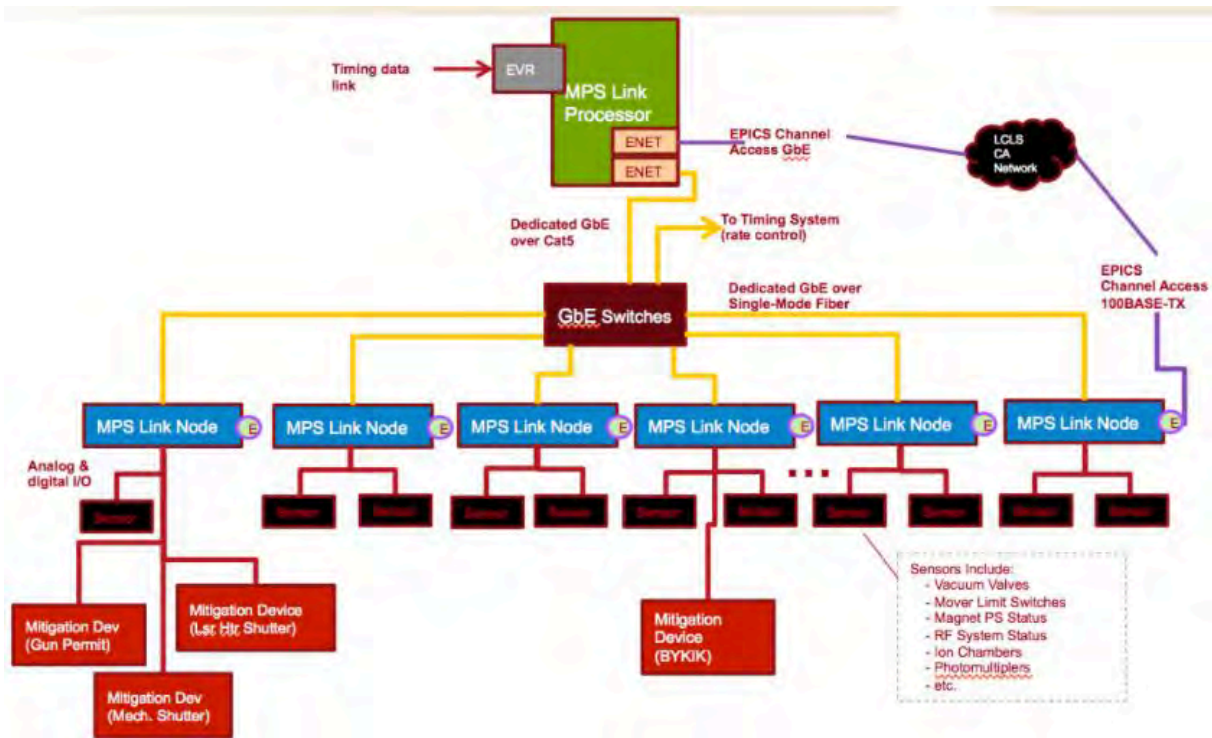
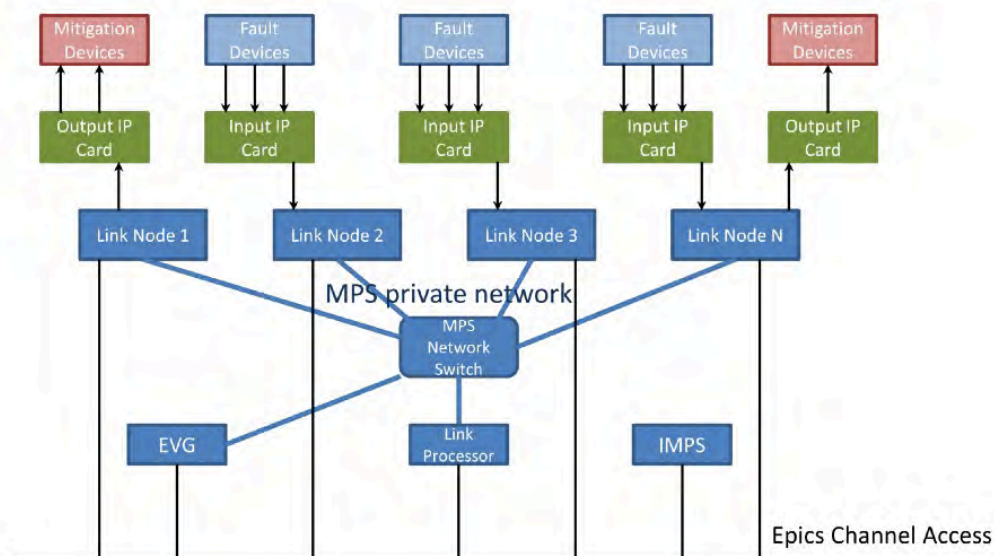


Fig. 6: Conceptual architecture diagram: CA, channel access; Dev, device; ENET, ethernet; EPICS, Experimental Physics & Industrial Control System; EVR, event receiver; GbE, Gb ethernet; LCLS, Linac Coherent Light Source; Lsr Htr, laser heater; Mech., mechanical; MPS, machine protection system; PS, power supply.



**Fig. 7:** Network architecture: EVG, event generator; IMPS, interface message processor system; IP, Internet protocol; MPS, machine protection system.

The link processor (a Motorola MVME 6100) has two copper Gb ethernet interfaces, a serial console port, and two peripheral component interconnect mezzanine card sites, along with an MPC7457 PowerPC processor that runs at 1.267 GHz, with 1 GB of RAM. The link processor's serial port is connected to a terminal server, a 1 Gb ethernet interface is used for high-speed communication with link nodes; the other is used for communication with the LCLS control system. It also sends synchronization and permit messages to link nodes. It faults all link node inputs to link nodes that provide a response within 8.3 ms.

The 32 LCLS link nodes are responsible for debouncing and latching digital inputs, digitizing analogue signals and comparing them with fault thresholds, and controlling the machine protection system mitigation devices. Link nodes are rack-mountable devices and occupy three rack units in a 19 inch rack. Built around the Xilinx Vertex four-field programmable gate arrays, each link node can be configured to support up to 96 digital inputs, 8 solid-state relay outputs, 4 TTL-compatible logic level trigger inputs, and 4 trigger outputs. One of each link node's two small form-factor pluggable slots is filled with a fibre-optic transceiver for high-speed communication with the link processor over the Gb ethernet. A full speed USB 1.1 port provides serial communication with the field programmable gate array while a separate DE-9 serial port gives access to the link node's EPICS (Experimental Physics & Industrial Control System) input-output controller serial port. The input-output controller serial ports are connected to terminal servers.

Four interface board slots allow signal conditioning to be placed between incoming signals and the link nodes' Industry Pack cards. Commercial off-the-shelf analogue-to-digital converter and digital-to-analogue converter Industry Pack cards are used to control and read back beam loss monitor high-voltage power supply voltages. A charge-integrating analogue-to-digital converter (QADC) Industry Pack card is used to digitize up to eight protection ion chamber or beam loss monitor signals, allowing each link node to monitor up to 32 analogue signals. The digitized signals are compared in the link node field programmable gate array against thresholds set by the link node's input-output controller via EPICS. Only the Boolean results of these comparisons are sent to the link processor for fault mitigation.

### 3.4 Communication

All time-critical data are transmitted over the machine protection system's dedicated Gb ethernet network using the user datagram or Internet protocol. The link processor uses a real-time protocol stack

originally created for the LCLS beam position monitor data acquisition system. The real-time protocol stack not only provides deterministic behaviour for the messaging, but also allows ordinary network hardware and software tools to be used to build and test the system, since no new protocols are introduced. On the link node side, the network stack is implemented in the field programmable gate array firmware. A stack of dedicated Gb ethernet switches connects the link nodes and the link processor. These switches queue and serialize concurrent data sent to the link processor and also handle the physical layer conversion of the link processor's copper and the link nodes' fibre Gb ethernet connections.

When the link processor is woken by the 360 Hz signal from the LCLS timing system, it broadcasts a synchronization message to all link nodes, requesting updated fault data, and providing the timing system's newest time-stamp. In response, the link nodes send the link processor a time-stamped status message containing all unacknowledged machine protection system device faults that have occurred since the previous synchronization message. The link processor copies the fault data to local buffers and returns the status message to the source link node. The link node uses this message as an acknowledgement of the faults that the link processor has received. All faults are latched in the link nodes and are cleared only when the link processor has acknowledged the fault and the fault itself has been cleared.

The link processor processes the faults using the currently running machine protection system logic and broadcasts a permit message to the link nodes. Link nodes allow the beam past their connected mitigation devices for 1/360 s if permitted. If a permit message is not received or if the beam is not permitted, link nodes stop the beam at their mitigation devices.

### **3.4.1 History server**

The link processor logs all fault and status messages to a machine protection system history server application, which stores the messages in an Oracle database in real time. Machine protection system messages are stored separately from the normal logging system so that no messages are lost; they are also forwarded to the normal message logging system, so that they can be correlated with other logged events. A machine protection system history viewer is available to the operators via the machine protection system graphic user interface.

The events logged are:

- device state changed;
- beam rate changed;
- destination changed;
- history servers notify link processor of their existence.

### **3.4.2 Faults bypass**

Device faults can be bypassed via an EPICS display by selecting a fault, choosing its bypass state, and supplying the bypass duration.

For example, an operator can choose to bypass a flow switch for one day by selecting the flow switch input, selecting its OK state, and giving a bypass duration of 24 h. All bypasses are logged and automatically timed by the machine protection system. The operator is alerted when the bypass time is reached, forcing the operator to re-evaluate bypasses.

## **3.5 Human-machine interface**

The human-machine interface is illustrated in Figs. 8 to 11.



Fig. 8: MAIL machine protection system: graphic user interface

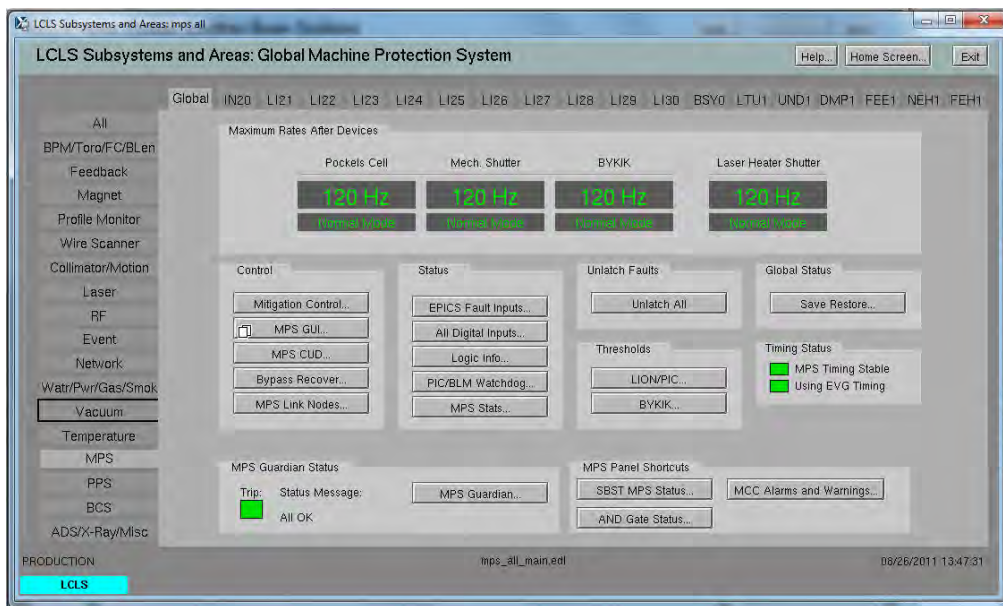


Fig. 9: Machine protection system: global panel

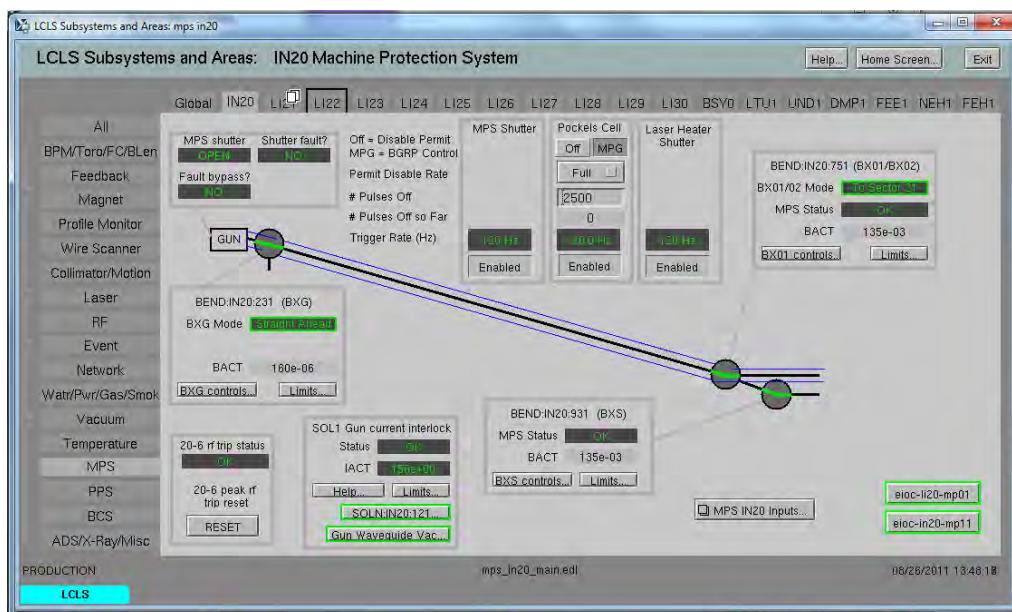


Fig. 10: Injector panel



Fig. 11: Injector inputs

### Acknowledgements

The author wishes to thank M. Boyes and F. Tao of SLAC for many fruitful conversations and their insights on machine protection systems for accelerators.

### References

- [1] ISO 14118 Safety of Machinery—Prevention of Unexpected Start-Up.
- [2] IEC 60204-1 Safety of Machinery, Electrical Equipment of Machines.
- [3] IEC 61508 Functional Safety of Electrical/Electronic/Programmable Electronic Safety-Related Systems.

## Bibliography

Centre for Chemical Process Safety, *Guideline for Safe Automation of Chemical Processes*, (AIChE/CCPS, New York, 1993).

SLAC, *Guidelines for Operations* (SLAC, Menlo Park, CA, 2008).

R. Langner, *Robust Control System Networks* (Momentum Press, New York, 2012).

W. Stallings, *Network and Internetwork Security* (IEEE Press, New York, 1995).

W.M. Goble, *Evaluating Control Systems Reliability* (ISA, Research Triangle Park, NC, 1992).

E. Marszal and E. Scharpf, *Safety Integrity Level Selection* (ISA, Research Triangle Park, NC, 2002).

D. Smith, *Reliability, Maintainability and Risk* (Butterworth-Heinemann, Burlington, MA, 2007).

W.M. Goble and H. Cheddie, *Safety Instrumented Systems Verification* (ISA, Research Triangle Park, NC, 2005).

C.A. Ericson II, *Hazard Analysis Techniques for System Safety* (Wiley-Interscience, Hoboken, NJ, 2005). <http://dx.doi.org/10.1002/0471739421>

H.E. Roland and B. Moriarty, *System Safety Engineering and Management* (Wiley-Interscience, Hoboken, NJ, 1990).

E. Schlosser, *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety* (Penguin, New York, 2014)

R.A. Stephans, *System Safety for the 21st Century* (John Wiley and Sons, Hoboken, NJ, 2004). <http://dx.doi.org/10.1002/0471662542>

K. Belt, *Certification Frequency* (SLAC Memorandum, Menlo Park, CA, 2012).

PICMG MTCA.0, revision 1.0, 2006-07-06, and revisions 2, 3, and 4.

*IEC 62381: Automation Systems in the Process Industry—Factory Acceptance Test (FAT), Site Acceptance Test (SAT) and Site Integration Test (SIT)*, 2006.

*ANSI/ISA 84.00.01* (IEC 61511 Mod).

*IEC 61511 Committee Draft*, 2012.

*ANSI/ISA TR84.00.03*.

## Beam Loss Monitors at LHC

*B. Dehning*

CERN, Geneva, Switzerland

### Abstract

One of the main functions of the LHC beam loss measurement system is the protection of equipment against damage caused by impacting particles creating secondary showers and their energy dissipation in the matter. Reliability requirements are scaled according to the acceptable consequences and the frequency of particle impact events on equipment. Increasing reliability often leads to more complex systems. The downside of complexity is a reduction of availability; therefore, an optimum has to be found for these conflicting requirements. A detailed review of selected concepts and solutions for the LHC system will be given to show approaches used in various parts of the system from the sensors, signal processing, and software implementations to the requirements for operation and documentation.

### Keywords

Machine protection; equipment protection; beam loss; dependability.

## 1 Introduction

After a LHC beam loss project study phase, a functional specification is compiled. The specification introduces the subject, first viewing the project globally by treating:

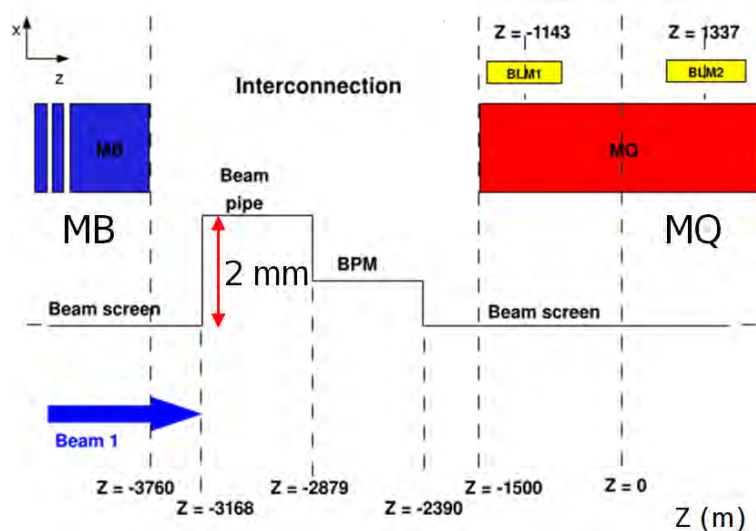
- location of monitors;
- time response;
- dynamic range;
- safety and reliability requirements.

The safety and reliability requirements need to be discussed at the system level, to define the overall quantitative requirements. The time response, dynamic range, safety, and reliability requirements limit the choice of sensors and define the acquisition chain. With the knowledge obtained in the project study phase, the following choices are made:

- sensor: ionization chamber;
- acquisition chain: distributed system with local and independent beam inhibit functionality.

A more detailed treatment of the global safety and reliability requirements has been covered in study groups and thesis projects. The subjects treated include:

- acquisition chain with:
  - parallel and voting for safety and reliability requirements;
  - radiation-tolerant electronics;
- fail-safe system;
- data flow path;
- management of settings;
- functional tests;



**Fig. 1:** Loss location considerations: aperture between a LHC bending magnet (MB) and a quadrupole magnet (MQ). The change in aperture is mainly controlled by the connection bellow and the beam position monitor (BPM) location. BLM, beam loss monitor.

- preventive actions;
- firmware updates;
- reliability software;
- human errors;
- documentation.

Several of these aspects will be discussed in this paper, and examples will be presented from the LHC beam loss monitoring system.

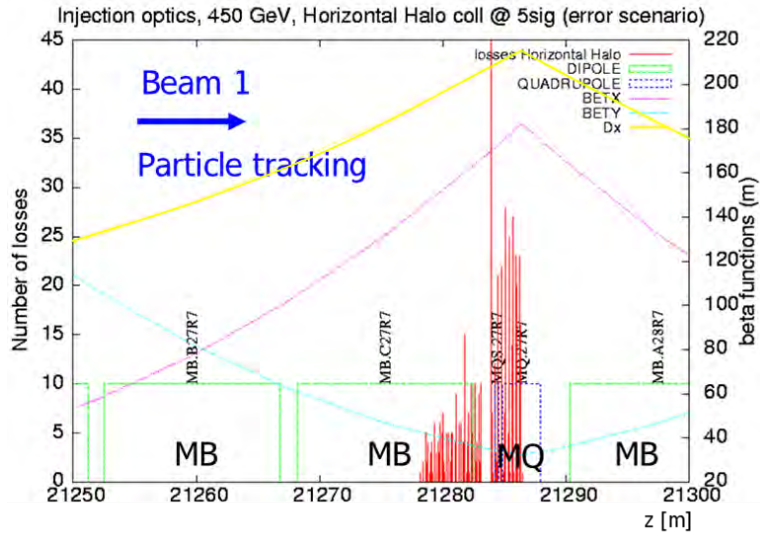
## 2 Global beam loss measurement requirements

For a beam loss protection system, the possible loss locations and therefore also the potential damage location are unknown parameters, to be addressed by particle tracking and particle shower simulations. In a second step, the optimal sensor locations are also determined by particle shower simulations. For the LHC, the considerations are illustrated in Fig. 1. The electrodes of the beam position monitors are retracted to be shielded by the nearby vacuum chamber walls against particle impacts, which could create electrical charges on the electrodes and disturb the beam position measurement. An aperture limitation results in a concentration of losses if off-orbit protons approach the aperture. At the LHC, this is the case for every transition between a bending and a quadrupole magnet. This can be visualized by the tracking simulation (Fig. 2), resulting in a maximum at the beginning of the quadrupole magnet. These loss locations are most probable, because:

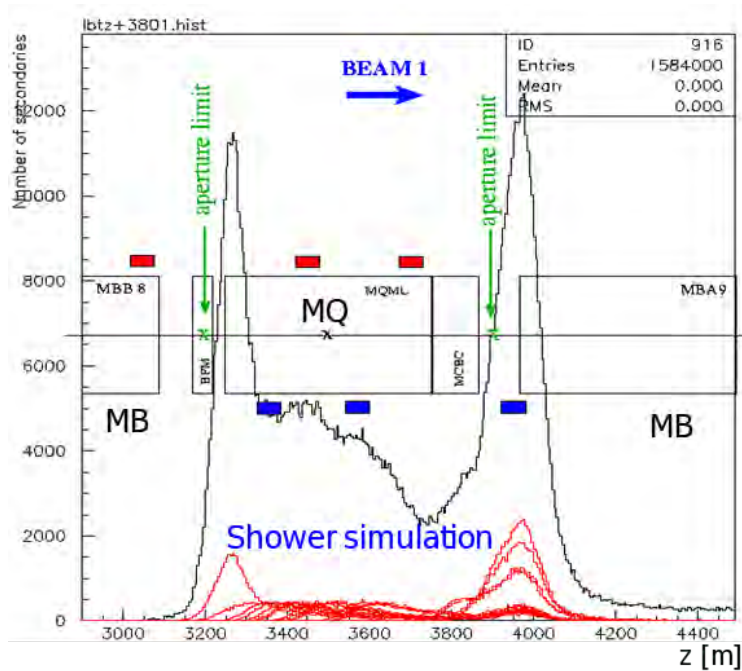
- the beta function, and therefore the beam size, is maximal;
- orbit bumps have a maximum at this location, because of the location of a dipole corrector magnet near to the quadrupole magnet;
- alignment errors are possible, causing an additional aperture limitation.

The shower particles initiated by lost protons can be best observed outside of the magnet yoke about a metre downstream of the proton impact location (see Fig. 3). A second maximum occurs at the downstream transition between the quadrupole and bending magnet, owing to the reduced material in the

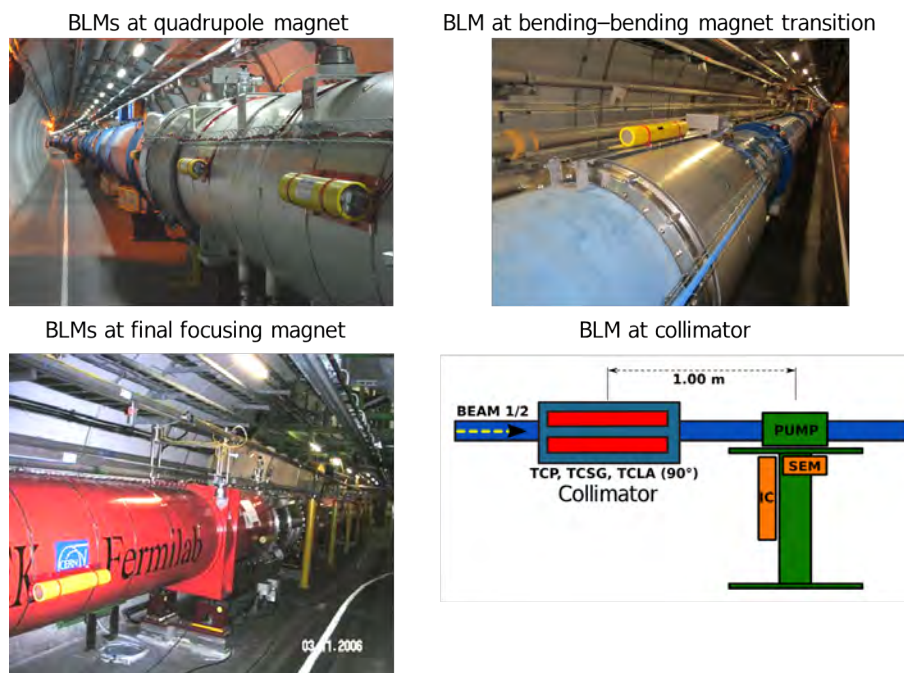




**Fig. 2:** Number of lost protons and beta function values with schematic of LHC regular cell as function of the location along the lattice. MB, bending magnet; MQ, quenching magnet.



**Fig. 3:** Number of secondary particles as function of location along the lattice. MB, bending magnet; MQ, quenching magnet.



**Fig. 4:** LHC tunnel photos with ionization chambers (yellow tubes) mounted on the outside of magnets and schematic of an ionization chamber near a collimator. BLM, beam loss monitor; IC, ionization chamber; SEM, secondary emission monitor.

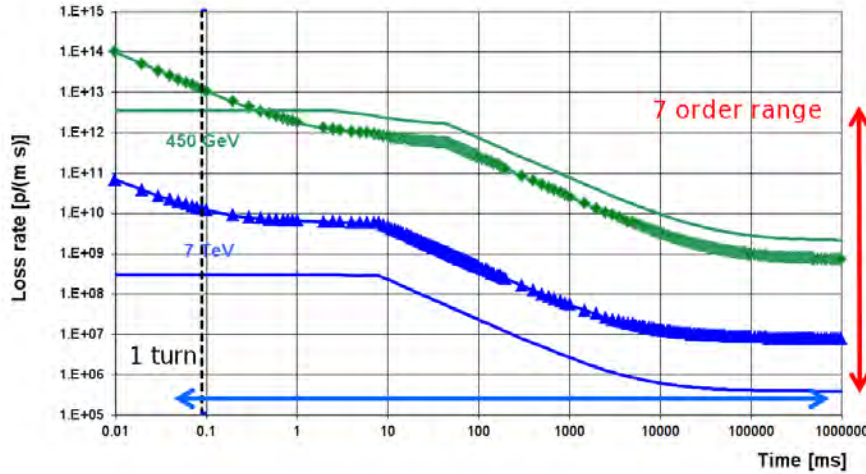
transition region. To make use of the high particle signal, resulting in the lowest statistical measurement error, the ionization chambers are located at or near to particle shower maxima (see Fig. 3, red and blue rectangular areas). A separation between the losses from beams 1 and 2 is given by the different locations of the shower particle maxima, owing to their opposite directions.

The LHC ionization chambers are cylindrical, with a sensitive volume of 1.5 l, covered by a yellow insulating tube and are mounted on the outside of the magnets or near collimators (see Fig. 4, bottom right, red and blue rectangular areas).

The limits of the time response and dynamic range requirements for LHC protection are mostly defined by the quench curves of the bending magnets. The quench levels of the magnets are orders of magnitude lower than the damage levels of the magnets. Magnet quenching is avoided, because of the gain in operational efficiency, by extracting the beam from the ring and therefore ending the deposition of heat in the coil before quenching can occur. In the case of a quench, the magnet coil is warmed up and the new cool down takes between 6 and 10 h. The allowed particle loss rate (see Fig. 5) in protons per metre per second is shown as the function of the loss duration. The characteristic superconducting magnet quench level curves are due to the quench margin of the superconducting cable filaments and the super fluid He cooling of the cables and the whole magnet coil. For short duration losses, the quench level is about four orders of magnitude higher than for steady-state losses and for both LHC nominal beam energies, of 450 GeV and 7 TeV, an order of two variation is seen.

The time resolution of the loss measurement system of 40  $\mu\text{s}$  is given by the duration of the extraction of the beam from the LHC, 89  $\mu\text{s}$ , and some signal propagations and synchronization considerations. The maximum duration is given by the reach of the steady-state quench level at about 80 s (see Fig. 5, blue arrow).

The maximal signal value is defined by the crossing of the 89  $\mu\text{s}$  line and the quench level at 450 GeV. Owing to an optimization process for the LHC acquisition electronics, the value has been chosen a little lower (see Fig. 5, vertical dashed black line (89  $\mu\text{s}$ ) and thin green line). The lower limit of



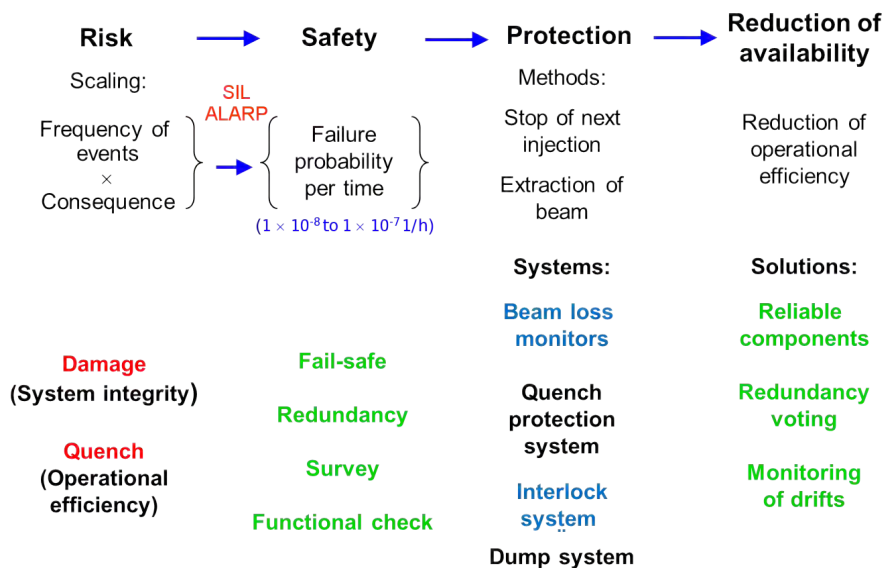
**Fig. 5:** Proton density rate as function of loss duration. Different curves indicate the functional dependence for different energies and the defined observation range. Red arrow, required proton density rate dynamic; blue arrow, duration dynamic.

the dynamic range is given by the steady-state quench level for 7 TeV and the need to observe losses, for accelerator tuning purposes, below the quench level (see Fig. 5, thin blue line, 80 s). These considerations led to a required signal dynamic of over seven orders of magnitude (see Fig. 5, red arrow). Operational experience required that the dynamic upper value be extended by two orders of magnitude for short-term losses in injection areas.

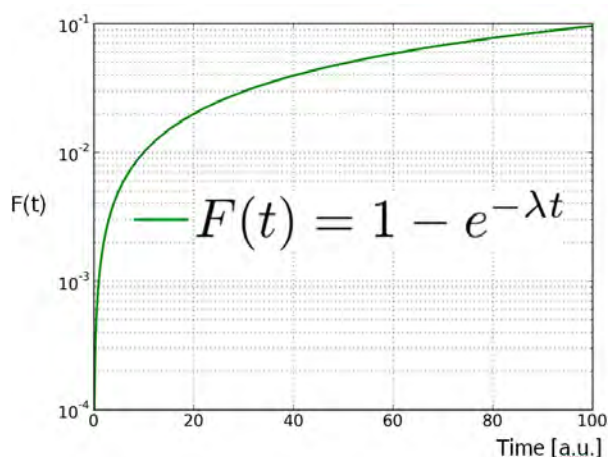
### 3 Safety system design approach

All considerations start with the recognition that the probable frequency and probable magnitude of a non-conformal behaviour could lead to a damage of the system integrity. The combined likelihood of frequency and magnitude determines the risk for a certain system (see Fig. 6, first column). The risk could be reduced by using a safety system providing protection, but increased complexity reduces the availability of the protected system (see Fig. 6, first row). To arrive at a quantitative demand for a safety level, the probable frequency of events and the probable magnitude of its consequence are utilized by the SIL (safety integrity level) approach [1] or the ‘as low as reasonably practicable’ (ALARP) approach.

For both approaches, a failure probability per time is estimated by calculating the risk of damage and the resulting downtime of the equipment [2]. A failure in the safety system itself should fall in a fail-safe state, with the consequence of reducing the operation efficiency. The main design criteria for the safety system are listed in the safety column of Fig. 6: fail-safe, redundancy, survey, and functional check. The protection column of Fig. 6 lists the methods for the protection of an accelerator: stop of next injection applicable for a one-path particle guiding system (linac, transfer line) and extraction of the beam for a multipath system (storage ring). The accelerator safety system consists of a beam loss measurement system, an interlock system, and a beam dump system. If superconducting magnets are used, some beam loss protection could also be provided by the quench protection system. The availability column of Fig. 6 lists the means used in the design of the safety system to decrease the number of transitions of the system into the fail-safe state. The effect of the number of components added to a system to increase the probability of a safe operation results in a reduction in the availability of the system. This negative consequence of the safety-increasing elements is partially compensated by the choice of reliable components, by redundancy, voting, and the monitoring of drifts of the safety system parameters.



**Fig. 6:** LHC protection system design approach (items in green are discussed in this paper). ALARP, as low as reasonably practicable; SIL, safety integrity level.

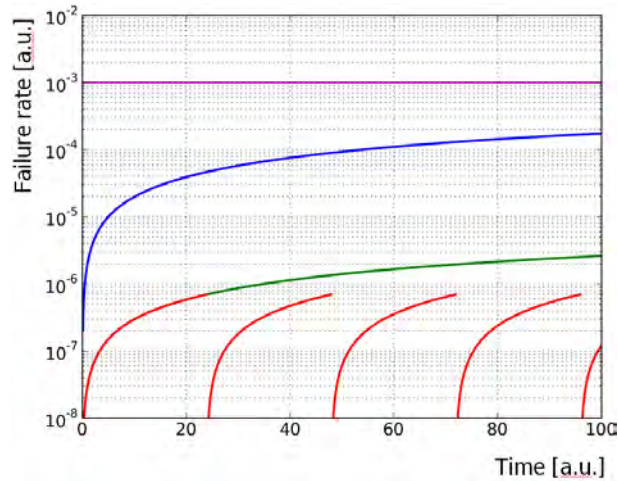


**Fig. 7:** Exponential failure probability

#### 4 Failure probability and failure rate reduction

To illustrate the available means of increasing safety, the system’s basic functional dependencies are discussed. An often-valid assumption is given by the exponential time dependence of the failure probability  $F(t)$  (Fig. 7). With increasing time, the probability of the occurrence of a failure in a system approaches 1. The failure rate,  $\lambda$ , is assumed to be time-independent (Fig. 8, magenta curve). In a next step, two systems with the same functionality are assumed to be working in parallel, to allow redundant operation. The failure rate,  $\lambda$ , decreases drastically for short times, but finally approaches the failure rate of a single system (Fig. 8, blue line).

It should be noted that the failure rate curve changes from time-independent to time-dependent behaviour. A further reduction in the failure rate could be reached by a survey of the system. With a system survey, some failure modes can be detected in advance and a repair can be planned (see Fig. 8, red and green line). This procedure results in a shift of the failure rate curve to lower values, which no longer approach the infinite times of the single system rate. Another strong reduction could be reached



**Fig. 8:** Failure rates of different systems as a function of time (arbitrary units). Magenta: single system. Blue: Two systems parallel. Green: Parallel systems with survey. Red: Parallel systems with survey and with regular test.

if the system could be regarded as new after a certain time period. The failure rate curve shows the time dependence of the surveyed system in the period  $t_0 = 0$  to  $t = t_1$  repeated after every time period (see Fig. 8, red lines). The conclusion that a system could be regarded as new after a certain time is justified if the system is subjected to a test. Functional tests will verify, on request, that the system has the defined functionality. In case of an internal system failure system, the very basic requirement is a fail-safe behaviour. Internal failure will not contribute to the unsafeness of the system but will contribute to its non-availability.

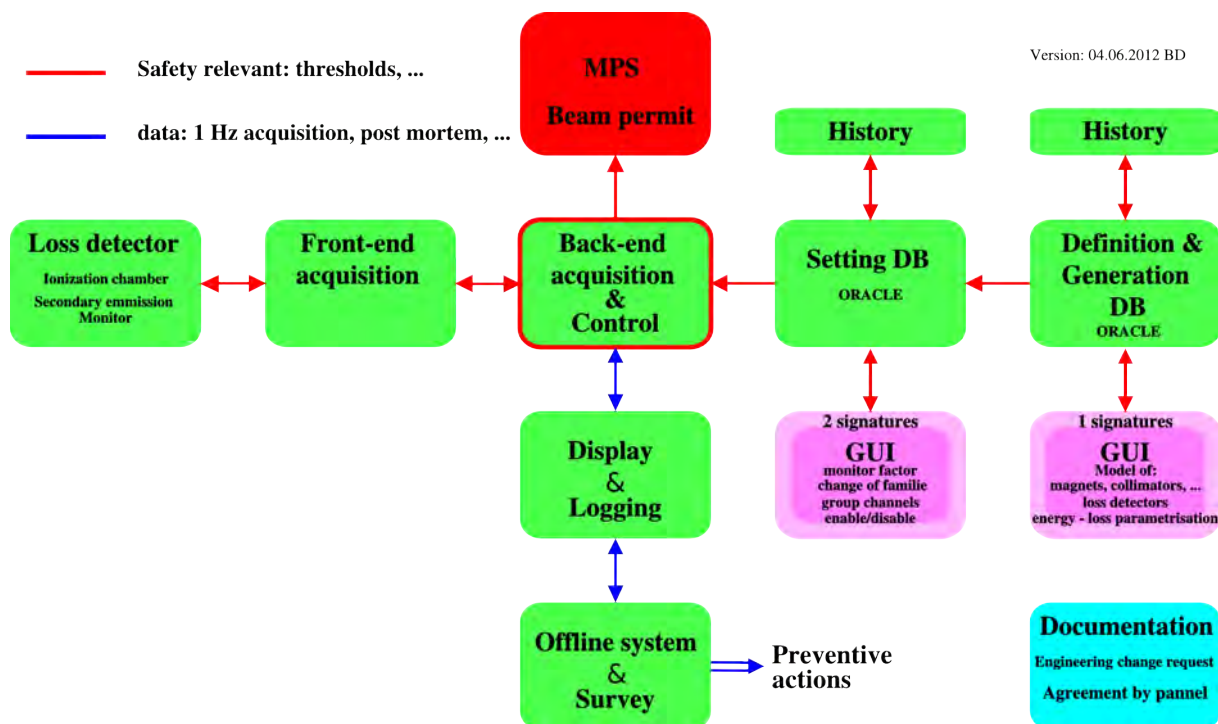
## 5 Protection system overview

As an example of a protection system, the CERN LHC beam loss monitoring (BLM) system will be used. The discussion will focus on protection, reliability, and availability aspects.

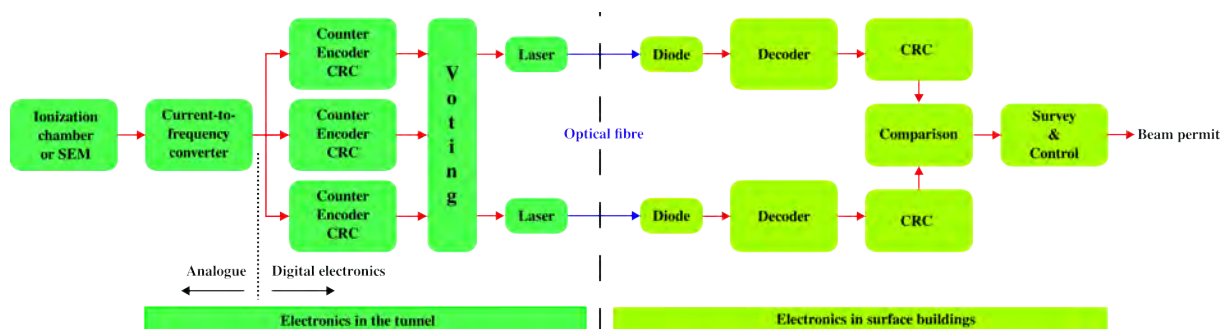
The main purpose of the BLM system is to convert particle shower information into electrical signals, which are then compared with limits. If the limits are exceeded, extraction of the LHC beam from the ring is initiated to stop the irradiation of equipment. In the case of the LHC, the protection function is often linked to the quench prevention of the superconducting magnets, since the threshold levels for beam extraction are lower (orders of magnitude) than for the damage protection of equipment [3].

The very first element of the protection system is the sensor that detects the irradiation of equipment. The conversion of the particle shower magnitude is done by ionization chambers [4] or secondary emission detectors [5] (see Fig. 9, left block). The front-end acquisition electronics convert the analogue detector signal into a digital signal and transmit the signal to the back-end acquisition and control unit, which is the decision-making centre of the whole system. The measured signals arrive here and are compared with the limits. In addition, beam permit signals are generated (see Fig. 9, red block), taking the information of the system settings (see Fig. 9, right-hand blocks) into account. The measurement data and all setting information are also distributed to the display and the logging databases (see Fig. 9, bottom blocks) from this unit. The control functionality is linked to the survey and test functionality, which are discussed later.

In the LHC, ionization chambers [4] and secondary emission detectors [5] are used. Their signals are digitized using a current-to-frequency converter [6, 7] (see Fig. 10, front-end acquisition unit in tunnel). Up to the end of the analogue signal chain, the signal is not redundant, because no technical solution has been found to the problem of splitting the detector signal while simultaneously allowing a large dynamic signal (nine orders of magnitude). To cope with this requirement for the analogue front-



**Fig. 9:** Information flow from the sensor up to the beam permit signal transmission. The red framed (back-end acquisition and control) unit is the local decision-making centre.



**Fig. 10:** CERN LHC beam loss measurement and protection system: CRC, cyclic redundancy check

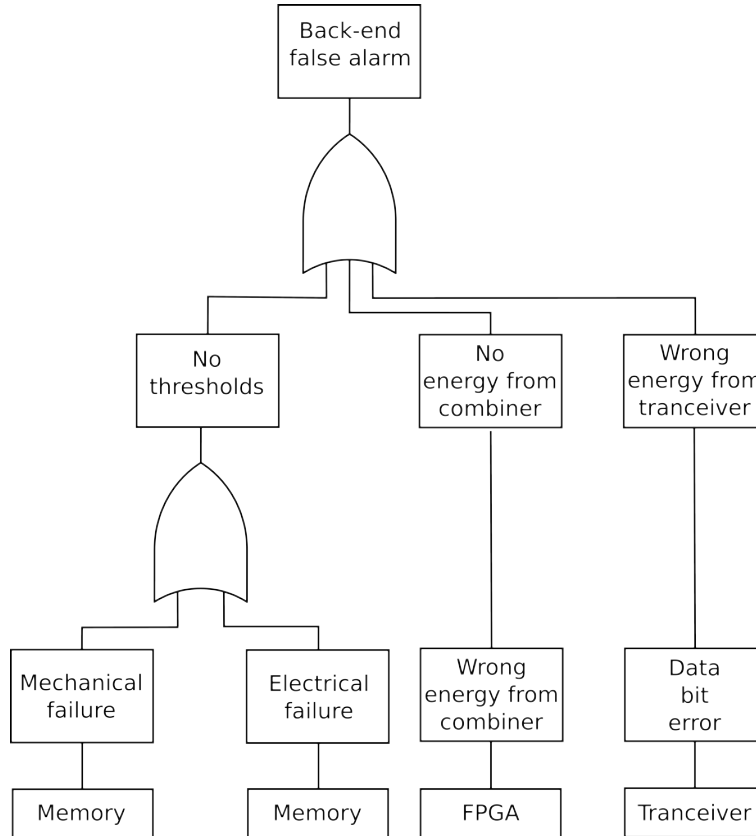
end unit, a low failure rate circuit concept has been chosen. To avoid the consequences of single event effects, and to increase the availability of a channel, the signal is trebled in the front-end logic. Two voting blocks are used to generate the signal transmitted over a redundant optical link. A redundant optical link has been chosen to increase the availability of the link, which is limited by the mean time between failures of the transmission laser.

The signals are decoded and cyclic redundancy checks (CRCs) are calculated for both signal chains (see Fig. 10, back-end acquisition unit at the surface). At the front-end unit, CRCs are also calculated and transmitted, to enable the CRCs of each line and also the CRCs for both lines to be compared. This procedure ensures high reliability and also maximizes the availability of the data link [8, 9].

The effect of the implementation of redundancy and trebling in the data transmission and treatment and the verification of loss-free data transmission are listed in Table 1. The most important technique for increasing the reliability of a system is given by a fail-safe design. In the case of an internal failure of a system, it should make the transition to a state that ensures the protection of the system. This could be

**Table 1:** Procedure and techniques to increase the reliability and availability of acquisition systems

|            | Comment position of monitor | Safety gain | Availability gain |
|------------|-----------------------------|-------------|-------------------|
| Fail-safe  | Active state = beam permit  | Yes         | No                |
| Voting     |                             | Yes         | Yes               |
| Redundancy |                             | Yes         | Yes               |
| CRC        | Cyclic redundancy check     | Yes         | No                |



**Fig. 11:** Image section of the false alarm generation fault tree of the LHC BLM system, showing the part describing the back-end acquisition unit.

done by assigning the active state to: ‘system is allowed to operate’. In case of an internal failure, e.g., if no power is supplied, the state will switch to a passive state and the system will be protected.

## 6 Fault tree analysis

The fault tree treatment of the system has been chosen to calculate, from the component level up to the system level, the damage risk, the false alarm, and the warning probability [10], taking into account the component failure, repair and inspection rates.

The false alarm slice of the fault tree (see Fig. 11) shows the signal chain for different false alarm generators (memory, beam energy from control unit (combiner), and energy tranceiver) of the back-end electronics [11]. The different inputs are linked with a Boolean ‘OR’ gate so that every single input generates, in the same way, a false alarm and, therefore, a downtime of the system and the LHC.

The results of the fault tree analysis have been essential for the design of the hardware and software, especially for the estimates of failure rates of the optical links and the propagated consequences of it up to the system damage and false rate probabilities. An optimization process has been instigated, to balance the probabilities of damage rates and false alarms. The failure rate calculations also lead to the

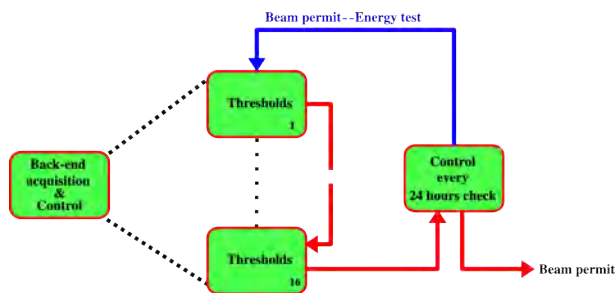


Fig. 12: Beam permit line functionality check

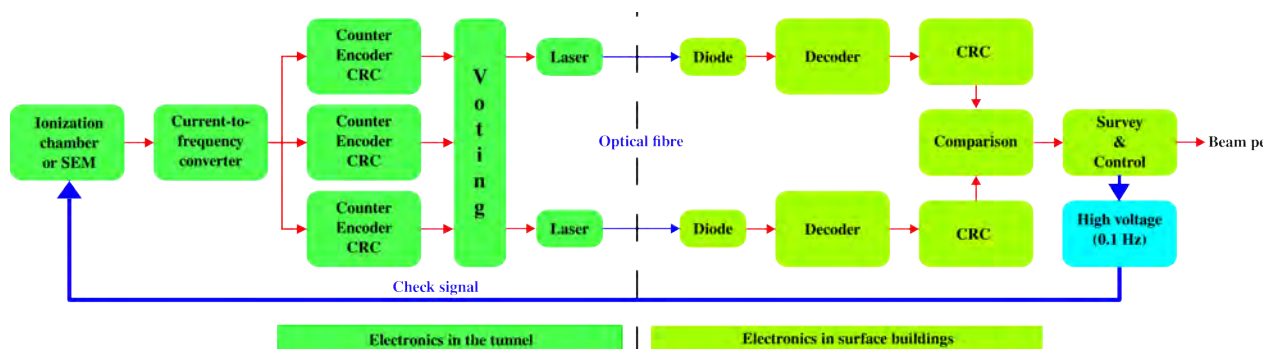


Fig. 13: Check of the whole acquisition chain

definition of functional tests and their frequencies. Failure modes are also defined for the limit values, detector names, channel assignments, and much more data needed by the system. Therefore, setting management and metadata verification tests are also treated in the fault tree analysis.

## 7 Functionality checks

As an example of a check, the signal distribution inside the VME crate for the beam energy and the beam permit line test is discussed [12, 13] (see Fig. 12). The test is initiated by a client, to allow optimal scheduling. The control unit (combiner card) holds a downtime counter requiring every 24 hours the execution of functional tests every 24 hours. If the tests are not completed in time, the downtime counter inhibits the beam permit immediately if no beam is circulating or when the beam present flag becomes false. For the tests, the whole system changes the status to 'mode' and, e.g., the control units send a request to inhibit the beam permit line to each acquisition card (threshold card) in sequence (see Fig. 12).

The test results are analyzed by the controller; if a false status is detected, a manual intervention is required, to repair the system before the test can be passed without a false status detected. The distribution of the beam energy levels between the controller and the acquisition card is tested by changing the energy levels in the test mode; this should result in the acquisition card returning the appropriate threshold settings for comparison with the settings sent.

In a second example, the test of the whole acquisition chain is presented [14, 15]. An electrical signal is introduced in the sensor by the capacitive coupling of the sensor electrodes and by a harmonic modulation of the applied high voltage supply (see Fig. 13). This test includes the complete signal chain, except for the ionization process in the ionization chambers and the secondary electron emission in the secondary emission monitor detectors. The conversion of the particle shower to an electrical signal in the detector is tested every few years with a radiative source placed outside the detector. The long intertest interval for this test is possible because the failure mode of a complete gas exchange with air (ionization chamber) or loss of the vacuum (secondary emission detector) of the detectors will still result in an



**Table 2:** Parameters deployed on each back-end unit (threshold comparator module)

| Parameters                           | Data 32 bit | Description                                       |
|--------------------------------------|-------------|---|
| Threshold values                     | 8192        | 16 channels $\times$ 12 sums $\times$ 32 energies |
| Channel connected                    | 1           | Generating (or not) a beam permit                 |
| Channel mask                         | 1           | 'Maskable' or 'unmaskable'                        |
| Serial A                             | 1           | Card's serial number (channels 1–8)               |
| Serial B                             | 1           | Card's serial number (channels 9–16)              |
| Serial                               | 2           | Threshold comparator                              |
| Firmware version                     | 1           | Threshold comparator's firmware                   |
| Expert names                         | 128         |   |
| Official names                       | 128         |   |
| DCUM                                 | 16          | Position of monitor                               |
| Family names                         | 128         | Threshold family name                             |
| Monitor coefficients                 | 16          | Monitor threshold coefficients                    |
| Last link-state advertisement update | 2           | Time stamp: master table                          |
| Last flash update                    | 2           | Time stamp: non-volatile memory                   |
| Flash checksum                       | 1           | CRC value for or from table integrity             |

appropriate signal, without loss of protection functionality. Also, this test is initiated and the results are analyzed by the back-end unit (survey and control) (see Fig. 13), allowing the beam permit line to be inhibited directly in the case of a negative result.

## 8 Setting management

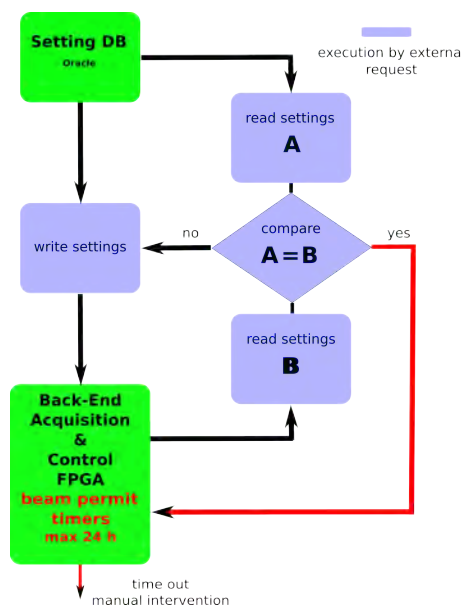
The system setting management controls the settings for the beam permit thresholds and also the settings used for system operation [16, 17]. These operational settings include hard and firmware information, to verify that the configuration stored in the database images the installed system. Table 2 illustrates the variety of the metadata needed to interpret the measured values or to check the configuration of the system. For example, a match between measured value, channel official names, channel expert names, DCUM (position of monitor), and monitor coefficient needs to be given and tested. To reduce the complexity of the metadata information chain (see Fig. 9, right blocks), a single path is defined for the metadata flow and the measurement values into the back-end unit. The back-end unit distributes the measurement values together with the metadata to ensure consistency and to have only one location where the data integrity needs to be tested. This concept is essential to reduce the number of possible failure modes for metadata corruption.

Having expressed the importance of a failure mode optimized metadata flow, the data check is achieved by comparing the data stored in a reference setting database (Oracle) with the data stored in the memory of the back-end electronics field-programmable gate arrays (see Fig. 14). Also, in this test, a downtime counter located in the back-end unit (survey and control) requests a comparison of the data stored at both locations every 24 h. If the test is not initiated, or if the test result is negative, the beam permit is inhibited. Since the comparison is made in a different software environment, the additional functionality required in the back-end unit is marginal, but it is necessary to test the comparison code from time to time.

### 8.1 Descriptive metadata

Metadata need to be generated and the option for required changes needs to be provided. To reduce human error, the graphical user interfaces (GUI) accessing the setting database (see Fig. 9, right block) need to be optimized by allowing for all data manipulation steps to include comparisons with previous data, checks on the magnitudes of changes, and several confirmation steps. The last confirmation steps require the electronic signatures of two independent persons.

The generation of sets of metadata required initially and for larger changes during the operation periods is done for the LHC system by a GUI for the database access. Generation of metadata, such as



**Fig. 14:** Comparison of descriptive metadata reference settings with settings in the back-end acquisition and control unit. The decision logic is indicated in the flow diagram. FPGA, field-programmable gate array.

limits for the beam abort thresholds, are parameterized and the calculation is made by code loaded into the database (Oracle) (see Fig. 9, rightmost block). The calculation made in the database environment, where database software changes and updates are made in a coherent manner, should ensure long-term maintainability [18].

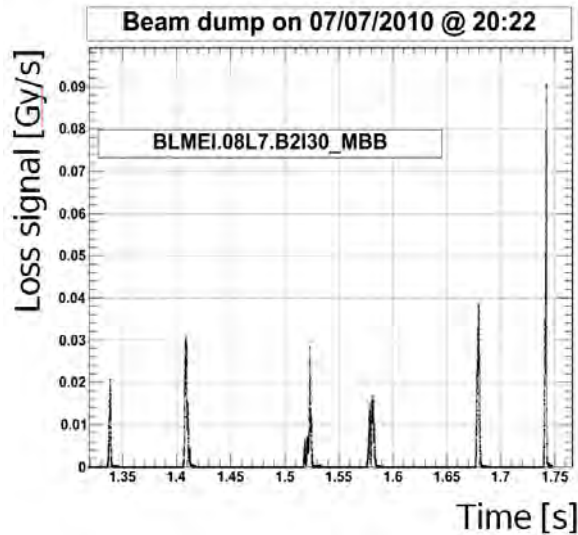
## 8.2 Documentation

In a complex system, designed for operation over decades, sufficient documentation is essential to describe the system for knowledge transfer. For a safety system, the function of the documentation is to avoid failure modes and failures. The design documentation, from the specification to documentation on operation and system changes, needs to be distributed for review, comment, and final approval by each client. At the LHC, standardized forms, electronic procedures, and signatures are in use to organize the process, e.g., an engineering change request outlines the motivation for a change, the description of the proposed change, and an estimate of the impact of the change on the functionality of the concerned system and other systems.

## 9 Snapshots of loss measurements triggered by events

The loss measurement recording rate has been set up at different speeds, with 40  $\mu$ s, 80  $\mu$ s, 80 ms, and 1.3 s integration times. The two first periods are event-triggered, to cope with the amount of data, while the latter periods are read out at 12 Hz and 1 Hz. The event-triggered measurements are used to analyze losses occurring at particular times during operation or depending on measurements and output analysis data acquisition freezing events. The 12 Hz measurements are used for the collimator positioning feedback system and the 1 Hz measurements are used for continuous observation of the accelerator status.

High-resolution data have been used not only for the detailed study of beam losses caused by dust events (see Fig. 15), but also to check for non-conformities of the acquisition system. When testing the system under extreme conditions, high loss levels with a large leading signal transition give an insight into system performance. The advantage of publishing different measurement signals is that it enables consistency checks to be performed. In the LHC, several clients are used to check the consistency of measurement data.



**Fig. 15:** Example of a particle loss triggered event recording. The trigger has been generated at 1.74 s. The measurements recorded before the trigger event reveal loss precursors. The losses are caused by collisions between the beam and dust particles.

## 10 Acquisition database

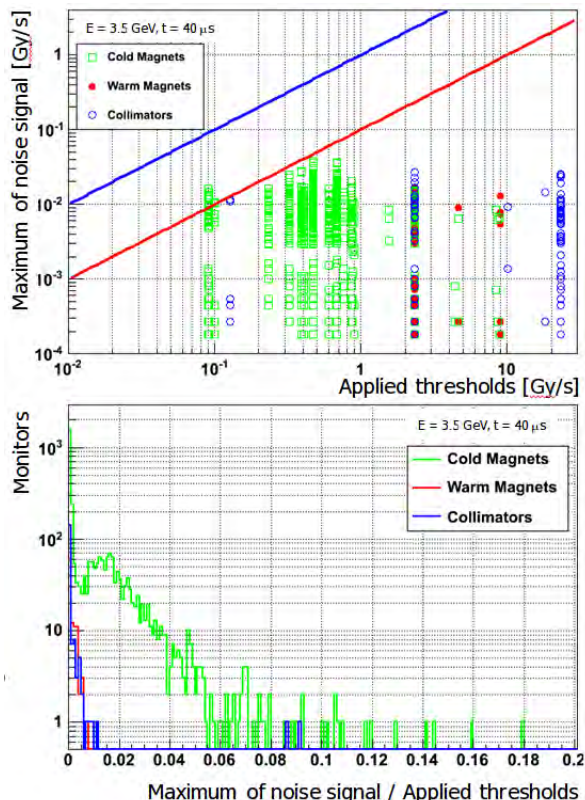
The storage and fast retrieval of measurement data and metadata is also essential for system checks. Besides the examples discussed previously, for which extended data storage were required, an extreme case is the check of noise amplitudes of the system (see Fig. 16). For a protection system with limits leading automatically to a beam abort and to accelerator downtime, there is a strong requirement to avoid false aborts caused by rare events (noise). This is extreme, because rare signals need to be retrieved from stored measurement data from acquisition periods lasting weeks. The measurements with the shortest integration periods of 40  $\mu\text{s}$  show the largest signal fluctuation, because signal averaging does not lead to a reduction in signal fluctuation. To reduce the amount of data to be stored, an on-line measurement data reduction algorithm has been implemented in the back-end unit. Only maximum values of the short integration times are stored for the 1 Hz read-out. This procedure reduces the quantity of data to be stored by over four orders of magnitude. In addition, a retrieval time optimized database structure has been implemented for this purpose.

## 11 Preventive action

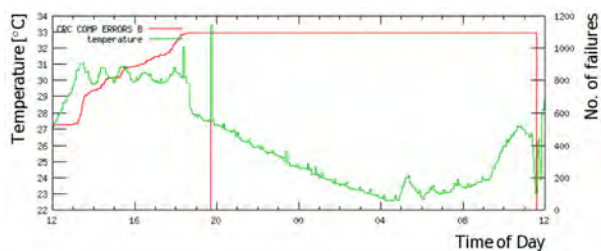
The discussion in Section 4 emphasized the reduction in failure rate achieved by surveying the system, to anticipate possible failure modes. In the LHC system, this survey task is realized by the daily retrieval of relevant database information and an automatic comparison with limits for initiating actions. Reports containing different levels of abstraction are produced daily and weekly. An example of this procedure is given by the survey of optical links. The links are redundant (see Fig. 10) and the calculations of different CRCs enable the differences between the CRC values to be recorded and correlated with board temperature variations (see Fig. 17). The limits for actions are set empirically, to minimize downtime and maintenance efforts.

## 12 Summary

A systematic design approach for machine protection systems will start with determination of the system failure rate. The failure rate magnitude could be based on well-established standards first developed for the design of military equipment, the aircraft industry, space missions, or nuclear power stations. The



**Fig. 16:** Noise level determination of all beam loss monitor channels. The LHC loss monitor channels are grouped by the observed loss, creating elements of cold and warm magnets and collimators. Top: Beam loss monitor noise signal taken with no beam circulating versus beam abort thresholds. The blue line indicates the threshold value and the red line the maximum noise goal set to avoid any noise false beam aborts. Bottom: Beam loss monitor spectrum normalized to the beam abort threshold.



**Fig. 17:** Optical link failures and printed circuit board temperatures versus time of day

effect of increasing complexity by adding protection functionalities and therefore reducing availability is best studied by reliability software packages [19]. The basic means of delivering a reduction in failure rate are provided by a system layout with parallel, redundant information, treated in combination with a regular survey of the system status and functional tests. A survey will allow preventive actions, to reduce the failure rate. For a protection system, a fail-safe design is essential so that protection is ensured in the case of a failure.

Functionality checks staged for all levels of the signal treatment are implemented for the LHC BLM system. The checks of the information exchange inside the VME crate and the analogue and digital signal chain have been discussed. Examples have been given to emphasize the importance of the metadata information flow. The combination of measurement and metadata as early as possible in

the signal chain is important for the reduction of failure modes and simplified test options. To attain low failure rates, rigorous metadata tests have to be implemented, to ensure metadata conformity. The generation of metadata and change options using a graphical interface also need to be analyzed in terms of failure modes, taking into account long-term usage and the maintainability of tests and validation procedures in the future. For the LHC, the most stringent requirement in avoiding human error is the request of two signatures to validate metadata changes. Although listed last, documentation tasks should be started first, including planning for reliability measures, and have to be continued as long as the system exists.

## References

- [1] International Electrotechnical Commission, IEC 61508. IEC, 2010.
- [2] G. Guaglio, Reliability of beam loss monitors system for the Large Hadron Collider, 11th Beam Instrumentation Workshop, Knoxville 2004 (AIP, 2004), vol. 732, p. 141, <http://hal.in2p3.fr/in2p3-00025196>
- [3] B. Dehning *et al.*, Overview of LHC beam loss measurements, 2011, p. THOAA03, <https://cds.cern.ch/record/1379469>
- [4] M. Stockner. Ph.D. thesis, Technische Universität Wien, 2006.
- [5] D. Kramer, Ph.D. thesis, Technical University of Liberec, 2008.
- [6] E. Effinger *et al.*, The LHC beam loss monitoring system's data acquisition card, 12th Workshop on Electronics for LHC and Future Experiments, Valencia, Spain, 2006, p. 108, <http://cdsweb.cern.ch/record/1027422>
- [7] E. Effinger *et al.*, Single gain radiation tolerant LHC beam loss acquisition card, Proc. DIPAC, Venice, Italy, 2007. p. 319. <http://accelconf.web.cern.ch/Accelconf/d07/papers/wepc06.pdf>
- [8] C. Zamantzas, *et al.*, An FPGA based implementation for real-time processing of the LHC beam loss monitoring system's data, San Diego, 2006, IEEE Nucl. Sci. Symposium Conf. Record (2006), vol. 2, p. 950, [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=4179157](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4179157)
- [9] C. Zamantzas, Ph.D. thesis, Brunel University, 2006.
- [10] G. Guaglio. Ph.D. thesis, Université Blaise Pascal, Clermont-Ferrand II, 2005.
- [11] Reliability software from Isograph – world leaders in reliability, maintenance and safety, <http://www.isograph.com>
- [12] C. Zamantzas *et al.*, Reliability tests of the LHC beam loss monitoring FPGA firmware, 14th Beam Instrumentation Workshop, Santa Fe, New Mexico, 2010, <https://cds.cern.ch/record/1268403>
- [13] B. Dehning *et al.*, Self testing functionality of the LHC BLM system, 10th European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators, Hamburg, Germany, 2011, p. 152, <https://cds.cern.ch/record/1375171>
- [14] J. Emery, *et al.*, First experiences with the LHC BLM sanity checks, Topical Workshop on Electronics for Particle Physics 2010, Aachen, Germany, 2010 [*J. Instrum.* **5** (2010) C12044. <http://dx.doi.org/10.1088/1748-0221/5/12/c12044>], <https://cds.cern.ch/record/1321592>
- [15] J. Emery *et al.*, LHC BLM single channel connectivity test using the standard installation, Beam Diagnostics and Instrumentation for Particle Accelerators, Basel, Switzerland, 2009, <https://cds.cern.ch/record/1183414>
- [16] E. Nebot Del Busto *et al.*, Handling of BLM abort thresholds in the LHC, 2nd International Particle Accelerator Conference, San Sebastian, Spain, 2011, p. WEPC170, <https://cds.cern.ch/record/1379461>
- [17] E. B. Holzer *et al.*, Generation of 1.5 million beam loss threshold values, 11th European Particle Accelerator Conference, Genoa, Italy, 2008, p. THPC147, <https://cds.cern.ch/record/1124306>

- [18] M. Nemcic, B.Sc. thesis, University of the West of England, Bristol 2012, [http://ab-div-bdi-bl-blm.web.cern.ch/ab-div-bdi-bl-blm/talks\\_and\\_papers/Nemcic](http://ab-div-bdi-bl-blm.web.cern.ch/ab-div-bdi-bl-blm/talks_and_papers/Nemcic)
- [19] S. Bhattacharyya, Ph.D. thesis, Ohio State University, 2012.

# Machine Protection and Interlock Systems for Circular Machines—Example for LHC

*R. Schmidt*

CERN, Geneva, Switzerland

## Abstract

This paper introduces the protection of circular particle accelerators from accidental beam losses. Already the energy stored in the beams for accelerators such as the TEVATRON at Fermilab and Super Proton Synchrotron (SPS) at CERN could cause serious damage in case of uncontrolled beam loss. With the CERN Large Hadron Collider (LHC), the energy stored in particle beams has reached a value two orders of magnitude above previous accelerators and poses new threats with respect to hazards from the energy stored in the particle beams. A single accident damaging vital parts of the accelerator could interrupt operation for years. Protection of equipment from beam accidents is mandatory. Designing a machine protection system requires an excellent understanding of accelerator physics and operation to anticipate possible failures that could lead to damage. Machine protection includes beam and equipment monitoring, a system to safely stop beam operation (e.g. extraction of the beam towards a dedicated beam dump block or stopping the beam at low energy) and an interlock system providing the glue between these systems. This lecture will provide an overview of the design of protection systems for accelerators and introduce various protection systems. The principles are illustrated with examples from LHC.

## Keywords

Machine protection; interlock system; high-power accelerator; beam loss; accident.

## 1 Designing a protection system for particle accelerators

The approach to the design of a machine protection system (MPS) starts with the identification of the hazards. A number of failures are identified that can change beam parameters and lead to the loss of particles that would hit the aperture and possibly damage equipment. A simple procedure to get started includes several steps.

1. There is a very large number of failures that can cause beam losses. The failures are classified in different categories.
2. The risk for each failure (or for categories of failures) is estimated.
3. The worst-case failures and their consequences are identified.
4. Prevention of a failure or mitigation of the consequences of a failure is worked out.
5. This allows to start with the design of systems for machine protection.
6. Back to item 1.

The design of the systems for machine protection starts during the early design phase of an accelerator. Since new hazards are identified during the life cycle of an accelerator, in particular during upgrades and modifications, new mitigation methods are developed; the process continues throughout the life cycle. It is required until end of operation.

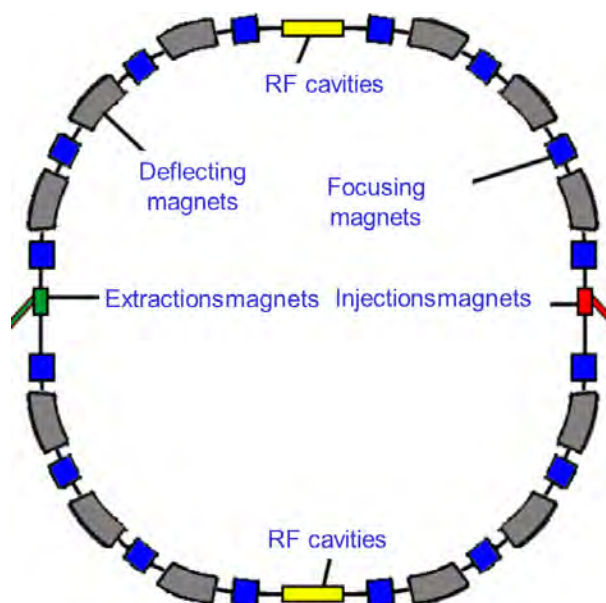


Fig. 1: Typical components of a circular accelerator

## 2 Circular accelerators and LHC

The main components of a synchrotron are deflecting magnets, magnets to focus the beam and correction magnets (Fig. 1). Pulsed magnets are required for injection and extraction. Power supplies provide the magnet current. Radio-frequency (RF) cavities accelerate the beam; the RF power is provided by the RF system. Other systems include beam instrumentation and the control system. The vacuum system ensures very low pressure for the beam circulating in the vacuum chamber.

The beams are injected at low energy and the energy is increased while ramping the magnetic field. The particles are accelerated by RF fields in RF cavities.

The layout of the LHC is shown in Fig. 2 [1]. It has eight arc sections and eight straight sections (or insertions) for experiments and accelerator systems. In four straight sections physics experiments are installed; in IR2 and IR8 these are together with the systems for injecting the beams coming from the SPS via two transfer lines. Protection systems are an essential part of the layout. Three insertions are used for elements related to machine protection: one insertion for the beam dumping system and two insertions for the collimation systems. Around the circumference more than 3600 beam loss monitors are installed. If one monitor detects beam losses above a predefined threshold, a signal is transmitted via an interlock system to the extraction kicker magnets in the beam dumping system and the beams are extracted towards the beam dump blocks.

A typical operational cycle is shown in Fig. 3. Beams are injected in batches with up to 288 bunches from the SPS to LHC at an energy of 450 GeV. It takes some 10 min to fill the two beams. Then acceleration starts and the energy ramp takes about 20 min. At top energy, the beams are brought into collisions and collide for many hours for the data taking of the physics experiments (from a few hours to some tens of hours). At the end of the fill and in case of a failure the beams are extracted towards the beam dump blocks.

Three phases of beam operation related to machine protection are defined: injection, operation with stored beams and the extraction of the beams.



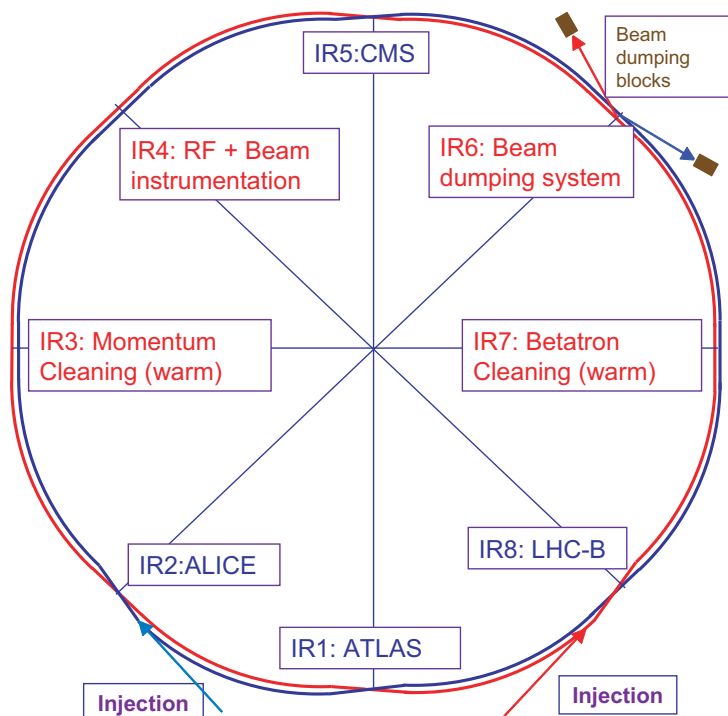


Fig. 2: Layout of LHC. Injection of the beams is via two 3 km long transfer lines from the SPS

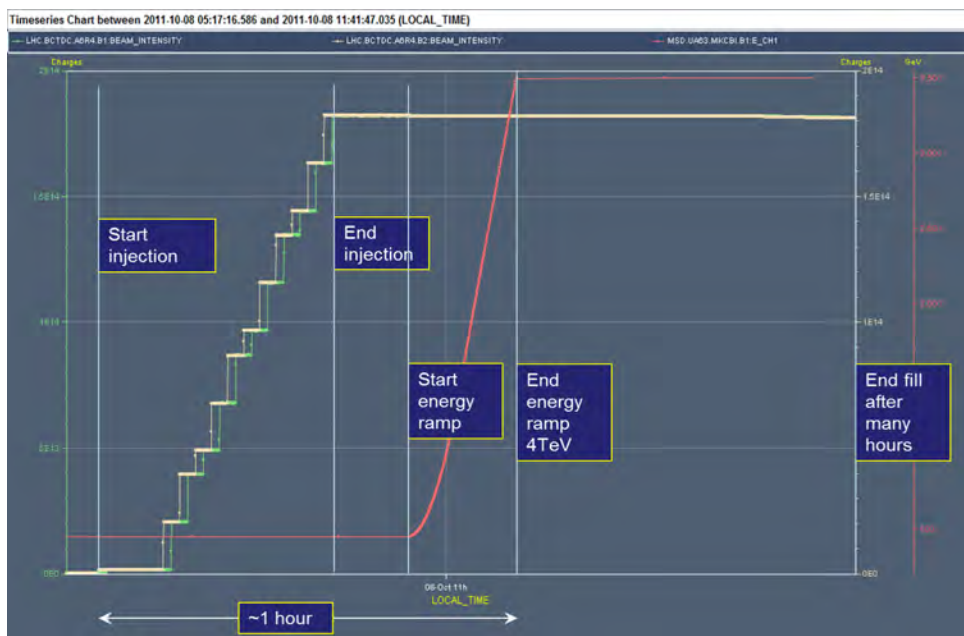


Fig. 3: Operational cycle of LHC, with injection at an energy of 450 GeV, energy ramp to 4 TeV (in 2012) and colliding beams.

### 3 Failures with an impact on the beam

#### 3.1 Classification of failures

In the first step different types of failures that can cause beam losses are identified and classified:

- hardware failure (trip of a power converter, magnet quench, AC distribution failure, object in vacuum chamber, vacuum leak, RF trip, kicker magnet misfire, etc);
- control failure (wrong data, wrong magnet current function, trigger problem, timing system failure, feedback failure, etc);
- operational failures (chromaticity/tune/orbit wrong values, etc);
- beam instability (due to too high beam current/bunch current/e-clouds, etc).

The most important parameters for a failure are:

- time constant for beam loss after the occurrence of the failure;
- probability of the failure occurring;
- damage potential in case no mitigation is applied.

An accurate understanding of the time constant is required, since this determines the reaction time of the machine protection systems. The risk defined as  $\text{risk} = \text{consequences} \times \text{probability}$  is another important input determining the required reliability for the protection systems. For very high risk the protection systems must be extremely reliable.

#### 3.2 Time constant for failures

The time constant for beam loss after a failure varies from nanoseconds to many seconds.

**Single-passage beam losses** in the accelerator complex have a time constant of a few nanoseconds to some tens of  $\mu\text{s}$ . In a circular accelerator such losses are related to failures of fast kicker magnets for injection and extraction. If other fast kicker magnets are present, for example for diagnostics, failures of such devices must also be considered. For failures of fast kicker magnets it is not possible to extract the beam or to stop the beam at the source, the particles will travel determined by the electromagnetic field along their path.

Single-passage beam losses are also an issue for any accelerator operating with pulsed beam. In between two pulses, equipment parameters can change (e.g. a magnet power supply can trip). During the following beam pulse, the beam would be mis-steered and can cause damage. This is typically the case for failures in a transfer line between accelerators (e.g. from SPS to LHC) or from an accelerator to a target station (target for secondary particle production or beam dump block). This is also an issue for linear accelerators operating with pulsed beams.

**Very fast beam losses** with a time constant in the order of 1 ms, e.g. multiturn beam losses in circular accelerators. Such losses can appear due to a large number of possible failures, mostly in the magnet powering system, with a typical time constant of about 1 ms to many seconds.

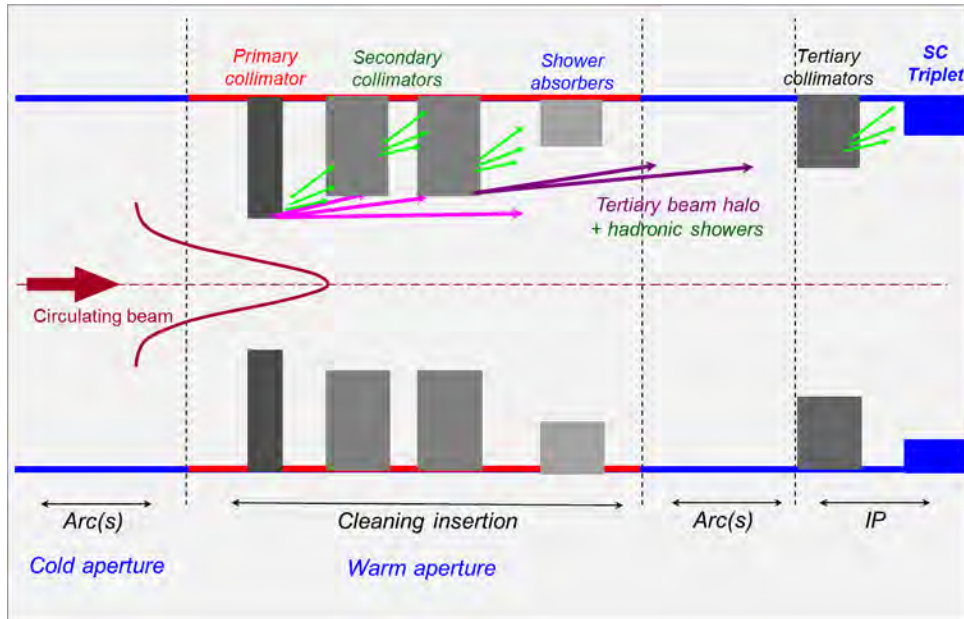
**Fast beam losses** with a time constant of 10 ms to seconds, due to many different effects. Beam instabilities in LHC are in general in this time range.

**Slow beam losses** take many seconds, e.g. due to non-optimized parameters, but also due to a failure.

Details for beam losses in circular and linear accelerators are presented in [2–4].

### 4 Particle distribution and aperture

The evaluation of the consequences of a failure requires the analysis of the trajectories of the particles due to the failure, in particular the location where they will touch the aperture.



**Fig. 4:** Illustration of the layout of the beam cleaning system in LHC (see [5])

In LHC, collimators are installed in two long straight cleaning insertions; they are always positioned to limit the aperture [5]. Primary collimators are set to a position closest to the beams. After a failure, the emittance and therefore the beam size might increase, the closed orbit might change or both happen at the same time. If the amplitude of the betatron oscillation of a particle increases, the particle will first hit a collimator and not the superconducting magnets or other parts of the accelerator. In general, particles are expected to hit a primary collimator. Scattered particles and showers from the collision of the protons with material are absorbed in secondary or tertiary collimators (see Fig. 4). A typical position of the primary collimator with respect to the beam centre is in the order of  $6\sigma$ , with  $\sigma$  defined as the root mean square beam size. Details of the collimation system are presented in [5].

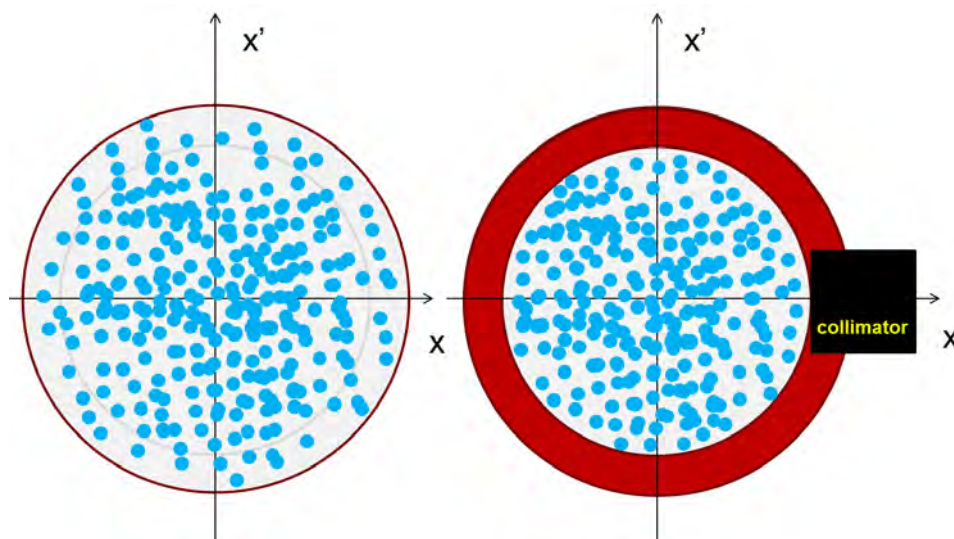
If a collimator is moved closer to the beam, all particles with amplitudes above the value defined by the collimator jaw position will be scraped away (Fig. 5). When the entire beam starts to move, e.g. in case of a magnet trip, the same happens and all particles with amplitudes larger than the position of the collimator will hit the collimator within a few turns.

A Gaussian particle distribution is assumed together with a collimator at a position corresponding to  $4\sigma$ . In case of a failure and a fast displacement of the beam by, say,  $1.7\sigma$ , all particles above an amplitude of  $2.3\sigma$  would hit the collimator jaw (see Fig. 5). If the energy stored in the beam corresponds to, say, 500 MJ, the energy loss would correspond to 35 MJ and the collimator would explode. For a collimator at  $5\sigma$  and the same fast displacement the energy loss is 2.2 MJ and for a collimator at  $6\sigma$  the energy loss is less than 0.1 MJ. The energy loss as a function of collimator setting in case of such failure is shown in Fig. 7.

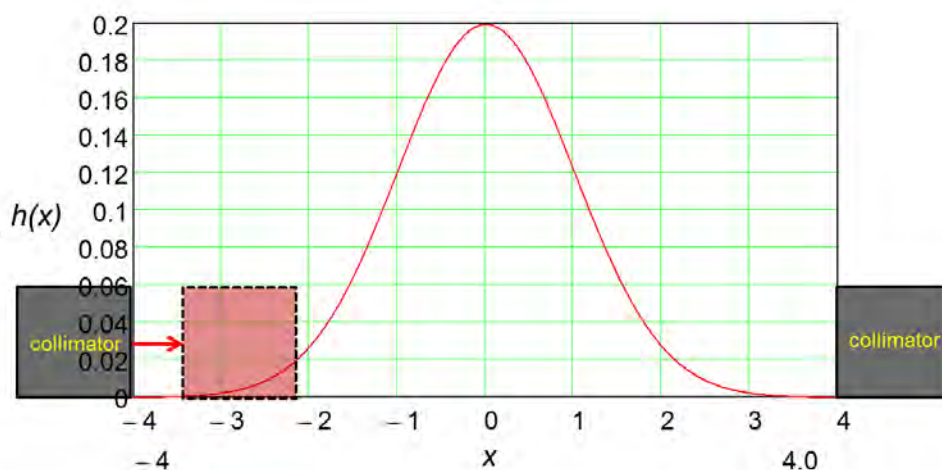
## 5 Mechanisms for creating beam losses: overview

**Magnetic fields:** In a circular accelerator magnetic fields are the dominant elements that determine the particle trajectories, either normal conducting or superconducting. Dipoles magnets provide the deflection of the particles, quadrupoles are installed for focusing, sextupoles for compensating the chromaticity and higher order multipole magnets for various reasons.

**Electric fields:** All circular accelerators use RF cavities with longitudinal electrical fields for acceleration and bunching of the beams. In some accelerators electrical fields are applied for transverse deflection.



**Fig. 5:** Illustration of the phase-space reduction by a collimator. On the left, the particle distribution in phase space is shown. On the right, a collimator is moved into the beam and all particles in the red circle are scraped.



**Fig. 6:** Gaussian beam distribution with a collimator position at  $4\sigma$ , moving by  $1.7\sigma$  to  $2.3\sigma$ . At  $4\sigma$ , about 99% of the particles survive; at  $2.3\sigma$ , this value is about 93%.

Examples are electrostatic separators to separate beams that have the same charge, e.g. at LEP and the SPS proton–antiproton collider. Transverse feedback systems are also frequently using devices that produce an electrical field. A new method for tilting the beam to deflect particles in a bunch as a function of the particle position within the bunch is using RF cavities (so-called crab cavities), e.g. for the B-factory at KEK and proposed for HL-LHC [6].

**Beam instabilities:** The bunch charge and the related electromagnetic fields can lead to beam instabilities. There are different types of instabilities acting on the particles within the bunch and on following bunches that can cause beam losses. At LHC the typical time constant for the growth of beam instabilities is in the order of several tens of ms to many seconds.

**Obstructions in the beam pipe:** There can be a piece of matter inside the beam pipe or the gas pressure can be much higher than the nominal value. The residual gas pressure in the vacuum pipe is in the order of  $10^{-6}$  to  $10^{-12}$  mbar. For beams that are circulating in the accelerator for many hours, the pressure

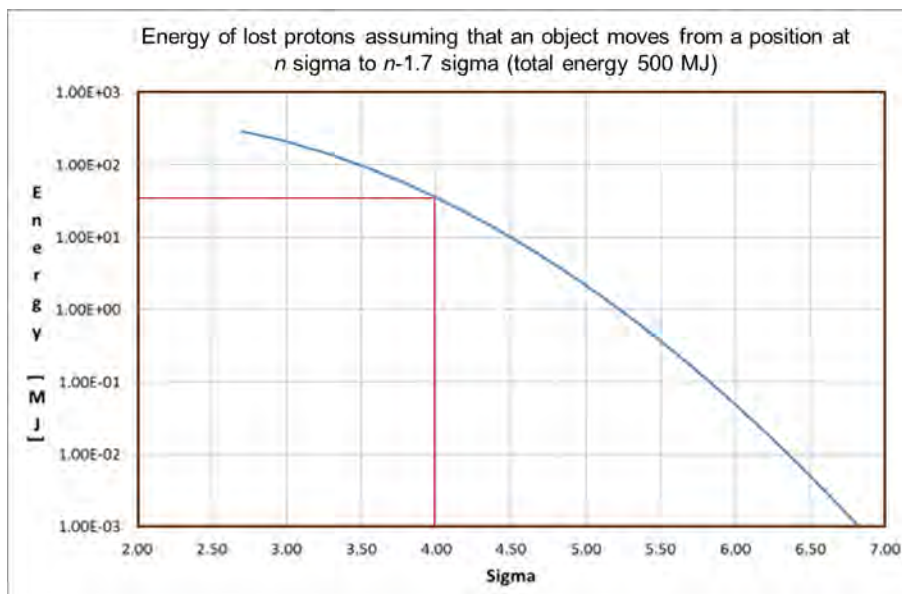


Fig. 7: Energy loss assuming that an object cuts into the beam tail by  $1.7 \sigma$

should be less than, say,  $10^{-8}$  mbar. In case of a vacuum leak or other effects (such as electron clouds) increasing the pressure, particles would collide with the gas preventing efficient operation. Beam losses risk quenching superconducting magnets and activating accelerator equipment. In most accelerators equipment is installed that can move into the beam pipe, such as vacuum valves and screens for the observation of the beam profile. If such elements are accidentally moved into a high-intensity beam, the beam will damage the equipment and hadron showers from the interaction of the beam with the obstructing material can cause further collateral damage.

## 6 Magnetic and electrical fields

When a dipole magnetic field in the accelerator slowly changes and deviates from the nominal field, the closed orbit changes. A change is considered to be slow when the orbit changes by  $1 \sigma$  in several turns. Fast-changing dipole fields (kicker magnets) introduce betatron oscillations. Quadrupole fields change the betatron tune and the optics and therefore the beam size around the accelerator. This might drive the beam onto resonances and cause beam instabilities. Sextupole fields change the chromaticity and might also drive the beam onto resonances and cause instabilities.

For LHC, a Ph.D. study [7] showed that a failure of normal-conducting dipole magnets close to the experiments in IR1 and IR5 has the fastest impact on the beam (excluding kicker magnets). The failure cases were defined and the trajectories of the particles after the failure were calculated. This allowed the calculation of the time between the onset of the failure and the particles touching the aperture.

### 6.1 Very fast beam losses: failures in normal-conducting magnet circuits

We assume that the magnetic field is proportional to the magnet current, a slightly pessimistic assumption since eddy current in the vacuum chamber will slow down fast changes of the magnetic field. The relevant magnet parameters are inductance  $L$  and resistance  $R$ ; the power converter parameters are the current  $I(t)$  and voltage  $V(t)$ . Magnets rarely fail; a magnetic field error is in general caused by a magnet current error. There are many failure modes for such error, due to a failure in the electrical supply, the power converter itself, after a quench, water cooling problems, controls or operation.

A power converter is a very complex device, but modelling a power converter failure for the purpose of machine protection considerations can be simplified. We assume nominal operation at constant

current with the nominal voltage  $V_{\text{nom}}$  and the resistance  $R$ . The current is given by Ohm's law:

$$I(t) = V_{\text{nom}}/R. \quad (1)$$

We consider some failure scenarios.

- The power supply trips and voltage goes to zero (this is the most likely failure, e.g. during a thunderstorm or due to other reasons).
- The control system requests the power supply to provide maximum voltage, possibly with opposite polarity  $V_{\text{fail}}$ . The effect on the current as a function of time can be larger than for a trip of the power converter.

With  $\tau = L/R$ , the current after the failure is given by

$$I(t) = I_{\text{nom}} \cdot \left( e^{-t/\tau} + \frac{V_{\text{fail}}}{V_{\text{nom}}} \cdot (1 - e^{-t/\tau}) \right). \quad (2)$$

A magnet with the field  $B$  and a length  $L$  deflects a particle with the energy  $E$  by an angle

$$\alpha = \frac{B \cdot L}{E} \cdot c \cdot e_0. \quad (3)$$

The change of the closed orbit as a function of the deflection angle  $\alpha$  in the horizontal plane is given by

$$x = \frac{\sqrt{\beta_1 \cdot \beta_2}}{2 \cdot \sin(\pi \cdot Q)} \cdot \alpha, \quad (4)$$

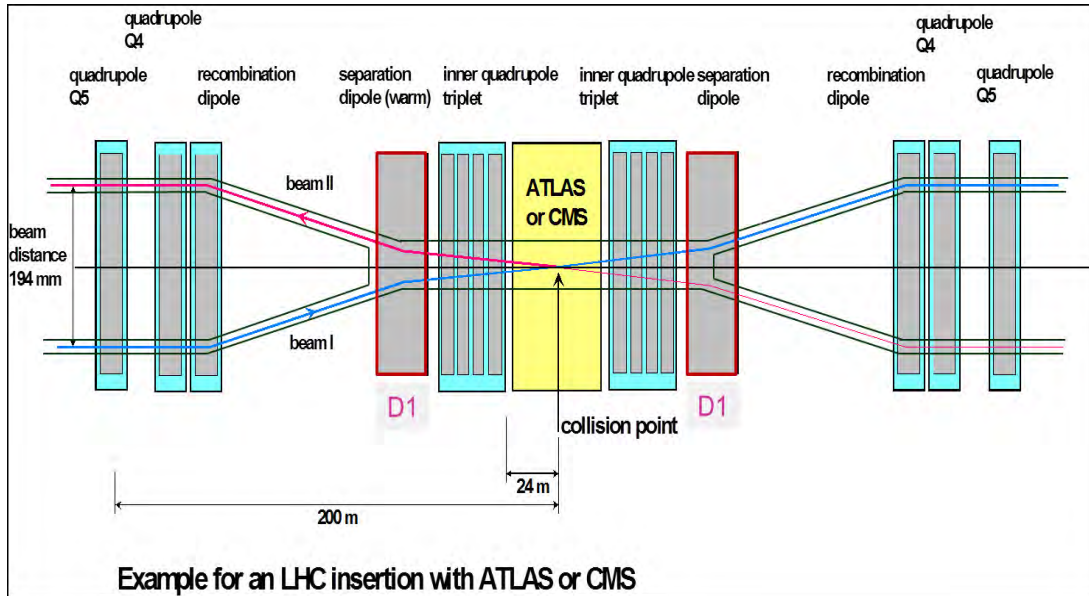
with  $\beta_1$  and  $\beta_2$  the values of the betatron function at the location of the magnet and the location of the observation point, respectively, and  $Q$  the betatron tune.

Resistive magnets have a high resistance and a low inductance compared to superconducting magnets and their current decay can be fast. If the magnet is installed at a position with high  $\beta$  function, the orbit change due to the failure is fast. Such magnets are installed in two of the LHC insertions to separate the beams (so-called D1 magnets, see Fig. 8). The LHC beams are brought together to collide in a common region. Over 260 m the beams circulate in one vacuum chamber with parasitic encounters (when the spacing between bunches is small enough). The D1 magnets separate the two beams.

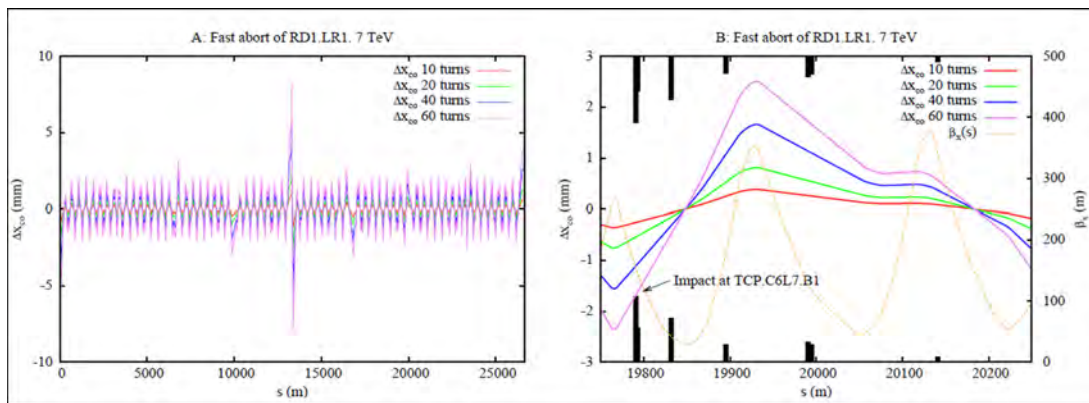
The inductance for 12 magnets powered in series is 1.7 H and the resistance 0.78  $\Omega$ . The nominal field at 7 TeV is 1.38 T; this yields a deflection angle of  $\alpha = 2.41$  mrad. With a time constant in case of a power converter trip of  $\tau = 2.53$  s, the magnetic field error after 10 turns (this corresponds to 89  $\mu\text{s}$ ) is  $\Delta B/B = 3.52 \times 10^{-4}$ . The deflection angle is  $\alpha = 0.848$   $\mu\text{rad}$ . The  $\beta$  function at the D1 magnets is 4000 m. The change of the beam position at a location with a  $\beta$  function of 100 m is 0.32 mm after a period of 10 turns. This change of the orbit corresponds to a change of 1.4  $\sigma$ , when assuming nominal emittance.

This type of failure as well as many other powering failures was simulated with the beam optics and particle tracking program MADX and the result is shown in Fig. 9.

Apart from failures of the extraction and injection kickers, the impact of a failure of the D1 magnet was identified as the fastest mechanism for creating beam losses. The orbit starts to move rapidly by 1  $\sigma$  in about 0.7 ms. In 10 ms the beam would move by 14  $\sigma$ , already outside of the aperture defined by the collimators. If such failure happens, the beam has to be extracted in a very short time. The probability for such failure during the lifetime of LHC is high. The consequences without protection would also be catastrophic, destroying the entire collimation system and possibly causing further damage. The protection needs therefore to be very fast and reliable. Basic parameters of the LHC MPS were designed



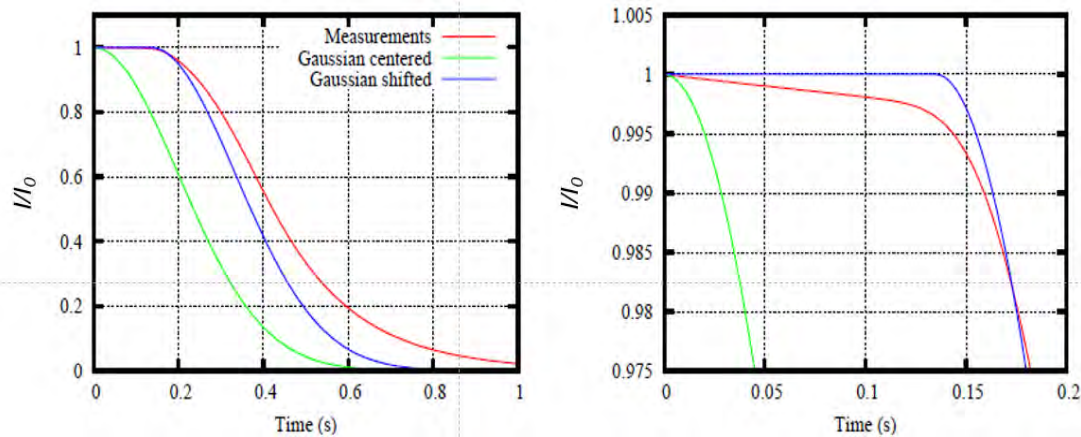
**Fig. 8:** Illustration of a LHC insertion with the normal-conducting D1 magnets to separate the beam left and right from the collision points.



**Fig. 9:** Change of closed orbit after a trip of the power converter for the D1 magnets. The left-hand graph shows the closed orbit around LHC and the right-hand graph the orbit in the collimation section [7].

to cope with this type of failure, in particular the reaction time of the MPS. The failed beam must be detected and the beam must be extracted in less than 1 ms. The detection of the failure is performed by several different systems (diverse redundancy).

- Detection of the failure of a wrong magnet current is challenging, since a fast detection of a current change on the level of  $10^{-4}$  is required. This is done with specifically designed electronics (FMCM = fast magnet current monitor) [8].
- Beam loss monitors detect losses when the beam touches the aperture and particles create a hadron shower (e.g. close to collimator jaws, but also other components) [9].
- In the future a fast beam current monitor will be used to measure the circulating current. In case of fast beam current changes exceeding a predefined threshold the beam will be extracted.



**Fig. 10:** Current decay for a quench in one of the LHC superconducting dipole magnets. The measured current shows a small linear drop until the entire magnet is quenched and then follows a Gaussian decay. The characteristics of the linear drop depend on the evolution of the quench in each particular case and on the reaction time of the magnet protection system. In this particular case, the analytical approach with a time constant of 200 ms yields a decay that is slightly faster than the measured one [7].

## 6.2 Fast beam losses: failures in superconducting magnet circuits

The parameters for superconducting magnets are very different from those of normal-conducting magnets. The inductance is high and the resistance low (determined only by the resistance of the normal-conducting cables between power converter and cryostat). In case of powering failure, the decay of the magnet field takes a long time (up to many hours). In case of a quench, the current decay depends on the mechanism that causes the quench. In general, the magnet starts to quench at a specific location and the quench spreads out. The resistance increases with time. Superconducting magnets require a magnet protection system that detects any quench, switches off the power converter and activates quench heaters. This signal is also provided to the interlock system that requests a beam dump.

There is no analytical equation for the current after a quench as a function of time. In Fig. 10, measurements from LHC magnets are shown. The current decay can be approximated by a Gaussian. In general, a quench is much less critical than the trip of the normal-conducting D1 magnet.

## 6.3 Failures of the transverse damper

The transverse damper is used to damp injection oscillations and instabilities. It also has several other applications: cleaning of the particle free abort and injection gaps, blowing up the beam for loss maps and aperture studies and generating beam losses for quench tests. In the future it will be used as a diagnostic tool to record bunch by bunch oscillations and for betatron tune measurement.

A worst-case calculation of the oscillation amplitude that the damper can produce follows. The damper creates an electrical field for deflecting the particles and can deflect the beam with an angle of  $\alpha = 2 \mu\text{rad}$  per passage at 450 GeV. In the worst case, the deflection of the transverse damper can add up coherently.

We assume that the betatron functions at the damper and at an observation point are  $\beta = 100$  m. The particles are deflected by a single kick to an amplitude of  $x_{d450} = \alpha \cdot \beta$  that corresponds to 0.2 mm, and at 7 TeV to 0.013 mm.

We assume a normalized emittance of  $\epsilon_n = 3.75 \times 10^{-6}$  m. The beam size at a location with a  $\beta$  function of 100 m yields  $\sigma = 0.24$  mm. In the worst case of coherent oscillation, an amplitude of  $1 \sigma$  is reached after 18 turns, slightly less critical than a failure of the D1 magnet.



## 7 Failures of fast kicker magnets

Fast kicker magnets are required for injection of beam into LHC and for extraction towards the beam dump block. Failures of kicker magnets cannot be mitigated by active protection systems. If the beam is mis-steered due to a kicker magnet failure, the only option is to install beam absorbers for capturing the beam. There are several failure modes for kicker magnets.

- The deflecting angle is wrong.
- The time when the kicker fires is wrong (too early, too late, too short, too long).
- The injection kicker deflects the beam when it is not intended to operate, e.g. when LHC is operating above injection energy.

Some failure modes can be avoided by interlocks and operational procedures. As an example, the injection kicker should never deflect the beam when the accelerator is not at injection energy; therefore, after starting the ramp the injection kicker is switched off. There are some failure modes that cannot be avoided, such as a flash-over of a kicker; such failures need to be mitigated in order not to damage equipment.

At injection, a batch of 288 bunches is injected always with the same energy of 450 GeV [2]. The injection elements and kicker and septum magnets must always have the same strength. The kicker magnets have a very short pulse, deflecting by a small angle. The septum magnet is a DC magnet that can be slowly pulsed, with no magnetic field acting on the circulating beam. The energy stored in a batch with 288 bunches injected into LHC is about 2 MJ; in case of accidental release the beam would seriously damage LHC equipment. Injection happens very frequently, in order to fill each of the two beams with up to 2808 bunches. An example for the elements of injection protection is shown in Fig. 11. Two failure cases are considered.

- The beam is transferred via the transfer line, but the injection kicker for deflecting the beam onto the closed orbit fails. The beam travels further and would damage equipment. An absorber is installed at a location with a betatron phase advance of 90 degrees to absorb the energy.
- The timing of the kicker pulse is wrong and the circulating beam is deflected. A second absorber is installed to absorb the energy of bunches that are deflected.

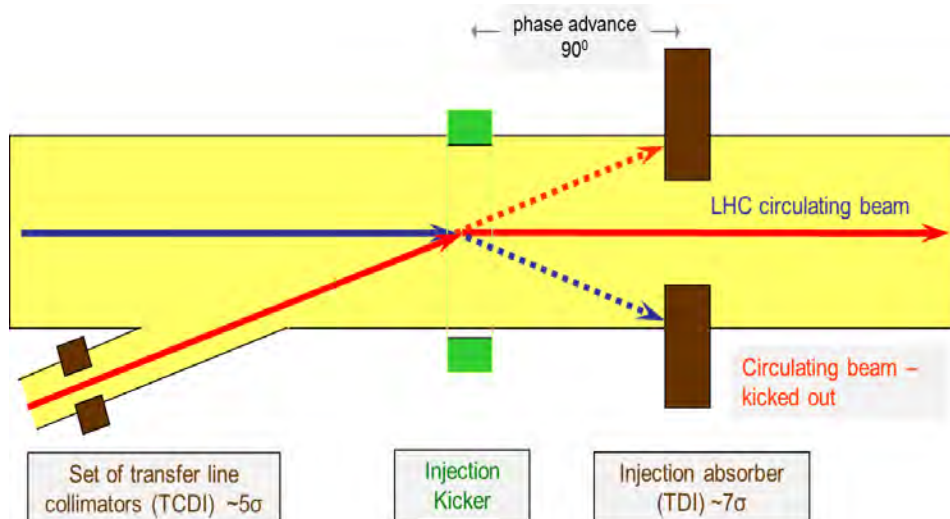
One of the most critical components of the machine protection system is the beam dumping system (see Fig. 12). One set of kicker magnets deflects the beam, septum magnets increase the deflection angle and a second set of kicker magnets dilutes the beam that travels through a 800 m long transfer line to the beam dump block.

Extraction can happen at any energy from 450 GeV to 7 TeV. The energy stored in the beam is up to 364 MJ. The strength of the kicker and septum field depends on the energy; the deflection angle must remain constant. The current of the main dipole magnets in four out of eight sectors of the LHC is measured and used to ensure the correct tracking of the kicker and septum magnet strengths. For a correct extraction, the kicker must fire synchronised with a particle free gap (during the kicker rise-time there are no particles), in order to deflect all bunches with the nominal angle. The duration of the kicker pulse must also be correct.

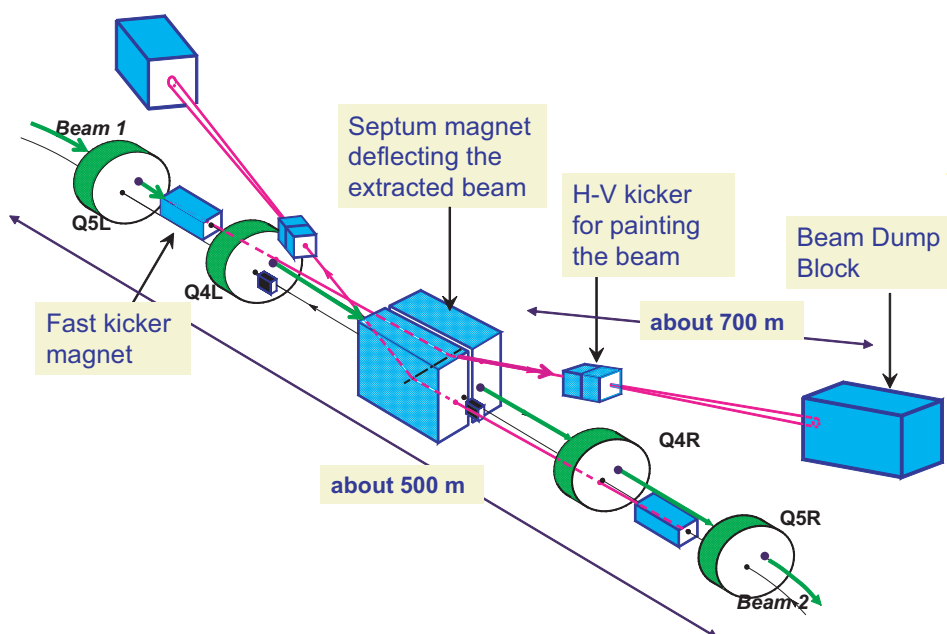
## 8 Failure modes for magnets to mis-steer the beam

There are many failure modes that can result in a wrong magnetic field:

- failure of a power converter, water cooling or quench of a magnet;
- wrong command entered by an operator (e.g. request for angle change of 0.01 mrad instead of 0.001 mrad);



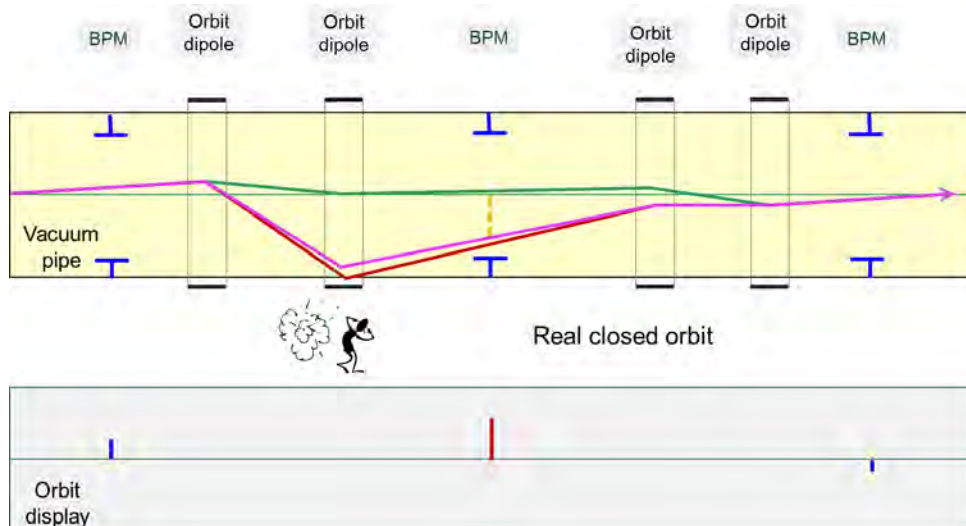
**Fig. 11:** Illustration of failures at injection and related protection absorbers. One failure is the injection kicker not deflecting the beam that would travel straight into an absorber protecting the aperture. Another failure is the kicker deflecting the circulating beam that would travel to an absorber on the opposite side.



**Fig. 12:** Layout of the beam dumping system

- timing event to start current ramp does not arrive or arrives at the wrong time;
- control system failure (data to power converter not sent or sent incorrectly);
- wrong conversion factor (e.g. from angle to power converter current);
- feedback system failure;
- failure of beam instrumentation, e.g. beam position monitor (BPM).

One failure that is not obvious but must be considered is due to a wrong functioning of a beam position monitor. We assume that one of the beam position monitors is faulty and always provides the



**Fig. 13:** The lower graph shows the reading of the orbit display, with one BPM always indicating the same wrong value. Initially, the orbit is correct. Then the BPM becomes faulty and the feedback system moves the beam by applying an orbit bump to correct for the offset until the beam touches the aperture.

same wrong value, e.g. 1 mm, independent of the beam position (see Fig. 13). In case an orbit feedback is operating, the feedback will try to correct the orbit at the position of the monitor by applying a closed orbit bump. The orbit deviation at this position will increase until the beam touches the vacuum chamber. Closed orbit bumps reduce the aperture and cannot be detected if a beam position monitor is faulty. An option to mitigate this fault is a software interlock reading the strengths of orbit corrector magnets.

Figure 14 shows that this really happens. The two graphs show orbit and strengths of corrector magnets at LHC for the vertical plane within an interval of 7 s; the data was taken in March 2011. In the first graph one BPM shows a large value and the feedback system started to correct the orbit with bumps. The bump was building up and finally the beam was lost.

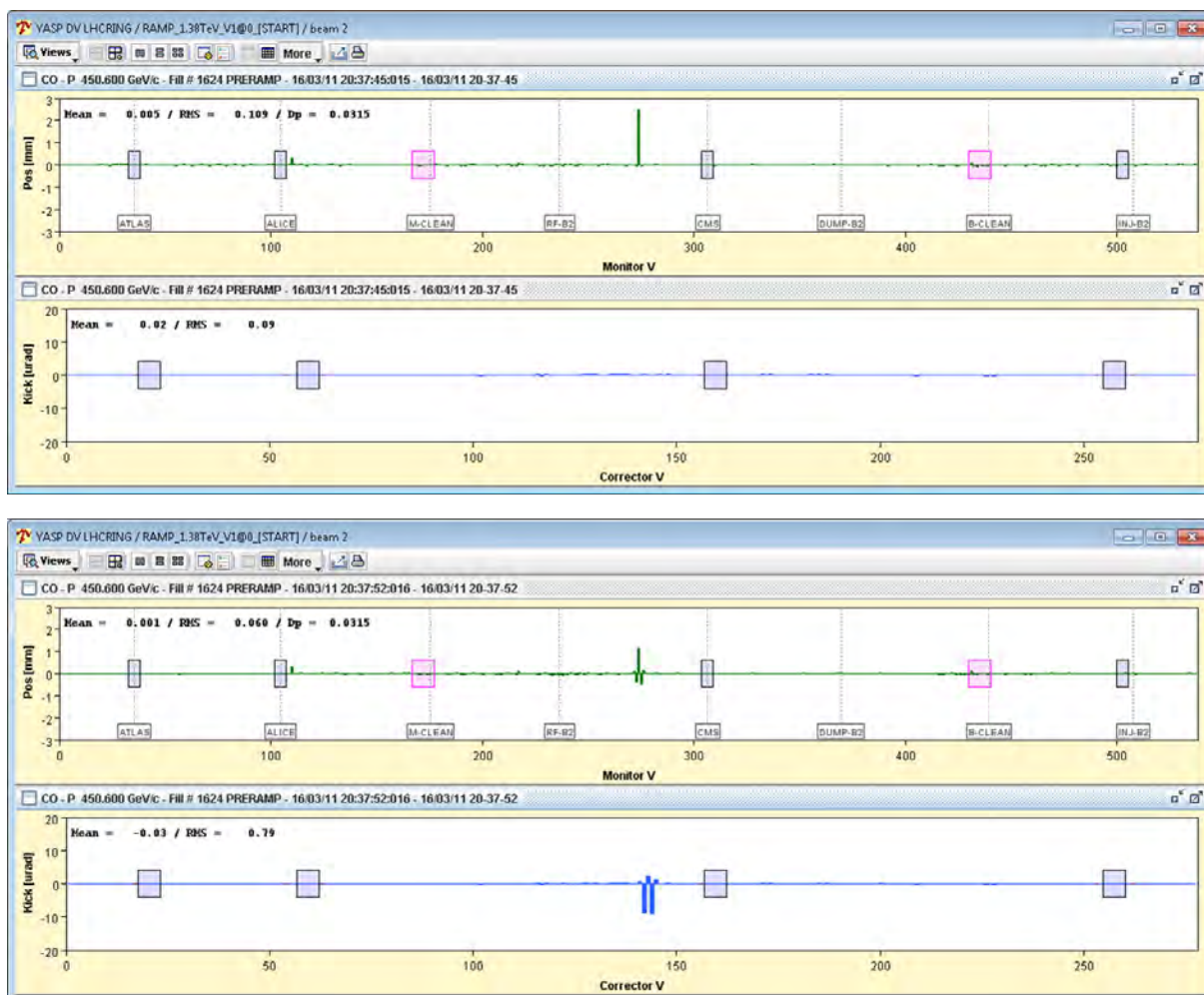
### 9 Objects that can block the beam passage

Around an accelerator there are many objects that can block the beam passage. Some equipment is designed to move into the beam pipe (movable devices), other equipment should never be in the path of the beam, but this might accidentally happen:

- vacuum valves, possibly a large number to isolate different accelerator sections;
- collimators and beam absorbers;
- beam instrumentation: screens for observation of the beam profile, mirrors to observe synchrotron light, wire scanners to measure the beam profile;
- experiments, e.g. so-called Roman pots, particle detectors to observe small angle scattered particles from the collision points that can move close to the beam.

Elements that should never be in the beam pipe:

- RF fingers for ensuring a continuous path for the image current across bellows. Such fingers can bend into the vacuum chamber;
- other material, e.g. left-over pieces in the beam pipe from activities that require opening the beam vacuum;



**Fig. 14:** Failure of a BPM at LHC. The upper graphs show the orbit before and after the failure; the lower graph shows the strengths of corrector magnets, with the bump building up. The data was taken with an interval of 7 s.

- elements that are getting into the pipe due to a failure (e.g. during cool down in a superconducting accelerator or during operation);
- gas above nominal pressure.

## 9.1 Wire scanners

Wire scanners are made out of a thin carbon or tungsten wire that moves through the beam. The particle shower or the secondary emission is recorded to measure the beam profile. With too high energy density of the beam the wire risks melting; protection of the instrument is required. Superconducting magnets in the vicinity risk quenching, leading to downtime of the accelerator. The energy density depends on the beam intensity and on the beam size. For hadron beams the size decreases with energy. Consequences of such failure are minor (instrument not available and risk of quenching a magnet). The probability that an operator sends the wire through a high-intensity beam is high. Since the risk is relatively small, a protection system with a low protection level is acceptable to prevent the wire passing through a high-intensity beam.

## 9.2 Collimators and beam absorbers

Collimators and beam absorbers are essential elements for machine protection (see [5]). Collimators protect, but cannot absorb a wrongly deflected high-intensity beam and therefore also need protection, in particular at injection and extraction. They should be at the correct position with respect to the beam. Usually the closest collimator jaws are at a position of about  $6\sigma$  from the beam centre. At LHC, the position of the collimators depends on the energy, since the beam size decreased during the energy ramp. Most collimators are therefore moved closer to the beam during the energy ramp.

A complex interlock system ensures that the collimators are correctly positioned. A timing event is used to start the collimator movement when the energy ramp starts. The position is verified by an independent method comparing the actual position of each collimator with the required position stored in a separate table for each energy. This can also be done when the optics changes and collimator position needs to be adjusted. Roman pots must always be outside the aperture defined by the collimator.

## 10 Consequences of beam loss

### 10.1 Beam losses and magnet quenches

Superconducting magnets produce high field with high current density in the superconducting cables. The superconducting state of a magnet occurs only in a limited area of temperature, magnetic field and transport current density, depending on the superconductor (for LHC, NbTi conductors are used). Lowering the temperature enables better usage of the superconductor by broadening its working range. The operating parameters of LHC superconductors (NbTi) are shown in Fig. 15. Most of the LHC magnets operate at a temperature of 1.9 K. When operating at low energy and therefore at low magnetic field, the temperature margin is relatively large and beam losses of some 100 J can be tolerated without quenching. For high energy, the margin is small and very small beam losses are sufficient to quench a magnet (see Fig. 16).

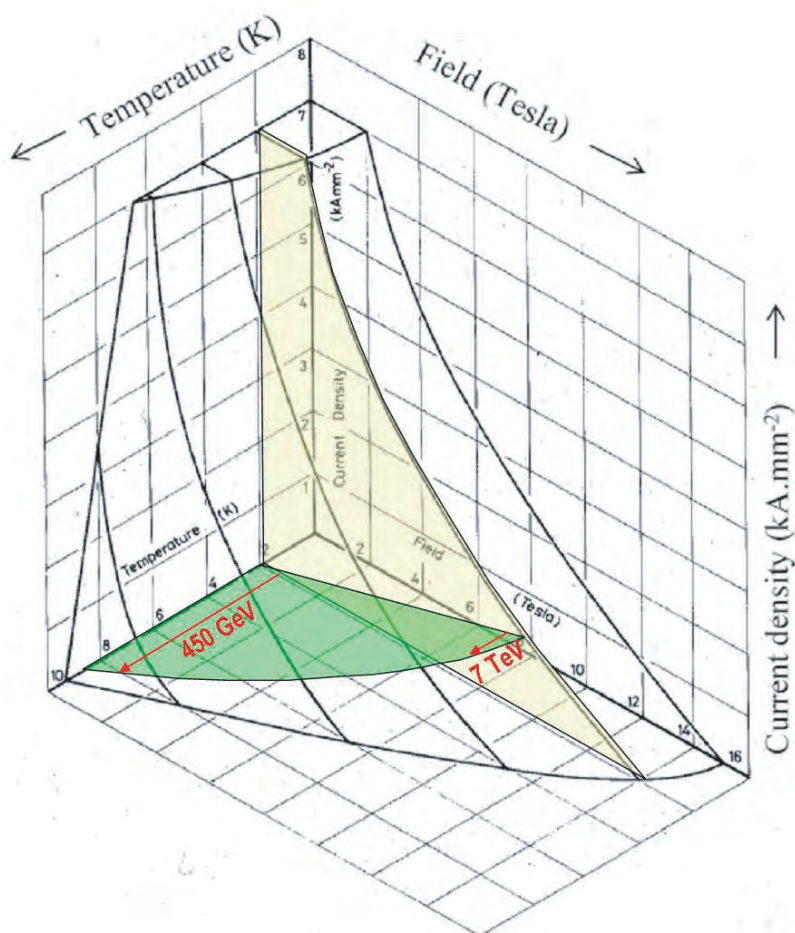
### 10.2 Wrong deflection during extraction

Several failures can lead to the beam deflected with non-nominal angle, e.g. if only one (out of 15) kicker magnets fires, or LHC operates at 7 TeV and the kicker magnets extract the beams with an angle corresponding to 450 GeV. Studies were performed to estimate the consequences of the full 7 TeV beam deflected into equipment such as a graphite collimator or a magnet [10]. A first indication is obtained by calculating the energy deposition in material with codes such as FLUKA [11]. The result is shown for a copper target in Fig. 17. The energy density is far above the melting and vaporization points of the material.

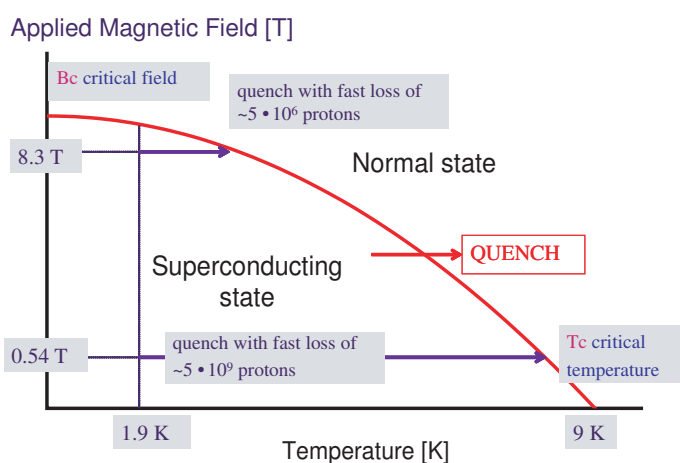
The correct calculation needs to take the time structure of the beam into account. So-called hydrodynamic tunnelling of beam through the target becomes important after the impact of some 10 bunches. The first bunches arrive, deposit their energy and lead to a reduction of the target material density. Bunches arriving later travel further into the target since the material density is reduced. This effect has been already predicted for SSC [12]. The calculation of hydrodynamic tunnelling is complex and performed in several steps. Typical parameters for the simulation are: 2808 bunches with  $1.1 \times 10^{11}$  protons,  $\sigma = 0.5$  mm and 25 ns bunch distance. The calculations predict a tunnelling depth of about 30 m for these parameters. Recently an experiment was performed to validate the simulation method [13].

### 10.3 Damage levels for particle beams

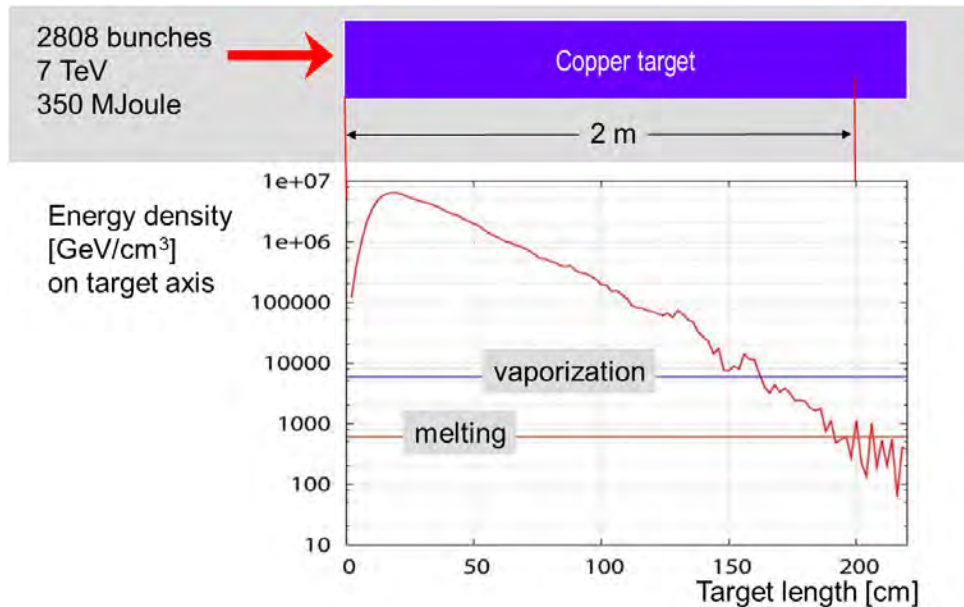
Commissioning of LHC starts with a low-intensity beam, in order to avoid any risk of damaging equipment. An obvious question is what beam parameters can lead to damage and what beam parameters are still safe. Relevant beam parameters are the particle type, the energy of the particles and the beam size. The answer to this question is far from being obvious. It depends not only on beam parameters, but also



**Fig. 15:** Operational margin for NbTi magnets with magnetic field, current density and temperature. To operate with a magnetic field corresponding to an energy of 7 TeV, the operating temperature must be below 2 K.



**Fig. 16:** Margin of a LHC superconducting magnet with respect to beam losses at injection and top energy. The absolute value of the energy deposition depends on the impact parameters; the numbers give the orders of magnitude.



**Fig. 17:** Energy density in a copper cylinder when a full 7 TeV LHC beam is deflected into a copper target

on the time distribution of the losses and the equipment that is exposed. In case of beam loss for more than a few ms cooling of the equipment has to be considered. Sensitive equipment such as particle detectors can be damaged by beam losses in the order of 1 J. Massive damage of equipment is not expected below some megajoules.

In the following, we discuss the criticality for high-energy proton beams and a beam size in the order of 1 mm. Figure 18 gives an idea of the current knowledge about damage levels. The data in the table is derived from past experience, either from accidents or damage experiments with particle beams. The data should be taken with care; much more research is required to establish reliable data for different beam and equipment parameters:

- TT40 experiment [2];
- SPS damage of UA2 [14];
- collimator damage experiments [5];
- quench tests at LHC [15];
- experience from SNS [4];
- TEVATRON accident [16].

When designing protection systems for a specific hazard, a conservative approach is required. However, a too pessimistic approach should be avoided. An accelerator should not be overprotected; this could lead to unnecessary investment and downtime during operation.

## 11 Machine protection and interlocks at LHC

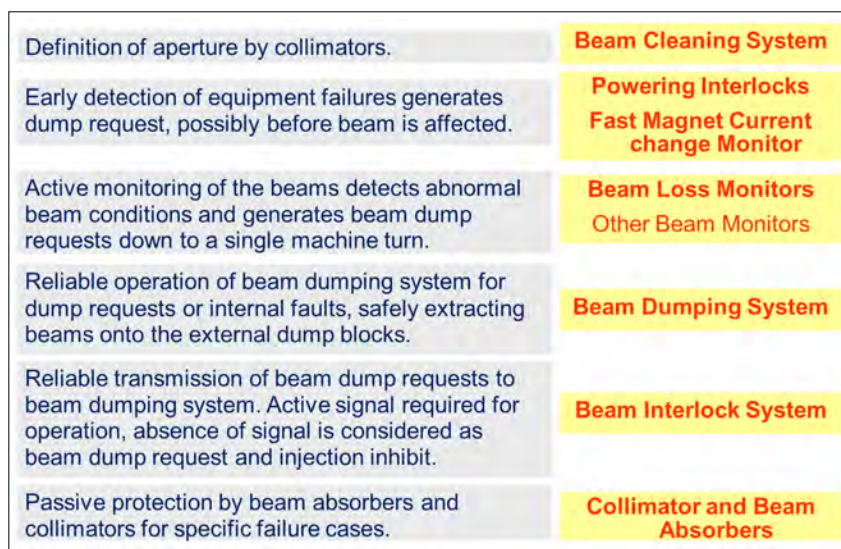
### 11.1 Strategy for machine protection

In this section we discuss the strategy adopted for machine protection of LHC and the related systems, illustrated in Fig. 19, together with the main subsystems for protection.

- Definition of aperture by collimators.
- Detect failures at hardware level and stop beam operation for critical failures.

| Fast beam impact, energy [J]<br>- beam size order of 1 mm -<br>high energy protons | Superconducting magnets           | Vacuum chamber and normal conducting elements   | Collimators (Graphite) | Collimators (Metal - W) | Superconducting cavities            |
|--|-----------------------------------|---|------------------------|-------------------------|-------------------------------------|
| 1-10   |                                   |   |                        |                         |                                     |
| 10-100   | Quench of a magnet (LHC at 7 TeV) |   |                        |                         |                                     |
| 100-1000   | Quench magnets                    |   |                        |                         | Effects on superconducting cavities |
| 1000-10000   | Quench magnets                    |   |                        |                         | Damage                              |
| 10000-100000   | Quench magnets                    |   |                        | Onset of damage         | Damage                              |
| 1E5-1E6  | Not known, damage likely          |   |                        | Damage                  | Damage                              |
| 1E6-1E7  | Damage                            | Damage of vacuum chamber with grazing beam incident – short exposure SPS-TT40 accident. | Onset of damage        | Damage                  | Damage                              |
| above 1E7  | Massive damage                    | Massive damage  | Massive damage         | Massive damage          | Massive damage                      |

**Fig. 18:** Indication for the quench and damage levels by high-energy protons. Fast beam impact is assumed, with a beam size of the order of 1 mm.



**Fig. 19:** Protection strategy for LHC

- Detect initial consequences of failures on the beam with beam instrumentation before it is too late.
- Transmit signal from instrumentation via a highly reliable interlock system to the extraction kickers and injection system.
- Stop beam operation by extracting the beams into beam dump block.
- Inhibit injection into LHC and extraction from the SPS.
- Stop beam by beam absorber/collimator for specific failures.

### 11.2 Machine interlocks at LHC

Figure 20 illustrates the interlock systems for LHC. The heart is the beam interlock system that receives beam dump requests from many connected systems. If a beam dump request arrives, a signal is sent to the beam dumping system to request the extraction of the beams. At the same time, a signal is sent to the injection system to block injection into LHC as well as extraction of beam from the SPS. A third signal



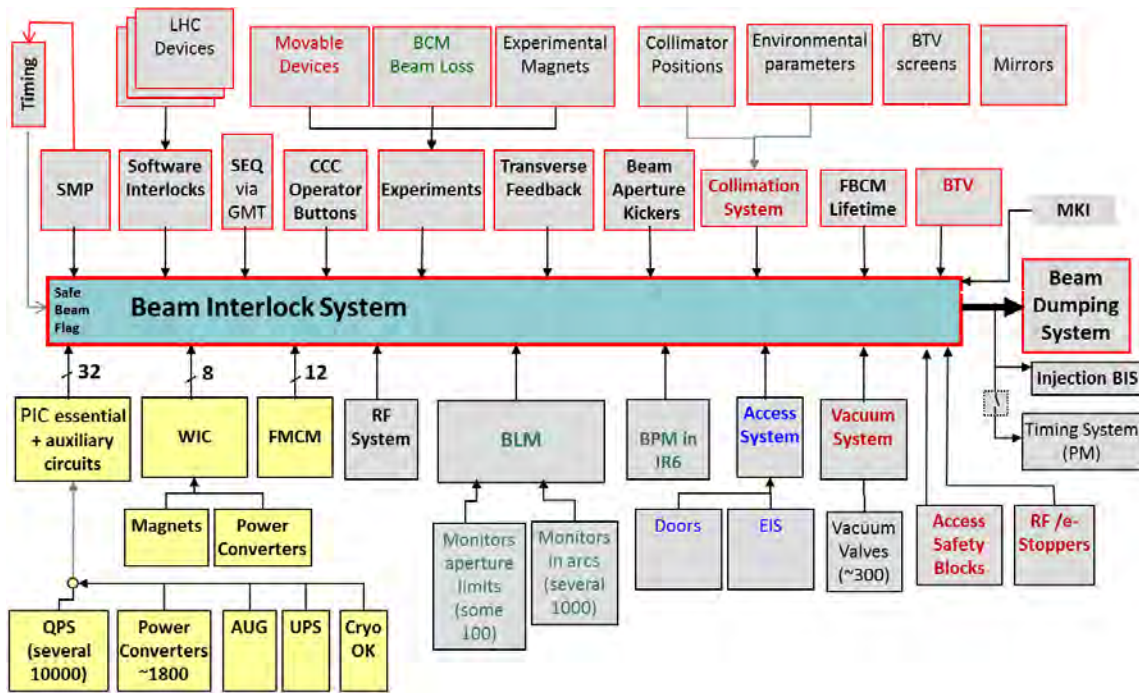


Fig. 20: Beam interlock system at LHC as well as connected systems

is provided for the timing system that sends out a request to many LHC systems for providing data that were recorded before the beam dump, to understand the reasons for the beam dump (typically beam loss, beam position, beam current, magnet currents, etc).

The most complex system of LHC is the superconducting magnets and powering system. The powering interlock system (PIC) ensures communication between systems involved in the powering of the LHC superconducting magnets. This includes power converters, magnet protection system, UPS (uninterruptible power supplies), emergency stop of electrical supplies (AUG) and the cryogenic system. As an example, in case a magnet quench is detected by the quench protection system (QPS), the power converter must stop. In total, there are several tens of thousands of interlock signals. When a failure is detected that risks stopping the powering of magnets, a beam dump request is sent to the beam interlock system. A second system manages interlocks from the normal-conducting magnets and their power supplies (WIC) that ensures protection of normal-conducting magnets in case of overheating.

The machine interlock system is strictly separated from interlocks for personnel safety such as the personnel access system; however, an interlock from the access system is sent to the beam interlock system.

As shown in Fig. 20, many other systems also provide beam dump requests in case of failure: beam loss monitors, other beam monitors, movable devices and LHC experiments.

## 12 Safety and protection integrity levels

As has been discussed in [14], the risk for a specific hazard depends on consequences and probability of an accident:  $\text{risk} = \text{consequences} \times \text{probability}$ .

For the design of protection systems in industry, standards are frequently used. One example of a widely used standard is IEC 61508, an international standard of rules applied in industry, for the Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems (E/E/PE or E/E/PES). For LHC, the standard was not strictly applied, but many ideas and principles were used, such as the safety integrity level (SIL) concept. Since the development of the LHC machine protection

|              |   |
|--------------|---|
| Frequent     | Hazard expected to become active at least once every 100 days   |
| Probable     | Hazard expected to become active at least once between every 100 days to 1000 days                                    |
| Occasional   | Hazard expected to become active at least once between every 1000 days to 10000 days                                  |
| Remote       | Hazard is not expected to become active in 10000 days (unlikely during lifetime of LHC)                               |
| Catastrophic | if the consequence requires more than 200 days of repair or the cost of the repair is greater than 50 MCHF.           |
| Major        | if the consequence requires between 20 to 200 days of repair or the cost of the repair is between 1 MCHF and 50 MCHF. |
| Severe       | if the consequence requires between 2 to 20 days of repair or the cost of the repair is between 100 kCHF and 1 MCHF.  |
| Minor        | if the consequence requires less than 2 day of repair or the cost of the repair is less than 100 kCHF.                |

**Fig. 21:** Definition of frequencies and consequences for hazards, defined for LHC

system did not strictly follow the IEC 61508 norm, the idea of a protection integrity level (PIL) was introduced [17].

The PIL level classifies hazards according to frequency and consequences, as in IEC 61508 for the safety integrity level (see Fig. 22). The design of the protection system for a specific hazard depends on the PIL level. The higher the PIL, the more effort is needed to demonstrate that the protection system is sufficiently robust.

For LHC, a definition for frequency and consequences was adopted as shown in Fig. 21. The definition of the consequences is not a unique table and here the numbers for LHC are shown that can be very different for other projects.

Some examples of hazards that were analysed during the design of the LHC protection systems are given.

- What is the risk when a probe (= low-intensity) beam is injected into the empty LHC ring and lost during the first turn? Low risk since the consequences of such event are minor: PIL 1.
- What is the risk that a power converter fails and the full LHC beam is deflected? High risk since the probability for such an event is high (probable) and without protection the consequences would be major or catastrophic for LHC: PIL 3 or PIL 4.
- What is the risk for a wire scanner accidentally moving through the high-intensity beam? The risk is low since the consequences are minor and the frequency occasional: PIL 1.
- What is the risk that the entire LHC beam will be deflected by a black hole generated in a proton–proton collision in ATLAS? Remote probability, not to be considered.

### 12.1 Design considerations for protection systems

There are a number of principles that should be considered in the design of protection systems, although it might not be possible to follow all these principles in all cases.

| Frequency  | Event        |        |       |              |
|------------|--------------|--------|-------|--------------|
|            | Consequences |        |       |              |
|            | Minor        | Severe | Major | Catastrophic |
| Frequent   | 2            | 3      | 4     | 4            |
| Probable   | 1            | 2      | 3     | 4            |
| Occasional | 1            | 1      | 2     | 3            |
| Remote     | 1            | 1      | 1     | 2            |

**Fig. 22:** Risk matrix defining the PIL level (protection integrity level), derived from the SIL definition

- Fail-safe design: in case of a failure in the protection system, protection functionalities should not be compromised. As an example, if the cable from the interlock system that triggers the extraction kicker of the beam dumping system is disconnected, the kicker must fire and operation must stop.
- Detection of internal faults: the protection system must monitor its internal status. In case of an internal fault, the fault should be reported. If the fault is critical, operation must be stopped.
- Remote testing should be an integral part of the design, for example between two runs. This allows regular verification of the correct status of the system.
- Critical equipment should be redundant (possibly diverse redundancy, with the same or similar functions executed by different systems).
- Critical processes for protection should not rely on complex software running under an operating system and requiring the general computer network.
- It should not be possible to remotely change the most critical parameters. If parameters need to be changed, the changes must be controlled, logged and password protection should ensure that only authorized personnel can perform the change.
- Safety, availability and reliability of the systems should be demonstrated. This is possible by using established methods to analyse critical systems and to predict failure rates.
- Operate the protection systems early on before they become critical, to gain experience and to build up confidence. This could be done before beam operation or during early beam operation when the beam intensity is low.
- It is inevitable to disable interlocks (e.g. during early phase of commissioning and for specific tests). Managing interlocks (e.g. disabling) is common practice and required in the early phase of operation. Keep track and consider masking of interlocks during the system design. Example for the realization at LHC: masking of some interlocks is possible, but only for low-intensity/low-energy beams ('safe beams').
- Avoid (unnecessary) complexity for protection systems, keep it simple.
- Having a vision to the operational phase of the system helps to defined the functionality.
- Test benches for electronic systems should be part of the system development.
- Most failures in electronics systems are due to power supplies, cables, mechanical parts and connectors. Redundancy of these parts will increase the availability.
- Careful testing in conditions similar to real operation is required.
- Reliable protection does not end with the development phase. Documentation for installation, commissioning, maintenance and operation of the MPS is required.
- The accurate execution of each protection function must be explicitly tested during commissioning.
- Requirements need to be established for the test interval of each function.
- All actions of the protection systems (e.g. beam dumps) need to be carefully analysed. This requires the presence of transient recording of all relevant systems.

### 13 Summary

In systems engineering, dependability is a measure of a system's availability, reliability and maintainability. This concept is a new challenge for accelerator laboratories and requires a different approach in engineering, operation and management.

Safety culture is how safety is managed in the laboratory and reflects attitudes, beliefs, perceptions and values that employees share in relation to safety. Safety culture at CERN has been developed over the last, say, 10 years. Even those who were sceptical about the need for a powerful protection system were convinced after the accident during the powering test in 2008 that such safety culture is needed.

At CERN for LHC, the experience with the systems for protecting equipment from beam accidents is excellent; there was no damage and no near miss. However, some non-conformities were detected that demonstrate that it is important to be vigilant.

Availability and safety are a trade-off relationship—when a given safety is met, the goal is to make the system as available as possible for providing beams to experiments. False beam dumps (beam dumps that are not strictly necessary for safe operation) need to be avoided to minimize downtime. The availability at LHC is not yet at a level that is acceptable for future operation and improvements are required.

From the experience with LHC there are a number of lessons to be learned for future accelerators to ensure safe operation with high availability (e.g. for accelerator-driven spallation, where operating with very high availability is the main challenge).

Machine protection is not equal to equipment protection and is not limited to the interlock system:

- it requires a thorough understanding of many different types of failures that could lead to beam loss;
- it requires a comprehensive understanding of all aspects of the accelerator (accelerator physics, operation, equipment, instrumentation, functional safety);
- it touches many aspects of accelerator construction and operation and includes many systems;
- it is becoming increasingly important for future projects, with increased beam power/energy density ( $\text{W}/\text{mm}^2$  or  $\text{J}/\text{mm}^2$ ) and increasingly complex machines.

### Acknowledgements

I wish to thank many colleagues from CERN, ESS and the authors of the listed papers for their help and for providing material for this paper.

### References

- [1] The LHC Study Group, The Large Hadron Collider: conceptual design, CERN/AC/95-05 (LHC) (1995).
- [2] V. Kain, Beam transfer and machine protection, these proceedings.
- [3] V. Kain, Beam losses in circular accelerators, these proceedings.
- [4] M. Plum, Beam dynamics and beam losses – linear machines, these proceedings.
- [5] S. Redaelli, Beam cleaning and collimation systems, these proceedings.
- [6] B. Yee-Rendon, R. Lopez-Fernandez, R. Calaga, R. Tomas, F. Zimmermann and J. Barranco, Fast crab cavity failures in HL-LHC, 5th Int. Particle Accelerator Conf., IPAC 2014, Dresden, Germany, 15–20 June 2014.
- [7] A. Gomez-Alonso, Ph.D. thesis, UPC Barcelona, CERN-THESIS-2009-02, 2009.

- [8] M. Werner, M. Zerlauth, R.R. Schmidt, V. Kain and B. Goddard, A fast magnet current change monitor for machine protection in HERA and the LHC, ICALEPCS 2005, Geneva, Switzerland, 2005.
- [9] B. Dehning, Beam loss monitors at LHC, these proceedings.
- [10] N.A. Tahir, J.B. Sancho, A. Shutov, R. Schmidt and A.R. Piriz, *Phys. Rev. ST Accel. Beams* **15** (2012) 051003. <http://dx.doi.org/10.1103/PhysRevSTAB.15.051003>
- [11] G. Battistone *et al.*, *AIP Conf. Proc.* **896** (2007) 31. <http://dx.doi.org/10.1063/1.2720455>
- [12] D.C. Wilson *et al.*, Hydrodynamic calculations of 20-TeV beam interactions with the SSC beam dump, Proc. PAC 1993, San Sebastian, Spain, 1993. <http://dx.doi.org/10.1109/pac.1993.309562>
- [13] J. Blanco *et al.*, An experiment on hydrodynamic tunneling of the SPS high intensity proton beam at the HIRdmat facility, Proc. HB2012, Beijing, China, 2012.
- [14] R. Schmidt, Introduction to machine protection, these proceedings.
- [15] B. Auchmann *et al.*, Testing beam-induced quench levels of LHC superconducting magnets in run 1, submitted to *Phys. Rev. Spec. Top. Accel. Beams*, June 2015.
- [16] N. Mokhov, Beam material interaction, heating and activation, these proceedings.
- [17] M. Kwiatkowski, Ph.D. thesis, Warsaw University of Technology, Faculty of Electronics and Information Technology, CERN-THESIS-2014-048, 2009.



## Protection of Hardware: Powering Systems (Power Converter, Normal Conducting, and Superconducting Magnets)

*H. Pfeffer, B. Flora, and D. Wolff*

US Particle Accelerator School, Batavia, IL, USA

### Abstract

Along with the protection of magnets and power converters, we have added a section on personnel protection because this is our highest priority in the design and operation of power systems. Thus, our topics are the protection of people, power converters, and magnet loads (protected from the powering equipment), including normal conducting magnets and superconducting magnets.

### Keywords

Magnets; superconducting; protection; quench; rectifier.

## 1 General protection techniques

In the protection of any of these topics, there are a number of general techniques that we use to help us achieve our protection goals. We will list them below and then point out when they are used in a variety of situations.

### 1.1 Redundancy

This means designing two independent paths leading to the desired protective action. For example, the overcurrent protection circuitry of a rectifier power converter might consist of a current transducer monitor circuit that turns off the thyristors in the power converter, as well as a shunt current monitor that opens the power-converter circuit breaker when the maximum current is exceeded.

A properly designed redundant protection system should have no elements that are common to both protective paths, since a failure of the common element might override both sources of protection.

### 1.2 Fail-safe design

This means that in the design of the protective circuitry, one should make sure that circuits are designed in such a way that any circuit failures that can be anticipated cause the system to turn OFF. The most common example of this is in the thermal protection of loads. We typically use a thermal switch that becomes an open circuit when the temperature is exceeded. The switch is connected to circuitry that turns off the power system when an open circuit is sensed. This avoids the anticipated problem of someone leaving the thermal switch disconnected from the system during installation or maintenance periods. If this does occur, the system cannot be turned on.

### 1.3 Response to power outages

As it is a virtual certainty that the a.c. power to the system will drop out during operation, it is important to design the electronics to respond in a safe and protective way to this situation. A common way to deal with this is to run the control system from the a.c. source of an uninterruptible power supply. This battery-backed system maintains control voltages intact until power elements can be safely de-activated.

## 1.4 Testing protection circuits

Well-designed protective circuits cannot be trusted until they are tested in place. This will involve opening doors to see that door-interlocks work, disconnecting or heating thermal switches, tripping the breaker supplying a.c. power to the system, etc. Redundant systems must be tested individually by bypassing each of the paths to see that the other path works.

## 1.5 Trouble-shooting aids

Usually when a power converter or magnet system trips off, this happens because an interlock has detected an improper situation and has acted correctly. Typically, it does not mean that the system is broken and in need of repair.

It is important to understand these trips and deal with potential problems before they lead to serious failures or endless annoyances.

To this end, it is important to ensure that all trip indications latch and identify themselves. Beyond this, modern day transient recorders that can record the sequence of events leading to a trip are an invaluable tool, especially in complex systems.

## 1.6 Self-contained protection

This means that the power system's internal controls must prevent incorrect commands from the accelerator control system from putting the power system in an unprotected state. For example, an operating current reference beyond the specified maximum current should be rejected or clamped.

# 2 Protection of people

## 2.1 Lock and tag out system

Our primary method of protecting people who will be working on a power system is our lock and tag out (LOTO) system. Before people are allowed to touch any element of the power system, they must first follow the LOTO procedure to turn the system off, lock out the sources of power, verify the absence of input power, discharge the energy stored in capacitor banks, and install ground clamps where necessary to ensure that nothing can become charged after the LOTO is completed. Installed locks are tagged with the name of the person who installed them and can only be removed by that person.

## 2.2 Interlocks

We use interlocks on power cabinet doors and in the accelerator tunnel to make sure that if someone unwittingly opens the door to a hazardous area, the power system will trip off. The main hazards will be removed from the equipment, but since the full LOTO has not been accomplished, the equipment will not yet be approved for access.

## 2.3 Captured key systems

In equipment with unusually high hazards, we often use a captured key system, which requires a person to lock off the source of power to the cabinet before gaining access to the key that will allow access to the cabinet. This does not take the place of the LOTO procedure; the LOTO procedure must be followed before one can work inside the system cabinet.



### 3 Protection of power converters

The most common types of power converter used in magnet systems are the rectifier power converter and the switch-mode power converter. In this paper, we will focus on protecting the rectifier power converter.

#### 3.1 Overcurrent protection in a.c. systems

As in a small bench supply, a high-power rectifier power converter must be protected from the overcurrent that will occur following an internal short circuit. This protection is typically provided by a circuit breaker capable of interrupting the high fault current and disconnecting the power converter from the power line (see Fig. 1).

There are two levels of fault current that must be handled. If the fault occurs on the secondary of the rectifier transformer, the fault current on the primary is limited by the ‘impedance’ of the rectifier transformer (essentially its leakage inductance) and is typically 20 times the maximum operating current.

If the fault occurs on the primary side of the rectifier transformer, the current is limited by the ‘impedance’ of the transformer (typically at the power substation) powering the feeder system to which the power converter is attached. This fault current is usually much higher than the first case. The power-converter circuit breaker must be specified to interrupt this fault current in a system that is properly ‘coordinated’.

#### 3.2 Overvoltage protection in a.c. systems

The rectifier transformer is typically protected by ‘surge suppressors’ located from the primary windings to ground, which limit the voltage if the feeder system is hit by lightning or if there is some other source of large transient voltages.

The transformer secondary is protected from transients that occur when the breaker opens under load by the use of RC (resistor–capacitor) snubber networks.

The transformer and circuit breaker are specified and tested according to industrial standards to be able to withstand high transient voltages (e.g., 110 kV impulse testing on 15 kV rated equipment).

#### 3.3 Protection in d.c. systems

The most important elements to protect are the thyristors (silicon-controlled rectifiers). These devices must be chosen to have ample voltage ( $2.5 \times$  operating voltage) and current ratings. Using the thyristor data sheets, we calculate the maximum temperature the thyristor junction will attain during the most severe operating mode. We select a device whose junction temperature we can keep below  $80^{\circ}\text{C}$ .

We also calculate the temperature variation of the junction in ramping systems and make sure that the temperature cycling is less than  $30^{\circ}\text{C}$  in systems whose cycle time is greater than 1 s. These two limits, which relate to the power-converter current, ensure a long lifetime for the thyristors.

The d.c. side of the power converter often has a passive filter to reduce ripple voltage on the rectifier output. We protect the filter choke, which is usually water-cooled in high current systems, with a thermal switch. The filter capacitors often have overpressure switches, which detect a failure and are used to interlock the power converter.

#### 3.4 Response to loss of power

We usually have an uninterruptible power supply to maintain the control voltages in the event of a power failure. This allows us time to bypass the current in the supply into the bypass thyristor (SCR) and then open the circuit breaker.

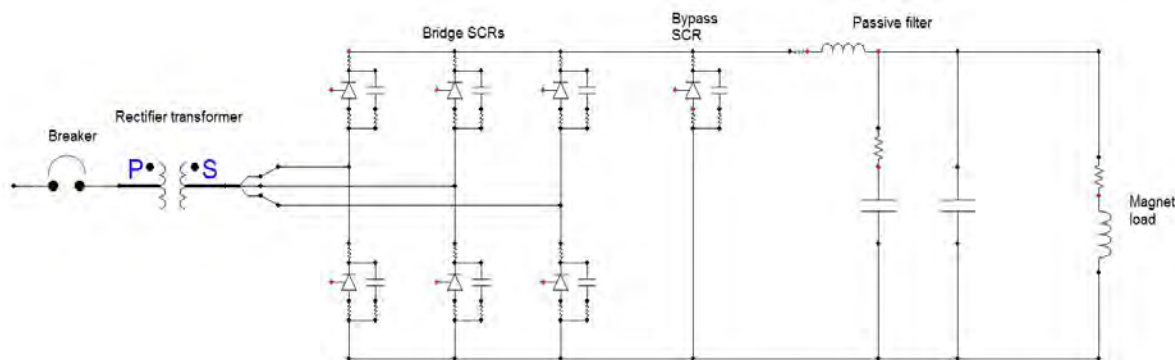


Fig. 1: Rectifier power-converter circuit

## 4 Protection of conventional magnets

### 4.1 Overcurrent protection

We normally use a direct current–current transformer (DCCT) to measure and regulate the load current coming out of the power converter. The same DCCT signal is compared with a trip threshold and bypasses the converter when the level is exceeded.

This protection is backed up with a shunt measurement that causes the circuit breaker to trip when the measured current exceeds the threshold, which is usually set slightly higher than the DCCT threshold.

For non-d.c. loads, it is sometimes necessary to design a circuit that trips the power converter based on the root mean square value of the current.

### 4.2 Voltage-to-ground protection

Magnets are manufactured with electrical conductors wound around an iron core, and isolated from the grounded core by a system of insulating material. The insulation system is designed to withstand up to a minimum voltage between the conductors and the core without breaking down.

The design and validation of the magnet insulation system must be coordinated with the worst-case voltages that magnets will experience while operating in their circuit.

The ‘ground fault circuit’ in a power converter has two functions: to detect unwanted current going to ground and to minimize the voltage-to-ground of the magnet load. We will consider the simple case of one power converter and one magnet as its load.

Figure 2 shows this simple case with two common forms of ground fault detector: the fused detector and the balanced high-impedance detector. In the former kind, the negative terminal of the power converter is held at ground with a low current fuse. If a short circuit to ground develops towards the positive terminal of the converter, it wants to pull the negative terminal below ground, but the fuse conducts to maintain the terminal at ground potential until there is sufficient current to blow the fuse. A circuit then detects that the fuse is blown and turns off the power converter.

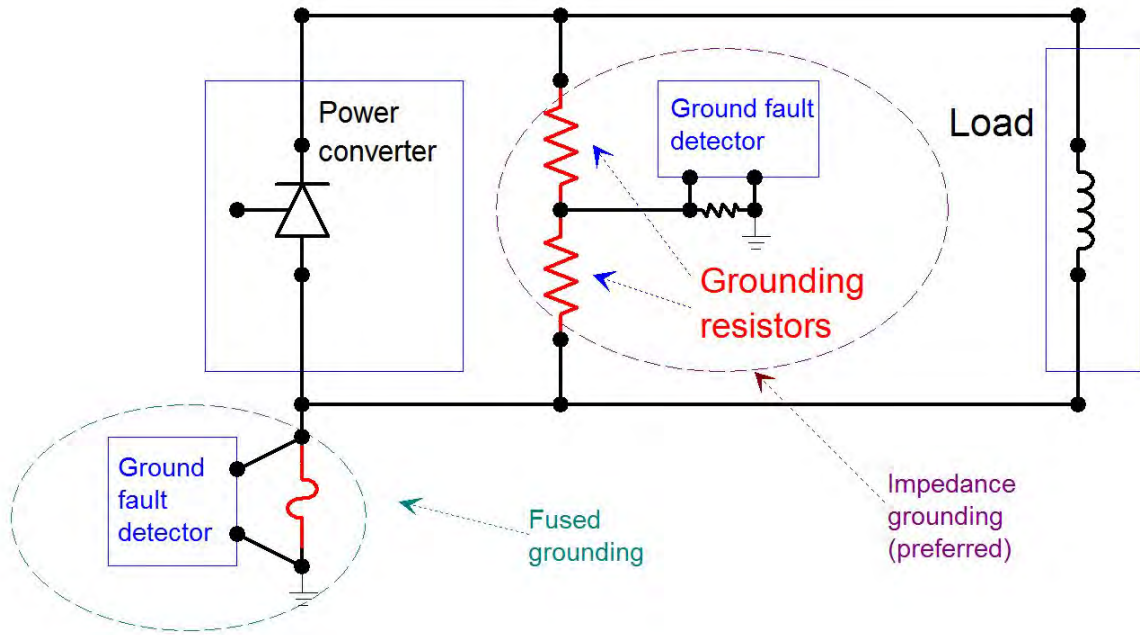


Fig. 2: Grounding schemes

This circuit has two drawbacks. If the power converter is a 1 kV unit, the part of the load connected to the positive terminal always sees the thousand volts to ground, stressing the magnet insulation at that level. Also, when a ground fault does develop, a high current may want to flow into the fuse and may exceed its interrupting rating. If this happens, this large current will flow through the faulted point for a long time and damage the point beyond recovery.

The balanced high-impedance grounding detector in Fig. 3 has grounding resistors of the order of a few kilo-ohms, so it cannot conduct high ground currents. This circuit also balances the voltage-to-ground, so that the maximum voltage that any part of the magnet sees in normal operation is  $\pm 500$  V. If a ground fault develops in this system, the summing point between the two grounding resistors, which normally stays close to ground, is deflected either a positive or negative direction, and this change in its voltage causes a trip of the power converter.

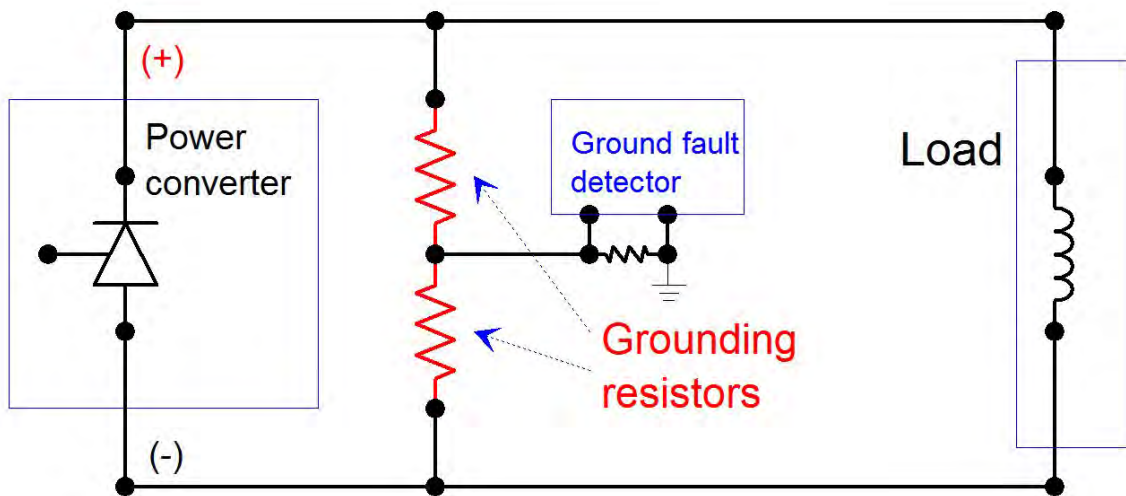


Fig. 3: Balanced high-impedance grounding

In this preferred scheme, although the maximum operating voltage-to-ground is half of the power-converter voltage, if a ground fault occurs near the negative terminal of the power converter, its positive terminal will be raised to the full voltage of the power converter before the ground fault circuit can react. If a second ground fault occurs there, this is called a ‘double ground fault’. This is very much to be avoided because it allows large, destructive currents to go through both ground points. Thus, the magnet insulation in this example must be designed to withstand more than the full power-converter voltage (with some margin), so that a double ground fault is highly unlikely.

## 5 Protection of superconducting magnets

Table 1 provides details for some example superconducting magnets. Circuits with superconducting magnets must have the same overcurrent and voltage-to-ground protection systems that are required for conventional magnet circuits. In addition, these circuits must have a quench protection system to protect the magnets when they lose their superconducting properties. We will start with some of the definitions and features of superconducting magnets.

### 5.1 Superconducting cables and magnets

- Magnets are most commonly wound with superconducting cable (niobium–titanium).
- Superconducting cable becomes superconducting (zero resistance) below a critical temperature (usually  $<10$  K). Low temperatures are established and maintained in a bath of liquid helium that is cooled in a cryogenic refrigeration system.
- Superconducting cable is made of several wire strands, each made of many superconducting filaments within a copper matrix.
- The cable has a ‘short sample’ maximum current, beyond which it will ‘quench’ and lose its superconducting properties.
- The copper matrix stabilizes the cable and provides an alternate current path for a short time when the superconductor ‘quenches’ or leaves its superconducting state.
- A magnet wound with the wire has a lower maximum current because magnetic fields within the magnet decrease the cable’s maximum conduction.
- A magnet’s maximum current can be increased by reducing its temperature. (Example: Fermilab’s Tevatron magnets went from 4000 A to 4400 A when the helium temperature was reduced by 0.75 °K at a cost of \$6 million).

**Table 1:** Different accelerator magnets with their respective operating limits and the cross-sectional areas of their cables (the MIIT limit column is explained later in the text).

| Magnet                | Short sample limit, kA | Maximum operating current, kA | Operating temperature, K | MIIT limit, A <sup>2</sup> s | Cable cross-section, mm <sup>2</sup> |
|-----------------------|------------------------|-------------------------------|--------------------------|------------------------------|--------------------------------------|
| LHC: main bend        | 13                     | 11.5                          | 1.85                     | $32 \times 10^6$             | 22                                   |
| LHC: main quadrupoles | 13                     | 12.1                          | 1.85                     | $32 \times 10^6$             | 22                                   |
| LHC: 600 A            | 0.6                    | 0.55                          | 1.85                     | $50 \times 10^3$             | 1                                    |
| Tevatron: dipoles     |                        | 4.4                           | 4.5                      | $7 \times 10^6$              | 10                                   |

## 6 Quenching

A magnet conducting current in superconducting mode at cryogenic temperatures can suddenly lose its superconductive state, usually beginning at a particular spot in the magnet cable, when something causes the temperature at that spot to rise above the critical temperature.

Once the initiating spot quenches, the heat generated from the resistance typically keeps it in the quenched state, and the quenched area spreads to nearby areas at a speed known as the ‘quench velocity’.

### 6.1 Causes of quenching

There are several well-known causes of quenching in accelerator magnet systems.

#### 6.1.1 Training

Training refers to the small slippage of one of the superconducting cables within the tightly clamped magnet structure. This slippage occurs when the magnet undergoes its initial powering. The slippage generates some heat, by friction, and the heat can cause a quench at that location. Usually, the cable slips into a more stable position and, once the magnet has undergone a few quenches as the current is increased to its maximum operating value, these quenches no longer occur; we say that the magnet has been ‘trained’.

#### 6.1.2 Excess $dI/dt$ —eddy currents

Fast current changes can induce local eddy currents within the superconductor cable itself. These can cause local heating, leading to a quench of the cable. Sometimes, the action of rapidly decreasing the current in a magnet system to protect an element in the circuit can cause other elements to quench from excess  $dI/dt$ .

#### 6.1.3 Particle beam heating

If the particle beam in a storage ring becomes unstable or the injection or extraction of the beam are not well controlled, part of the beam can hit the superconducting cable and cause a quench.

#### 6.1.4 Cooling system problems

If the helium warms to above the critical temperature, quenches will occur.

#### 6.1.5 Exceeding the short sample limit

Normally, the powering system is set up to protect against this happening, but sometimes the protection does not work correctly.

#### 6.1.6 Spontaneous quenches—unknown origins

Mystery quenches do happen in complex systems. This is why it is important to have good monitoring systems to provide the best chance of understanding each quench.

### 6.2 Heating of the initiating spot—MIITs

The initiating spot starts to heat first, and is normally the hottest place in the quenching magnet. Keeping its ultimate temperature below a damaging level (450 K) is critical in protecting the quenched magnet.

A simplified way of thinking about the temperature rise at the initiating spot is to imagine it as a length of copper wire weighting  $M$  (g) with resistance  $R$  ( $\Omega$ ) and specific heat  $C$  (J/(g K)). Then the adiabatic temperature rise of the spot will be:

$$\Delta T = R \frac{\int I^2 dt}{MC} = \frac{R}{MC} \int I^2 dt .$$

Note that the temperature rise in this calculation is independent of the wire length.

You can calculate an integral of  $I^2$  that will raise the temperature of the initiating spot from 10 K to 450 K. This integral is called the ‘MIIT’ limit of the cable. Usually this integral is in ‘millions of amp squared seconds’, hence the term ‘MIIT’.

This simplified example assumes constant values of  $R$  and  $C$ . In real life, the resistance of copper varies by about a factor of 100 between cryogenic and room temperatures and the specific heat varies by about a factor of 300. The calculation of realistic MIITs is not so different from the ideal because both  $R$  and  $C$  increase with increasing temperatures, and thus tend to compensate each other.

The MIITs calculation is crucial for determining the two primary factors of the quench protection system:

- How quickly the current in the magnet must be reduced once the quench is detected;
- How quickly a quenched must be detected once the initiating spot quenches.

The allowable MIIT (see Table 2) in a magnet and the maximum operating current in its circuit set a maximum time-scale for the reduction of the current once a quench has been detected. Some examples of this relationship follow.

**Table 2:** MIIT limits and operating currents for three accelerator circuits: kIIT, ‘thousands of amp squared seconds’; MIIT, ‘millions of amp squared seconds’.

| Accelerator circuit      | MIIT limit | Operating current, kA |
|--------------------------|------------|-----------------------|
| Tevatron dipole          | 7 MIIT     | 4                     |
| LHC dipole               | 32 MIIT    | 12                    |
| Tevatron correction coil | 3.2 kIIT   | 0.05                  |

The time-scales for the reduction of the currents in these circuits are:

- For a Tevatron dipole running at 4 kA (16 MIIT/s), the current in the quenching magnet must be substantially reduced within 0.44 s (7 MIIT ÷ 16 MIIT/s).
- For an LHC dipole running at 12 kA (144 MIIT/s), the current in the quenching magnet must be substantially reduced within 0.22 s (32 ÷ 144).
- For a Tevatron correction element running at 50 A (2.5 kIIT/s), the current in the quenching magnet must be substantially reduced within 1.28 s (3.2 ÷ 2.5).

### 6.2.1 *Methods to reduce number of MIITs following quench detection in magnet circuits*

- Reduce power-converter voltage to zero if cable resistance is enough to limit the MIITs (example: Tevatron extraction quadrupole loops).
- Reduce power-converter voltage and use energy extraction circuit to insert a resistance sufficient to limit the MIITs (example: Tevatron main quadrupole correction loop).
- Reduce power-converter voltage and fire ‘Heaters’ (example: Tevatron low-beta magnets).

### 6.2.2 *Methods for limiting MIITs after quench detection*

#### 6.2.2.1 *Method a*

Reduce power-converter voltage to zero if cable resistance is sufficient to limit the number of MIITs (e.g., Tevatron extraction quadrupole loops). See Fig. 4.

System data:

Four magnets at 0.46 H each = 1.84 H ,

Cable  $R = 2.45 \Omega$  ,

$L/R = 0.75 \text{ s}$  ,

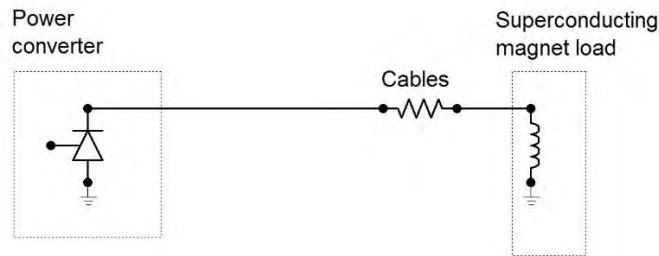
$I_{\max} = 50 \text{ A}$  ,

For exponential decay,

$$\text{IIT} = \frac{1}{2} (I_{\text{pk}}^2 \tau) ,$$

$$\text{IIT} = \frac{1}{2} \times 50 \times 50 \times 0.75 \text{ s} = 0.94 \text{ kIIT} ,$$

which compares well with the 3.2 kIIT limit.



**Fig. 4:** Limiting MIITs after quench detection, method a

#### 6.2.2.2 Method b

Reduce voltage and use energy extraction circuit (insert a resistance, as in the Tevatron main quadrupole correction coil loops). See Fig. 5.

System data:

90 magnets at 0.46 H each = 41 H ,

Cable  $R = 5.8 \Omega$  ,

Cable  $L/R = 7.14 \text{ s}$  ,

Dump  $R = 20 \Omega$  ,

$L/R = 1.6 \text{ s}$ , including dump  $R$  ,

$I_{\max} = 50 \text{ A}$

$$\text{IIT} = \frac{1}{2} \times 50 \times 50 \times 7.14 \text{ s} = 8.9 \text{ kIIT without dump } R ,$$

$$\text{IIT} = \frac{1}{2} \times 50 \times 50 \times 1.6 \text{ s} = 2.0 \text{ kIIT with dump } R .$$

The result with the dump resistance compares well with the 3.2 kIIT limit.

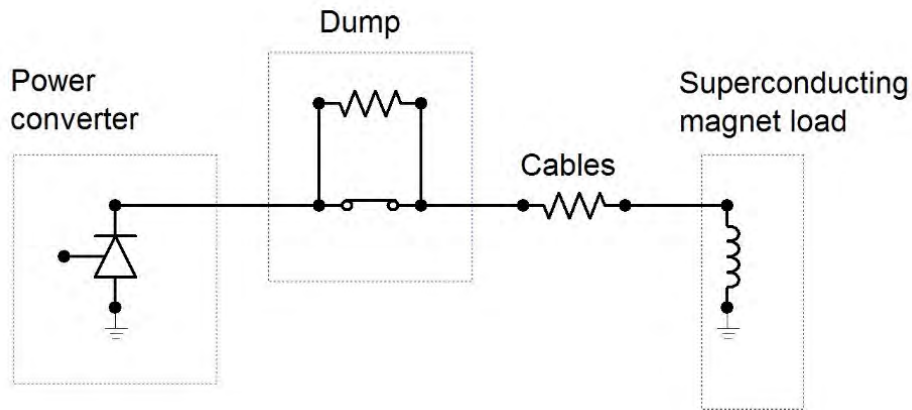


Fig. 5: Limiting MIITs after quench detection, method b

6.2.2.3 Method c

Reduce voltage and fire heaters (e.g., Tevatron low-beta magnets). See Fig. 6.

System data:

Two magnets at 46 mH each = 92 mH,

Cable  $R = 3 \text{ m}\Omega$ ;

Cable  $L/R = 30 \text{ s}$ ,

Magnet  $R = 375 \text{ m}\Omega$  equivalent with heater firing,

Magnet, with heater firing,  $L/R = 0.25 \text{ s}$ ,

$I_{\max} = 5 \text{ kA}$ ,

$\text{IIT} = \frac{1}{2} \times 5000 \times 5000 \times 30 = 375 \text{ E6} = 375 \text{ MIIT}$

With heater firing,

$\text{IIT} = \frac{1}{2} \times 5000 \times 5000 \times 0.25 = 3.13 \text{ E6} = 3.13 \text{ MIIT}$

This result compares well with the 3.2 MIIT limit.

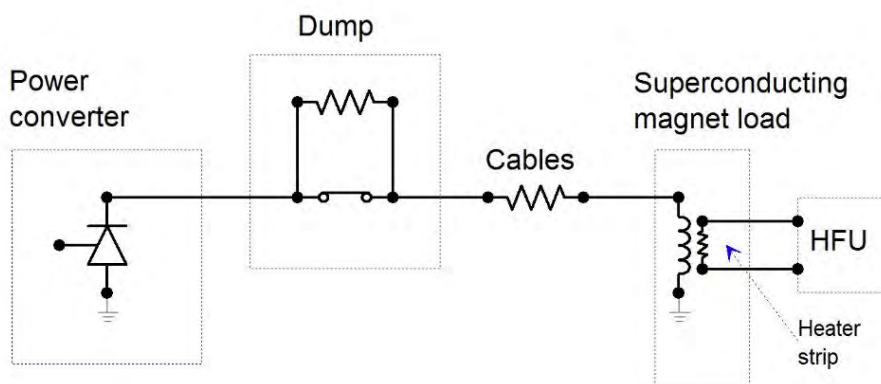


Fig. 6: Limiting MIITs after quench detection, method c: HFU, heater firing unit



Heaters are steel strips that are pressed against the outer windings of a magnet. They are designed to have a reasonably low thermal insulation but sufficient electrical insulation to avoid arcing to the magnet winding. When a quench is detected, the heater firing unit is triggered and discharges the energy in a capacitor bank into the steel strip. This energy is sufficient to initiate a quench in a large fraction of the magnet cabling. The increase in resistance of this large volume of quenched cabling is sufficient to reduce the magnet current before it reaches its MIIT limit. Details of two examples of heater circuit are given in Table 3.

**Table 3:** Two examples of heater circuit

| Circuit         | Capacitance, mF | Voltage, V | Energy, kJ | Strip resistance, Ω | Discharge time, ms |
|-----------------|-----------------|------------|------------|---------------------|--------------------|
| Tevatron dipole | 6.6             | 450        | 0.67       | 20                  | 18                 |
| LHC dipole      | 7               | 900        | 2.8        | 12                  | 84                 |

Although heaters are an effective way to induce a large-scale quench in a superconducting magnet, the technique requires extra strips to be built into the magnet and the use of a high-energy discharge circuit. Most concerning in this approach is the possibility of heaters failing and, in the worst case, shorting the magnet winding to ground. Although heaters have certainly failed, we are not aware of cases in which the magnet was shorted.

A new induced-quenching technique is being developed at CERN by E. Ravaioli. It is called coupled loss-induced quenching, and it involves discharging a capacitor bank across the terminals of the magnet, causing a high-frequency ringing and inducing a  $dI/dt$  quench.

### 6.2.3 The special case of series magnet strings with large quantities of stored energy

The accelerator world often contains extended systems with many magnets and large quantities of stored energy.

**Table 4:** Two examples of extended multi-magnet systems

| Accelerator       | Number of magnets | Maximum current, kA | Total inductance, H | Energy, MJ |
|-------------------|-------------------|---------------------|---------------------|------------|
| Tevatron ring     | 776               | 4.4                 | 30                  | 290        |
| LHC dipole sector | 154               | 11.5                | 15.4                | 1018       |

It is impractical to remove this much energy from the magnet systems in a fraction of a second, so an approach using heater firing, ‘bypassing’, and ‘energy extraction’ has been used. When a quench is detected in one of the magnets, the quench protection system takes the three actions:

- fire the heater firing unit on the quenching magnet;
- establish a bypass path for the main circuit current to go around the quenching magnet while its own current decays within a fraction of a second;
- open switches to insert dump (energy extraction) resistors so that the magnet circuit current can decay on a time-scale of several seconds.

The time constant of the dump is coordinated with the number of MIITs that the bypass path can absorb without overheating. The time constants used in the two cases listed in Table 4 are:

$$\text{Tevatron} = 12 \text{ s ,}$$

$$\text{LHC dipole sector} = 100 \text{ s .}$$

Fig. 7 is a composite showing the quench protection systems of the LHC and the Tevatron. Both systems use heater firing units, as described before.

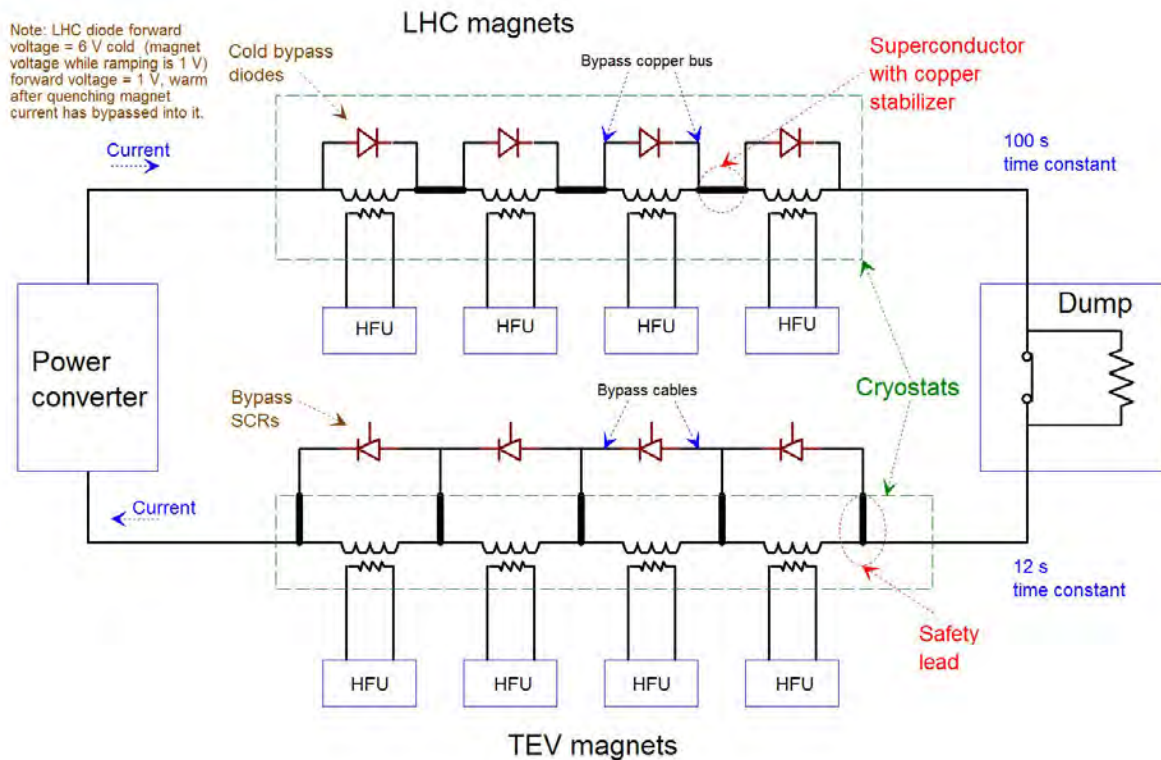
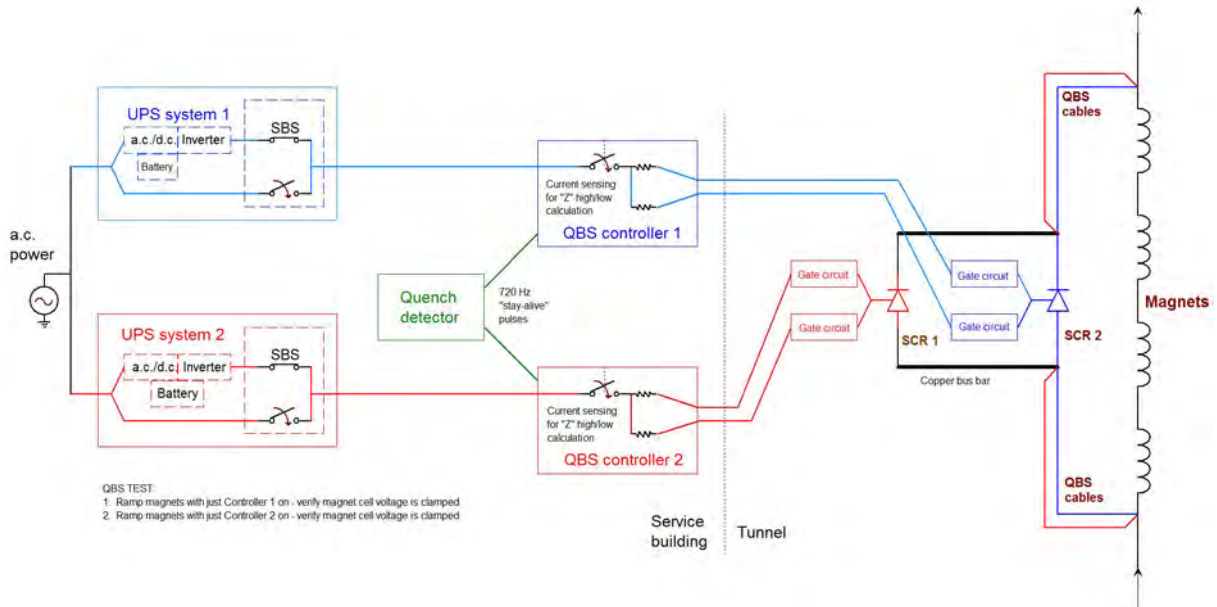


Fig. 7: Tevatron and LHC quench protection systems

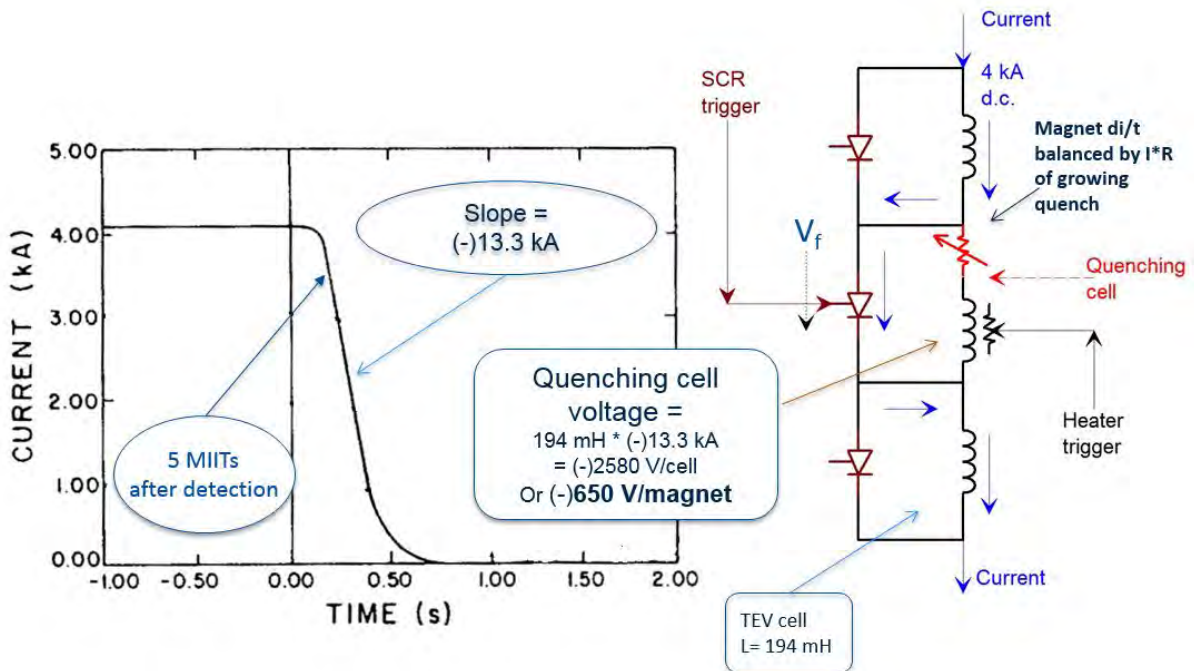
In the Tevatron, the bypass circuit goes through a silicon-controlled rectifier (thyristor) that is triggered only when a quench is detected. The current in the quenching magnet exits through copper ‘safety leads’, is carried to the external silicon-controlled rectifier bypass switch via copper cables, and returns on the other side of the quenching magnet in the same way. The limiting elements of the bypass circuit are the safety leads, which are made as small as possible to limit heat leaks from the cryogenic system to the outside. These leads can handle the peak magnet current (4.4 kA) during a 12 s exponential decay without overheating.

Figure 8 shows the redundant features of the Tevatron bypass switch system. It features parallel silicon-controlled rectifiers, parallel firing circuits, and parallel uninterruptible power supply systems.



**Fig. 8:** Tevatron quench bypass system, emphasizing redundant architecture. QBS, quench bypass switch; SBS, standby power supply; SCR, silicon-controlled rectifier; UPS, uninterruptible power supply.

Figure 9 shows what happens to the current in a quenching Tevatron magnet in a string with a bypass silicon-controlled rectifier. We have measured the decay time of the magnet current as less than 0.2 s. This indicates that the resistive drop of the heater-driven magnet builds up to 650 V. The current is driven out of the magnet with an amp squared seconds integral of 5 MIITs.



**Fig. 9:** What happens to the current in a quenching Tevatron magnet in a string with a bypass silicon-controlled rectifier when the heaters are fired and the current is bypassed.

In the LHC, the bypass circuit goes through the 'cold bypass diode', which starts conducting when the voltage across the quenching magnet exceeds 6 V. At this point, the diode's self-heating rapidly transforms it into a normal diode that conducts at 1 V. It then bypasses the quenched magnet current

through a copper bus. This ‘copper stabilizer bus’ is a 1.5 cm× cm copper bus with a slot inside it that carries the superconductor cable between magnets. This stabilizer bus must carry the bypass current if a section of interconnecting bus quenches. The limiting elements of the bypass circuit are the copper stabilizer buses. They can handle the peak magnet current (12 kA) during a 100 s exponential decay without overheating.

Table 5 gives a summary of MIITs that are deposited in various systems once the energy extraction circuits have been triggered.

**Table 5:** MIITs that are deposited in various systems once the energy extraction circuits have been triggered

|  | MIITs    | Maximum  |
|--|----------|----------|
| Tevatron extraction quadrupole           | 0.9 kIIT | 3.2 kIIT |
| Tevatron main quadrupole correction loop | 2.0 kIIT | 3.2 kIIT |
| Tevatron low-beta magnets                | 3.2 MIIT | 5.2 MIIT |
| Tevatron dipoles                         | 5.0 MIIT | 7.0 MIIT |

## 7 Quench detection

To trigger the energy extraction circuits to limit the MIITs in a magnet, we must detect that a quench is occurring in that magnet.

What are we detecting? Basically, the extra  $IR$  ‘resistive’ voltage that should not be there in a superconducting load ( $R$  = resistance of quenching cable as quench propagates).

How much time do we have? The time between the initiation and detection of a quench,  $T$ .

- Remember, MIITs start accruing from moment that the initiating spot quenches.
- Once the quench is detected and the protection system responds, a certain number of (after energy extraction) MIITs will be deposited.
- Therefore, the maximum time  $T_{\max}$  is given as:

$$T_{\max} = \frac{(\text{Maximum MIIT} - \text{MIIT after energy extraction})}{I^2}.$$

What voltage must we detect in the allowed time? This depends on the quench velocity within the magnet. We need to make calculations and make measurements to determine what sensitivity our quench detection system must have. Two examples will illustrate the process. Table 6 provides details for some quench detection sensitivities.

### 7.1 Example A: The quadrupole correction loop

Max = 3.2 kIITs; after energy extraction = 2.0 kIITs,

$$\Delta T = \frac{\text{IITs(rating)} - \text{IITs(after energy extraction)}}{I(\text{before quench})^2},$$

$$\text{Time to detect, } \Delta T = \frac{3.2 \text{ kIITs} - 2.0 \text{ kIITs}}{50^2} = 0.48 \text{ s.}$$

#### 7.1.1 How much voltage? How quickly does resistive voltage increase?

Experiments were performed on dipole correction elements during which a heater was fired to cause a quench condition while 50 A was being conducted through the element. The coil voltage reached 10 V within approximately 0.25 s. This compares well with the calculated detection time 0.48 seconds.

Therefore, a 10 V detection threshold would be sufficient. We were able to operate with a threshold of 4 V without nuisance trips.

### 7.2 Example B: The main Tevatron loop

After energy extraction, 5 MIIT are deposited when running at 4 kA.

#### 7.2.1 How much time do we have?

$$\text{Time to detect, } \Delta T = \frac{(7 \text{ MIITs} - 5 \text{ MIITs})}{4^2} = 0.125 \text{ s} .$$

#### 7.2.2 How quickly does resistive voltage increase?

A complicated calculation involving quench velocity in single wires, reveals that we must choose a threshold of 0.5 V. In practice, we aim for a stringent goal and see how close we can come, within the constraints of periodic and random noise. We must have a large enough margin to avoid nuisance trips.

**Table 6:** Summary of various quench detection sensitivities

| System                         | Trip threshold, V | Averaging time, ms |
|--------------------------------|-------------------|--------------------|
| Tevatron quadrupole correctors | 4.0               | 10                 |
| Tevatron main dipoles          | 0.5               | 50                 |
| LHC main dipoles               | 0.1               | 10                 |
| LHC 600 A circuits             | 0.4               | 200                |

## 8 How do you detect resistive voltage?

### 8.1 Compare voltage across similar magnets in series

#### 8.1.1 Example A: Tevatron correction quadrupole loop

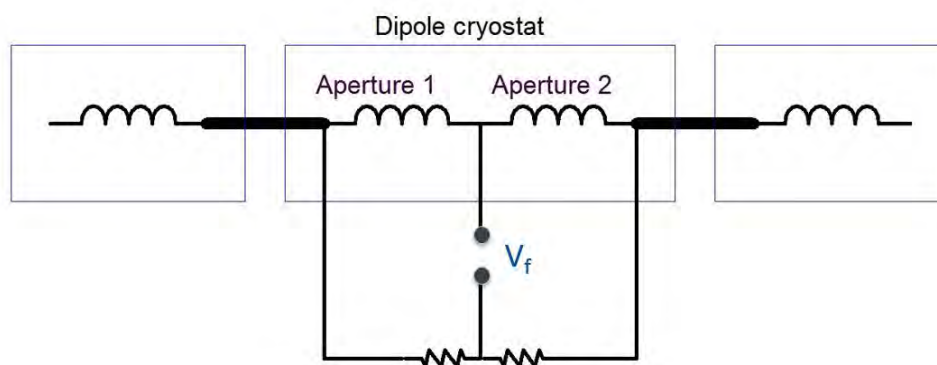
Compare the voltage across one set of 45 magnets with that across the other set of 45 magnets (carefully, using centre tap). Look for a 4 V difference.

#### 8.1.2 Example B: Tevatron main loop

Compare voltage across four cells (five magnets each). Look for a 0.5 V difference.

#### 8.1.3 Example C: LHC 13 kA dipole circuits

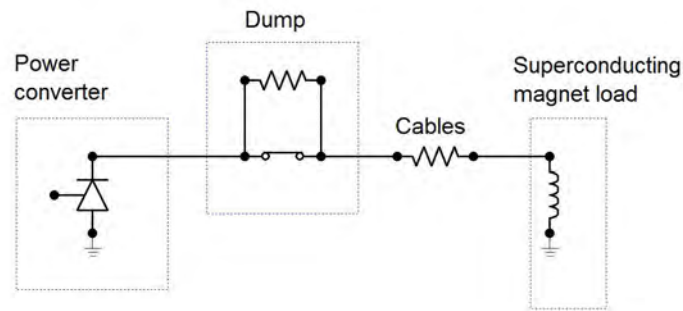
This is illustrated in Fig. 10. Look for a 0.1 V difference.



**Fig. 10:** Detecting resistive voltage in an LHC 13 kA dipole circuit

## 8.2 Compare $L \times dI/dt$ or magnet with measured magnet voltage

### 8.2.1 Example: LHC 600 A corrector circuits



**Fig. 11:** Comparing  $L \times dI/dt$  or magnet with measured magnet voltage for an LHC 600 A corrector circuit

This is illustrated in Fig. 11. Subtract:

$$V_{\text{mag}} - L \frac{dI}{dt}.$$

Look for a 0.4 V difference.

Find the quench voltage,  $V_q$ :

$$V_q = V_{\text{mag}} - Lm \frac{dI}{dt}.$$

This method is sensitive to noise on the measured current signal and to the complex impedance of the magnet. (The LHC is still upgrading these quench detection systems.)

## 8.3 Comparing quench detection method

The method of comparing similar magnet voltages with each other typically allows for lower quench detection thresholds:

- no  $dI/dt$  noise issues;
- no complex magnet impedance issues.

Comparing only two magnet voltages with each other introduces a vulnerability to symmetrical quench growth in both magnets. The LHC encountered this in the main dipole bus and mitigated the possibility by installing an additional system, which compared four magnets with each other.

## 9 Unexpected problems

### 9.1 Large Hadron Collider

Bad solder joint in the copper stabilizer used in the main dipole bus splices (Fig. 12) could not support the current when the superconducting cable in the splice quenched. The resulting arc deposited huge amounts of energy into the cryogenic system, causing excessive pressure, and leading to a ‘helium leak’.

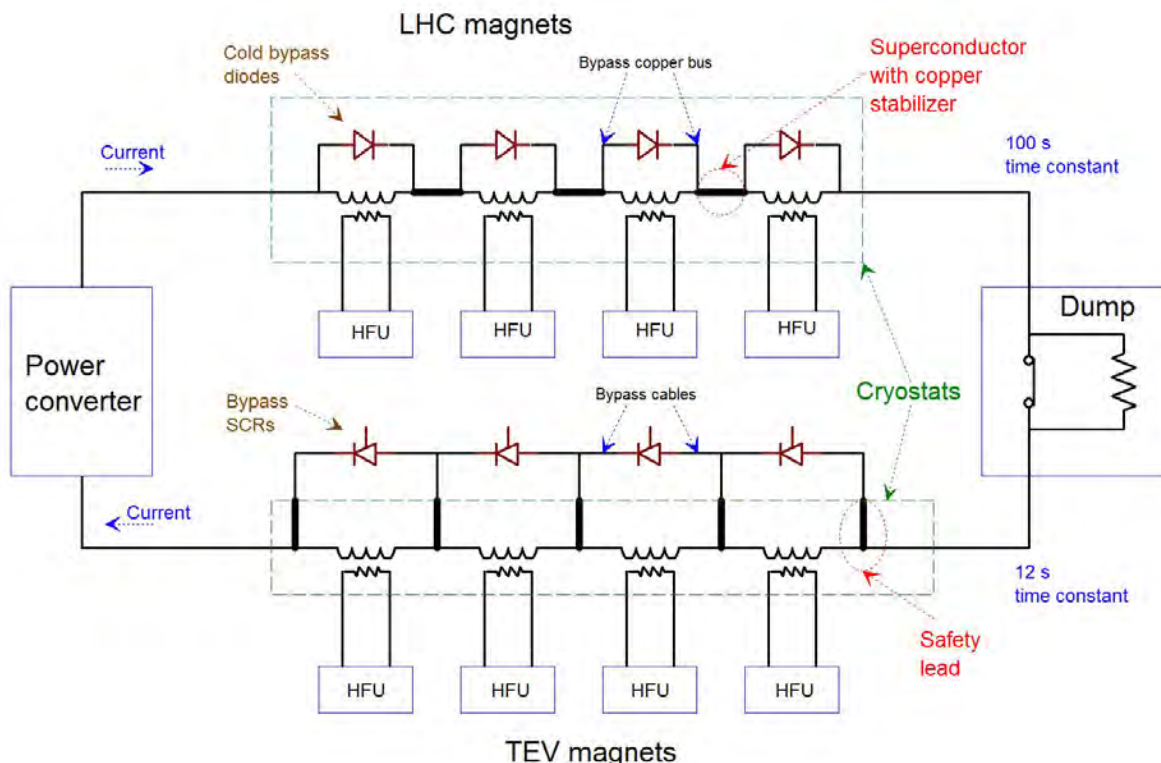


Fig. 12: Arrangement of LHC and Tevatron magnets: HFU, heater firing unit; SCR, silicon-controlled rectifier

### Definitions

**Cable short sample curve:** The two-dimensional curve in magnetic field and current space formed by the intersection of the critical surface and a plane of constant operating temperature, where the current density is integrated over the cross-section of a specific cable. This curve is measured with a ‘short sample’ of the cable placed in different magnetic fields while the current is slowly increased, until a quench occurs.

**Critical surface, cable short sample curve, magnet short sample limit:** These terms, which are directly related but distinctly different, are often referred to by slightly different or abbreviated names. Starting with the most general term, ‘critical surface’, each term is increasingly more specific and less general.

**Critical surface:** The three-dimensional surface in temperature, magnetic field, and current density space under which a specific conductor remains superconducting. The points where this surface intersects the three axes are called the critical points;  $T_c$ ,  $B_c$ , and  $J_c$ , respectively.

**DCCT:** Direct current–current transformer

**Dumping:** Process of inserting resistors into a circuit consisting of superconducting elements to remove stored energy

**HFU:** Heater firing unit

**Lower critical field:** The magnetic field at which the magnetic flux starts to penetrate a type 2 superconductor

**Magnet short sample limit:** The current where the magnet (peak field) load line intersects the cable short sample curve

**Magnet short sample margin:** The difference between the operating current and the magnet short sample limit

**Magnet temperature margin:** The temperature elevation necessary to diminish the magnet short sample margin to zero

**MIIT:** The exact (adiabatic) relationship between MIITs and temperature depends on only two things, the intrinsic conductor material properties and the cross-sectional area squared:

$$A^2 D \int_{T_0}^T \frac{c(T)}{\rho(T)} dT = \int_0^\infty I(t)^2 dt = \text{MIIT} .$$

**MOV:** Metal oxide varistor, a non-linear device for controlling overvoltages

**Power converter:** Any device that converts one form of voltage or current to another form. In this context, usually a power supply that converts an incoming a.c. line to d.c.

**QBS:** Quench bypass switch

**Quench:** Sudden runaway loss of superconductivity, driven by the heat of normal conduction, driven by the loss of superconductivity, driven by the heat of normal conduction, etc.

**SCR:** Silicon-controlled rectifier, a solid state switch, in which applying a voltage to the ‘gate’ will switch the device from an open circuit to a diode

**Superconductivity:** A phenomenon of exactly zero electrical resistance and expulsion of magnetic fields occurring in certain materials when cooled below a characteristic critical temperature

**Type 1 superconductors:** This category of superconductors mainly comprises metals and metalloids that show *some* conductivity at room temperature. They require incredible cold to slow down molecular vibrations sufficiently to facilitate unimpeded electron flow.

**Type 2 superconductors:** Except for elements vanadium, technetium, and niobium, the type 2 category of superconductors comprises metallic compounds and alloys. They achieve higher  $T_c$  than type 1 superconductors by a mechanism that is still not completely understood. Current wisdom holds that it relates to the planar layering within the crystalline structure.

**Upper critical field:** The magnetic field (usually expressed in teslas (T)), which completely suppresses superconductivity in a type 2 superconductor at 0 K (absolute zero).

## Bibliography

E. Ravaioli *et al.*, *IEEE Trans. App. Supercond.* **24**(3), 0500905.

<http://dx.doi.org/10.1109/TASC.2013.2281223>

D. Wolff, Quality control and maintenance considerations in equipment design, ADDP-EE-2004 (2004).

[http://www-bdees.fnal.gov/world/misc/ADDP\\_EE\\_2004\\_rev\\_1.pdf](http://www-bdees.fnal.gov/world/misc/ADDP_EE_2004_rev_1.pdf)



## Protection of Accelerator Hardware: RF systems

*S.-H. Kim*

Oak Ridge National Laboratory, Oak Ridge, TN, USA

### Abstract

The radio-frequency (RF) system is the key element that generates electric fields for beam acceleration. To keep the system reliable, a highly sophisticated protection scheme is required, which also should be designed to ensure a good balance between beam availability and machine safety. Since RF systems are complex, incorporating high-voltage and high-power equipment, a good portion of machine downtime typically comes from RF systems. Equipment and component damage in RF systems results in long and expensive repairs. Protection of RF system hardware is one of the oldest machine protection concepts, dealing with the protection of individual high-power RF equipment from breakdowns. As beam power increases in modern accelerators, the protection of accelerating structures from beam-induced faults also becomes a critical aspect of protection schemes. In this article, an overview of the RF system is given, and selected topics of failure mechanisms and examples of protection requirements are introduced.

### Keywords

Radio-frequency system; machine protection; breakdown; accelerating structure; beam-induced fault.

## 1 Introduction

Particle acceleration using time-varying electromagnetic fields is called radio-frequency (RF) acceleration. Radio-frequency system frequencies for particle accelerators range from 10 MHz to 30 GHz, either in continuous-wave or pulsed operation. The RF power per unit station ranges up to a few megawatts in continuous-wave machines and 100 MW in pulsed machines. Many state-of-the-art technologies are involved in producing the RF systems, such as vacuum science, high-voltage technology, surface physics, advanced materials, and high speed controls. As superconducting RF technologies become a choice for modern machines, cryogenics, superconducting RF science, and ultra-clean processing play important roles for RF systems.

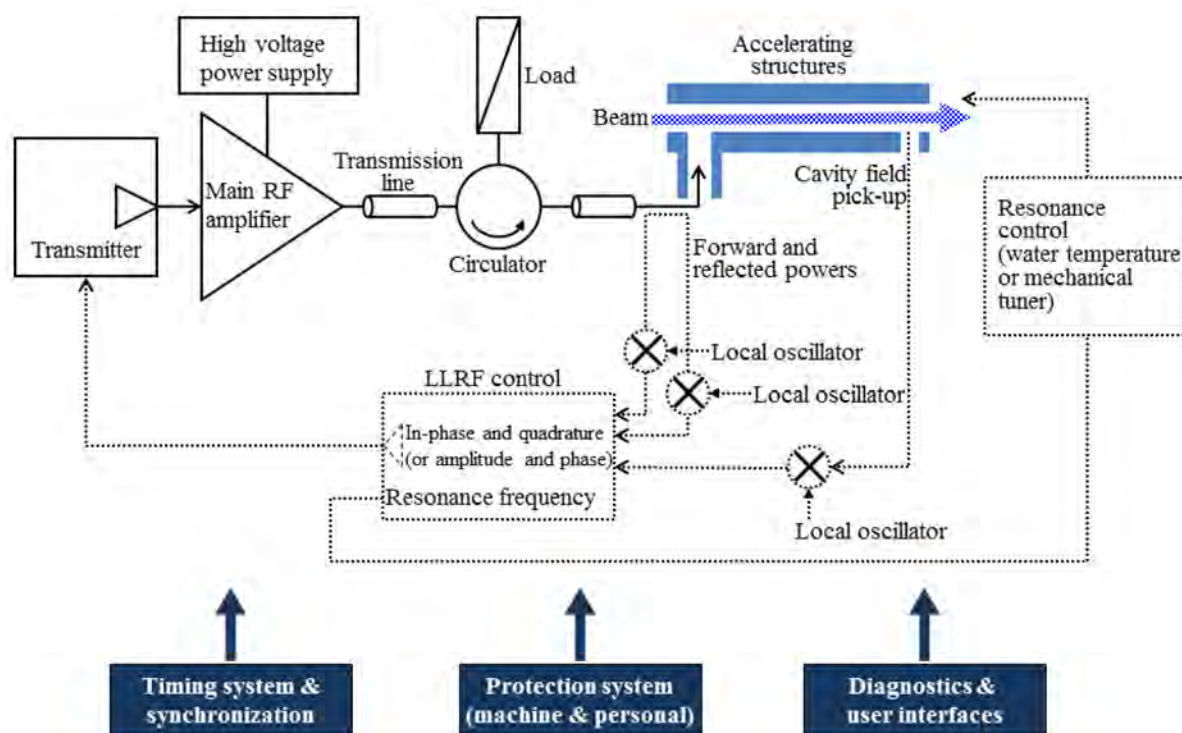
A typical layout of an RF system for a particle accelerator is shown in Fig 1. The high-power chain is depicted with solid lines and the low-power part with dashed lines. For proton or ion acceleration, various types of accelerating structure are needed because each type of accelerating structure has a limited acceptance of particle velocity. Figure 2 shows different types of accelerating structure using normal conducting and superconducting technologies. As the beam power or energy stored in the beam increases in modern accelerators, the protection of accelerating structures from the beam becomes more important in machine protection schemes. Also, the accelerating gradients required for future machines are increasing, and there are questions over the breakdown limits, in terms of accelerating gradient, pulse length, trip rate, and expected lifetime in connection with material type, processing method, and processing history.

The goal of hardware protection is to keep the system in a reliably operable condition until the end of the equipment's lifetime. The concepts for this would be classified in two categories; setting the normal operating parameters to ensure their lifetime and minimizing abnormal operating conditions that

could damage the system. There are many mechanisms that could cause damage and lead to catastrophic failures, e.g.:

- water leaks into the system components;
- air leaks into the vacuum boundary spaces;
- sharp edges in high-electric-field regions;
- dirty RF surfaces;
- resonant electron multipacting;
- condensed or trapped gas on RF surfaces;
- large reflected power to RF sources;
- over-power (voltage or current) to system components;
- beam bombardments on RF surfaces: activation, errant beam, mis-steered beam.

The specific requirements for the system protection vary depending on equipment-specific characteristics, such as machine type, beam power, beam energy, or beam pulse length. Since the hardware protection system mostly deals with abnormal conditions, it is essential to have a good understanding of the physics of individual equipment or components, and their interplay with the system. This understanding is important not only for normal operating conditions but also for any possible upset conditions.



**Fig. 1:** Typical RF system for typical particle accelerators

One of the most threatening consequences in RF systems is the generation of surface damage, which could be caused by several different mechanisms. For example, arcing or discharge could occur in a system; interaction with large quantities of available RF power or stored energy in the system could result in severe arcing conditions and consequent surface damage. Another example might be a mis-steered beam that directly strikes the accelerating structure and could also cause surface damage. If an event is mild, a system will be brought to a normal condition through conditioning processes. However,

when surface damage happens at a location of high electric field, ceramic surfaces, welding or brazing joints, or aggressive multipacting regions, the process could be irreversible. The severity of arcing events is determined by local physical conditions. In many cases, readings from diagnostic devices, such as vacuum gauges, beam loss monitors, or temperature sensors, do not reveal actual local conditions. Without a thorough understanding of the failure mechanism, it may not be possible to detect a precursor of the threatening events. When systems are over-protected, beam availability will be reduced. Thus, the protection system should be designed to avoid or minimize these threatening events, while maintaining beam availability.

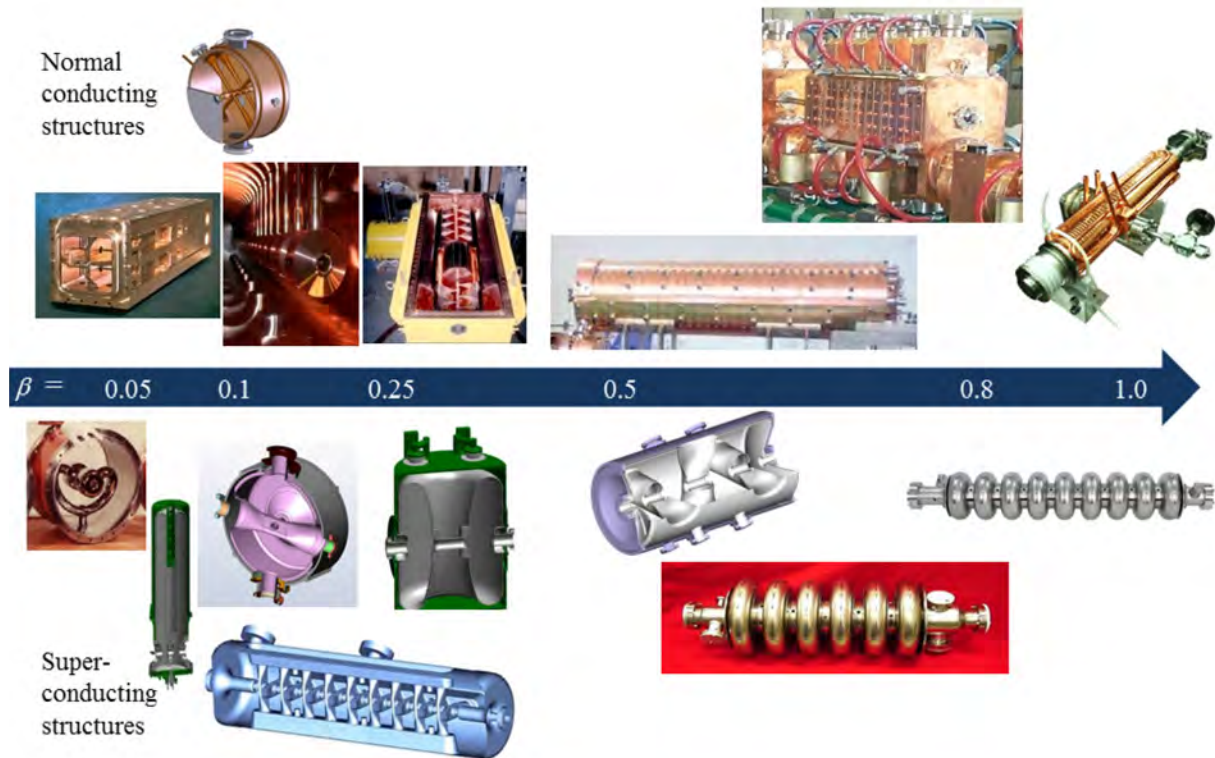


Fig. 2: Accelerating structures for particle acceleration

## 2 Breakdown

### 2.1 Vacuum breakdown mechanism

There has been a great deal of work to understand mechanisms for the vacuum breakdown, e.g. in high-voltage, plasma, particle accelerator, or vacuum science. In a high vacuum, the electron mean free path is much longer than the distance between electrodes or the length of RF structures, which means that there is no formation of electron avalanches in space for the initiation of breakdown, as in gas glow discharge. The details of physics in vacuum breakdowns are not yet well understood; however, several models developed by various authors are available. There are differences between RF and d.c. breakdowns. In an RF field, electrodes (cathode and anode) are not defined, as in d.c. fields. One can consider surfaces where electrons are emitted as cathodes and surfaces hit by electrons as anodes in RF breakdown. In this context, cathode and anode surfaces in an RF field reverse every half cycle. There is also a difference related with the duty factor. However, RF and d.c. breakdown have similar underlying physics. Studies to explain vacuum breakdown are focused on mechanisms that trigger a breakdown, such as the release of electrons and gases into the vacuum space. Thus, breakdown models can be classified in terms of breakdown initiation. Selected mechanisms are summarized in the following subsections.

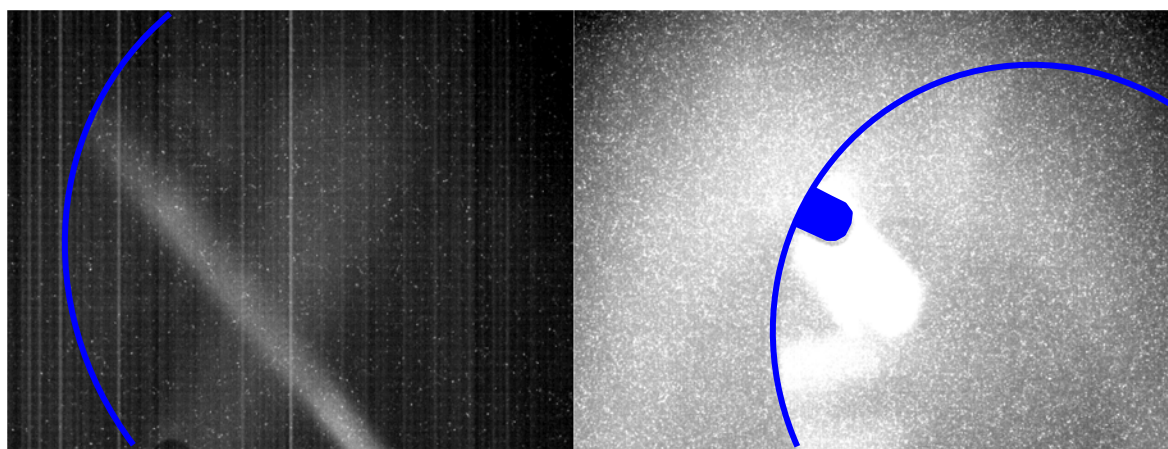
### 2.1.1 Field emission model

As the electric field at a surface is increased, the potential barrier at the surface that bounds electrons inside the material becomes narrower. At a very high field, this potential barrier is so narrow that some electrons can pass through it. This effect is known as the tunnelling effect and results from the quantum mechanical nature of electrons. The resulting electron emission is called field emission, and is also known as cold emission. The Fowler–Nordheim equation predicts the current density from field emission,

$$j \propto \frac{(\beta E_s)^2}{\phi} \exp \left[ \frac{-a \phi^{3/2}}{\beta E_s} \right],$$

where  $\beta$  is the field-enhanced factor determined by the emitter geometry,  $E_s$  is the surface electric field,  $\phi$  is the work function of the surface, and  $a$  is the constant.

Peak surface electric fields of accelerating structures are normally higher than 10 MV/m and up to a few hundred megavolts per metre. X-ray measurements show the existence of electrons and their accelerations in normal operating conditions. This current is often called the dark current and this effect is one of electron loading in the RF systems. In superconducting RF cavities especially, it could limit the achievable accelerating gradient, reduce the quality factor, and quench local spots or the whole system. If field emission passes a threshold, breakdown happens, with sharp increases in the emissions of X-rays and electron currents. Figure 3 shows images on a phosphor screen installed at the end of superconducting RF cavities during testing. The bright spots in the image indicate X-rays and electron bombardments from field emission. As long as the field-emitted electron currents remain in pre-breakdown condition, the system is stably operable. As the dark current increases, and if the acceleration of field-emitted electrons is efficient, the vacuum pressure could be increased, and eventually affect operational stability.

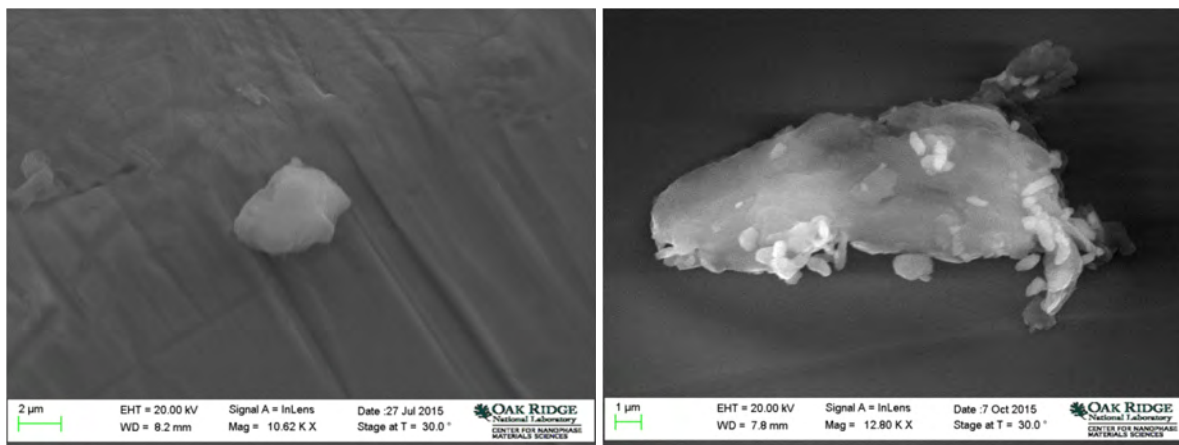


**Fig. 3:** Phosphor screen images of X-rays and electrons from field emission

The theoretical field emission threshold is about 1 GV/m; however, field emission is observed even as low as 10 MV/m in many RF systems, presumably owing to other enhancing factors. Models for the field emission enhancement are:

- protrusion-to-protrusion;
- absorbed gases and other contaminants;
- activation of field emitter at elevated temperature;
- dielectric layer.

It is quite difficult to quantify the characteristics of field emitters, owing to their complex nature, with factors such as size, shape, species, charge status, binding status on the bulk material, temperature, or process history to be taken into account. The emitters are statistically distributed over surfaces and the control of emitters during fabrication, processing, and operation is quite challenging. Figure 4 shows examples of field emitters observed on samples in a study of particulate control during the surface process for superconducting RF cavities.



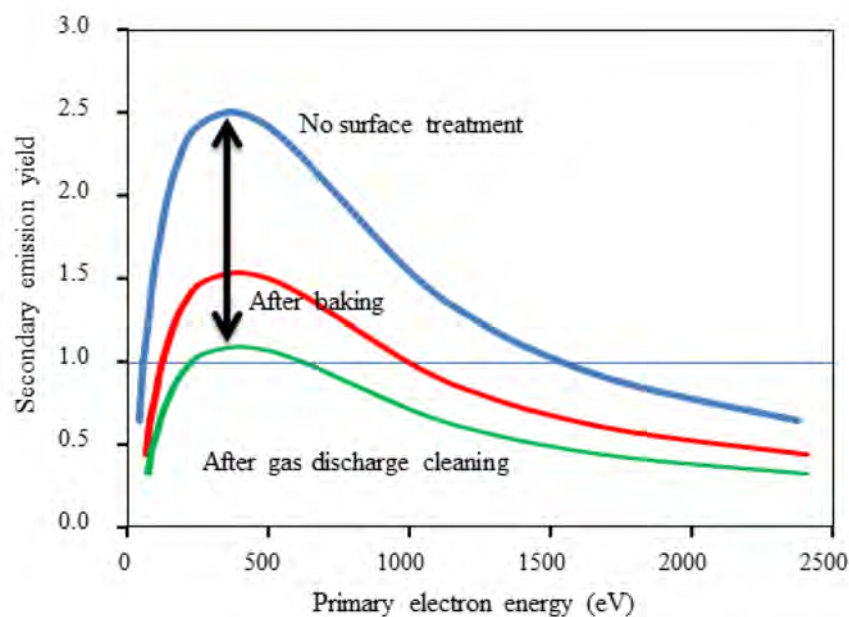
**Fig. 4:** Scanning electron microscopy images of potential field emitters observed on accelerator material samples (EHT, acceleration voltage; Signal A, electron detector mode; WD, working distance; Mag, magnification).

The field emission model for the initiation of vacuum breakdown assumes the existence of field emitters on surfaces. One mechanism for vacuum breakdown by field emission is that field-emitted electrons hit other surfaces, or anodes, causing local temperature increases and resulting in gas desorption. The desorbed gases can be ionized by the primary electrons. The ions can then increase the emission of primary electrons, owing to space charge formation, or generate secondary electrons by hitting surfaces. The process can continue and eventually lead to vacuum breakdown, in a condition similar to Townsend gas discharge. This model of d.c. breakdown is called the anode heating mechanism, and describes a breakdown mechanism with an avalanche of ionization around the anode. The other mechanism of vacuum breakdown by field emission is called the cathode heating mechanism. Field emission provides a pre-breakdown current at the field emitter. This current could heat up the emitter region, by resistive heating. When the emitter reaches a critical condition, it will melt and vaporize. This heating can increase the field emission current density. Under this condition, very dense local plasma is formed and craters, called cathode spots, are produced [1].

### 2.1.2 Multipacting model

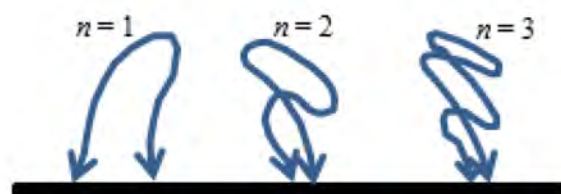
Another major electron-loading effect in RF systems can come from multiple impacts of electrons, called multipacting. Multipacting can occur when both of the following conditions are met: (1) electron motions meet a resonant condition through interaction with an electromagnetic field in vacuum and (2) the secondary electron emission process develops more than one electron per one electron impact.

Secondary emission is a determining parameter for multipacting conditions and is characterized by the secondary emission yield, which is, the statistically measured number of emitted electrons per an incident electron. The secondary emission yield depends not only on material type but also surface treatment history, such as baking, gas discharge cleaning, coating, and impurity contents [2]. Figure 5 illustrates how the secondary emission yield changes depending on surface cleaning or treatment.



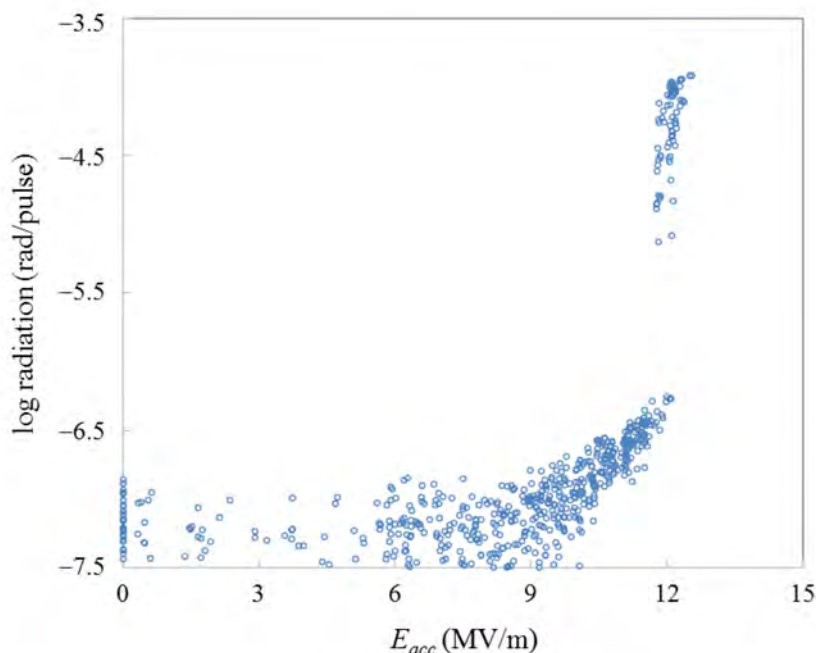
**Fig. 5:** Illustration of secondary emission yield changes

The resonance condition for multipacting is determined by the RF field profiles and field level. Figure 6 shows examples of electron trajectories under interactions with electromagnetic fields. If the time interval between impacts is  $nT$  or  $(n - 1/2)T$  (where  $n = 1, 2, 3, \dots$  and  $T$  is the RF period), depending on multipacting type, for so-called one point and two point multipacting, emitted electrons will have similar starting conditions to the primary electron condition. Under this circumstance, and when the secondary emission yield is larger than unity, an electron avalanche will happen. Because of this condition, the multipacting barrier typically has bands. As indicated from the secondary emission yield, high-energy electrons do not play a role in the multipacting barrier.



**Fig. 6:** Electron trajectories in resonance conditions

Multipacting usually happens at low-electric-field regions in RF structures. Low-electric-field regions can be in RF windows, irises, couplers, beam pipes, or equators of cavities and are prone to multipacting barriers. When an acceleration gap is much shorter than the wavelength, multipacting can happen around the acceleration gap, where the electric field is high. In pulsed machines, multipacting can occur during the RF ramp-up and decay periods of each pulse, if the multipacting barrier exists below an operating field. During the process of multipacting, the non-resonant portion of electrons will escape the multipacting trajectory and can be accelerated in accelerating structures. These electrons heat the surfaces and can generate X-rays, in the same way as electrons from field emission, and give rise to thermal instability, especially in superconducting RF cavities. Figure 7 shows an example of X-rays from multipacting. From this plot it is not clear whether the X-rays result from field emission or multipacting. By monitoring changes of radiation waveform in pulsed operation as a function of field level, multipacting can be verified relatively easily.



**Fig. 7:** X-rays by non-resonant electrons from multipacting

In many cases, the multipacting barrier can be processed out by careful RF conditioning. However, if severe multipacting develops, the multipacting region will heat up, and trapped or condensed gases will be released from the surface. This condition can eventually cause vacuum breakdown and a material defect can occur. This kind of breakdown has been observed in almost every RF systems in vacuum, such as electron tubes, couplers, waveguides, RF windows, and accelerating structures. To avoid a severe multipacting barrier, careful analysis should be done from the design stage. As mentioned earlier, surface processing plays an important role; one must keep the surface clean and avoid contaminants, perform careful RF conditioning, bake surfaces, apply discharge cleaning where possible or surface coatings with low-secondary-emission-yield material, and apply d.c. biasing to shift multipacting barriers.

### 2.1.3 Particle exchange model

This model assumes that a charged particle comes out of a surface, which is always possible statistically, then hits other surfaces of the electrode and liberates particles. Oppositely charged particles are accelerated back to the initial surface of the electrode. If this process becomes cumulative, it could trigger vacuum breakdown. This mechanism is based on avalanching of mutual secondary emission of ions and electrons, and also photons and absorbed gases on surfaces.

### 2.1.4 Clump model

This model assumes that a loosely bound particle cluster is on the surface. In a high voltage, the cluster is then charged and accelerated and impacts other surfaces, releasing gases and vapours, and triggers vacuum breakdown.

### 2.1.5 Kilpatrick criterion

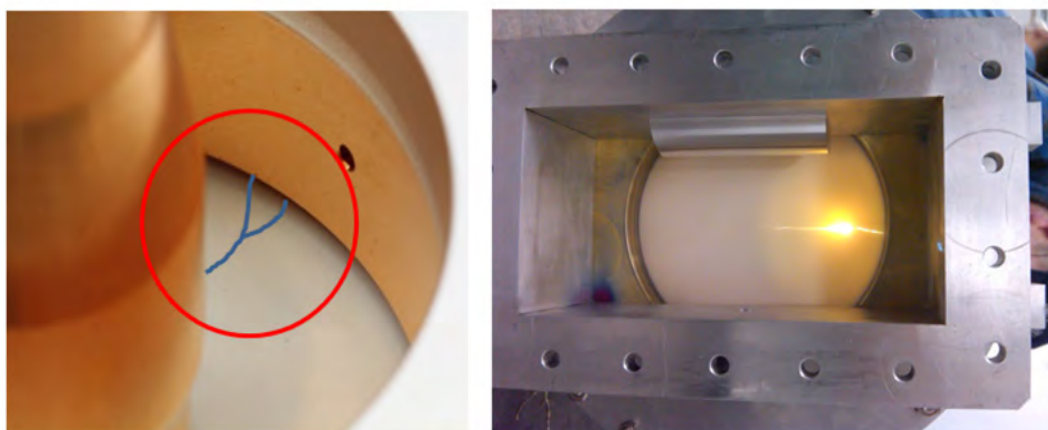
In 1950s, W. Kilpatrick developed the criterion on RF breakdown from experimental data [3]. The Kilpatrick empirical formula relates the breakdown electric field with RF frequency;

$$f = 1.64 E_s^2 \exp\left(-\frac{8.5}{E_s}\right),$$

where  $f$  is the RF frequency in MHz and  $E_s$  is the breakdown electric field at the surface in MV/m. The experimental data for this criterion were generated when clean vacuum systems were not available. Modern accelerating structures are designed for and run at higher fields than this criterion; however, it is still in use as a kind of figure of merit. It is interesting to note that some much later analyses using anode heat mechanism [4] and experimental data [5] show similar dependencies on RF frequency. However, there is no good explanation for this similarity. Conversely, it has been reported [6] that no increase of breakdown field is observed at frequencies higher than 10 GHz.

## 2.2 Dielectric breakdown

Dielectric materials are widely used in various components of RF systems to separate electrodes or form a boundary between vacuum and air. Failures of dielectric materials in RF systems are among the most frequent and severest events. Figure 8 shows examples of dielectric material failures. Dielectric materials have intrinsic breakdown limits, and actual operating conditions are typically set well below these intrinsic limits. However, electrical breakdown also occurs well below the intrinsic limit. Possible enhancing factors for dielectric breakdown would be charged-particle bombardment or non-uniformity of the dielectric material. These enhancing factors provide initiation of breakdown, generate field concentration, and can lead to non-uniform charge build-up. Sources of non-uniformity are voids, impurities, inhomogeneity of the material, non-ideal boundary junctions, absorbed gases, and wrinkles. A selection of mechanisms practically important for RF systems is briefly reviewed next.



**Fig. 8:** Left: 550 kW, 805 MHz coaxial power coupler RF window crack. Right: 5 MW, 805 MHz RF window failure.

- Chemical deterioration: Dielectric materials facing air or other gases could undergo oxidation or hydrolysis processes that give rise to surface cracks or changes in material properties. Operation at an elevated temperature would accelerate these unwanted chemical processes. Even dielectric materials under vacuum could have chemical deterioration through particle bombardments. For example, an electronic activity such as multipacting on an alumina surface results in oxygen defects and leads to severe breakdown [7].
- Treeing: When a dielectric fails, sometimes a visible conducting path or pinhole across the dielectric is observed. The leakage current passes through the conducting path, leading to sparks. The spread of spark channels is called ‘treeing’, since it looks like the branches of a tree. Treeing is a pre-breakdown phenomenon and a damaging process from partial discharges. At the edge of a treeing there is a high field concentration. When a system runs for a long time under electrical stress, the treeing progresses through the dielectric, finally leading to failure. Initiation of this process comes from inhomogeneity of the surface, contaminants on the surface, or charge build-up from other sources.



- Breakdown due to voids: Dielectric materials have voids within the medium or at junction boundaries. Voids are usually under vacuum or filled with gases whose electric permittivity is smaller than that of the dielectric. The electric field strength is, therefore, higher in the voids. When the field strength in the voids is higher than the voltage hold-off capability, local breakdown occurs. In this case, the effect of local discharges and their developments in the voids is similar to treeing.
- Surface flashover: Surface flashover typically occurs at a much lower electric strength than volume breakdown. Experimental measurements show that the surface of a dielectric in a vacuum becomes positively charged under a high-voltage d.c. by electron multiplication through a secondary emission process [8] or an electron cascade in a thin surface layer [9]. Once electron multiplication is developed, travelling electrons form a pre-breakdown current and release gases from the surface by electron-stimulated desorption. The desorbed gases are ionized by electron collisions and breakdown occurs along the surface. It is broadly agreed that the initiation of flashover starts at a so-called triple junction point (Fig. 9). In practice, voids or small gaps exist at the joint, where the electric field strength is much higher than the majority of the bulk area. Similar phenomena could be initiated with charged-particle bombardments that are produced somewhere else.
- Mechanical failure due to thermal shock: An arcing event generates a thermal gradient in a short period of time that results in mechanical stress due to a localized thermal expansion. If the event is large enough, the mechanical stress can reach a rupture condition.

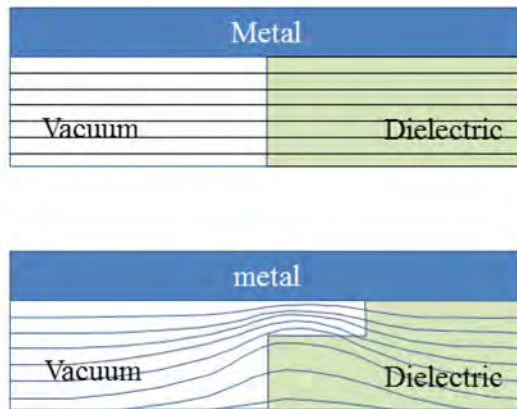


Fig. 9: Electric field profile at the triple junction (top, ideal case; bottom, practical case)

### 2.3 Vacuum

As discussed earlier, in a vacuum condition, released gases from the surface play a particularly important role in breakdown. Mechanisms for gas release can be classified as follows.

- Desorption is the process of gas release from the surface and is the final state of all outgassing mechanisms. The binding energy of adsorbed gases is much less than an electron volt. Thus, the desorption rate increases at an elevated temperature and a much higher desorption rate can be achieved with energetic charged-particle and photon bombardments.
- Vaporization is the phase transition of a material into a gas. Materials with higher vapour pressure can evaporate in a vacuum or at an elevated temperature.
- Dissolved gases in bulk materials, such as metals and ceramics, move to the vacuum surface. This process is called diffusion. Hydrogen has a high mobility in the bulk material.
- Adsorbed gases from the outside surface diffuse through the bulk material and are then desorbed from the surface in a vacuum. This process is called permeation.

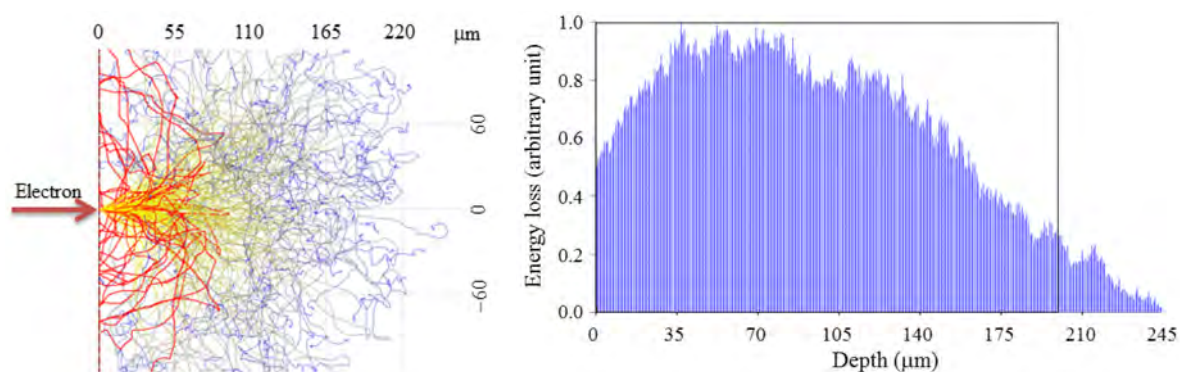
Reducing outgassing is the critical process for stable operation of electron tubes, accelerating structures, power couplers, and high-power RF windows. Polishing reduces the effective surface area and capacity of adsorption. Baking and vacuum firing are the most widely used processes for reducing outgassing [10, 11]. Baking in vacuum at moderate temperatures (150°C–300°C) reduces the outgassing rate, especially for water. This moderate baking in vacuum is not sufficient to remove hydrogen from the bulk material. High temperature vacuum firing, however, drastically reduces the hydrogen outgassing rate. During the baking or vacuum firing, H<sub>2</sub>, CO, CO<sub>2</sub>, and CH<sub>4</sub> gases are usually released along with water. To further clean the surface, gas discharge cleaning or electron beam showering is used.

### 3 Beam related issues

One of the most important purposes of hardware protection is to reduce equipment activation levels to as low as reasonably achievable, to allow hands-on maintenance and protect machine equipment from damage due to uncontrolled beam strikes. Many groups have studied allowable uncontrolled beam loss and it is generally agreed that an average beam loss of 1 W/m is a reasonable limit for hands-on maintenance [12, 13]. As demands for beam power and beam power density increase beyond previously experienced levels, the operation of high-intensity machines below this limit will be very challenging. Currently, there is a great deal of effort with computer simulation, wide dynamic range diagnostics, and fast beam abort schemes in response to upset conditions, to reduce uncontrolled beam losses. Since other sessions in this course cover details of beam dynamics and equipment activation issues, in the following other possible sources of beam-induced damage to accelerating structures will be introduced.

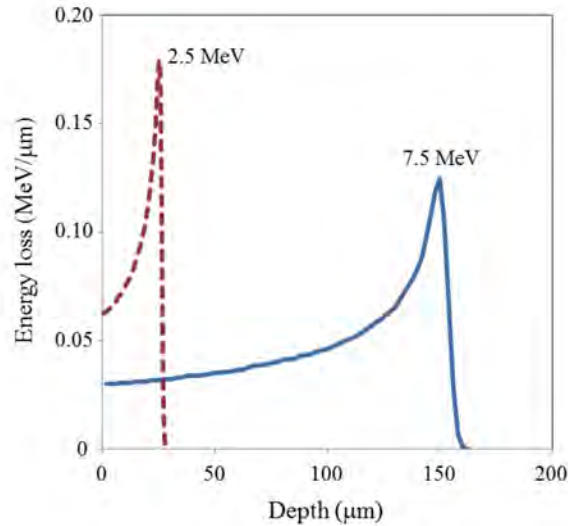
#### 3.1 Mis-steered beam

When a beam is steered by mistake or as a result of a magnet fault, it can strike the surface of the accelerating structure. The stopping power, which is the loss of particle energy per unit path length, for electrons is rather uniform, while that for ions has a peak before stopping, known as the Bragg peak. Figure 10 shows trajectories and the stopping power of electrons in a simulation of the Spallation Neutron Source (SNS) linac beam dump window. The electron energy in this example is ~1 MeV, stripped from the H<sup>-</sup> ion. Figure 11 shows an example of Bragg curves for 2.5 and 7.5 MeV protons.



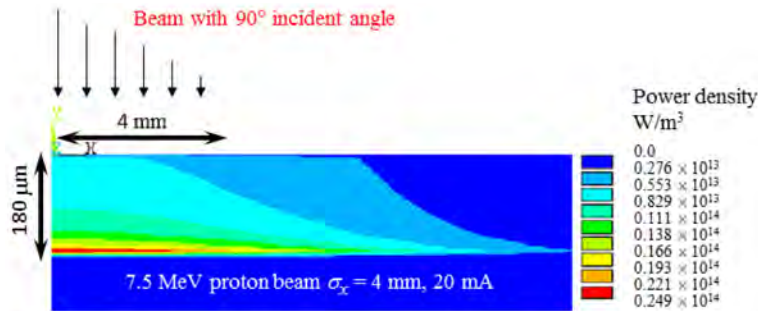
**Fig. 10:** Example of (left) electron trajectories and (right) energy loss per unit length in nickel alloy

When particles lose their energy while passing through matter, a good portion of energy will be converted into kinetic energy; hence, the local temperature will increase. It is important to understand the dependencies of beam parameters on damage to accelerator structures. Thus, the machine protection system for aborting the beam should be designed accordingly. In the following analysis, a guide to determine the beam abort speed is discussed, using the SNS beam parameters.

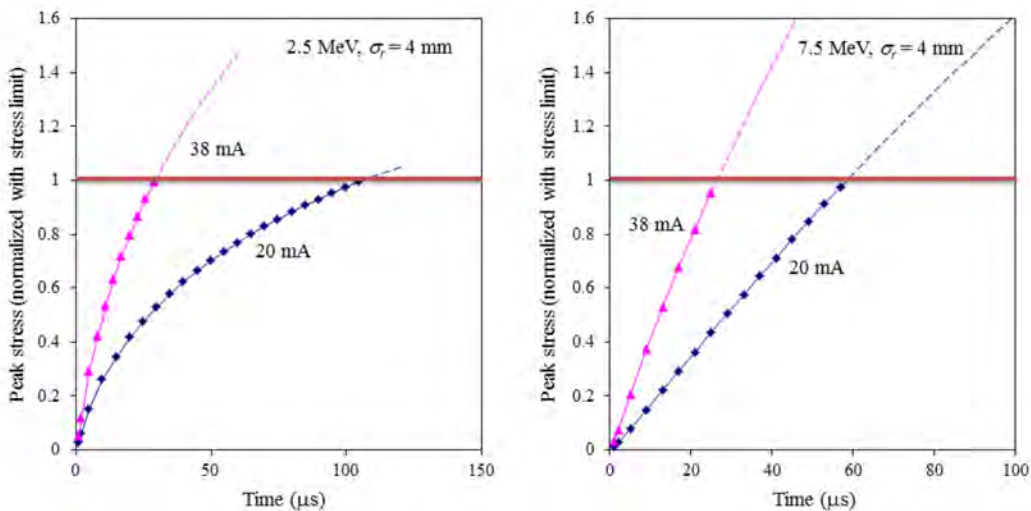


**Fig. 11:** Example of energy loss per unit length of proton in copper

Figure 12 shows an example of the thermal load from the round Gaussian beam with  $\sigma = 4$  mm at 20 mA, with 7.5 MeV protons. Since the heating effect by the beam is localized, a very steep thermal gradient will occur and result in thermal stress. If the stress level exceeds a mechanical stress limit, such as the yield point or tensile strength, mechanical defects could occur. Usually, this mechanical stress threshold comes much earlier than material melting. Figure 13 shows the peak stress development in copper as a function of beam strike duration. In this example, the peak temperature is around 200°C when the peak von Mises stress from the thermal gradient reaches the tensile limit.

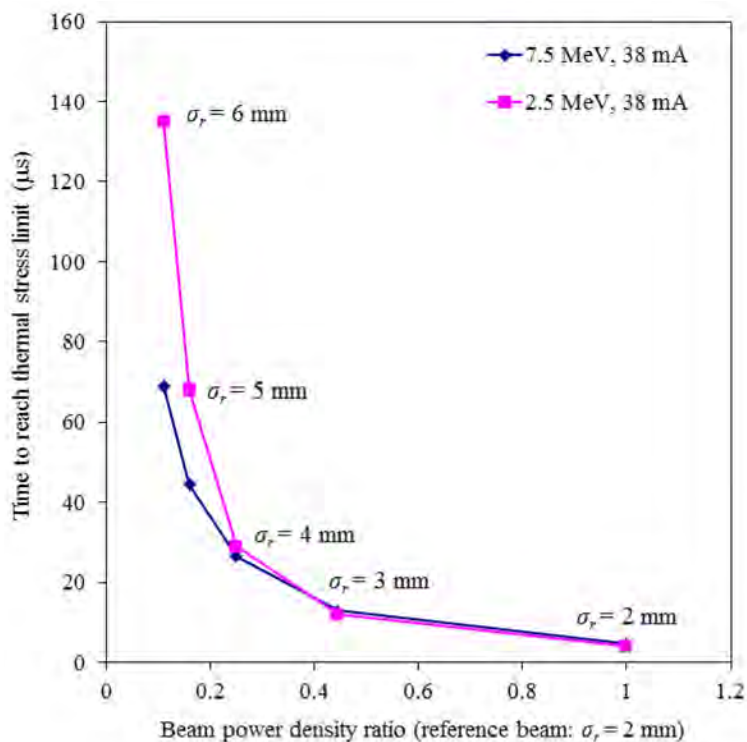


**Fig. 12:** Thermal load profile in copper from beam strikes

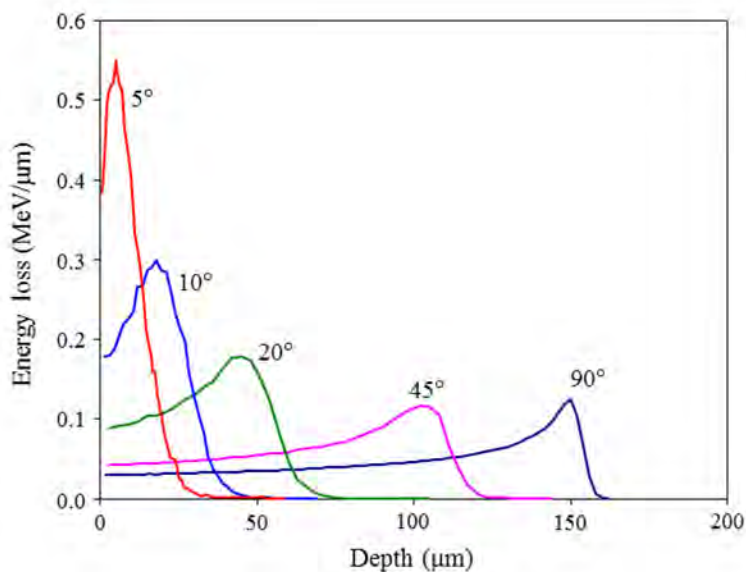


**Fig. 13:** Stress development from different beam conditions at 90° incident angle

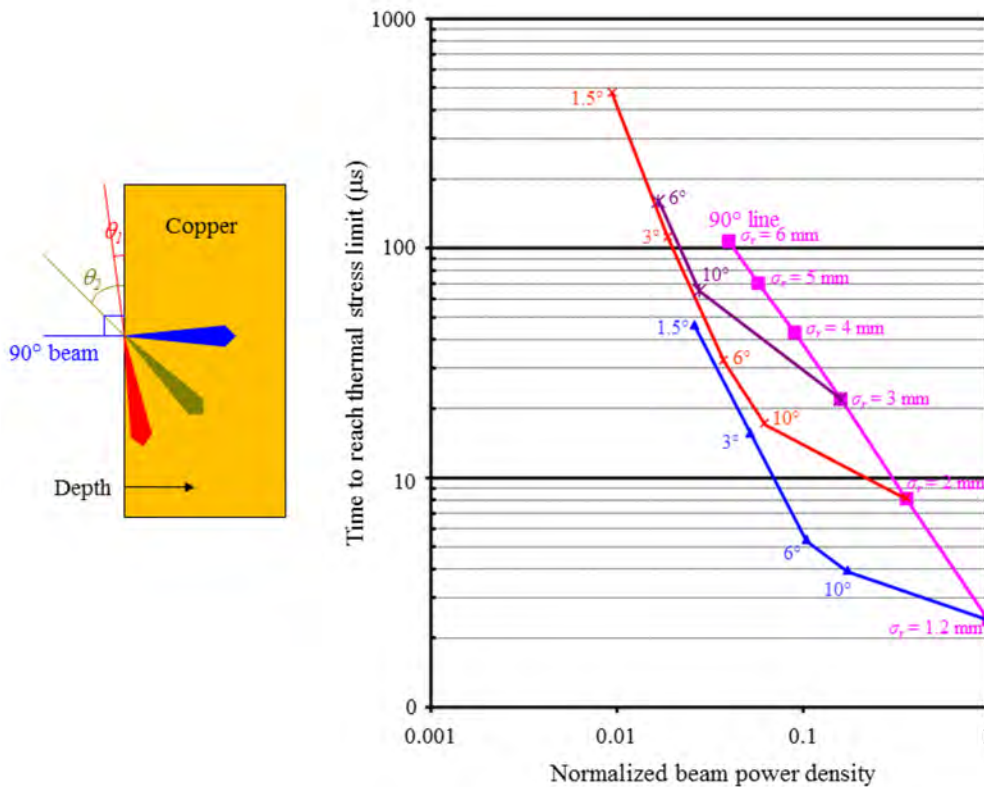
Using the same model, one can explore stress limits under various beam conditions, as shown in Fig. 14. In practice, when a mis-steered beam condition occurs, the beam is most likely to strike at a grazing angle on the accelerating structure. Figure 15 shows an example of the energy loss per unit length at various incident angles. The energy loss per unit length at grazing angles increases as the incident angle reduces but the effective beam size also increases rapidly. The depth is measured along the direction normal to the surface. Beam dynamics simulations predict that the worst probable case for the SNS accelerator not more than  $3^\circ$ . Figure 16 shows an example of the allowable duration for a mis-steered beam for grazing angle incidents compared with the  $90^\circ$ -incident cases.



**Fig. 14:** Time to reach the mechanical stress limits for various beam condition at  $90^\circ$  incidence. In this example, a beam with  $\sigma_r = 2$  mm is used as a reference nominal beam size.



**Fig. 15:** Energy loss per unit length for various beam incident angle for protons with 7.5 MeV



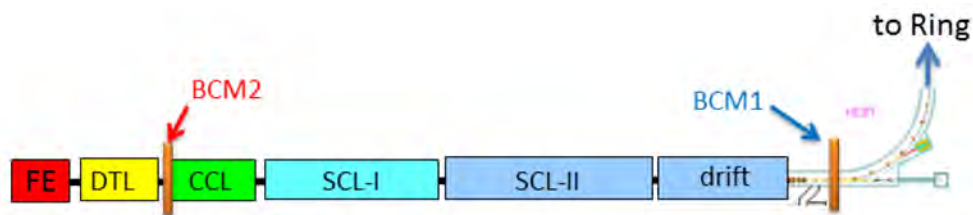
**Fig. 16:** Time to reach mechanical stress limits from 7.5 MeV, 26 mA beam at various beam sizes and incident angles.

### 3.2 Errant beam

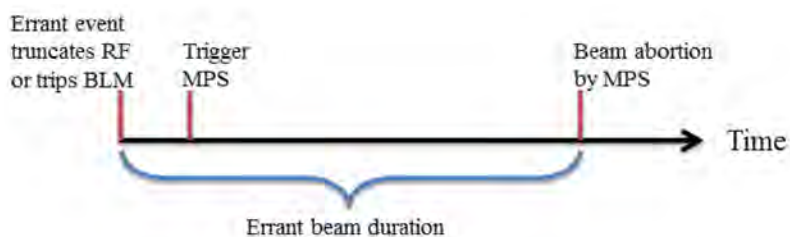
The term ‘errant beam’ hereafter refers to an off-energy beam generated anywhere in the accelerator, and transported downstream in a fault condition; this is different from the uncontrolled beam loss in a normal operating condition. Since the errant beam is an off-energy beam, it is mostly lost while transported through the linac, resulting in beam trips in the linac. During an errant beam condition at the SNS linac, the beam loss region ranges from several to ten cryomodule lengths or longer. Two mechanisms are identified as sources of errant beams at the SNS: (1) RF truncations in upstream RF structures and (2) abnormal beam generation in the front end, which includes the ion source, the low energy beam transport system, and the RF quadrupole.

The SNS linac (Fig. 17) is composed of the front-end system, 6 drift tube linacs, 4 coupled cavity linacs, and a superconducting linac that houses 81 superconducting RF cavities in 23 cryomodules. Each RF structure has its own fast interlocks, which truncate an RF pulse at an upset condition, such as a vacuum burst, discharge, arc, or any RF signal beyond a predefined threshold. These interlocks trigger the machine protection system, which aborts the beam. There is a delay time for the machine protection system to abort the beam, as illustrated in Fig. 18. When an interlock of any RF structure truncates its RF pulse at an upset condition, the beam from the start of the RF truncation up to beam abortion by the machine protection system will be an off-energy beam. The second type of errant beam arises from the front-end system, such as a lower-than-nominal current beam pulse, a partial beam pulse or another abnormal beam arising from high-voltage arcing, unstable plasma, or mis-triggered RF for plasma generation. The beam under these conditions also becomes an off-energy beam while it is accelerated, since the RF adaptive feed forward is only for the nominal beam pulse. In this case, beam loss monitors are the first indicators of the event. The machine protection system delay is the same as for the RF truncation case. The original requirement of the machine protection system delay at the SNS is 30  $\mu$ s.

The SNS machine protection system current is providing ‘beam abortion’ in 10–25  $\mu\text{s}$  for these errant conditions.

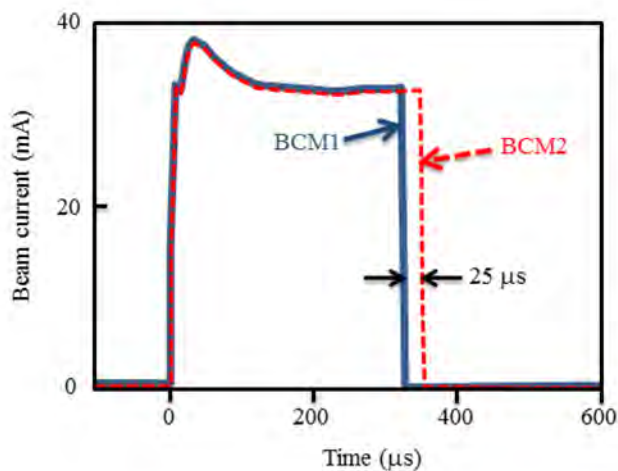


**Fig. 17:** Layout of SNS LINAC and positions of beam current monitors for errant beam monitoring. BCM, beam charge monitor; CCL, coupled cavity linac; DTL, drift tube linacs; FE, front end; HEBT, high-energy beam transport; SCL, superconducting linac.

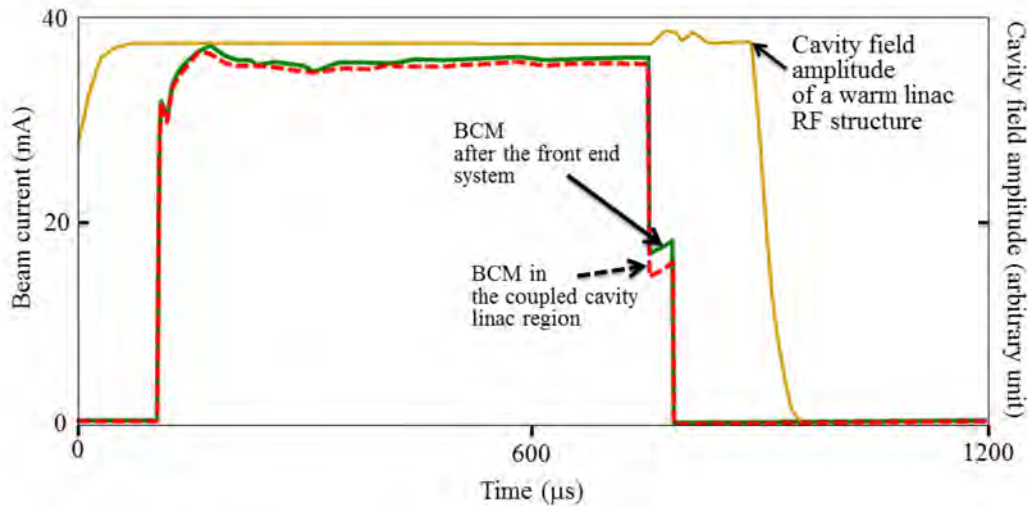


**Fig. 18:** Machine protection system delay time until abortion of beam. BLM, beam loss monitors; MPS, machine protection system.

Figure 19 shows an example of signals from beam current monitors 1 and 2 at an errant beam condition caused by RF truncation in one of the drift tube linac structures. In this example, the whole beam was lost for about 25  $\mu\text{s}$  in the superconducting linac. An example of errant beam events from the front end is shown in Fig. 20. The beam current waveforms after the front end and before the superconducting linac are shown. The other waveform is the cavity field waveform in one warm linac RF structure. In this example, the beam current dropped in the middle of a pulse from the front end. Notice that cavity fields of all linac RF structures do not provide a flat field, owing to the nature of the adaptive feed forward. That portion of the beam became an off-energy beam and was lost in the linac. As mentioned earlier, beam loss monitors trigger the machine protection system and the machine protection system shuts the beam off.



**Fig. 19:** Examples of beam current monitor (BCM) signals in the event of RF truncation in a warm linac structure



**Fig. 20:** Example of errant beam generated in the front-end system. Beam current monitor signals after the front end and in the coupled cavity linac region, and cavity field amplitude of a warm linac RF structure.

The severity of the events depends on the locations where errant beams are lost and the frequency of events that are associated with the conditioning status of errant beam sources (warm linac and front end). Usually, in the first few weeks after a long maintenance shut-down, errant beam events are more frequent. The average number of errant beam events in the past at SNS was about 30–40 times a day out of five million pulses a day. Most errant beam events result in beam loss monitor trips or sometimes small vacuum excursions. However, when similar events recur, there is a chance that an errant beam might evaporate gases and a following RF interaction could create an environment for severe discharge or arcing. The energy of one mini pulse (about 1  $\mu\text{s}$  beam) at the SNS is about 24 J. Since the beam loading in the SNS superconducting RF cavities is high, the available RF power is large enough to create a dangerous discharge. Unwanted consequences from errant beam events might be additional gas or particulate contamination, surface damage that leads to RF performance degradation, and component damage, such as ceramic windows and ceramic feed-through failures. The window crack shown in the left side of Fig. 8 was caused by arcing from errant beam events. Various suggestions have been put forward to minimize the number of errant beam events, such as careful conditioning for the front end and warm linac, routine maintenance of vacuum systems, and continuous adjustment of operating parameters for warm linac structures. Presently, the frequency of errant beam events is about 10–15 times per day. In addition, a new dedicated protection system to abort the beam within 6  $\mu\text{s}$  has been developed and will be used for operation in the near future [14].

#### 4 Summary

Breakdown conditions as indicated in this section depend on various factors, which affect the performance of a system. For example, apparently random failure events and large scattering performances, which are commonly encountered even in the same types of system, result from the irregularity of physical and chemical properties. Therefore, it is important to take into account all possible factors in an organized way. Some factors leading to breakdown are definitely determined during design and fabrication, for example, material surface finish, process history, baking, shape, electromagnetic field configuration, pulse length, and RF frequency. Other factors are due to changes during operation, such as residual gas species, partial pressure, contaminants, radiation, circuit characteristic, temperature, history of operation, and beam conditions. Because breakdown is affected by combinations of these factors, there is no unique mechanism for a specific breakdown event.

Surface properties strongly depend on the previous operation history. It is usually found that breakdown conditions in a vacuum can be substantially improved by allowing repeated mild

breakdowns or the passage of appreciable pre-breakdown current. This behaviour is called conditioning, and is a necessity in most high-power RF systems. This conditioning process should be developed in a controlled way in conformity with experimentally observed phenomena. Conditioning changes surface conditions in similar ways to breakdown mechanisms; sputtering of surface protrusions, release and relocation of trapped or condensed gases by electron bombardments, etc.

Most high-power RF systems operate with considerable energy or power, to create discharge. In many cases there is no clear boundary between conditioning and irreversible damage processes since threatening breakdown can develop quickly. If the damage is large enough, it will reduce the breakdown field, leading to high pre-breakdown current, loss of vacuum, mechanical damage, and, eventually, catastrophic failure. The amplitude of each discharge during the conditioning process can be limited by proper interlocks, such as vacuums, arc detectors, X-ray detectors, or forward RF power thresholds.

For large-scale accelerators, RF equipment protection now involves complex systems that consist of slow interlocks, fast interlocks for beam abortion, sequences to verify equipment status ready for beam operation, and logic circuits to resume normal operation automatically. The RF protection system is fully integrated with other machine protection systems, global timing systems, diagnostics systems, and control systems in modern accelerators. Fault condition input to the main machine protection system to abort beam operation should be classified based on a good understanding of the causes and consequences of a potential failure mechanism.

## References

- [1] J.E. Daalder, *J. Phys. D* **12**(10) (1979) 1769. <http://dx.doi.org/10.1088/0022-3727/12/10/019>
- [2] N. Hilleret, The secondary electron yield of technical materials and its variation with surface treatment, Proc. European Particle Accelerator Conf., Vienna, Austria, 2000, p. 217.
- [3] W.D. Kilpatrick, *Rev. Sci. Instrum.* **28**(10) (1957) 824. <http://dx.doi.org/10.1063/1.1715731>
- [4] M.D. Karetnikov, *Particle Accel.* **57** (1997) 189.
- [5] J.W. Wang and G.A. Loew, RF breakdown studies in copper electron linac structures, SLAC PUB-4866 (1989).
- [6] I. Wilson, High-gradient testing breakdown and surface damage, 9th Int. Workshop on Linear Collider (LC02), SLAC, CA, 2002.
- [7] Y. Saito, Breakdown phenomena in vacuum, Proc. Int. Linear Accelerator Conf., Ottawa, Canada, 2012, p. 575.
- [8] R.A. Anderson and J.P. Brainard, *J. App. Phys.* **51**(3) (1980) 1414. <http://dx.doi.org/10.1063/1.327839>
- [9] N.C. Jaitly and T.S. Sudarshan, *J. App. Phys.* **64**(7) (1988) 3411. <http://dx.doi.org/10.1063/1.341496>
- [10] T.D. Kirkendall *et al.*, *J. Vac. Sci. Tech.* **3**(4) (1966) 217. <http://dx.doi.org/10.1116/1.1492477>
- [11] E. Hoyt, Effect of surface treatment and baking on the outgassing characteristics of 304 stainless steel pipe, SLAC-TN-64-5 (1964).
- [12] N.V. Mokhov and W. Chou, Proc. 7th ICFA Mini-workshop on High Intensity High Brightness Hadron Beams, Wisconsin, USA (1999)
- [13] A.H. Sullivan, *A Guide to Radiation and Radioactivity Levels Near High Energy Particle Accelerators* (Nuclear Technology Publishing, Ashford, 1992).
- [14] W. Bloklund and C. Peters, A new differential and errant beam current monitor for the SNS accelerator, Proc. 2nd Int. Beam Instrumentation Conf., Oxford, UK, 2013, p. 921.



# Machine Protection and Operation for LHC

*J. Wenninger*

CERN, Geneva, Switzerland

## Abstract

Since 2010 the Large Hadron Collider (LHC) is the accelerator with the highest stored energy per beam, with a record of 140 MJ at a beam energy of 4 TeV, almost a factor of 50 higher than other accelerators. With such a high stored energy, machine protection aspects set the boundary conditions for operation during all phases of the machine cycle. Only the low-intensity commissioning beams can be considered as relatively safe. This document discusses the interplay of machine operation and machine protection at the LHC, from commissioning to regular operation.

## Keywords

Operation; machine protection; beam loss; LHC.

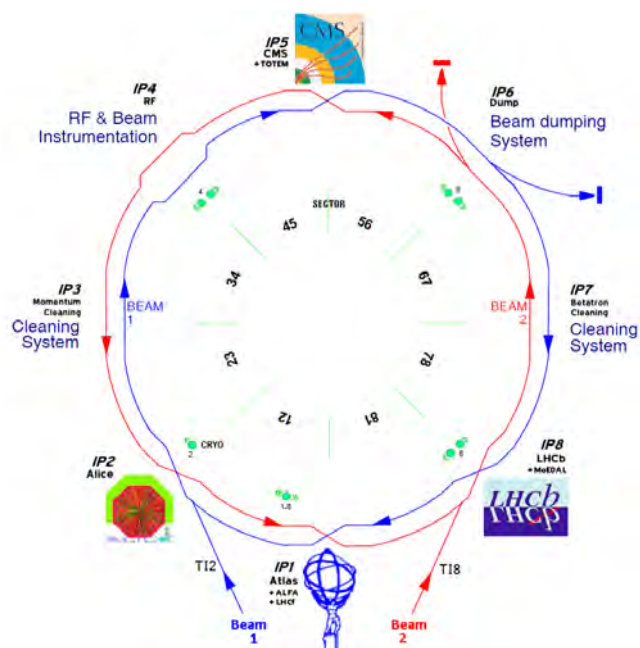
## 1 Introduction

The Large Hadron Collider (LHC) is the last in the series of hadron colliders after the ISR (Intersecting Storage Ring), SPS (Super Proton Synchrotron), Tevatron, HERA and RHIC (Relativistic Heavy Ion Collider). The machine elements are installed on average 100 m below the surface in the 26.7 km long accelerator tunnel that housed the Large Electron Positron collider (LEP) between 1989 and 2000 [1, 2]. The ring consists of eight arcs and of eight long straight sections (LSSs). The large particle physics experiments ALICE, ATLAS, CMS and LHCb are installed at interaction points (IPs) in the middle of four LSSs, while the other LSSs house the collimation (or beam cleaning) system, the radio-frequency (RF) system, the beam instrumentation and the beam dumping system. The layout of the LHC is shown in Fig. 1.

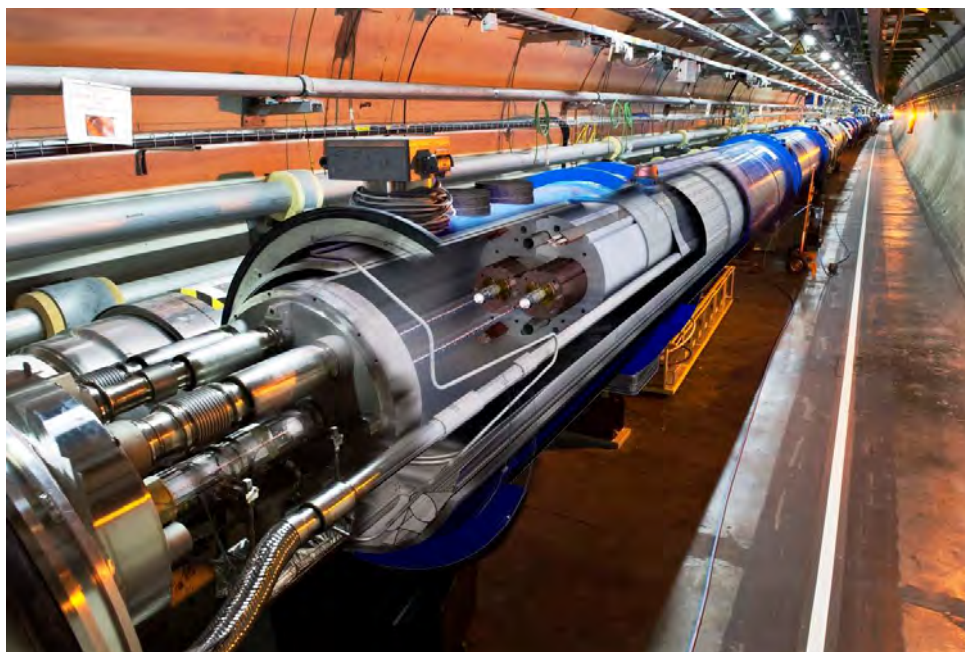
A dipole field of 8.3 T is required to bend hadrons with a momentum of 7 TeV/c per unit charge in the tunnel; this is 60% higher than in previous accelerators. Such a magnetic field strength is achieved with superconducting dipole magnets made of NbTi. With a 2-in-1 magnet design the two rings fit inside the 3.8 m diameter LEP tunnel; see Fig. 2. Both rings are accommodated in a single cryostat and the distance between the two vacuum chambers is only 19 cm. The two proton or ion beams circulate in opposite directions in two independent vacuum chambers. Each dipole magnet is 14.3 m long; the associated cryostat is 15 m long. Besides the 1232 dipole magnets that constitute around 85% of each arc, the magnet lattice also includes quadrupole magnets that focus the beam, sextupole magnets to correct chromatic effects and octupoles to stabilize the beam. A total of 8000 superconducting magnets are used to control the two beams.

Eight continuous cryostats with a length of 2.7 km each cool the superconducting magnets to their operating temperature of 1.9 K. After cool down the LHC cryostats contain 130 tons of liquid helium and around 37 000 tons of material are cooled to that temperature. The magnets and the cooling system based on superfluid helium form by far the longest refrigerators on Earth. The cryogenic system has to be extremely reliable; in 2012 the system achieved an overall uptime of 95%.

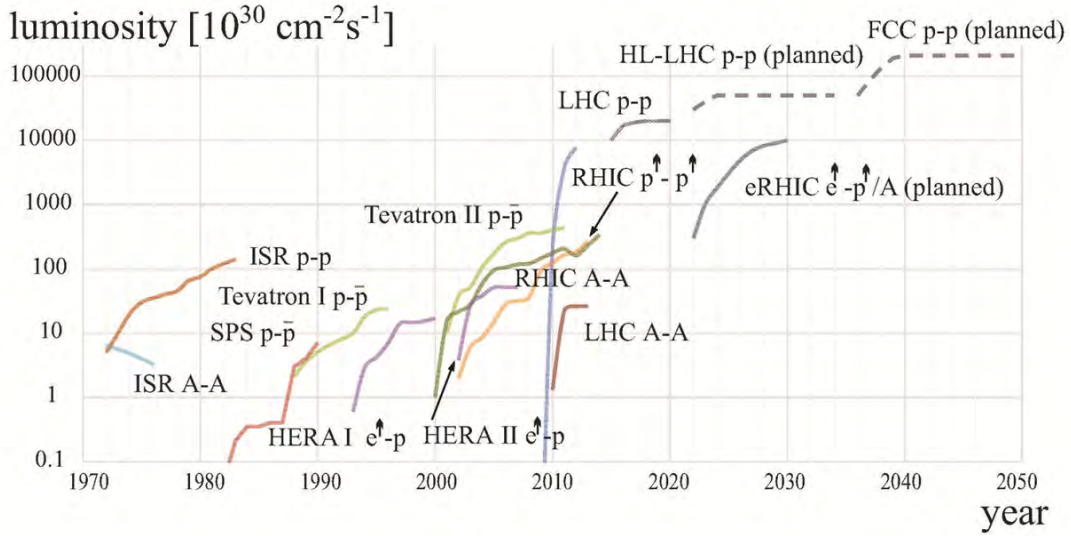
Around 1600 power converters provide current to the magnets; for the main circuits the peak currents reach 13 kA. The magnetic energy stored in each arc cryostat is around 1 GJ. This energy has to be safely extracted in case one of the magnets quenches, i.e. performs a transition from the superconducting to the normal-conducting state [3]. Large dump resistors that are capable of absorbing the energy are automatically switched into the main circuits in case of a quench.



**Fig. 1:** Layout of the LHC with the eight interaction points labelled IP1 to IP8. The experiments ATLAS, ALICE, CMS and LHCb are installed in IP1, IP2, IP5 and IP8, respectively. Beam 1 is injected close to IP2 and circulates clockwise, Beam 2 is injected close to IP8 and circulates counter-clockwise. The two beams exchange position between outside and inside of the ring at every experiment to ensure that the path length is the same for both beams. The two beam dumps are located around IP6.



**Fig. 2:** View of the LHC tunnel with an artistic cut through a LHC dipole magnet, highlighting the twin magnet coils as well as the two vacuum chambers.



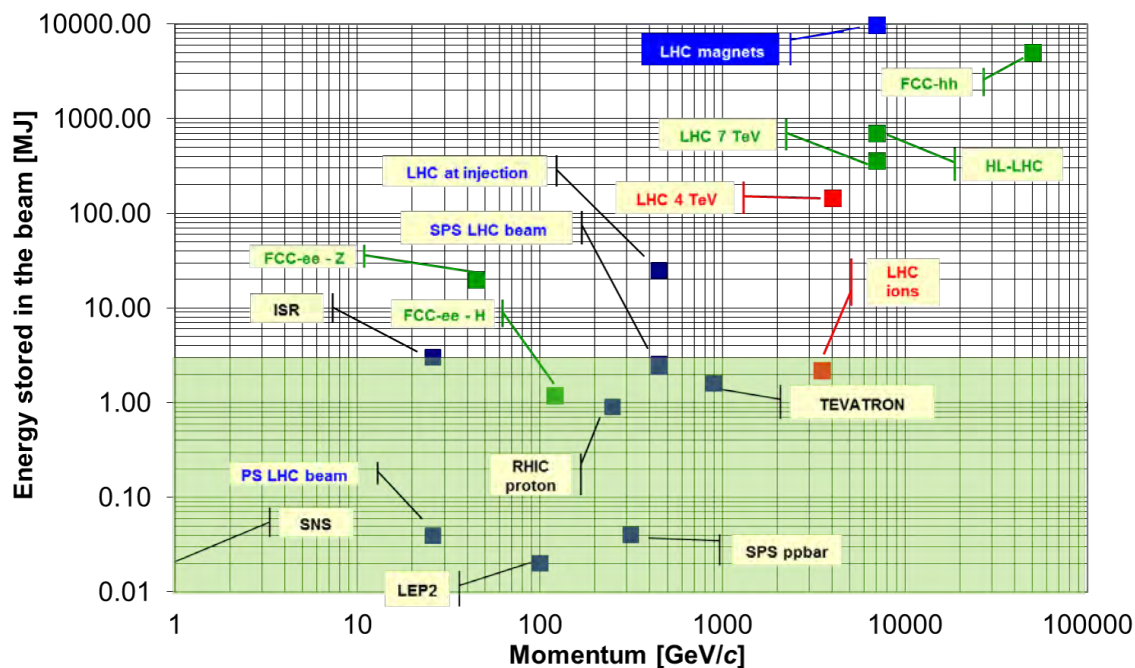
**Fig. 3:** Evolution of the peak luminosity of hadron colliders since 1970 (image courtesy of W. Fischer, BNL). Future projects like the high-luminosity LHC (HL-LHC) [5, 6], eRHIC and a possible 100 km circumference collider (FCC) [7] are also indicated.

The performance of a collider is characterized by its luminosity  $\mathcal{L}$ . The event rate of a physical process with cross-section  $\sigma$  (with unit of an area [ $\text{m}^2$ ]) is given by  $\sigma \times \mathcal{L}$ . The luminosity may be expressed as

$$\mathcal{L} = \frac{k f N^2}{4\pi\sigma_x\sigma_y}, \quad (1)$$

where  $f$  is the revolution frequency (11.24 kHz for the LHC),  $k$  is the number of bunches,  $N$  is the number of particles per bunch and  $\sigma_x$  and  $\sigma_y$  are the horizontal and vertical beam sizes at the collision point. The highest luminosity is achieved with the smallest possible beam cross-section, a large number of bunches and a high bunch population. Up to 2808 bunches can be filled and accelerated in each LHC beam; the minimum distance between bunches is 25 ns or 7.5 m. Each proton bunch consists of up to  $1.7 \times 10^{11}$  protons. The bunches have a typical length of 7 to 10 cm. At the interaction point the transverse rms sizes of the colliding beams are around  $20 \mu\text{m}$  [4]. Figure 3 presents the evolution of hadron collider luminosity over time. The LHC pushes the energy frontier by a factor of seven and the luminosity frontier by a factor of 25; the luminosity gain is mainly obtained with very high beam intensities.

The LHC dipole magnets were produced by three industrial firms and the last dipole magnet was delivered to CERN in November 2006. Each magnet was trained on CERN test benches to a magnetic field of 8.7 T, approximately 5% above the design target. A few training quenches were typically required to reach the nominal field of 8.3 T. Training quenches are due to the release of extremely small amounts of frictional energy (10–100 nJ) due to coil movements when the magnetic field is increased. In June 2007 the first arc (sector) of the LHC was cooled down and ready for commissioning and in April 2008 the last dipole magnet was lowered into the LHC tunnel. One of the essential components of the commissioning phase was the testing of the LHC superconducting magnets and the associated powering and protection equipment. In early 2008 it became apparent that the LHC dipole magnets had to be re-trained to their nominal field; the first magnet quenches appeared at fields corresponding to beam energies of around 5.5 TeV. A training campaign on one arc revealed that the number of required re-training quenches increased rapidly with the magnetic field. The estimated number of quenches required to reach 6.5 TeV is around 140, confirmed during the re-commissioning in 2015, while for 7 TeV the expected number



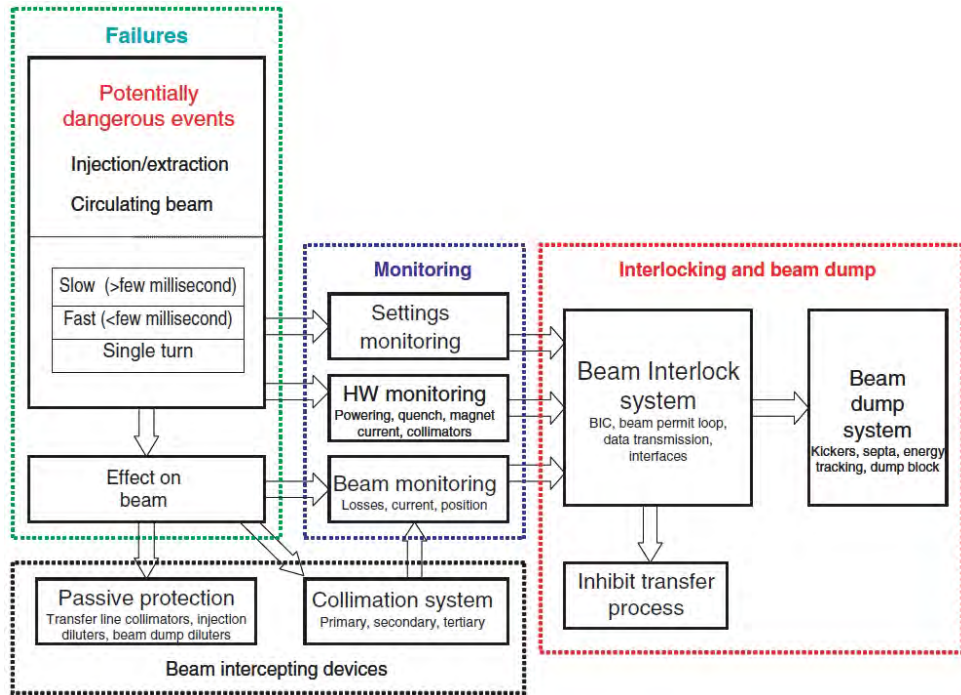
**Fig. 4:** Stored energy of the beams as a function of the momentum for various accelerators. The area shaded in light green corresponds to the state of the art before the startup of the LHC. All accelerators operate with hadrons except the lepton colliders LEP2 and FCC-ee (the CERN 100 km collider study).

can be as high as 1000. Since such a training campaign would have required a long time, it was decided to lower the energy for the commissioning and first operation phase to 5 TeV [8].

On 10 September 2008 beams were circulating for the first time in both LHC rings. The startup was however brought to an abrupt halt on 19 September 2008 when a defective high-current soldering between two magnets triggered an electric arc that released around 600 MJ of stored magnetic energy [9]. The accelerator was damaged over a distance of around 700 m; 53 magnets had to be replaced or repaired. The beam vacuum was polluted with dust and soot over 2 km. The repair, improvement and re-commissioning of the LHC lasted until November 2009. More details on the incident that did not involve beams are presented in the Appendix. During the repair a systematic soldering problem affecting roughly 15% of all the high-current (13 kA) cable joints was discovered. Since this problem could not be solved immediately, the beam energy had to be limited to 3.5 TeV until a complete repair campaign could be performed. The LHC therefore operated in 2010 and 2011 at this energy, before the energy was increased to 4 TeV in 2012 and 2013 because no magnet was quenched at 3.5 TeV during beam operation, thus lowering the risk of problems with the cable joints. The energy will be pushed to 6.5 TeV and above in 2015 after the repair campaign of 2013 and 2014 [10, 11].

## 2 Machine protection at the LHC

At 7 TeV each nominal LHC beam stores an energy of 360 MJ. This corresponds to the energy content of 80 kg of explosives and is a hundred times higher than previously achieved accelerator records. The energy stored in the beams of various accelerators is shown in Fig. 4. Machine protection systems (MPSs) with unprecedented safety levels are therefore required to operate the LHC [12, 13]. This is achieved with a combination of active protection by equipment and beam parameter monitoring, as well as with passive protection by a large number of collimators as indicated in Fig. 5.

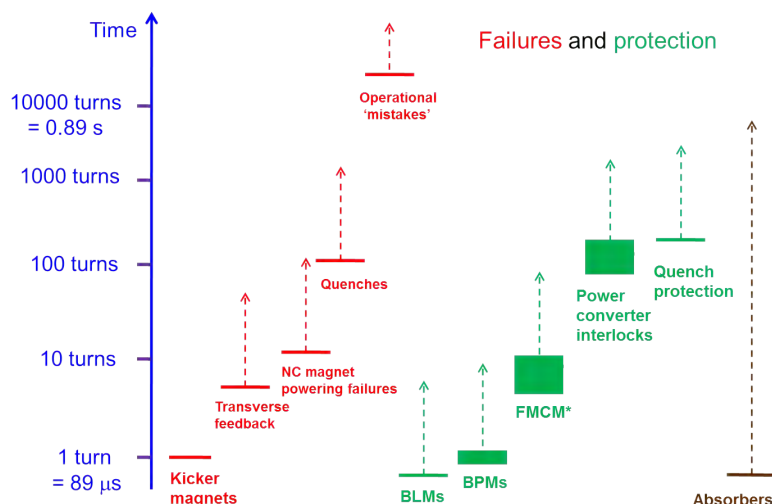


**Fig. 5:** Schematic outline of the main functions of the LHC MPS. Protection against failures is achieved by active monitoring followed by a dump action or by passive protection with absorbers.

The function of the LHC MPS is to protect the LHC accelerator, its injection and extraction transfer lines, as well as its experiments [14] against a large spectrum of failures [15]. Failures may be detected by monitoring equipment or beam parameters, for example power converter currents and state or beam loss rates. In parallel their impact is mitigated by passive absorbers in the form of collimators and absorbers. In the case where abnormal conditions are detected, interlock signals are sent to the the beam interlock system (BIS) [12, 13] that will trigger a beam dump or stop injection of the beams. A summary of the time-scale of failures and of the reaction time of MPS components is shown in Fig. 6.

The LHC BIS interfaces the sources of interlocks (clients) and the beam dumping and injection kicker systems. There are currently 189 inputs from client systems. The list of connected systems includes:

- the powering interlocks for normal-conducting and superconducting circuits;
- the fast current change detection of electrical circuits with fast failure time constants;
- the electrical circuits of the main experiment magnets;
- the interlock signals from the experiments (excessive rates);
- approximately 4000 beam loss monitors (BLMs) [18];
- the beam position monitors (BPMs);
- the beam screens;
- the collimators and absorbers (positions and temperatures);
- the movable experimental detectors;
- the RF system;
- the beam dumping system (state and trigger unit);
- the injection and aperture kicker systems;
- the personnel access system and its associated beam stoppers;



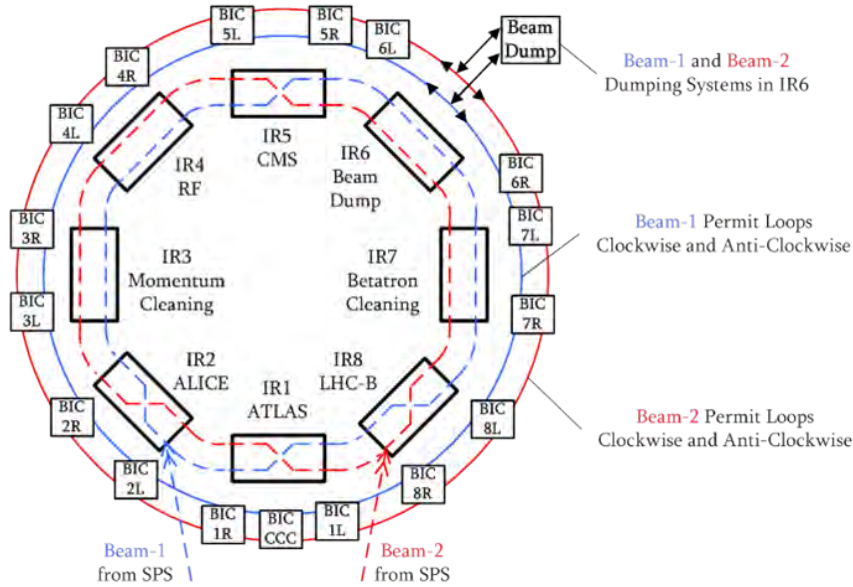
**Fig. 6:** Time-scales for failures (red) and for the reaction of the active protection systems (green) at the LHC. The symbol FMCM stands for fast magnet current change monitor, a device to detect fast current changes in normal-conduction magnet circuits with very short time constants [16]. The symbol NC stands for normal-conducting magnet.

- the operator inhibit buttons;
- the vacuum valves;
- the programmable beam dump.

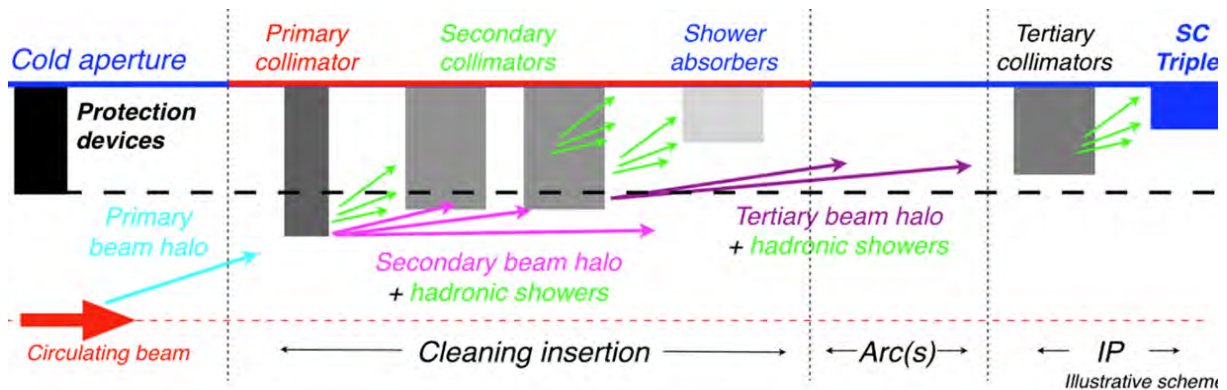
Beam interlock controller (BIC) modules are installed in each LHC access point and provide the interface between the client systems and the BIS; see Fig. 7. The BIC units are connected to each other and to the beam dumping system through permit loops. There are two permit loops per beam; the signals in the loops propagate clockwise in one loop and counter-clockwise in the other loop. This arrangement ensures always the fastest possible transmission of a dump request to the dumping system. At the LHC the dump delay can reach around three turns or  $270 \mu\text{s}$ ; this delay includes propagation of an interlock on the permit loops, synchronization of the beam dump with the beam abort gap (see below) and one turn to extract the entire beam.

The permit loops are connected directly to the LHC injection interlock system, which is made from the same building blocks (BIC modules). The LHC injection interlocks act on the LHC injection kicker and on the SPS extraction interlock system. This ensures that no beam may be extracted from the SPS injector when the beam permit loops are not armed [17].

The superconducting magnets in the LHC lose their superconducting state (quench) when they are heated with about 10 mJ per cubic centimetre. This is to be compared with the stored energy of many 100 MJ. To prevent unavoidable particle losses from the beam hitting the vacuum chamber within the magnets, possibly leading to sufficiently large energy deposition to trigger a quench [19], a collimation system [20] must intercept the beam losses with very high efficiency. Contrary to previous hadron colliders that used collimators only for experimental background conditions, the LHC cannot operate without its collimation system. Approximately 100 LHC collimators are installed in two long straight sections as shown in Fig. 1. To be able to absorb the energy of the 7 TeV hadrons, the LHC requires a multistage collimation system that intercepts the particles in a four-step process as outlined in Fig. 8, covering the entire six-dimensional phase space, catching particles with large transverse oscillation amplitudes and energy errors [20]. Each collimator is composed of two parallel, typically 1.2 m long material blocks (jaws) that define a variable gap through which the beams circulate. The smallest gap widths are roughly 2 mm.



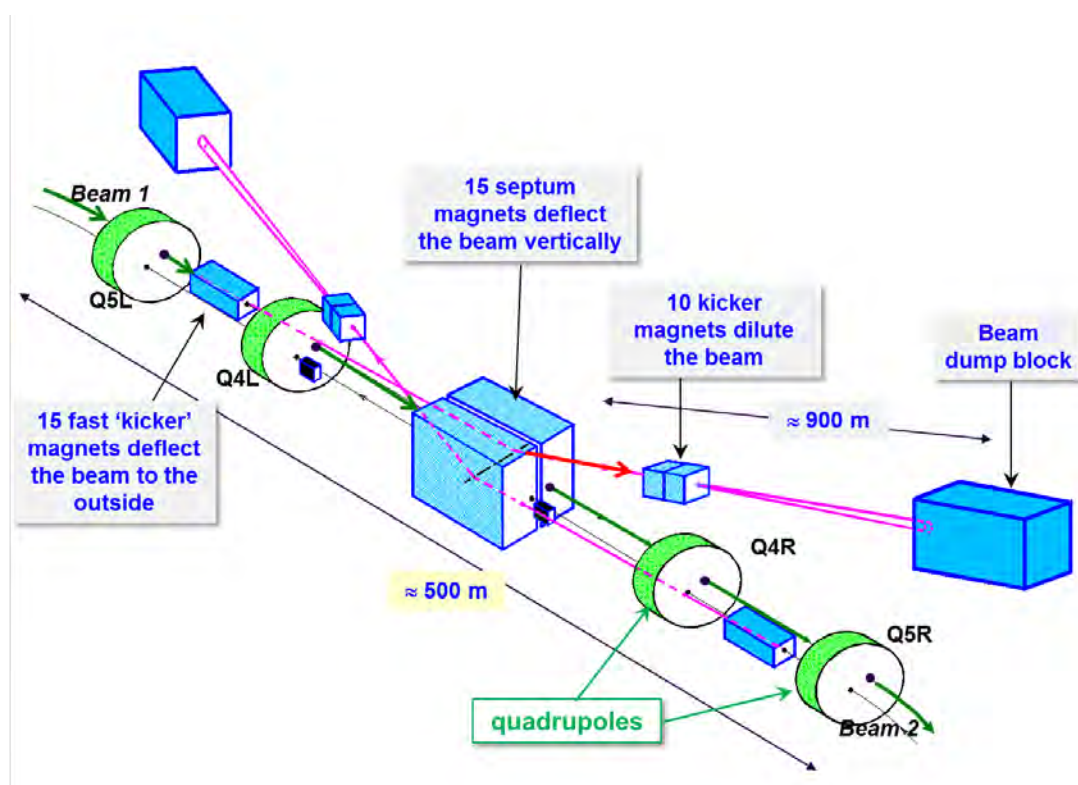
**Fig. 7:** Schematic layout of the LHC beam interlock system with 17 beam interlock controllers (BICs) distributed around the ring (image courtesy of B. Todd, CERN). The BIC modules are connected to beam permit loops that interface with the LHC beam dumping system.



**Fig. 8:** Principle of multistage beam collimation (cleaning) at the LHC

Depending on their location in the ring and on their role, the collimator jaws are made of fibre-reinforced carbon, strongly absorbing copper and tungsten blocks. The system worked perfectly so far, also thanks to excellent beam stability and machine reproducibility. Around 99.99% of the protons that were lost from the beam were intercepted. A single set-up of the collimation system was required per year. From the machine protection perspective, the LHC collimator must fulfil a dual role, halo collimation (beam cleaning) and passive protection of the accelerator component.

The LHC beam dumping system (LBDS) is a complex system composed of 15 fast kicker magnets that deflect the beam to the outside of the ring, 15 normal-conducting septum magnets that deflect the beam vertically out of the plane of the ring, 10 dilution kicker magnets that spread the beam over the beam dump surface and finally the beam dump block [1, 2, 12, 13]. The layout of the LBDS is shown in Fig. 9. The 10 m long carbon dump block is the only LHC element capable of absorbing the nominal beam. The beam is swept over the dump surface to lower the power density. Without the sweep the beam could drill a hole with a depth of a few metres into the dump block through hydro-dynamic tunnelling [22].



**Fig. 9:** Layout of the LHC beam dumping system (LBDS). For each beam 15 fast kicker magnets deflect the beam out of the ring into septum magnets that deflect the beam further into the vertical plane. On their way to the beam dump, fast dilution kickers spread out the bunches on the surface of the dump block (courtesy of M. Gyr).

A  $3 \mu\text{s}$  long particle-free gap in the beam (beam abort gap) provides a time window for the LBDS dump kickers to raise to their nominal field. The dump kickers must be accurately synchronized to the beam abort gap to avoid spreading beam across the aperture during the kicker rise time. Possible failure modes are:

- the abort gap fills with beams (RF fault, debunching, injection error);
- the kicker synchronization fails or a kicker fires spontaneously (not synchronized) – so-called *asynchronous beam dumps*.

The asynchronous dump is the *ultimate unavoidable failure*: the LHC must be protected from this failure *passively* by absorbers. Two large absorbers in front of the extraction septum and in front of the first superconducting magnet downstream of the dump kickers and septa protect the LHC against damage from asynchronous dumps and from residual beam in the abort gap. Dump kicker powering, synchronization and triggering are designed to exclude out-of-sync triggers with high reliability. The spontaneous trigger is estimated to occur roughly once per year. So far none has been observed during high-intensity operation, but the system operated at reduced high voltage (4 TeV instead of 7 TeV), which also lowers the probability of spontaneous kicker firing.

A direct link between the LBDS and the injection system ensures that the LHC injection kicker magnets cannot be triggered with a timing that would inject beam into the abort gap. The beam population in the abort gap is monitored using synchrotron light. When the particle density in the abort gap exceeds a pre-defined threshold, the transverse feedback system of the LHC is used to excite those particles until they hit the collimators.



### 3 Interlock masks for commissioning and operation

Already during the design phase of the LHC MPS, the need for masking interlocks was recognized [12, 13]. Some flexibility is always required for low-intensity commissioning and setting up phases where the risk of damage is also much lower. To avoid masking interlocks by raising thresholds and opening tolerance windows for many parameters (with the risk of errors during the reversal), the concept of masking interlocks when *the beam is safe* was introduced at an early stage. Such a safe beam should not be able to damage accelerator components. The corresponding beam intensity limit depends on the beam energy and on the beam emittance. It also depends on the material that is considered for damage. At the LHC, copper was selected as reference material. Given the large difference between the energy required to quench a superconducting magnet and the energy required to damage the same magnet, it is clear that even a safe beam will be able to quench one or more magnets.

In the early design phase of the LHC machine protection system, around the years 2000–2004, not much was known about equipment damage by high-energy and high-intensity beams. In order to benchmark damage by a high-energy LHC-type beam, a dedicated experiment was set up in a SPS transfer line [23]. In this controlled experiment beams of 450 GeV protons with LHC bunch structure (25 ns spacing) and nominal LHC beam emittances were directed into a special target composed of a sandwich of materials (copper, stainless steel and zinc). From this experiment the damage limit for copper was established at  $2 \times 10^{12}$  protons. It should be noted that the damage limit for stainless steel was more than a factor of four higher. The observations were in good agreement with rather simple estimates for the damage limit based on shower simulations to define the deposited energy and heating of the materials up to the melting point.

FLUKA simulations [24] were used to extrapolate the experimental results from 450 GeV to the LHC operating energy of 7 TeV. The simulations predicted the following scaling law for the safe beam intensity  $I_{SB}$  with beam energy  $E$ :

- larger energy deposition implies a scaling  $\propto 1/E$ ;
- shrinking of the transverse emittance (beam area) implies a scaling  $\propto 1/E$ ;
- increase of the shower length provides some dilution  $\propto \log(E)$ .

The following effective scaling law with energy (at fixed normalized emittance) was obtained:

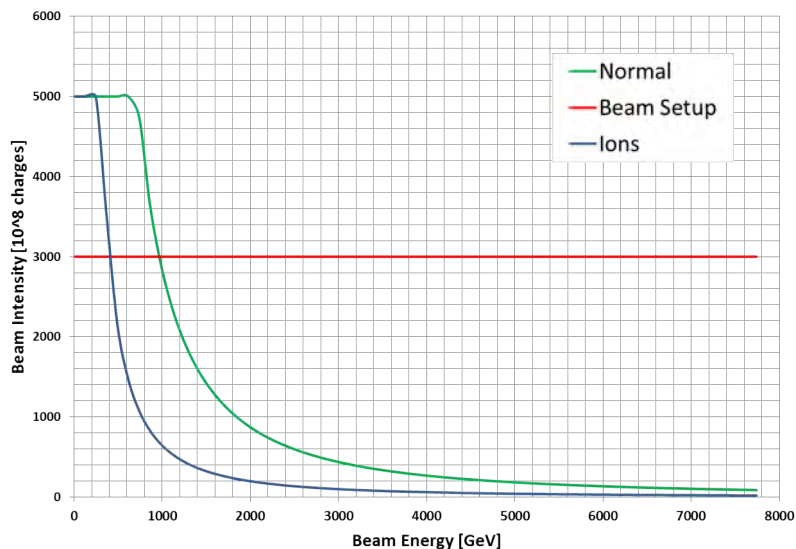
$$I_{SB} \propto E^{-1.7}. \quad (2)$$

This equation is implemented inside the LHC safe machine system (SMP) and is indicated as the *Normal* set-up beam flag (SBF) in Fig. 10. At injection the intensity limit was set initially to  $10^{12}$  charges, but was later lowered to  $5 \times 10^{11}$  because the beam emittances were roughly a factor of two smaller than expected in the LHC design. The SMP system is connected to reliable Beam Current Transformers (BCTs) and energy sources (based on the dipole fields with four-fold redundancy). It generates the SBF that is distributed over the LHC machine timing system to the BIS. The SBF states are:

- true = set-up beam: maskable BIS input signals can be masked;
- false = unsafe beam: no BIS input signals may be masked.

The beam interlock system is configured to allow masking certain classes of interlocks (maskable) when the SBF is true.

During initial operation it was realized that the definition of the SBF was too restrictive for beam commissioning that required higher bunch charges than expected (nominal bunches with  $10^{11}$  protons) due to systematic effects in the LHC beam instrumentation, mainly the beam position monitoring system. As a consequence, relaxed SBF equations were introduced to be able to mask certain interlocks with 2–3 nominal bunches at top energy as shown in Fig. 10. During commissioning periods a MPS expert can relax the equation of the SBF within the SMP system for certain tests.



**Fig. 10:** Limiting beam intensity (in  $10^8$  charges) as a function of energy for the set-up beam flag. The *Normal* SBF corresponds to the scaling of Eq. (2). The *beam set-up* equation is a relaxed version used for beam commissioning and MPS set-up. For ion beams the limits are scaled by the ratio  $Z/A$  of the charge per nucleon.

#### 4 Commissioning and organization

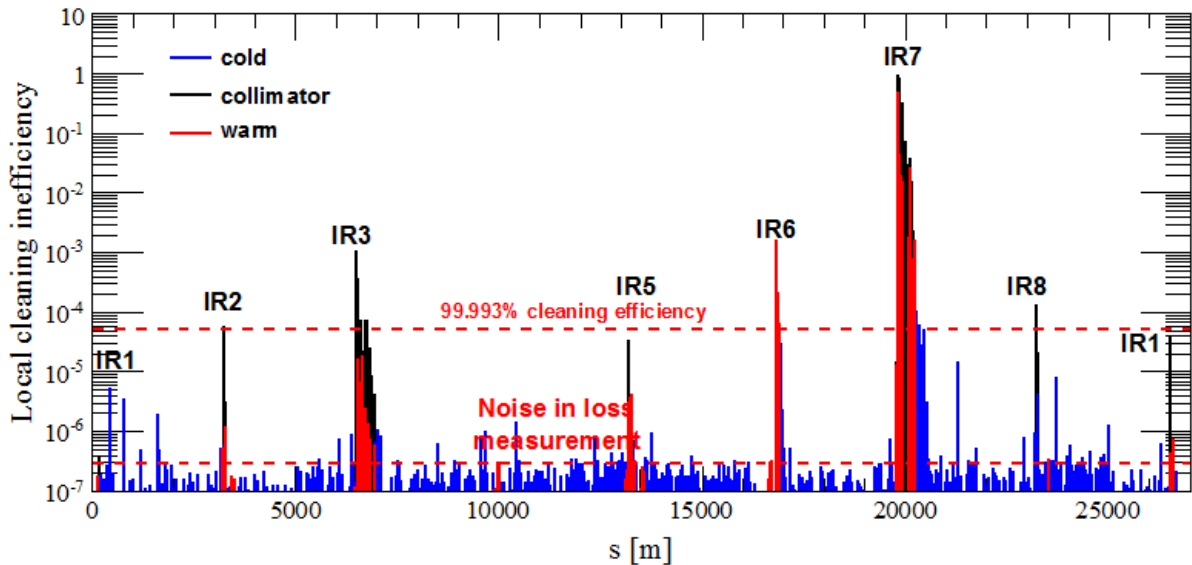
The MPS activities of the LHC were organized since the year 2000 inside a Machine Protection Working Group (later renamed to Machine Protection Panel, MPP). This group steered the design and followed up on implementation, issues and performance of the LHC MPS. All groups involved in MP activities were represented in the MPP.

An executive MP body, the restricted MPP (rMPP), was created when beam commissioning started. Each core MP system has one representative within the rMPP. The rMPP takes decisions related to MPS (example: BLM threshold changes) and steers the intensity ramp up of the machine. Its recommendations are submitted to the CERN management; in almost all cases the recommendations were accepted.

Before the machine startup, procedures were developed for the commissioning of the machine protection subsystems. The procedures contain test descriptions and frequency of tests (after stop or intervention). The procedures were then translated into a series of individual tests to be performed on the machine with and without beam, if required also at different intensity steps.

Machine protection tests are currently documented and tracked on a web page. One MPS expert of the commissioning team checks that all tests required for a given commissioning phase have been performed by the experts. It is in general the system expert that executes the tests for his system and not an independent person. This simple mechanism for tracking the commissioning will be improved in the future. The new concept with test tracking and electronic expert signatures is already in place for the commissioning of the LHC electrical circuits and magnets. The *AccTesting* framework [25] is based on pre-defined and agreed test sequences. Tests that are ready for execution can be scheduled for execution; test sequences are blocked until tests are analysed and signed by experts. The results are tracked and stored; no test step can be forgotten. Unfortunately due to lack of resources the test tracking was not yet implemented for the LHC MPS, but it is foreseen in the near future.

Passive protection of the LHC aperture with collimators and absorbers is a key ingredient for operating the LHC safely at high intensity. All failures affecting the machine on a global scale (orbit, optics, emittance, perturbations) must be intercepted by a protection device. The LHC machine setting



**Fig. 11:** Example of the beam loss distribution during a loss map performed at a beam energy of 4 TeV. The vertical scale gives the losses normalized to the peak loss at the primary collimator. The beam travels from left to right. Blue markers correspond to BLMs located on superconducting elements, red markers to room-temperature elements and black markers to collimators.

up involves:

- a well-corrected orbit (rms below 0.5 mm);
- a well-corrected optics (betatron function beating below 20% at injection and below 10% with colliding beams);
- a good knowledge of the aperture bottlenecks (after orbit and optics correction).

The machine aperture is measured after orbit and optics corrections have been finalized. The global aperture is measured at injection energy (typical aperture limit is at  $\approx 12 \sigma$ , where  $\sigma$  is the beam size with a nominal normalized emittance of  $3.5 \mu\text{m}$ ). Local apertures are measured at injection (injection region, beam dump section and around the four experiments) and at top energy with fully squeezed beams (only the four experiments). At top energy the aperture depends on the optics at the different collision points. The aperture limit is generally located in the low-beta quadrupoles next to the collision points.

During machine set-up, all collimators and absorbers are aligned around the closed orbit with appropriate retractions. For good performance the orbit must be reproducible at the level of  $50 \mu\text{m}$ . The machine set-up (orbit, optics, aperture and protection devices) is then validated by a campaign of loss maps and simulated asynchronous beam dump tests.

- During a *loss map* the transverse beam emittance is blown up until losses are observed on the collimators and absorbers; an example is shown in Fig. 11. The loss distribution provides a validation of the collimator alignment and hierarchy [20]. Initially the emittance was blown up by crossing the third-order resonance in the horizontal or vertical plane. This technique was however not always very reproducible; sometimes the losses were too small, sometimes they were massive and a large fraction of the beam was lost. From 2012 onwards emittance blow up was obtained by noise excitation using the transverse feedback system. This provided fine control over the losses and blow up could be applied to individual bunches [21].
- For a simulated *asynchronous test dump* a low-intensity beam (typically 1–3 bunches) is debunched by switching of the RF system. After a few minutes the particles drift into the abort gap,

the density of particles in the gap being monitored by a dedicated device based on synchrotron light. When the population in the abort gap is sufficient, a dump is triggered. The beam present in the region of the abort gap mimics the effect of an asynchronous dump. The loss distribution along the ring provides a validation of the dump protection alignment.

## 5 Intensity ramp

The entire LHC cycle, from injection to collisions, is always set up with low-intensity beams. The maximum intensity is three bunches with nominal intensity ( $\approx 1.1 \times 10^{11}$  protons). The set-up is made with less than 1 permill of the nominal intensity, which represents a challenge for the beam instrumentation, since there should not be a significant bias of measurements between the set-up and the nominal full-intensity beam (for example for the beam position monitoring).

The intensity increase is steered through the restricted Machine Protection Panel (MPPr), which defines the intensity steps and the requirements to proceed with the intensity increase. The plan for the first learning year in 2010 foresaw three phases:

- low-intensity beams for commissioning and early experience. This phase was followed by an internal review of the MPS performance;
- an intensity ramp up to a stored energy of 1–2 MJ followed by 4 weeks of operation at that stored energy. Such a stored energy corresponded to state of the art in 2010. This phase was followed by an external review of the MPS performance;
- increase of the stored energy into the regime above 10 MJ.

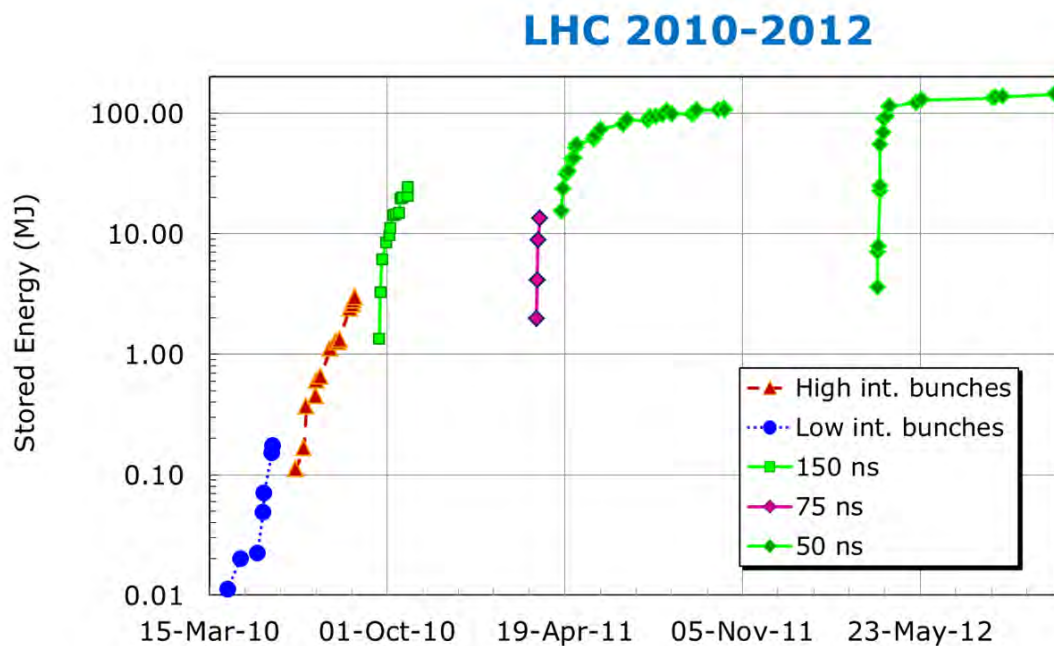
Figure 12 shows the evolution of the peak stored energy for one LHC beam between 2010 and 2012. The slow ramp up of 2010 is clearly visible. With the experience that was accumulated in 2010, the stored energy was ramped up above 100 MJ in 2011 over 11 intensity steps, up to a maximum of 1300 bunches. The 2011 intensity ramp up took around 9 effective weeks; the rate was dictated by operational (and not MP) issues as soon as around 600 bunches were stored [4]. Losses and the need to adjust of BLM thresholds, vacuum element heating by the beam, beam stability, etc slowed down the pace. Finally, the 2012 intensity ramp up took just 2 weeks with seven intensity steps, as can be seen in Fig. 13.

## 6 Beam losses at the LHC

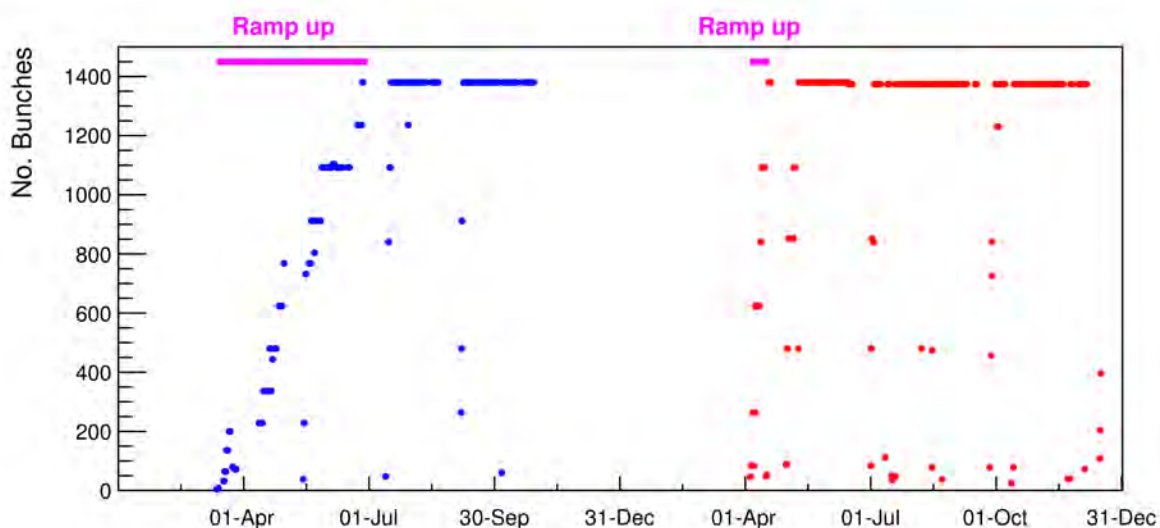
### 6.1 Witness beam

Injection into a synchrotron that has no circulating beam has the same MP issues as a linac [17]. Injection of an intense beam can represent a serious risk or require very important monitoring efforts (all power converters etc) to ensure that the beam is not lost directly on the aperture during the first turn, before the MPS is able to react. To overcome this issue, the concept of *witness* beam (or *beam presence*) was introduced for the LHC [26]. The principle is that only a probe bunch (typically  $10^{10}$  protons, max  $10^{11}$  protons) may be injected into an *empty* ring. Such a low intensity cannot provoke damage to components. A high-intensity injection requires a minimum beam intensity to be circulating; this is the best guarantee that the conditions are reasonable to avoid failures happening on the first turn, before the MPS is able to react. At the LHC the concept is based on a highly reliable and redundant intensity measurement. A flag indicating the beam presence (true/false) is transmitted to the extraction interlock system of the SPS injector where it is combined with a flag indicating that the SPS beam is has a probe intensity (max  $10^{11}$  protons).

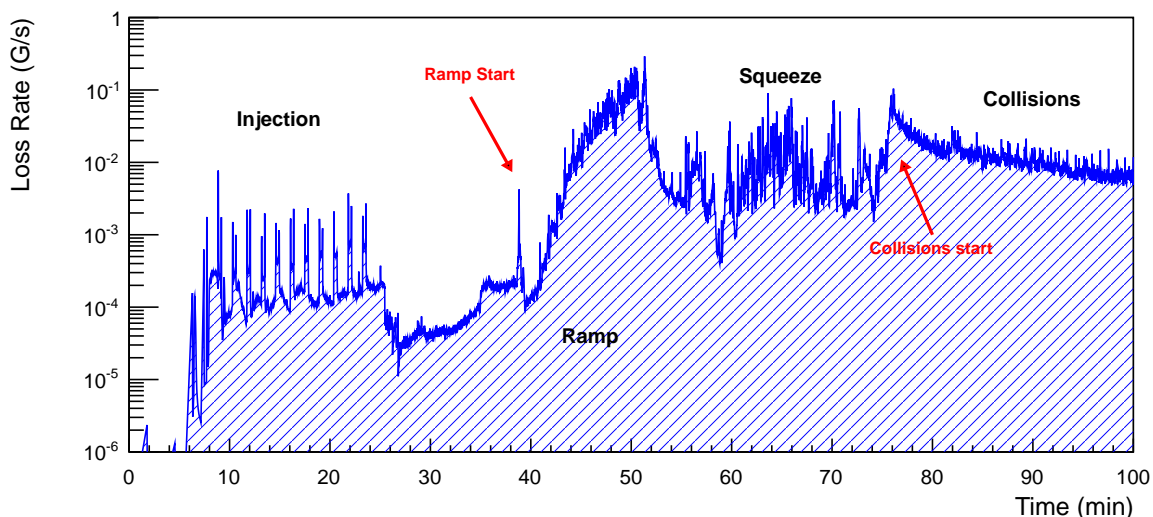
Despite storing up to 140 MJ at 4 TeV, not a single superconducting magnet was quenched at the LHC with circulating beam, despite quench thresholds of only a few tens of mJ. On the other hand, many magnets were quenched at injection, mainly due to (expected) injection kicker failures (seven events in 2012). The beam (roughly 2 MJ) is safely absorbed in injection dump blocks, but the shower leakage can quench magnets over a distance of around 1 km.



**Fig. 12:** Evolution of the peak stored energy in one LHC beam during Run 1. The different colours detail the beam structure in single bunches (first two periods) or in trains with bunch spacings of 150, 75 and 50 ns.



**Fig. 13:** Intensity ramp up of the LHC beams, expressed here through the number of bunches in 2011 and 2012



**Fig. 14:** Beam loss rates averaged over 1 s at the primary collimator along a typical LHC cycle in 2012. During the squeeze phase the optics around the experiments is changed to reduce (squeeze) the betatron function at the IP.

## 6.2 Beam loss in the cycle

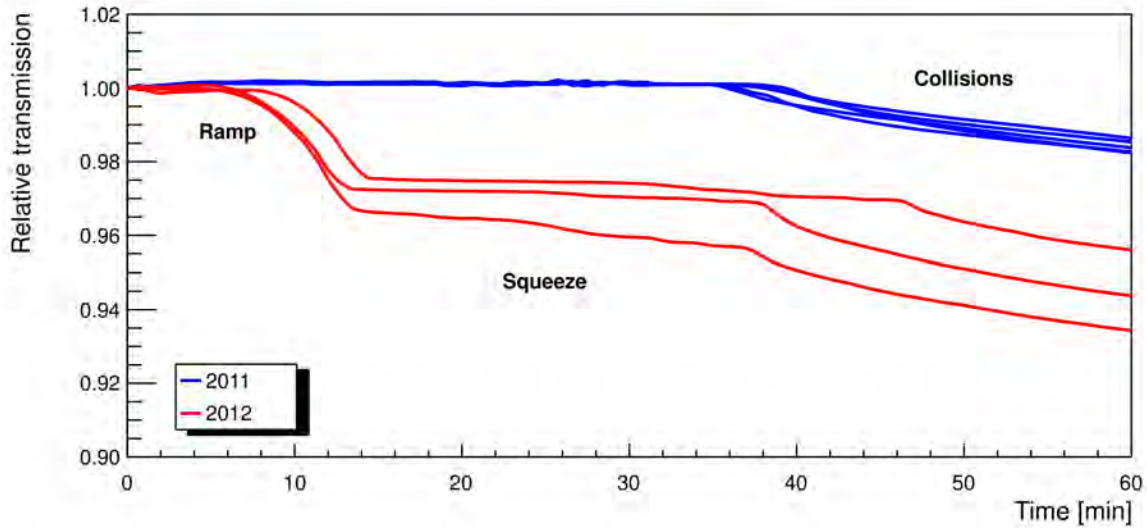
At the LHC characteristic beam losses are observed in the various phases of the machine cycle. Those losses are part of regular operation and must be tolerated, even if one tries to minimize them. The normal losses are:

- injection losses (tails of the injected beams, injection oscillations, de-bunched beam);
- start of ramp losses (uncaptured beam loss);
- scraping on collimators (gap changes, orbit and tune shifts);
- losses from the beam halo when beams start to collide (beam–beam effect);
- losses due to the beam burn-off that are proportional to luminosity and performance.

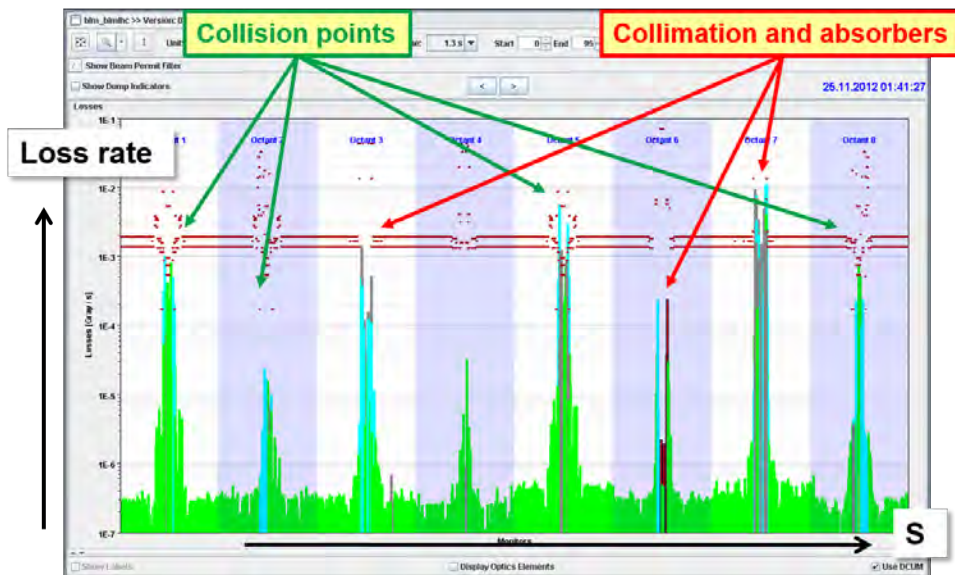
The different types of losses are shown in Fig. 14 for a typical LHC cycle. The importance of the primary collimator opening appears clearly if the total beam intensity transmission presented in Fig. 15 is compared between 2011 and 2012. In 2012 the collimators were set closer to the beam in order to protect a smaller aperture, which allowed smaller betatron functions to be reached at the IP and therefore smaller beam sizes at the collision points, yielding a 60% higher luminosity [27]. The gap width has a strong impact on beam transmission and losses in the cycle.

- In 2011 the intensity losses during the ramp and squeeze phases are so small that they cannot be measured with the current transformers. The losses are completely dominated by collisions. The primary collimator gaps correspond to an opening of  $\pm 7.5 \sigma$ , where  $\sigma$  is the real beam size.
- In 2012, on the other hand, there are already significant losses during the ramp when the collimator gaps are closed to their high-energy setting. There are noticeable losses in the optics squeeze due to an increased sensitivity to even small orbit jitter at the level of  $50 \mu\text{m}$ . The primary collimator gaps correspond to an opening of  $\pm 5.2 \sigma$ , where  $\sigma$  is the real beam size.

The regular loss distribution during collisions is shown in Fig. 16. The beam loss rates near the experiments are almost as high as at the collimators. This is due to the fact that around the experiments the loss monitors record collision debris, the physics processes at very small angles that are not covered by the experiments. Their signals are proportional to the collider luminosity.

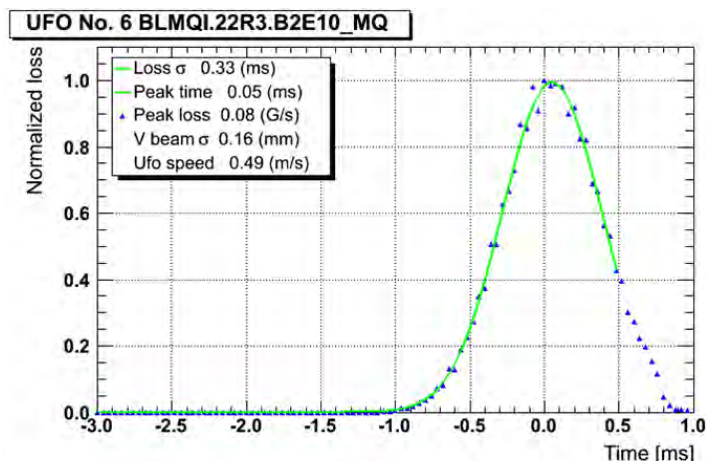


**Fig. 15:** Intensity transmission through the LHC cycle including the energy ramp and the betatron squeeze, the operation phase when the IP betatron functions are reduced to their value in collision.



**Fig. 16:** Distribution of steady-state beam loss around the LHC ring with stable colliding beams

The thresholds of the LHC loss monitors are set to prevent quenches for monitors installed on superconducting elements and to prevent damage for room-temperature elements (for example collimators) [19]. In both cases some safety margin is desired. The initial thresholds were set before LHC operation started based on rather coarse quench level estimates coupled to GEANT, FLUKA and MARS simulations [24]. During the first years of operation the thresholds were progressively adapted (many were increased) based on experience. Initially the thresholds on collimators were set to limit the average power loss significantly below the peak design power of 500 kW. The thresholds were only increased to match the nominal power loss once the operation of the LHC was well controlled and understood. Quench tests with wire scanners (nice point-like particle source), orbit bumps and short and high losses in the collimation area were used to determine more accurately the quench limits [19].



**Fig. 17:** Time evolution of the beam loss during a UFO event. The vertical scale represents the beam loss relative to the highest loss during the UFO event.

### 6.3 UFO losses

Very fast beam loss events (time-scale of a millisecond and below; see Fig. 17) mainly in superconducting regions have been the *surprise of LHC operation*. Those fast losses were nicknamed UFOs (unidentified falling objects) when it became clear that the UFO-type beam losses were due to small objects falling into the beam, the subsequent interaction of the beam particles leading to the showers and the beam losses [28, 29]. The BLM signals are consistent with small (tens of  $\mu\text{m}$  diameter) dust particles ‘entering’ the beam. The vast majority of UFO events lead to losses below dump threshold, but around 20 beam dumps were triggered by such UFO-type events every year between 2010 and 2012. UFO events localized within the injection kickers could be traced to aluminium oxide dust present in large quantities on the inner surface of the ceramic vacuum chamber of those elements. A cleaning campaign of the kickers was made during the long shutdown in 2013–2014. UFOs occur mainly with high-intensity beams, but there is conditioning with beam; the event rate drops from 10 to 2 per hour over a year as shown in Fig. 18. To improve the sensitivity of the LHC arc BLMs to UFO events, two out of six BLMs installed around the arc quadrupoles were re-located to the dipoles during the shutdown [30]. The monitoring and protection (quench prevention) capabilities of the BLMs are significantly improved with this re-location for the coming 6.5 TeV run. While the beam losses will increase due to the higher beam energy, the quench thresholds of the magnets will come down by a factor of 3–4, making the situation at 6.5 TeV more critical.

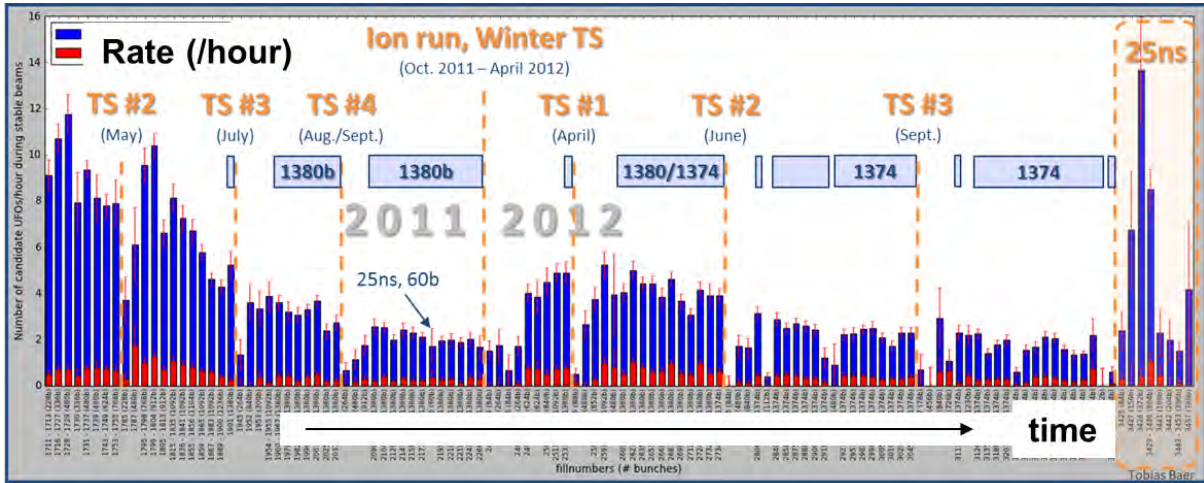
## 7 Diagnostics and control

The three main requirements for a modern MPS are:

- protect the machine: the highest priority is obviously to avoid damage to the accelerator;
- protect the beam: complex protection systems reduce the availability of the accelerator; the number of ‘false’ interlocks stopping operation must be minimized. This implies sometimes a trade-off between protection and operation;
- provide the evidence: clear (post-mortem) diagnostics must be provided when the protection systems stop operation or when something goes wrong (failure, damage, but also ‘near misses’).

Once the MPS components have been commissioned, it is essential that the protection functionality is maintained during operation. Automated checks of the MPS components as pre- or post-flight checks





**Fig. 18:** Evolution of the hourly rate of UFO events in 2011 and 2012. A slow conditioning is observed during each run (adapted from [28, 29]).

can ensure that the MPS functionality is not degraded. For colliders with long cycle times there are two types of checks that fit well into the cycle, namely pre-flight checks before injection and post-flight checks on data collected during a fill or during the beam dump (post-mortem data). Such tests can come in multiple forms, for example the verification of MPS-related settings such as interlock thresholds, configuration checks of the beam interlock systems, automated analysis of the faults and MPS reaction chain and automated analysis of the dump system action. At the LHC the BLM system integrity checks must be performed at least once per 24 h period (or after the following beam dump if that interval is exceeded) [31]. The integrity checks are performed by a high-voltage modulation that is analysed to detect ‘dead’ channels.

At the LHC the MPS is so critical that for every beam dump post-operation checks (POCs) are performed on the beam dump system post-mortem data (equipment and beam signals). Data collection and analysis are triggered automatically after each dump. The analysis covers internal beam dump system signals and external beam information like dumped beam intensities, losses and beam positions in and around the dump channel. The analysis ensures that all signals are correct and that there is no loss of redundancy; the beam dumping system can then be considered ‘as good as new’. Machine operation is stopped if the beam dumping system POCs fail; an expert verification is required to re-start beam operation. This concept was so successful that it was extended to LHC injection in the form of automated checks of each injection quality.

Already during the design of the LHC MPS, post-mortem (PM) diagnostics was identified as a key component to understand the root causes of beam dumps [32]. All key LHC systems implement circular PM data buffers that are frozen and read out when a beam dump is triggered. The PM trigger is transmitted over the LHC machine timing system. The PM system also inserts an automatic entry in the LHC electronic logbook. Sampling frequencies of the data in the PM buffers range from ms or turn level to tens of milliseconds and are adapted to each system. Synchronization is critical to make sense of the data and define the event sequence; therefore time stamping is made at the source of the data. The LHC post-mortem data volume is currently around 200 MBytes. Since the analysis of the large volume of LHC PM data can be tedious, automatic analysis tools have been implemented to support operators and MPS experts in their work. The LHC PM server and graphical user interfaces are based on the JAVA language with standard interfaces to extract raw data and provide analysed data [33]. This allows many persons to contribute to the analysis modules. After a beam dump, injection into the LHC is blocked until the PM data is collected, pre-analysed (automatic) and signed by the operation crews. If an automated

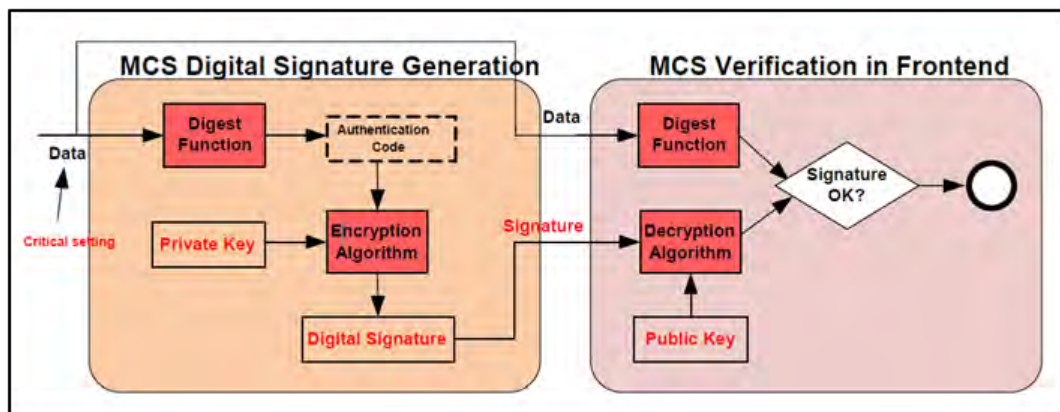


Fig. 19: Principle of management of critical settings (MCS) with digital signature of the data

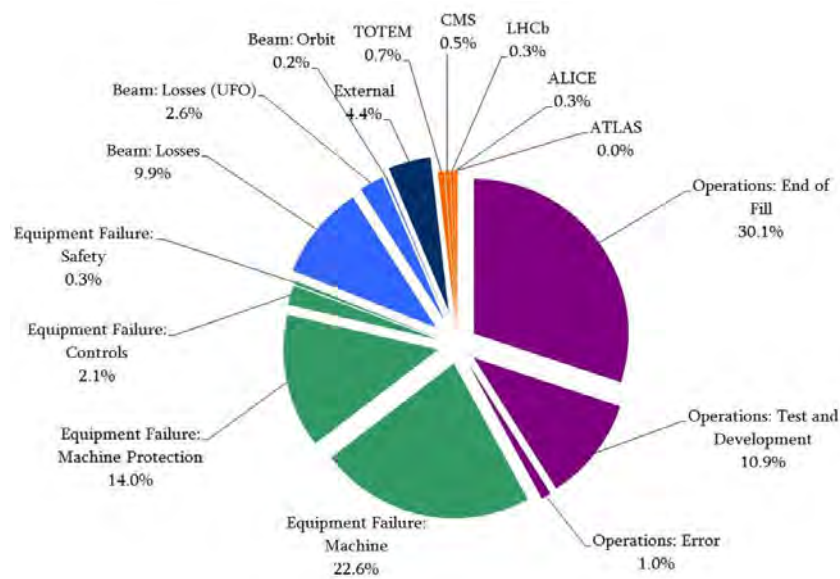
analysis identifies a critical problem, injection can only be released by a MPS expert. To detect any possible long-term trend MPS experts re-analyse all PM events collected above injection energy within a few days. This also provides a view that is independent of the operation teams.

## 7.1 Settings control

Depending on its size, complexity and energy range an accelerator has a large volume of MPS settings in the form of BLM thresholds, current references and tolerances, etc. The management of settings associated to the MPS poses special problems, since access to the settings must be limited to authorized experts, yet it must be possible to inspect and download the settings from a central database to the front-end systems of the MPS. For the LHC access restrictions have been put in place based on *role-based access control* (RBAC) [34] and on digital signatures. To protect MPS settings the concept of *management of critical settings* (MCS) [35] was developed, a settings protection system that is fully embedded in the controls middleware and settings management. A setting that is defined as *critical* has an associated digital signature. Only a user that has the appropriate RBAC role (MPS expert, BLM expert, etc) is able to generate the digital signature. The digital signature is generated at the moment when a setting is changed by an authenticated expert. The setting and the associated digital signature are transmitted together to the front-end computer: a critical setting is only accepted with a valid digital signature, see Fig. 19.

## 7.2 Software interlocks

The LHC MPS has inputs from many systems that operate independently and there is only very limited information exchange between the various systems. To implement interlocks on a global machine scale with correlation of data between many LHC systems, a software interlock system (SIS) was developed [36, 37]. This system is able to collect information from any control device of the LHC and its injectors. The LHC SIS can dump one or both beams and inhibit LHC injection. It may perform global scale analysis among systems, correlate injector and LHC ring data for injection protection, etc. New interlock tests can be implemented rapidly to protect against unexpected issues. The LHC SIS is by design rather slow (reaction time of the level of seconds) but it is able to detect anomalies that could lead to problems in the future or prevent unnecessary beam dumps at injection. The SIS is based on a JAVA core server, with a large data buffer for currently up to around 2500 devices and settings. For each device data a timeout and no-data policy is defined. Over 5000 tests are defined with the LHC SIS; tests can be defined as simple value comparisons or complex JAVA logic.



**Fig. 20:** Distribution of beam dump causes during the 2012 LHC proton run (adapted from [38]). The statistics include only beam dump above injection energy.

## 8 Machine availability

Peak performance in terms of luminosity is not the most important LHC parameter; the integrated performance is much more critical. To optimize the integrated time with collisions it is important to maximize the time spent colliding the beams and to minimize the time to re-establish collisions at the end of a fill (turn-around time). The fraction of time that the LHC spends colliding beams stably for the experiments amounted to 30–35% of the total scheduled operation time [4]. The average duration of collisions was only 6 h while the optimum duration would be in the range of 8 to 12 h. The reason why the fills are so short is because during high-intensity operation only one out of three fills is dumped by the operation crews; see Fig. 20. The other two-thirds of the fills are dumped by the LHC MPS [38]. Roughly 14% of the beam dumps are due to the failures of MPS subsystems (‘false’ dumps), with the following distribution of the MPS subsystems:

- quench protection system (radiation to electronics): 65%;
- BLM system: 13%;
- beam dumping system: 12%;
- software interlock system: 5%;
- powering interlock system: 2.5%;
- beam interlock system: 1.5%.

A reliability working group predicted the rate of false dumps and the safety of the LHC MPS for 7 TeV operation before the LHC was switched on. This can now be compared with observations, but it must be observed that so far LHC operated at 4 TeV while the predictions were made for 7 TeV, which is not completely equivalent for certain systems. The observations are roughly in line with predictions, but some failures do not match completely; in particular, radiation to electronics was not included in the initial predictions [39, 40].

## 9 Machine experiments

Machine experiments can be very exciting, but also risky periods for an accelerator. Frequently the machine is operated at some ‘distance’ from standard conditions. For example collimator settings, orbit and optics may be changed. At the LHC every experiment is categorized according to the foreseen changes to the machine and to the beam intensity. Experiments using intensities above the SBF limit must be prepared with a detailed written description of the changes to machines and the test procedures. In many cases the analysis of the document helped improve the efficiency of the experiment by spotting ‘impossible’ things. This encourages experimenters to think about options with smaller MP footprint, for example lower intensity.

## 10 Summary

The LHC MPS has been very successful in protecting the LHC with over 100 MJ of stored beam energy while the LHC was operated at 4 TeV. No component was damaged by a failure leading to beam loss; the MPS therefore fulfilled its job. As expected, operation of the LHC was significantly constrained by MP due to the high stored energy and very low quench levels. Nevertheless, operation was rather smooth and the intensity ramp up from well below 1 MJ during the commissioning phase to over 100 MJ stored energy took only 2 weeks in 2012. In 2015 the beam energy will be increased to 6.5 TeV: the stored energy of the beams will increase by a factor of 2 to 3, while the quench levels of the magnet drop by a factor of 3 to 5. This will be a new challenge for LHC operation, while at the same time the focus is shifting more and more towards high(er) machine availability.

## References

- [1] L. Evans, *The Large Hadron Collider: A Marvel of Technology* (EPFL Press, Lausanne, 2009).
- [2] O. Brüning et al. (eds), LHC Design Report, CERN-2004-003 (2004). <http://dx.doi.org/10.5170/CERN-2004-003-V-1>.
- [3] H. Pfeffer, Protection of hardware: powering systems (PC, NC and SC magnets), these proceedings.
- [4] J. Wenninger et al., Operation and configuration of the LHC in Run 1, CERN-ACC-NOTE-2013-0041 (2013).
- [5] L. Rossi, The Large Hadron Collider of CERN and the roadmap toward higher performance, CERN-ACC-2014-0225.
- [6] R. Schmidt et al., Machine protection challenges for HL-LHC, Proc. 5th International Particle Accelerator Conference, IPAC14, Dresden, Germany, <http://www.jacow.org/>.
- [7] F. Zimmermann, M. Benedikt, D. Schulte and J. Wenninger, Challenges for highest energy circular colliders, Proc. 5th International Particle Accelerator Conference, IPAC14, Dresden, Germany, <http://www.jacow.org/>;  
Future circular collider study, <https://espace2013.cern.ch/fcc/Pages/default.aspx>.
- [8] F. Bordry, Status of the Large Hadron Collider (LHC), Proc. 2008 European Particle Accelerator Conference, EPAC08, Genoa, Italy, 2008, <http://www.jacow.org/>.
- [9] J. Wenninger, Status of the LHC, Proc. 2009 Particle Accelerator Conference, PAC09, Vancouver, Canada, 2009, <http://www.jacow.org/>.
- [10] F. Bordry et al., The first long shutdown (LS1) for the LHC, Proc. 4th International Particle Accelerator Conference, IPAC13, Shanghai, China, 2013, <http://www.jacow.org/>.
- [11] J.Ph. Tock et al., Consolidation of the LHC superconducting circuits: a major step towards 14 TeV collisions, Proc. 4th International Particle Accelerator Conference, IPAC13, Shanghai, China, 2013, <http://www.jacow.org/>.
- [12] R. Schmidt et al., *New J. Phys.* **8** (2006) 290. <http://dx.doi.org/10.1088/1367-2630/8/11/290>
- [13] R. Schmidt et al., Machine protection and interlock systems for LHC, these proceedings.

- [14] R.B. Appleby et al., *Phys. Rev. ST Accel. Beams* **13** (2010) 061002. <http://dx.doi.org/10.1103/PhysRevSTAB.13.061002>
- [15] V. Kain, Beam dynamics and beam losses for circular machines, these proceedings.
- [16] M. Werner et al., A fast magnet current change monitor for machine protection in HERA, Proc. ICALEPS 2005, Geneva, Switzerland, <http://www.jacow.org/>.
- [17] V. Kain, Beam transfer and machine protection, these proceedings.
- [18] B. Dehning, LHC beam loss monitor system, these proceedings.
- [19] B. Auchmann et al., *Phys. Rev. ST Accel. Beams* **18** (2015) 061002. <http://dx.doi.org/10.1103/PhysRevSTAB.18.061002>
- [20] S. Redaelli, Beam cleaning and collimation systems, these proceedings.
- [21] B. Salvachua et al., Handling 1 MW losses with the LHC collimation system, Proc. 5th International Particle Accelerator Conference, IPAC14, Dresden, Germany, <http://www.jacow.org/>.
- [22] A. Bertarelli, Beam induced damage mechanisms and their calculation, these proceedings.
- [23] V. Kain et al., Material damage test with 450 GeV LHC-type beam, LHC-Project-Report-822 (2005).
- [24] N. Mohkov and F. Cerutti, Beam material interaction, heating and activation, these proceedings.
- [25] A.A. Gorzawski et al., The AccTesting framework: an extensible framework for accelerator commissioning and systematic testing, Proc. ICALEPS 2013, San Francisco, CA, USA, <http://www.jacow.org/>.
- [26] R. Schmidt and J. Wenninger, LHC injection scenarios, LHC Project Note 287 (2002).
- [27] S. Redaelli et al., Operation of the betatron squeeze at the LHC, Proc. 4th International Particle Accelerator Conference, IPAC13, Shanghai, China, <http://www.jacow.org/>.
- [28] T. Baer, Very fast losses of the circulating LHC beam, their mitigation and machine protection, CERN-THESIS-2013-233.
- [29] T. Baer et al., UFOs in the LHC: observations, studies and extrapolations, Proc. 3rd International Accelerator Conference, IPAC12, New Orleans, LA, USA, <http://www.jacow.org/>.
- [30] B. Auchmann, BLM threshold strategy, LHC Performance Workshop, Chamonix, 2014, <http://indico.cern.ch/event/315665/>.
- [31] J. Emery et al., LHC BLM single channel connectivity test using the standard installation, Proc. European Workshop on Beam Diagnostics and Instrumentation for Particle Accelerators DIPAC09, Basel, Switzerland, <http://www.jacow.org/>.
- [32] E. Ciapala, F. Rodriguez Mateos, R. Schmidt and J. Wenninger, The LHC post-mortem system, LHC Project Note 303 (2002).
- [33] R. Gorbonosov et al., Plug-in based analysis framework for LHC post-mortem analysis, Proc. International Conference on Accelerator and Large Experimental Physics Control Systems ICALEPS 2013, San Francisco, CA, USA, <http://www.jacow.org/>.
- [34] P. Charrue et al., Role-based access control for the accelerator control system at CERN, Proc. International Conference on Accelerator and Large Experimental Physics Control Systems ICALEPS 2007, Knoxville, TN, USA, <http://www.jacow.org/>.
- [35] W. Sliwinski et al., Management of critical machine settings for accelerators at CERN, Proc. International Conference on Accelerator and Large Experimental Physics Control Systems ICALEPCS 2009, Kobe, Japan, <http://www.jacow.org/>.
- [36] J. Wozniak, V. Baggiolini, D. Garcia Quintas and J. Wenninger, Software interlock system, Proc. International Conference on Accelerator and Large Experimental Physics Control Systems ICALEPS 2007, Knoxville, TN, <http://www.jacow.org/>.
- [37] L. Ponce, J. Wenninger and J. Wozniak, Operational experience with the LHC software interlock system, Proc. International Conference on Accelerator and Large Experimental Physics Control

- Systems ICALEPS 2013, San Francisco, CA, USA, <http://www.jacow.org/>.
- [38] B. Todd, Performance and availability of MPS 2008–2012, Proc. 2013 MPP Workshop, CERN-ACC-2014-0041.
- [39] R. Filippini et al., Reliability analysis of the LHC beam dumping system taking into account the operational experience during LHC Run 1, Proc. International Conference on Accelerator and Large Experimental Physics Control Systems ICALEPS 2013, San Francisco, CA, USA, <http://www.jacow.org/>.
- [40] J. Uythoven, Dependability calculations prior to 2008 and operational experience during first LHC run, <https://indico.cern.ch/event/277684>.
- [41] M. Bajko et al., Report of the task force on the incident of 19th September 2008 at the LHC, CERN-LHC-PROJECT-Report-1168.

## Appendix A: Powering incident in 2008

Since the LHC powering incident in 2008 was the most severe damage that ever happened to an accelerator, a brief description of it is appended here even though this incident happened *without beam* [9].

In the morning of 19 September 2008 the last commissioning step of the main dipole circuit (154 magnets) of sector 34 (between IP3 and IP4 in Fig. 1) was started, a ramp to 9.3 kA which corresponds to a beam energy of 5.5 TeV. During the ramp an electrical fault developed in the powering busbar interconnection at a current of 8.7 kA. A resistive voltage appeared and increased to 1 V after less than 0.5 s, leading to the power converter trip. The current started to decrease in the circuit and the energy discharge switch opened, inserting dump resistors in the circuit. In this sequence of events, the quench detection, power converter and energy discharge systems behaved as expected. No resistive voltage appeared on the dipoles of the circuit, individually equipped with quench detectors and the quench of a magnet has been excluded as initial event.

Within the first second, a main electrical arc and multiple smaller secondary arcs developed and punctured the helium enclosure, leading to release of helium into the insulation vacuum of the cryostat. After a few seconds the beam vacuum also degraded. The spring-loaded relief discs on the insulation vacuum enclosure opened when the pressure exceeded atmospheric, thus relieving the helium to the tunnel. They were however unable to maintain the pressure rise below the nominal 0.15 MPa absolute, thus resulting in large pressure forces acting on the vacuum barriers separating neighbouring subsectors (a subsector corresponds to two 107 m long cells), which damaged them as can be seen in Fig. A.1. The forces displaced dipoles in the affected zone from their cold internal supports and knocked the short straight section (SSS) cryostats housing the quadrupoles and vacuum barriers from their external support jacks at three positions, in some locations breaking the anchors in the concrete floor of the tunnel. The displacement also damaged the connections to the cryogenic distribution line. The main damage zone extends over approximately 700 m.

About 2 tons of helium were rapidly released to the tunnel, producing a cloud which triggered oxygen deficiency hazard detectors and tripped an emergency stop, thus switching off all electrical power from sector 34. Before restoration of electrical power enabled the actuation of cryogenic valves, another 4 tons of helium were lost at lower flow rates. The total loss of inventory thus amounts to about 6 tons out of 15 tons initially in the sector.

A post-mortem analysis of cryogenic temperature data revealed a significant temperature anomaly in sector 34 during a powering step to 7 kA performed a few days before the incident. A steady temperature increase of up to 40 mK occurred in the cryogenic cell of the incident. The excess power in the incident cell corresponds to an unaccounted resistance of around 220 n $\Omega$ . Given the location of the primary electrical arc, the most likely hypothesis for the cause of the incident is a problem of the busbar joint. The structure of such a joint is shown in Fig. A.2. The joints are brazed but not clamped and the nominal joint resistance is 0.35 n $\Omega$ . The incident could be reproduced in simulation assuming a bad electrical and thermal contact of the copper stabilizer at the joint due to lack of solder or poor-quality brazing [41].

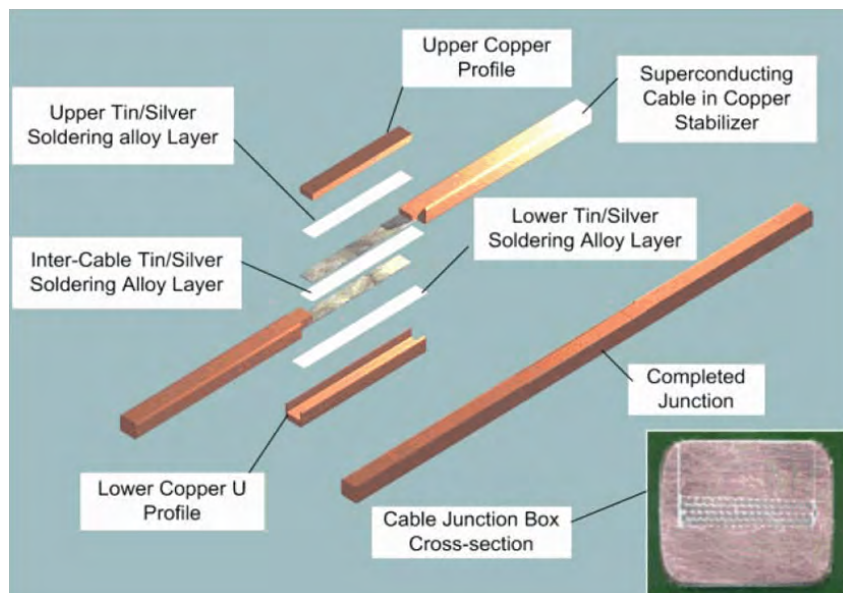
A total of 53 magnets, 39 dipoles and 14 quadrupole SSSs had to be removed from the tunnel and brought to the surface for cleaning and repair. Most of them were replaced with spare magnets. All magnets have been thoroughly re-tested before re-installation in the tunnel.

Both arc beam vacuum chambers were contaminated by soot from electrical arcs and chips of multilayer insulation over roughly 80% of their length as can be seen in Fig. A.3. Contamination by chips of multilayer insulation has been found over long distances away from the position of the original incident. These chips are deposited mostly on the beam screen surface, from where they are removed by in situ cleaning.

Following the incident and a review of the magnet and busbar protection system, a Quench Protection System (QPS) upgrade was launched to protect all busbar joints of the arc main dipole and main



**Fig. A.1:** A damaged interconnect between a quadrupole and a dipole magnet



**Fig. A.2:** Schematic of the main dipole busbar joint. The superconducting cable is embedded in a copper stabilizer. At the joint the two busbar ends are inserted with solder into a copper profile and welded.



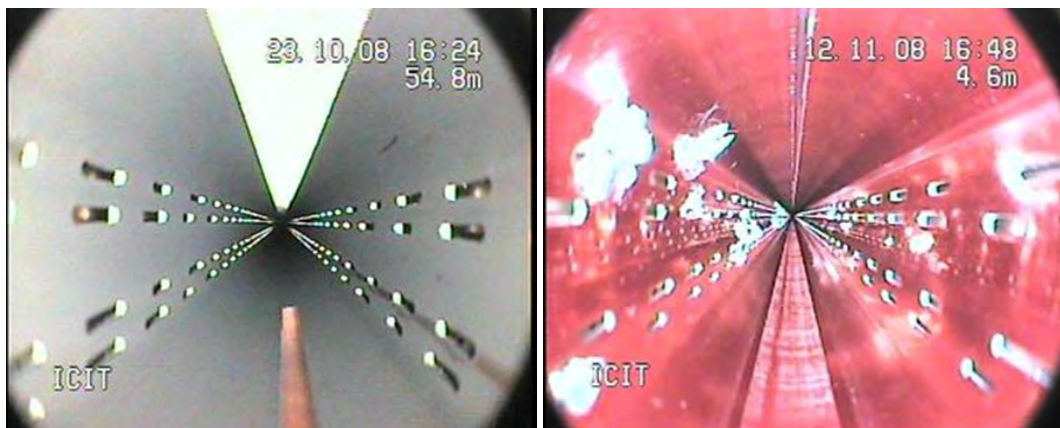


Fig. A.3: Example of a vacuum chamber beam screen covered with soot from the incident

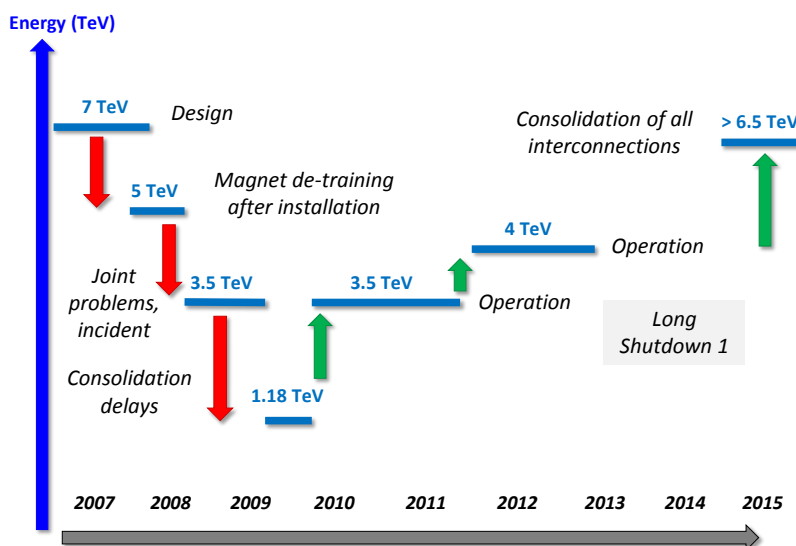


Fig. A.4: Evolution of the LHC beam energy from design to the present day

quadrupole circuits. The required voltage tabs were available, but a large volume of electronics had to be developed. The busbar protection was finally operational in the spring of 2010 when the LHC was commissioned at 3.5 TeV. The evolution of the LHC beam energy over time is shown in Fig. A.4, in 2015 the energy will be pushed to 6.5 TeV after the consolidation campaign of the 2 year long shutdown.



## Beam Cleaning and Collimation Systems

*S. Redaelli*

CERN, Geneva, Switzerland

### Abstract

Collimation systems in particle accelerators are designed to dispose of unavoidable losses safely and efficiently during beam operation. Different roles are required for different types of accelerator. The present state of the art in beam collimation is exemplified in high-intensity, high-energy superconducting hadron colliders, like the CERN Large Hadron Collider (LHC), where stored beam energies reach levels up to several orders of magnitude higher than the tiny energies required to quench cold magnets. Collimation systems are essential systems for the daily operation of these modern machines. In this document, the design of a multistage collimation system is reviewed, taking the LHC as an example case study. In this case, unprecedented cleaning performance has been achieved, together with a system complexity comparable to no other accelerator. Aspects related to collimator design and operational challenges of large collimation systems are also addressed.

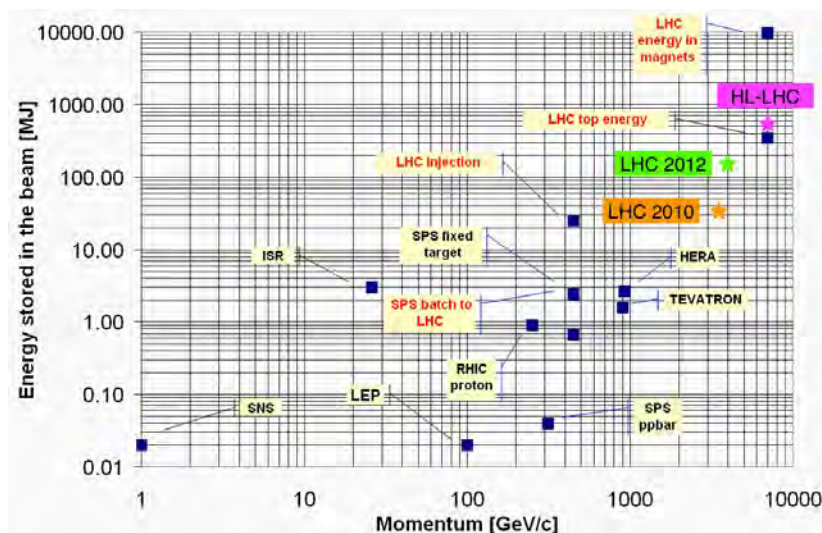
### Keywords

Beam collimation; multi-stage cleaning; beam losses; circular colliders; Large Hadron Collider.

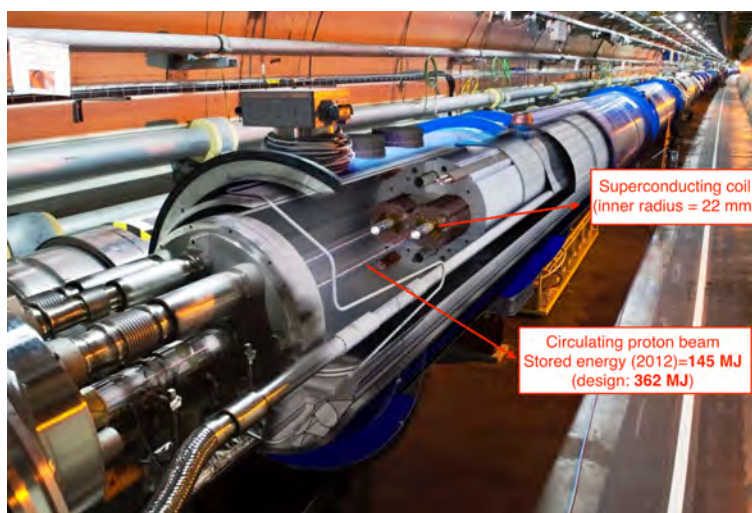
## 1 Introduction

The role of beam collimation systems in modern particle accelerators has become increasingly important in the quest for higher beam energies and intensities. For reference, the beam stored energy of recent and future particle accelerators is shown in Fig. 1, which includes the design (362 MJ) and achieved (150 MJ) values of the CERN Large Hadron Collider (LHC) [1], as well as the 700 MJ goal for its high-luminosity upgrade (HL-LHC) [2, 3]. High-power accelerators simply cannot operate without adequate systems to control unavoidable losses in standard beam operation. The operation and physics goals of recent superconducting, high-energy hadron colliders, such as the Tevatron [4], the Relativistic Heavy-Ion Collider [5], and the LHC, could not be fulfilled without adequate beam collimation. With the LHC, the design complexity and the performance of beam collimation has achieved unprecedented levels. This is required to ‘clean’ beam losses efficiently before they reach the small apertures of superconducting magnets. As illustrated in Fig. 2, the inner aperture of LHC magnets sits only a few centimetres apart from the circulating beams, which carry a total energy more than a billion times larger than that necessary to perturb the operation of superconductors.

In this document, the design of collimation systems for hadron accelerators is presented, with a special focus on the requirements and design aspects of high-energy and high-intensity machines. The general scope of a collimation system is to dispose, safely and in a controlled way, of beam losses that would otherwise occur at sensitive locations or on accelerator equipment that is not designed to withstand beam losses. In practice, this general definition finds its concrete implementations depending on the specific design goals required for a given accelerator. For example, collimation requirements are different for ‘warm’ high-power machines, where loss localization is crucial, than for ‘superconducting’ accelerators, where operation efficiency is ensured by keeping losses in cold magnets below quench limits. The roles of collimation systems in accelerators are discussed in Section 2. In Section 3, some basic notation is introduced and the inputs to collimation design from machine aperture and beam loss mechanisms are discussed. The design of a multistage collimation system is outlined in Section 4.



**Fig. 1:** Livingston-like plot of beam stored energy achieved and planned for different present and future particle accelerators. Courtesy of J. Wenninger.



**Fig. 2:** The LHC dipole in the tunnel, showing the cross-section of the magnet cold mass. The inner horizontal and vertical half dimensions of the dipole beam screen are 22 mm and 17 mm, respectively.

The second part of this document is focused on the presentation of the LHC collimation system as a case study. In Section 5, the system layout is reviewed and operational challenges for beam collimation are introduced, presenting the solutions deployed at the LHC. The LHC collimator design is discussed in Section 6 and the collimation performance achieved in LHC run I [6], at energies of up to 4 TeV and stored beam energies of 150 MJ, is reviewed in Section 7. The lecture is concluded with a brief review of advanced collimation concepts that are being considered for upgrading the present LHC collimation and for implementation in future accelerators under study. This is presented in Section 8.

## 2 Roles of collimation systems in particle accelerators

Typical roles of collimation systems are summarized in the following.

- **Cleaning of betatron and off-momentum beam halos:** Unavoidable beam losses of halo particles must be intercepted and safely disposed of before they reach sensitive equipment. The required cleaning performance depends on the design of the accelerator. The most challenging

requirements arise for superconducting accelerators, where loads from beam losses must remain below the quench limits of superconducting magnets. For example, the LHC design beam stored energy of 362 MJ has to be compared with typical quench limits of a few tens of  $\text{mW}/\text{cm}^3$  [7].

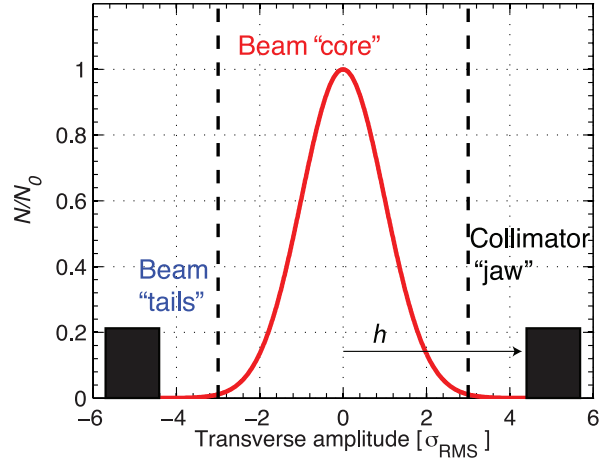
- **Passive machine protection:** Collimators are the closest elements to the circulating beam and represent the first line of defence in various normal and abnormal loss cases, as discussed in several companion lectures at this school. Owing to the damage potential of hadron beams, this functionality has become one of the most critical aspects of the operation of accelerators [8, 9], as well as a crucial input to the design of collimators that must withstand design failures.
- **Cleaning of collision products:** In colliders, this is achieved with dedicated movable collimators located in the outgoing beam paths of each high-luminosity experiment, to catch the products of collisions: direct collision debris and beam particles that emerge from the collision points with modified angles and energy.
- **Optimization of the experiment background** (i.e., minimization of halo-induced noise in detector measurements): this is one of the classical roles of collimation systems in previous colliders. Beam tail scraping or local shielding at the detector locations can reduce spurious signals in detectors (see, for example, a recent report [10]).
- **Concentration of radiation doses:** for high-power machines, it is becoming increasingly important to be able to localize beam losses in confined and optimized ‘hot’ areas rather than having a distribution of many activated areas along the machine. This is an essential design requirement for collimation systems, to allow easy access for maintenance in the largest fraction of the machine.
- **Local protection of equipment for improved lifetime against radiation effects:** Dedicated movable or fixed collimators are used to shield equipment locally. For example, passive absorbers are used in the LHC collimation inserts to reduce total doses, and to warm dipoles and quadrupoles that would otherwise have a short lifetime in the high-radiation environment foreseen during the nominal LHC operation. The exposure of radiation to equipment might not pose immediate limitations to operation of a machine but its optimization is crucial to ensure long-term reliability.
- **Beam halo scraping and halo diagnostics:** Though rarely a driving design criterion, the possibility to scan the beam distribution actively can be a very useful functionality of a collimation system. Collimator scanning in association with sensitive beam loss monitoring systems proved a powerful method of probing the population of beam tails [11, 12], which are otherwise too small, compared with the beam core, to be measured by conventional emittance measurements. Thanks to their robustness, the LHC primary collimators can be efficiently used to scrape and shape the beams, as in Ref. [13]. Full beam scraping also provides precise, though destructive, measurements of beam sizes.

A collimation system might typically fulfil several roles. For example, the concentration of radiation losses or the reduction of experimental background are natural by-products of a very efficient beam collimation design. Conversely, before designing a collimation system, it is important to identify the driving requirements for its design in a specific accelerator. For the LHC, the driving design criterion is halo cleaning, which must be excellent to operate the machine below the quench limit of the superconducting magnets at maximum beam energy. It is interesting to note that the present LHC beam collimation [1, 14] is quite special in that it fulfils all the roles listed, thanks to a careful design that has been extended beyond the cleaning functionality. The price to pay for this performance is the unprecedented complexity, which poses important operational challenges, as discussed in Section 7.

### 3 Inputs to collimation design from aperture and beam loss mechanisms

#### 3.1 Basic definitions for collimation and beam halo

Particles with transverse amplitudes or energy deviations significantly larger than those of the reference particle are referred to as *beam halo particles*. One can distinguish between *betatron* and *off-momentum*



**Fig. 3:** Gaussian distribution, which is typically adequate to model the particle distribution of the beam core (red line). Overpopulated tails may be intercepted by collimator jaws, which constrain particle motion at a given transverse betatron amplitude.

halos, which are formed in the case of larger-than-nominal transverse emittance or energy error, respectively. The transverse amplitude of a particle  $i$  around a closed orbit,  $z \equiv (x, y)$ , can be expressed as a function of the longitudinal curvilinear coordinate  $s$  for the Twiss parameters  $\beta_z(s)$ ,  $D_z(s)$ , and  $\phi_z(s)$  as

$$z_i(s) = \sqrt{\beta_z(s)\epsilon_{z,i}} \sin[\phi_z(s) + \phi_{z,i,0}] + \left(\frac{\delta p}{p}\right)_i D_z(s), \quad (1)$$

where  $\epsilon_{z,i}$  is the single-particle emittance,  $(\delta p/p)_i$  is the energy error, and  $\phi_{z,i,0}$  is an arbitrary phase. The r.m.s. size of the beam at location  $s$  is then given by

$$\sqrt{\beta_z(s)\epsilon_z + \left(\frac{\delta p}{p}\right)^2 D_z^2(s)}, \quad (2)$$

where  $\epsilon_z$  and  $\delta p/p$  are the r.m.s. transverse emittance and energy spread of the beam. The notation to express the machine aperture and the collimator settings will use, unless specified otherwise, the *betatron beam size*,

$$\sigma_z(s) = \sqrt{\beta_z(s)\epsilon_z}, \quad (3)$$

which takes into account only the contribution to the beam size from the betatron motion. Collimator settings might then be given in normalized units as

$$n_\sigma = \frac{h}{\sigma_z}, \quad (4)$$

where  $h$  is the distance in millimetres between the collimator jaw and the circulating beam (e.g., the half gap of a two-sided collimator centred around the beam, as shown in Fig. 3).

The distinction between halo and core particles is, to a certain extent, arbitrary. For Gaussian distributions, one may define as halo particles those with amplitudes above three r.m.s. deviations of the Gaussian (see dashed lines in Fig. 3), i.e., with emittances larger than  $9\epsilon_z$ . For a beam with perfect two-dimensional Gaussian distributions in the  $(z, z')$  plane, about 1.1% of the total beam particles have amplitudes above  $3\sigma_z$  and 0.03% above  $4\sigma_z$ , respectively. Particles this far out from the beam core are rarely of any use for accelerators and are more likely to cause nuisances (beam losses, irradiation of components, background in detectors, etc.). For off-momentum halos, a similar definition could be adopted. For other purposes in circular accelerators, one might consider as halo the particles outside the

RF bucket that are lost when beams are accelerated or in the presence of synchrotron radiation (which is non-negligible at the LHC).

Beam collimation is achieved by placing blocks of material, the *collimator jaws*, close to the circulating beams, to constrain the betatron amplitudes of stray particles outside the core. This is shown schematically by the black boxes in Fig. 3. Collimation of off-momentum tails might be achieved in a similar way as for betatron tails, by placing collimators at locations of high dispersion, where the particle's energy shift results in a transverse offset, as in the second term on the right-hand side of Eq. (1).

How close to the beam a collimator should be depends on various aspects that will be elaborated on in the rest of this paper. In particular, it will be shown that the outer limit for a collimator setting depends on the available machine aperture that needs to be protected and on the cleaning performance that needs to be achieved. The inner limit depends essentially on how collimators perturb beam stability through an increase in the machine impedance. Tighter settings also typically lead to higher beam losses and tighter positioning tolerances against orbit movements and optics errors. Thus, collimators should not be set closer to the beams than is strictly necessary.

### 3.2 Collimation cleaning inefficiency

The cleaning performance of a collimation system is measured by the *collimation efficiency*, a figure of merit that expresses the fraction of halo particles ‘caught’ by the system over the total lost from the beam. A perfect beam collimation provides 100% cleaning, i.e., there is no beam loss at sensitive equipment. Alternatively, the *cleaning inefficiency*,  $\eta_c$ , can be introduced as the relative fraction of beam that ‘leaks’ to other accelerator components,  $A_{\text{lost}}$ , compared with what is intercepted and safely disposed of by the collimators,  $A_{\text{coll}}$ :

$$\eta_c = \frac{A_{\text{lost}}}{A_{\text{coll}}} . \quad (5)$$

The relevant measure of ‘beam loss’, indicated as  $A$  in this equation, has to be identified for the specific design criteria that the collimation system addresses.

The LHC beam collimation requirements are driven by the challenge to keep beam losses below the quench limits of the superconducting magnets. In this case, the inefficiency  $\eta_c$  is defined as the number of protons lost as a fraction of the total number of particles absorbed by the collimation system. The *local cleaning inefficiency*,  $\tilde{\eta}_c \equiv \tilde{\eta}_c(s)$ , is defined as a function of the longitudinal coordinate  $s$  as the fractional loss per unit length,

$$\tilde{\eta}_c = \frac{\eta_c}{L_{\text{dil}}} = \frac{N(s \rightarrow s + \Delta s)}{N_{\text{abs}}} \frac{1}{\Delta s} , \quad (6)$$

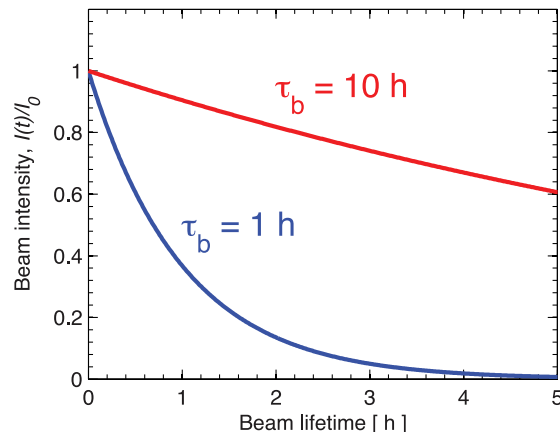
where  $N(s \rightarrow s + \Delta s)$  is the number of particles lost over the distance  $\Delta s$ , i.e., in the longitudinal range  $(s, s + \Delta s)$ , and  $N_{\text{abs}}$  is the number of particles absorbed by the collimation system.

This definition has the advantage that it can be directly compared with the quench limits of superconducting magnets if a proper *dilution length* is chosen. Indeed, for the LHC it was estimated [15] that the quench limits in units of proton lost per unit length,  $R_q$ , are

$$R_q^{\text{inj}} = 7.0 \times 10^8 \text{ protons}/(\text{m} \cdot \text{s}) \text{ (450 GeV)} , \quad (7)$$

$$R_q^{\text{top}} = 7.6 \times 10^6 \text{ protons}/(\text{m} \cdot \text{s}) \text{ (7 TeV)} , \quad (8)$$

for beams at injection (inj) and top (top) energies, respectively. These approximate figures were used early in the LHC design phase and in first collimation performance estimates [16]. Although nowadays detailed simulation tools and more adequate models are available to compare peaks of energy deposition in the magnet coils directly against quench limits of superconducting cables (see, for example, Ref. [7]), the formalism introduced here is very useful to introduce challenges for collimation design, as discussed next.



**Fig. 4:** Relative reduction of beam current versus time,  $I(t)/I_0$ , for beam lifetime values of 1 h and 10 h

### 3.3 Beam lifetime and loss modelling

There are various mechanisms that lead to losses in particle accelerators, as also discussed in companion lectures at this school [8, 9]. One can distinguish between *regular* and *abnormal beam losses*, referring to unavoidable losses that occur during standard operation, as opposed to losses caused by failures of accelerator systems or by wrong beam manipulations. For both categories, losses might occur over a broad range of time-scales, from a fraction of a single turn to tens of seconds.

In circular colliders, a main source of loss comes from the collisions of the opposing beams that cause burn-off of beam particles. Other sources of loss are interactions with residual gas, intrabeam scattering, beam instabilities of various types (single-bunch, collective, beam–beam, etc.), the noise of feedback systems used to stabilize beams, transverse and longitudinal resonances, include RF noise. Other losses inherent to operation phases of the accelerator, such as capture losses at the beginning of the ramp, injection and dump losses, losses during dynamics changes of the operational cycle (orbit drifts, optics changes, energy ramp, etc.), are referred to as ‘operational losses’ [17].

Ignoring, for the moment, very fast loss scenarios and their impact on collimator design [8, 9], let us consider the requirements for beam collimation in the presence of *diffusive losses*. In this case, the transverse increase of particle action per turn is much smaller than one sigma of the r.m.s. distribution. Rather than treating each loss mechanism in detail, losses are modelled by considering the beam lifetime.

The time-dependent circulating beam intensity,  $I(t)$ , can, for most practical purposes, be modelled by an exponential decay function whose time constant,  $\tau_b \equiv \tau_b(t)$ , defines the *beam lifetime* as

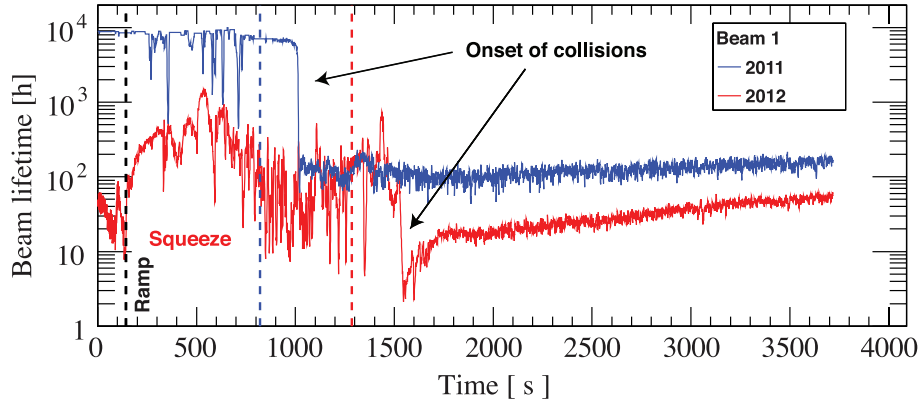
$$I(t) = I_0 e^{-t/\tau_b}, \quad (9)$$

for an initial beam current  $I_0$ . After a time  $\tau_b$ , the total beam current is reduced to about 37%. Example profiles of relative beam intensity versus time,  $I(t)/I_0$ , are shown in Fig. 4 for lifetime values of 1 h and 10 h. In a linear approximation, beam loss rates,  $dI/dt$ , are inversely proportional to  $\tau_b$  and can be calculated as

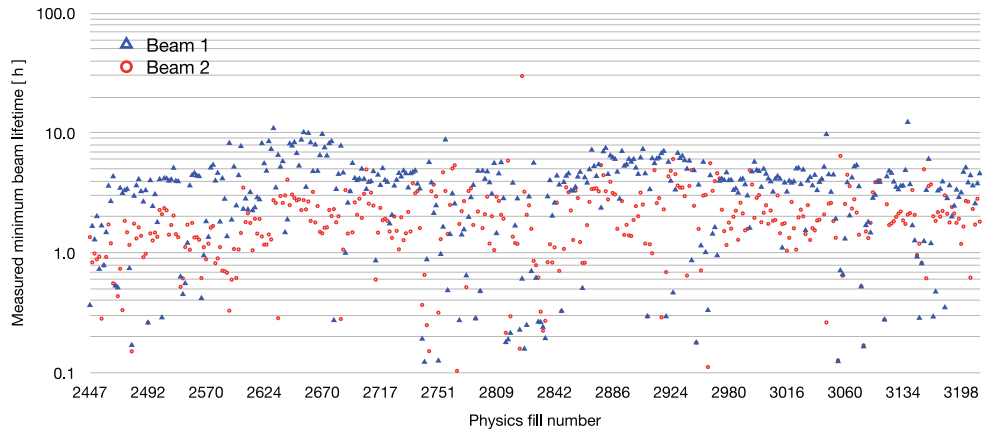
$$-\frac{1}{I} \frac{dI}{dt} = \frac{1}{\tau_b}. \quad (10)$$

It is important to emphasize that  $\tau_b(t)$  is indeed a function of time and is not constant through the operational cycle. The sources of beam losses introduced previously—operational losses and other accelerator physics mechanisms—occur at different times and might become apparent as drops of beam lifetime at given times in the cycle. An example of measured  $\tau_b$  during LHC fills for physics is shown in Fig. 5. In 2012, proton beams were accelerated to 4 TeV, whereas in 2011 the maximum energy was 3.5 TeV. The machine configuration and TCP settings were different in these runs.





**Fig. 5:** Measured beam lifetime LHC during two typical LHC physics fills in 2011 at 3.5 TeV (blue) and in 2012 to 4 TeV (red), as a function of time in the cycle. The ramp and squeeze durations were different, so the onset of collisions (reduction in  $\tau_b$  indicated by black arrows) started at different times. The primary halo cut in the betatron cleaning insertion changed from  $5.7\sigma_z$  to  $4.3\sigma_z$  for  $\epsilon_z = 3.5 \mu\text{m}$ . Courtesy of B. Salvachua.



**Fig. 6:** Minimum lifetime measured during the squeeze of physics fills in 2012 as a function of fill number (to be considered as arbitrary unit). Courtesy of B. Salvachua.

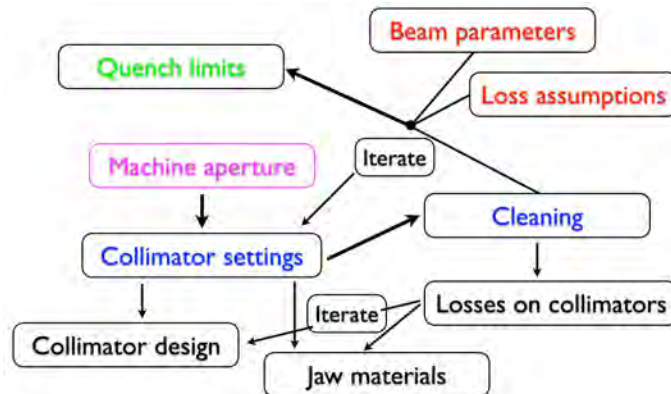
A collimation system must be designed to cope with the maximum expected rates of beam loss. This is determined by the *minimum allowed beam lifetime*,  $\tau_b^{\text{min}}$ , throughout the operational cycle, most notably during phases at maximum energy (flat-top, squeeze, collision preparation, and physics data recording) when the beam stored energy is largest. The design value used to specify the LHC collimation system is  $\tau_b^{\text{min}} = 0.2 \text{ h}$  for up to a maximum time of 10 s [15]. The minimum lifetime measured during the squeeze process in 2012 is shown in Fig. 6. Values of  $\tau_b$  below 1 h were recorded on a regular basis, with several cases even below 0.2 h. The same behaviour at higher beam intensity and energy, as expected in 2015, will cause frequent beam dumps, with a severe impact on LHC operation.

## 4 Design of a multistage collimation system

### 4.1 Design requirements and work flow

The LHC case of beam collimation in the presence of high-energy and high-intensity proton beams is considered here. For the collimation system to fulfil the required cleaning goals, it must be ensured that:

- (1) the aperture bottlenecks of the accelerators are geometrically shielded such that, for all loss scenarios, primary beam losses hit first collimators;



**Fig. 7:** The various ingredients and competencies required to design a complete collimation system

- (2) the total energy carried by the beam, i.e., out-scattered beam particles and the secondary products of beam particles' interactions with the collimator matter, is absorbed within the collimation region, with tolerable leakage to sensitive equipment, notably to cold magnets, for any relevant loss scenarios;
- (3) the collimators themselves and other equipment installed in the collimation regions must withstand, without damage, beam losses for different design scenarios;
- (4) the contribution to machine impedance from the collimator materials approached to the beam must be tolerable and ensure that the high-intensity beams remain stable.

This last aspect is particularly critical when it comes to designing collimators and absorbers. For more details, see a companion lecture [18]. The complete design of a complex system like that of the LHC requires several steps and iterations between different domains that go well beyond the field of accelerator physics. Figure 7 shows relevant steps towards a complete design of a collimation system. As indicated, several iterations are required.

The understanding of the machine aperture and the required cleaning determines a specification of the collimator settings that can be used to calculate the collimation cleaning. With the input of loss assumption and machine parameters, the first cleaning estimates are compared with the quench limits. This initial design phase is worked out in this section. Closing this loop provides a first conceptual design, which is then used for detailed collimator design. Cleaning simulations also provide distributions of losses on the collimators that are used, in an iterative process, to specify adequate collimator design and jaw material choices. This aspect will not be discussed in this document (see Ref. [18] for more detail).

## 4.2 Beam cleaning specifications

The total design beam intensity of the LHC beams is  $I_{\text{tot}} = 2808 \times 1.15 \times 10^{11}$  protons, i.e.,  $3.2 \times 10^{11}$  protons, where  $n_b = 2808$  is the number of bunches and  $I_b = 1.15 \times 10^{11}$  protons is the bunch population. At the minimum allowed lifetime of 0.2 h, this corresponds to a proton loss rate of  $4.4 \times 10^{11}$  protons/s. At 7 TeV, the beam stored energy is 362 MJ and loss rates approach 500 kW. By expressing the quench limits in the approximate formulation of Eq. (8), one can derive a specification for the local cleaning inefficiency in a cold magnet as

$$\tilde{\eta}_c \leq \frac{1}{10000} [1/\text{m}]. \quad (11)$$

Although simulation tools are now available to combine particle tracking and energy deposition simulations, so as to evaluate precisely the energy lost in the superconducting magnet coils, it is very useful to

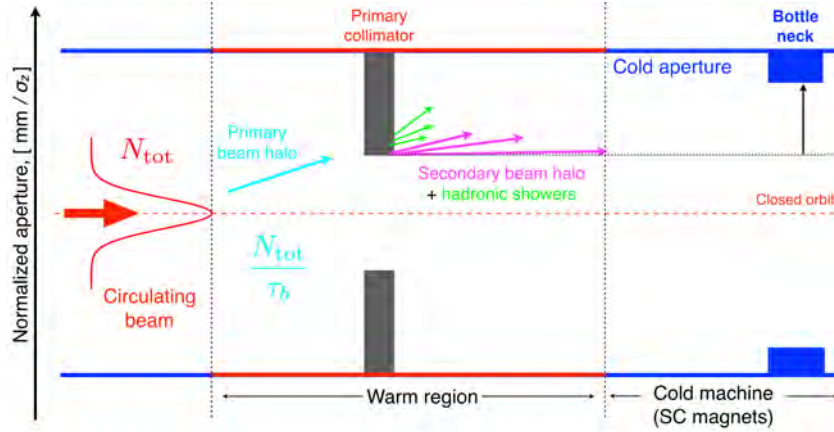


Fig. 8: Single-stage collimation system: SC, superconducting

follow this approximate approach. This formalism provides a powerful tool for designing a collimation system. The number of protons lost per unit length can be simulated with a fast and accurate set-up [19], which provides an essential design optimization tool. Final validation of collimation solutions then follows, using more sophisticated tools that also involve energy deposition simulations [20].

### 4.3 Machine aperture and collimator settings

To design a betatron cleaning system, one must first compute the available aperture of the accelerator. Let us assume that the circulating beam sees an isolated *aperture bottleneck*,  $A_{\min,z}$ , in the transverse plane  $z$ . This is defined as the smallest normalized transverse aperture at any location around the ring. For convenience, the aperture is normalized by the local betatron beam size  $\sigma_z$ . The limiting location is calculated as

$$\hat{A}_{\min,z} = \min \left[ \frac{A_z(s)}{\sigma_z(s)} \right], \quad (12)$$

where the minimum is calculated for all locations  $s$  around the ring and  $A_z(s)$  is the distance in millimetres between the circulating beam and the mechanical aperture. This quantity can be measured directly in an accelerator [21, 22]. While, operationally, it is convenient to express  $A_{\min,z}$  for a given beam emittance, what matters is actually the *aperture acceptance*  $A_z(s)/\sqrt{\beta_z}$ . A cold aperture bottleneck is indicated in Fig. 8 as a blue box, and projects into the nominal aperture.

The minimum aperture calculated during the LHC design phase [16] for both planes and beams at injection and top energy are listed in Table 1. Calculations relied on a conservative approach [23] that ensured adequate margins during beam commissioning. While measurements during LHC run I [24] indicated that the LHC aperture is indeed larger than assumed, such conservative figures are considered in this lecture for the collimation design.

### 4.4 Single-stage collimation

Designing a collimation system involves finding an optics solution and an arrangement of collimators that ensure that losses in cold magnets remain below the quench limits for all design loss rates. In an ideal machine without beam losses, there would be no need for beam collimation, if the minimum machine aperture were at a safe distance from the beam core. In practice, various beam loss mechanisms cause outwards drifts of halo particles, which eventually hit the aperture if there is no mechanism to intercept them. The deposited energy at this location would then depend on the primary beam loss rates,  $N_{\text{tot}}/\tau_b$ .

One could build a simple single-stage collimation system by placing a primary collimator (TCP; ‘target collimator, primary’) that intercepts beam losses. Preferably, collimators are placed in a warm region, as far as possible from superconducting magnets. The collimator jaws must be set at a transverse

**Table 1:** Minimum horizontal and vertical apertures at injection (450 GeV) and top energy (7 TeV,  $\beta^* = 0.55$  m) for warm and cold elements, as estimated in the LHC design phase [16].

|               | 450 GeV |      | 7 TeV |      |
|---------------|---------|------|-------|------|
|               | Warm    | Cold | Warm  | Cold |
| <b>Beam 1</b> |         |      |       |      |
| Horizontal    | 6.8     | 7.9  | 28    | 8.9  |
| Vertical      | 7.7     | 7.8  | 8.3   | 8.4  |
| <b>Beam 2</b> |         |      |       |      |
| Horizontal    | 6.7     | 7.7  | 28    | 8.1  |
| Vertical      | 7.7     | 7.6  | 8.7   | 8.8  |

aperture below that of the machine bottleneck,  $\hat{A}_{\text{TCP}} \leq \hat{A}_{\text{min},z}$ . This simple system would work if the TCP were a black absorber that could stop all the primary particles at their first passage through the jaw. Also note that, because of the mixing of positive and negative amplitudes of halo particles from the betatron motion, a single-jaw collimator suffices to protect the aperture against slow diffusive losses. (For standard losses, impact parameters in the submicrometre range are expected [25]. At this scale, particles do not see the full jaw length at their first passage because of jaw flatness and surface roughness errors. This increases the inefficiency of a single-stage cleaning system, as more turns are required before particles accumulate enough interactions with the TCP.)

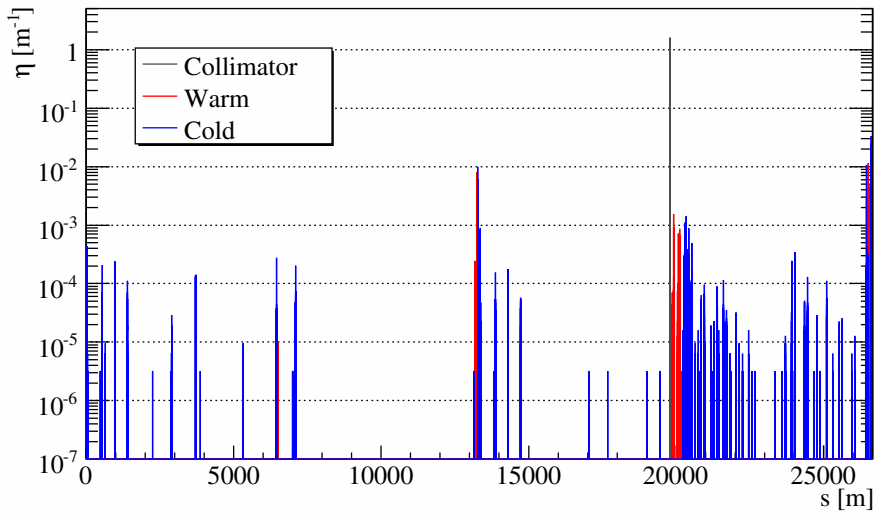
The single-stage system of Fig. 8 does not provide sufficiently efficient halo cleaning. The halo protons that are out-scattered before being absorbed by the jaw material leave the collimator at larger normalized amplitudes and modified energies. These particles populate the so-called *secondary beam halo*, which risks being lost in the machine before interacting again with the collimator in subsequent turns. In addition, the products of hadronic and electromagnetic showers are not contained in the collimator volume and might reach sensitive elements without additional downstream collimators or absorbers.

The cleaning performance of the single-stage system described here was simulated under the assumption that a horizontal TCP is installed in the current LHC betatron cleaning insert. The tools in Ref. [19] allow one to calculate the number of halo protons lost in the collimators and machine aperture. Simulations properly model the proton tracking through the magnetic elements and the scattering in the collimator materials. In Figs. 9 and 10, the predicted local cleaning inefficiency of Eq. (6) is given as a function of the longitudinal coordinate  $s$ .

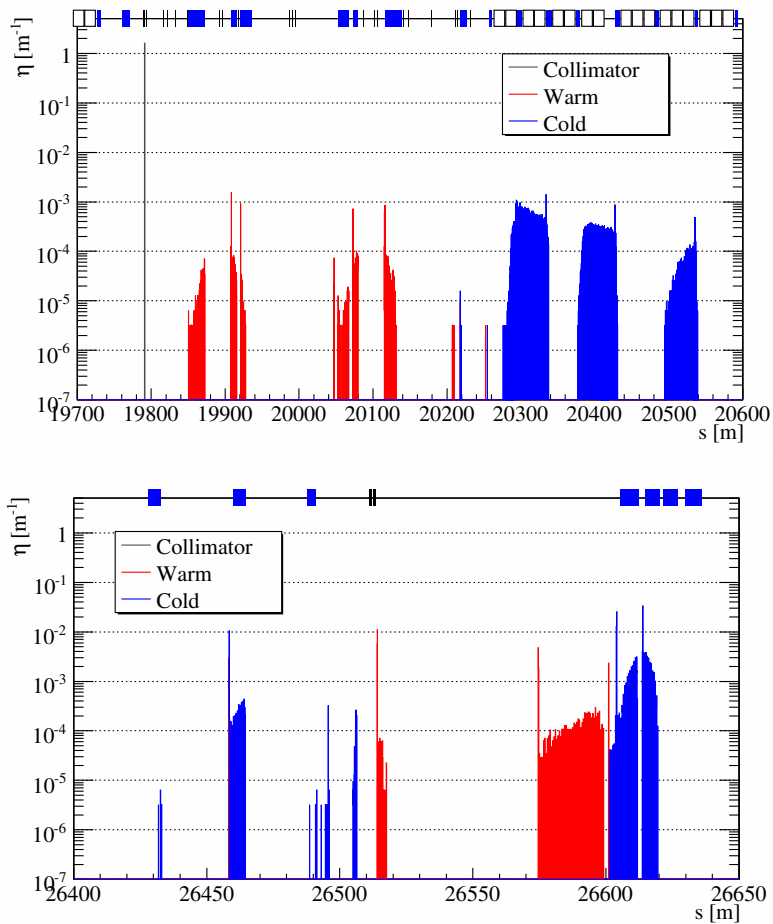
In these cleaning inefficiency plots, black peaks indicate losses at collimators (only one TCP in this case), blue peaks indicate losses at cold magnets and red peaks indicate losses at warm elements. It can be seen, by looking at Fig. 10, which shows zoomed plots around various interaction points, that cold losses reach cleaning inefficiency levels of up to 0.01 /m. This estimate, which is made for a perfect machine without errors, and which does not take into account the energy deposited by hadronic showers, indicates losses at least two orders of magnitude higher than the value specified in Eq. (11). One can therefore conclude that a single-stage collimation system is inadequate for high-intensity superconducting machines, such as the LHC.

#### 4.5 Multistage collimation

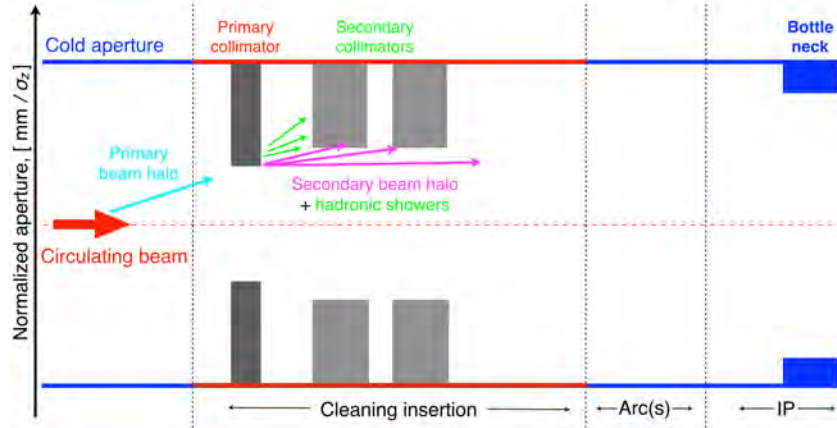
The performance of a single-stage cleaning system can be improved with additional collimators downstream of the TCP to catch the secondary halo particles, as shown in Fig. 11. These are called secondary collimators (target collimators, secondary; TCSs) and are typically longer than TCPs, to maximize the absorption of particles out-scattered at the TCPs. On the one hand, the TCS aperture must be larger



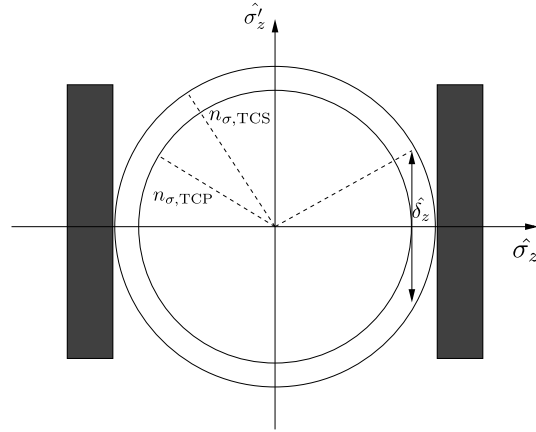
**Fig. 9:** Simulated cleaning inefficiency at the LHC for a single-stage collimation system achieved with one horizontal primary collimator (TCP) located at the beginning of the LHC warm betatron cleaning insert. The position of the existing primary collimators, i.e.,  $s = 19.8$  km, is used. Courtesy of D. Mirarchi.



**Fig. 10:** Enlargement of Fig. 9 in the regions immediately downstream of the cleaning insertion (top) and upstream of the ATLAS experiment (bottom). Courtesy of D. Mirarchi.



**Fig. 11:** Two-stage beam collimation system, obtained by adding a set of secondary (TCS) collimators to the single-stage cleaning system of Fig. 8. IP, interaction point.



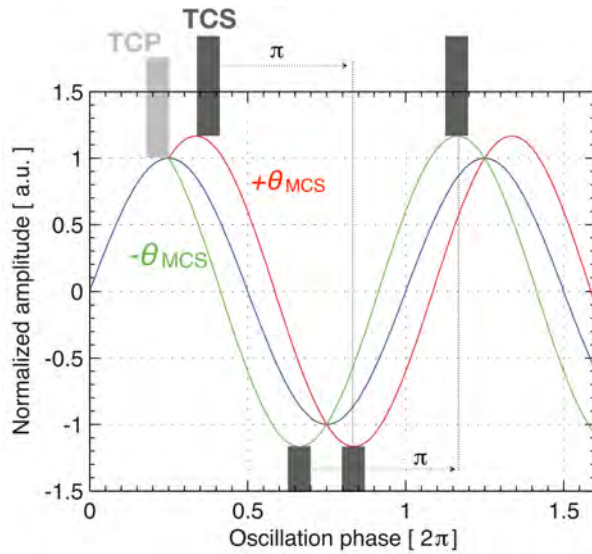
**Fig. 12:** Normalized phase space with the circumferences radii  $n_{\sigma,TCP}$  and  $n_{\sigma,TCS}$ . A normalized kick  $\hat{\delta}'$ , as in Eq. (13), is necessary for halo particles impinging on the TCP to reach the TCS aperture.

than that of the TCP, to ensure that the *collimation hierarchy* is respected without the risk of a TCS becoming closer to the beam than the TCP, which would result in a single-stage system similar to the one discussed earlier. On the other hand, the TCS aperture should be small enough to maximize its efficiency in catching the particles out-scattered by the TCPs. From Fig. 12, one can calculate the kick of particles impinging on the TCP necessary to reach the amplitude of the TCS as

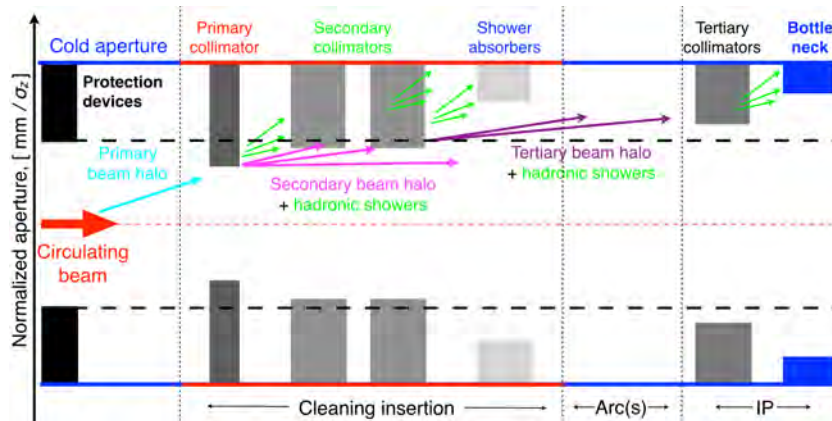
$$\hat{\delta}' = \frac{\delta'}{\sigma'} = \sqrt{n_{\sigma,TCS}^2 - n_{\sigma,TCP}^2}, \quad (13)$$

where  $\sigma' = \sqrt{\epsilon/\beta}$  is the r.m.s. divergence. Such a kick is typically accumulated after multiple passages through the TCP. For a given TCS–TCP retraction, the longitudinal positions of the TCS collimators must be optimized to intercept secondary halo particles. This is illustrated in Fig. 13 for a one-dimensional case. This condition is respected at betatron phase advances, where the multiple Coulomb scattering angle translates into maximum offsets in the collimation plane.

The problem of optimum phase locations for a two-stage collimation system is worked out in detail in Ref. [26]. Finding a solution is more complicated than appears in Fig. 13 because scattering occurs in all directions. A one-dimensional model is thus not adequate. However, it can be demonstrated that an arrangement of primary and secondary collimators in three planes (horizontal, vertical and skew) can be found to ensure satisfactory multiturn cleaning [26].



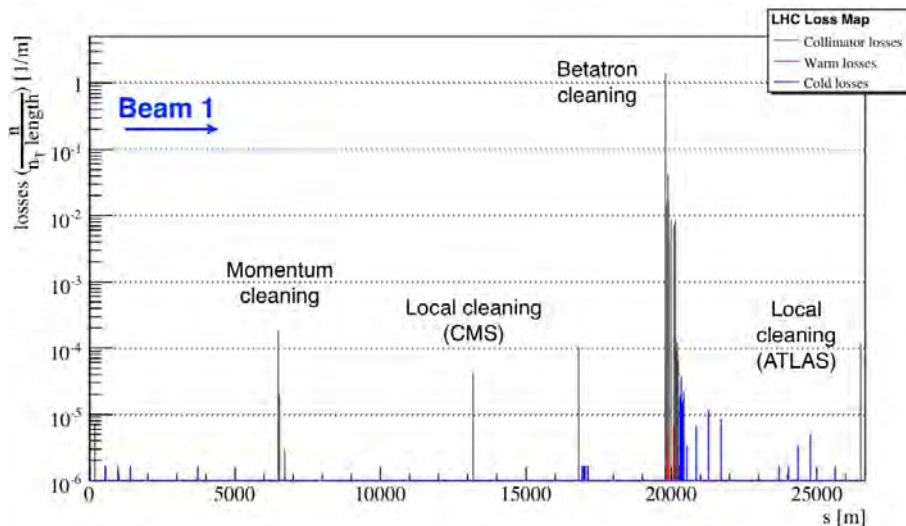
**Fig. 13:** Qualitative definition of optimum locations for secondary collimators in a two-stage system, in which TCSs must intercept beam particles out-scattered at the primary collimators. In this one-dimensional model, two phase locations exist, where the amplitudes caused by multiple Coulomb scattering are a maximum for the two signs of the scattering angle,  $\pm\theta_{MCS}$ .



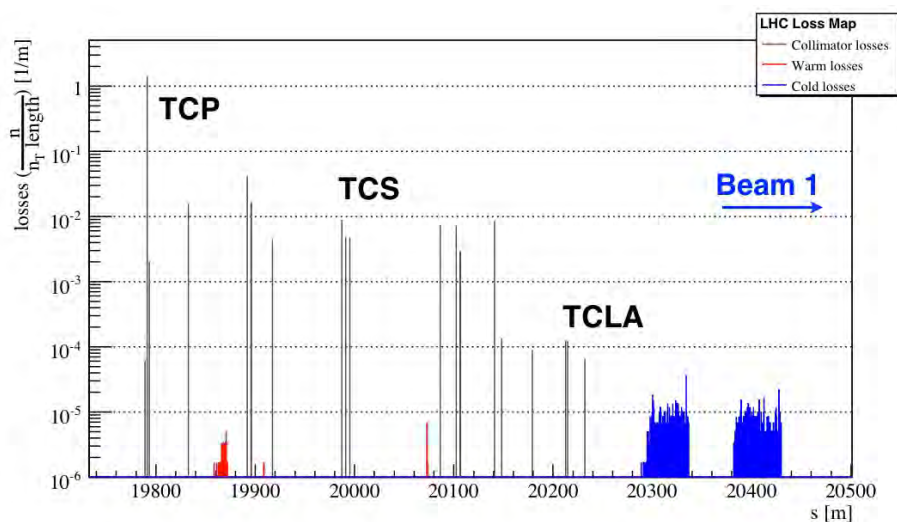
**Fig. 14:** Key elements of the LHC multistage collimation system: IP, interaction point

Detailed performance analysis of a two-stage cleaning process for the LHC was conducted in the design phase [16]. While this scheme can ensure efficient shielding of the LHC aperture from transverse halo losses, it is not sufficient to absorb products of hadronic showers before they reach cold magnets downstream of the cleaning insert. Moreover, a two-stage system localized in a single insertion is not adequate for the local protection of critical bottlenecks that might be exposed to losses, notably the triplet magnets around the experiments that become critical during the squeeze. The collimation system of the LHC has therefore evolved into a *multistage collimation system* that includes, in addition to TCPs and TCSs, tertiary collimators (target collimators, tertiary; TCTs) in front of critical bottlenecks, shower absorbers in the warm cleaning inserts, and protection devices in the dump region, to shield the machine in case of dump kicker failures. The LHC multistage collimation system is shown in Fig. 14.

The cleaning performance of the final LHC collimation system [14] is shown in Fig. 15. While the system is described in detail in the next section, the simulations are shown here for a direct comparison with the single-stage system. The insertion regions (IRs) where the largest losses occur are the betatron (IR7) and momentum (IR3) cleaning, ATLAS (IR1) and CMS (IR5). This simulation is for beam 1 (B1),



**Fig. 15:** Local cleaning inefficiency as a function of  $s$  for the final collimation system of the LHC run I. Loss distributions are simulated for the LHC beam 1 at 7 TeV for a perfect machine, with the collision optics squeeze to  $\beta^* = 0.55$  m in IR1 and IR5. Courtesy of D. Mirarchi.



**Fig. 16:** Enlargement of the IR7 region of the cleaning inefficiency plot of Fig. 15. Labels indicate the approximate locations of the three families of collimator in IR7. TCLA, target collimator long absorber; TCP, target collimator (primary); TCS, target collimator (secondary).

nominally 7 TeV, in collision conditions. An enlargement of the loss map around the betatron cleaning insert is shown in Fig. 16. For a perfect machine, cold losses are now below  $\sim 10^{-5}$ . The highest peaks are localized in the dispersion suppressor regions downstream of IR7.

## 5 The LHC collimation system

The LHC collimation system was designed to handle proton beams of a stored energy of 362 MJ and is now being upgraded to cope with the design HL-LHC goal of about 700 MJ per beam. A complex and distributed system is needed to achieve the excellent halo cleaning required to operate the LHC below quench limits. In this section, the collimation layout is presented and the collimator design is



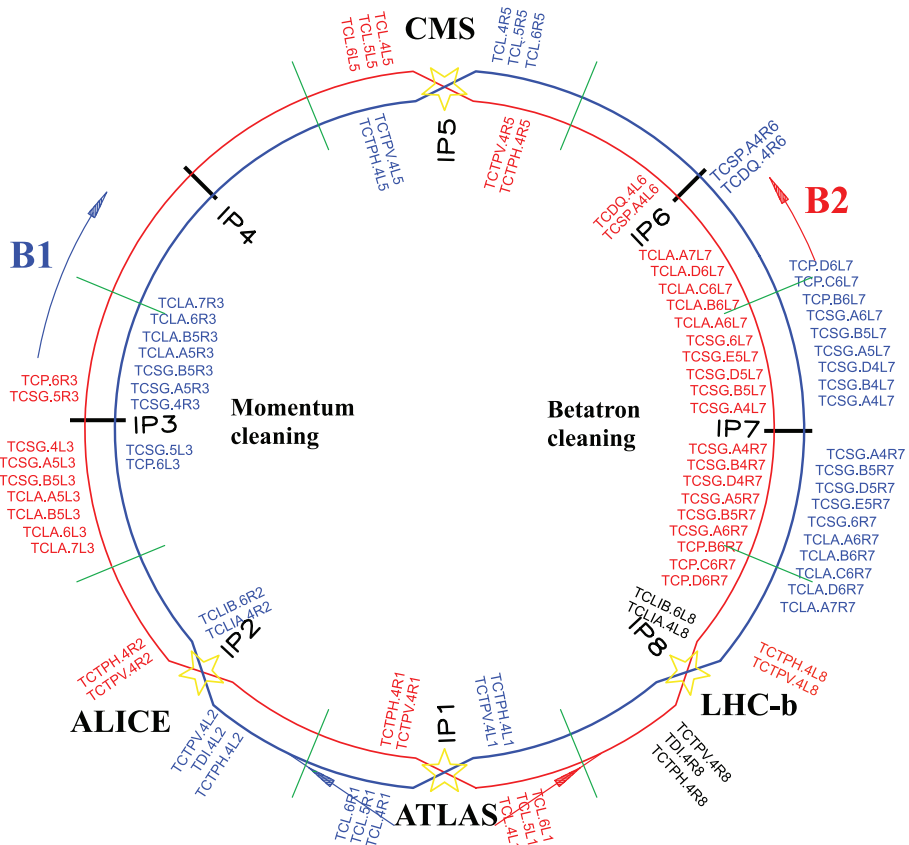


Fig. 17: Layout of the LHC, showing the collimator locations around the ring

reviewed. Operational challenges for the collimation at the LHC are then introduced, presenting the solutions produced to set the system up for optimum performance in all operational phases.

### 5.1 LHC ring collimation layout

Figure 17 shows the LHC layout and the positions of the collimators around the ring. A list of collimator types, with a description of their functionality (primary, secondary, etc.) and key collimator properties is given in Table 2. Including the dump protection block (target collimator dump quadrupole, TCDQ) and the injection protection collimator (target dump, injector TDI), the system deployed for the 2015 LHC operation comprises 110 movable collimators installed in the LHC ring and its transfer lines.

Halo collimation is achieved by the multistage cleaning system introduced in Section 4. This comprises three stages in IR3 (momentum cleaning) and IR7 (betatron cleaning), where the primary collimators (TCPs), closest to the beam, are followed by secondary collimators (TCSs) and active absorbers (TCLAs). For optimal performance, the particles in the beam halo should first hit a TCP, and the TCSs should only intercept secondary halo particles that have been already scattered by, and escaped from, upstream collimators. The TCPs and TCSs, which are the closest collimators to the beam and hence intercept large beam losses, are made of a carbon fibre composite (CFC) to ensure high robustness. These collimators are also more likely to be hit by the beam if there is a failure. The TCLAs catch tertiary halo particles scattered out of the TCSs, as well as showers from upstream collimators. The TCLAs are made of a tungsten alloy, in order to stop as much as possible of the incoming energy. However, they are not as robust as the CFC collimators and should therefore never intercept primary beam losses. The setting hierarchy is chosen to ensure that this condition is respected in all operation state.

In addition to the dedicated inserts in IR7 and IR3, there are collimators in most other IRs. A pair of tertiary collimators (target collimators, tertiary, pick-up; TCTPs), made of a tungsten alloy, are

**Table 2:** List of movable LHC collimators for run II. CFC, carbon fibre composite; H, horizontal; S, skew; V, vertical.

| Functional type              | Name | Plane | Number | Material   |
|------------------------------|------|-------|--------|------------|
| Primary IR3                  | TCP  | H     | 2      | CFC        |
| Secondary IR3                | TCSG | H     | 8      | CFC        |
| Absorbers IR3                | TCLA | H,V   | 8      | W alloy    |
| Primary IR7                  | TCP  | H,V,S | 6      | CFC        |
| Secondary IR7                | TCSG | H,V,S | 22     | CFC        |
| Absorbers IR7                | TCLA | H,V   | 10     | W alloy    |
| Tertiary IR1/2/5/8           | TCTP | H,V   | 16     | W          |
| Physics debris absorber      | TCL  | H     | 12     | Cu/W alloy |
| Dump protection              | TCSP | H     | 2      | CFC        |
|                              | TCDQ | H     | 2      | C          |
| Injection protection (lines) | TCDI | H,V   | 13     | CFC        |
| Injection protection (ring)  | TDI  | V     | 2      | C          |
|                              | TCLI | V     | 4      | CFC        |
|                              | TCDD | V     | 1      | CFC        |

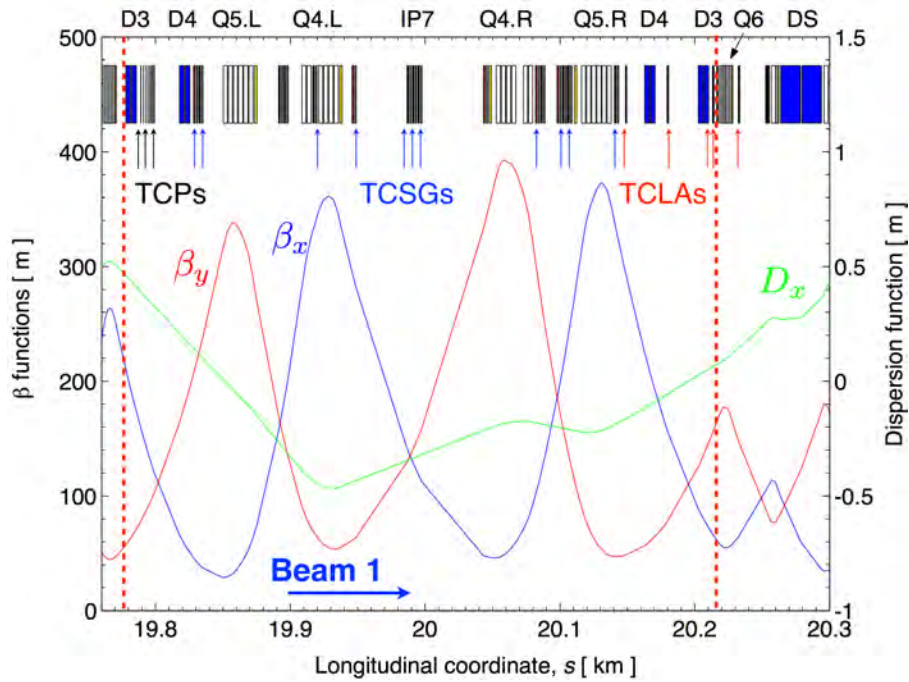
installed in both beams about 150 m upstream of the collision points for all experiments, one TCTP in the horizontal plane (TCTPH) and one in the vertical (TCTPV). They provide local protection of the quadrupole triplets in the final focusing system, which are the limiting cold apertures during physics operation. They are also important for decreasing the experimental background. Downstream of the high-luminosity experiments, ATLAS and CMS, there are three TCLs (target collimator, long) per beam, to intercept the collision debris. Furthermore, at the beam extraction in IR6, dump protection collimators are installed as a protection against miskicked beams in the case of extraction failures. Similarly, there are injection protection collimators in IR2 and IR8.

During the long LHC shutdown in 2013 and 2014, 18 new collimators based on a beam position monitor design [27], in which beam position monitor pick-ups are embedded in the jaws to measure the beam position at the collimator location, have been installed. They replaced the TCSGs (target collimator, secondary, graphite) in IR6 and the tertiary collimators in all experiments, as these locations are considered more critical for orbit control, in order to enhance LHC performance [28]. These collimators are called TCTP and TCSP, where ‘P’ stands for pick-up.

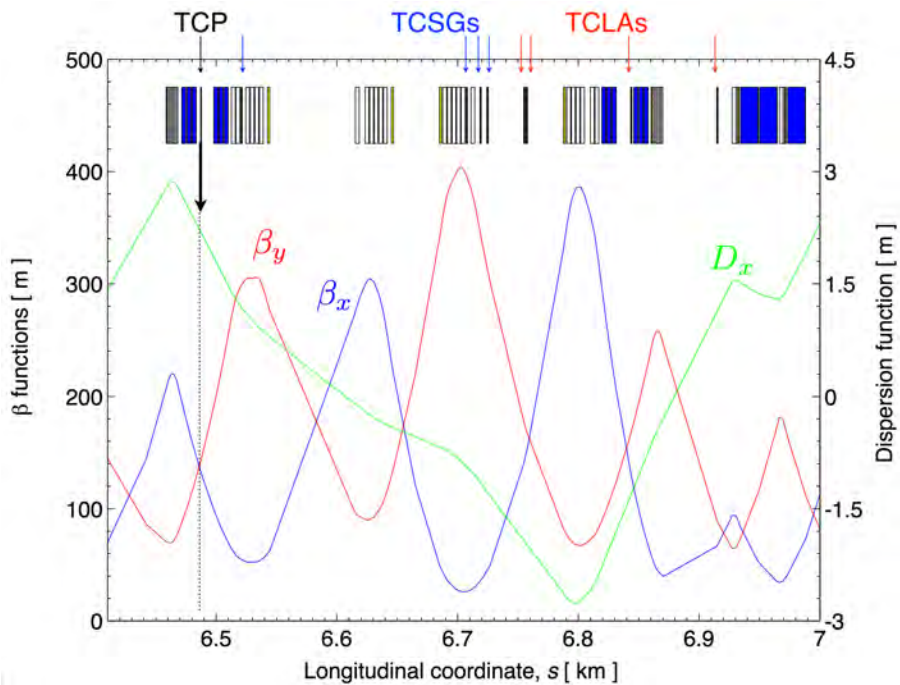
## 5.2 Optics and layout of cleaning inserts

The optics and layout of the betatron and momentum cleaning inserts are shown in Figs. 18 and 19, respectively. In both inserts, four *dog-leg* dipoles, called D4 and D3, are placed symmetrically on either side of the ‘IP7’, and are used to enlarge the beam–beam separation from 194 mm to 224 mm, making more transverse space for collimators. The two D4 magnets also delimit the  $\approx 500$  m long warm insert, which comprises the warm quadrupoles Q4 and Q5. The Q6 quadrupoles on either side of the D4 dipoles are the first superconducting magnets before the beam enters the cold arc.

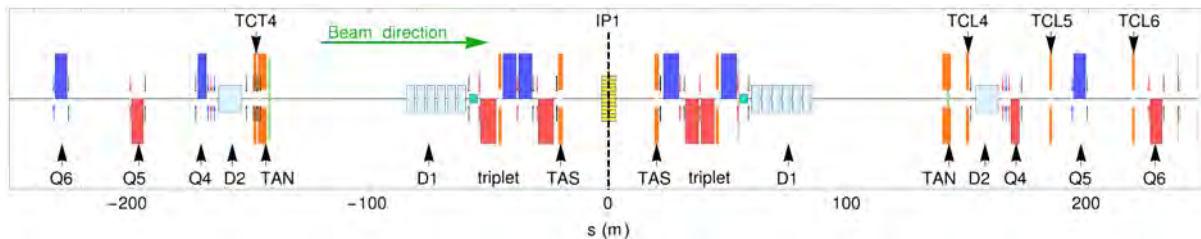
In IR7, three primary collimators intercept horizontal, vertical, and skew halos. They are located in the region between the D3 and D4 dipoles, i.e., on the upstream side of the warm insertion for each beam. This maximizes the length of the warm section downstream of the primary loss location. A similar implementation is adopted in IR3, where, however, only one horizontal TCP is needed, placed at a location with large normalized dispersion,  $D_x/\sqrt{\beta_x}$ , to intercept particles with energy deviations. Momentum cleaning in one plane is sufficient, as at the LHC, vertical dispersion is negligible. The IR3



**Fig. 18:** Betatron ( $\beta_x, \beta_y$ ) and dispersion ( $D_x$ ) functions as a function of  $s$  in the LHC betatron cleaning insertion IR7. The main layout elements are also shown: quadrupoles (white boxes), dipoles (blue), and collimators (black). Vertical arrows indicate the installed collimators: 3 TCPs, 11 TCSGs; 5 TCLAs. Vertical red dashed lines indicate the limits of the warm regions (Q6 magnets at either side of IP7 are the first cold magnets). D, dipole magnet; IP, interaction point; L, left; R, right; Q, quadrupole magnet; TCLA, target collimator long absorber; TCP, target collimator (primary); TCSG, target collimator (secondary, graphite).



**Fig. 19:** Betatron ( $\beta_x, \beta_y$ ) and dispersion ( $D_x$ ) functions as a function of  $s$  for B1 in the LHC momentum cleaning insertion IR3. Vertical arrows indicate the installed collimators: 1 TCP, 4 TCSGs; 4 TCLAs. The main layout elements are also shown: quadrupoles (white boxes), dipoles (blue), and collimators (black). TCLA, target collimator long absorber; TCP, target collimator (primary); TCSG, target collimator (secondary, graphite).



**Fig. 20:** Layout elements around IR1 (ATLAS) in the 2015 configuration of the LHC collimation system. D, dipole; IP, insertion point; Q, quadrupole; TAN, target absorber (neutral); TAS, target absorber (secondary); TCL, target collimator (long); TCT, target collimator (tertiary). Courtesy of R. Bruce.

primary collimator needs to be at larger transverse betatron amplitudes than those of the IR7, to decouple the functionalities of the two inserts. Typical transverse betatron amplitudes expressed in units of  $\sigma_x$  as in Eq. (3) are 2.5–3 times larger than in IR7, to ensure that IR3 does not act as a betatron system for particles with small energy errors.

The collimators of IR3 and IR7 are indicated in Figs. 18 and 19 by black boxes. Eleven TCS collimators are used in IR7, whereas four are used in IR3, since collimation occurs in one plane only. Five active absorbers (TCLAs) are used in IR7 and four in IR3. These devices, of types TCP, TCSG, and TCLA (see Table 2) are all two-sided collimators. Even if a one-sided collimator might be sufficient for a multiturn cleaning process, two-sided collimators are crucial for precise alignment of the circulating beams.

The layout of IR1 (ATLAS) is shown in Fig. 20. A pair of horizontal and vertical TCTPs protect the triplet from incoming beam losses. Three TCL-type physics debris absorbers protect the magnets downstream of the IR from collision products. The other high-luminosity experiment, CMS in IR5, has an equivalent layout. For IR2 (ALICE) and IR8 (LHCb), there is no need for a TCL collimator because the lower luminosity values do not put the matching sections at risk of quenching.

In addition to the movable collimators, 10 passive absorbers are also mounted in front of the most exposed warm magnets of each collimation insert: the D3 magnets downstream of the TCPs and the first modules of the Q5 and Q4 quadrupoles. These fixed-aperture collimators, called TCAPs, dramatically reduce the radiation doses to magnet coils, increasing their lifetimes by a factor of 10 or more (chapter 18 of [1]).

### 5.3 Operational challenges and beam-based set-up

#### 5.3.1 LHC operational cycle and recap. of machine configurations

The main phases of the LHC operational cycle, which is periodically run to prepare for periods of physics data acquisition (‘stable beam’ mode), are injection, energy ramp, betatron squeeze, and preparation of collisions (‘adjust’ mode). The squeeze, in which the optics around the interaction points are changed to reduce the colliding beam sizes, has so far been performed at constant flat-top energy. In this phase, the betatron function is enlarged at the inner triplets as required to reduce the  $\beta^*$  values, i.e., the beta functions at the collision points.

The LHC design value of  $\beta^*$  for the high-luminosity points IP1 (ATLAS) and IP5 (CMS) is 55 cm for a beam energy of 7 TeV, limited by the available triplet aperture. During LHC run I, a  $\beta^*$  value of 60 cm was achieved at 4 TeV. The first year, 2015, of LHC run II started with a  $\beta^*$  value of 80 cm at 6.5 TeV, to ease recommissioning after the 2 year shutdown [29] but it is planned to move to a  $\beta^*$  value close to 40 cm in 2016. These excellent results were achieved thanks to a better aperture than had been anticipated during the LHC design phases, which was also better than the one used to specify various LHC systems. For the scope of this lecture, it is still useful to review the system design by starting from the design values.

### 5.3.2 Collimation settings strategy in the LHC operational cycle

The LHC aperture was reviewed in Section 3; see Table 1. With an injection stored energy of 22 MJ, i.e., not only above the quench limit but also significantly above the damage limit of metals [18], beam collimation is required in every phase of the LHC operational cycle, from injection to collision. Particularly challenging are the dynamic phases (energy ramp, betatron squeeze, and change of orbit configurations), when collimator movements must be synchronized precisely with other accelerator systems, such as power converters and RF units. This operation mode imposes tight constraints on the collimator control design.

At the injection, distributed aperture bottlenecks are expected in the arcs, as the magnet aperture was designed to fit the beams at injection [1]. At 7 TeV, the arc aperture is no longer critical because the betatron amplitudes are damped at larger beam energies. The aperture is now limited by the triplets, where  $\beta$  functions of up to  $\approx 4500$  m are required to achieve small beam sizes at the interaction points. By design (see Table 1), the normalized apertures,  $\hat{A}_{\min,z}$ , are actually similar for the two extreme cases. Thus, even if the accelerator physics motivations are different, similar collimator settings are deployed at injection and in physics conditions. This involves moving collimators to follow the shrinking beam envelope.

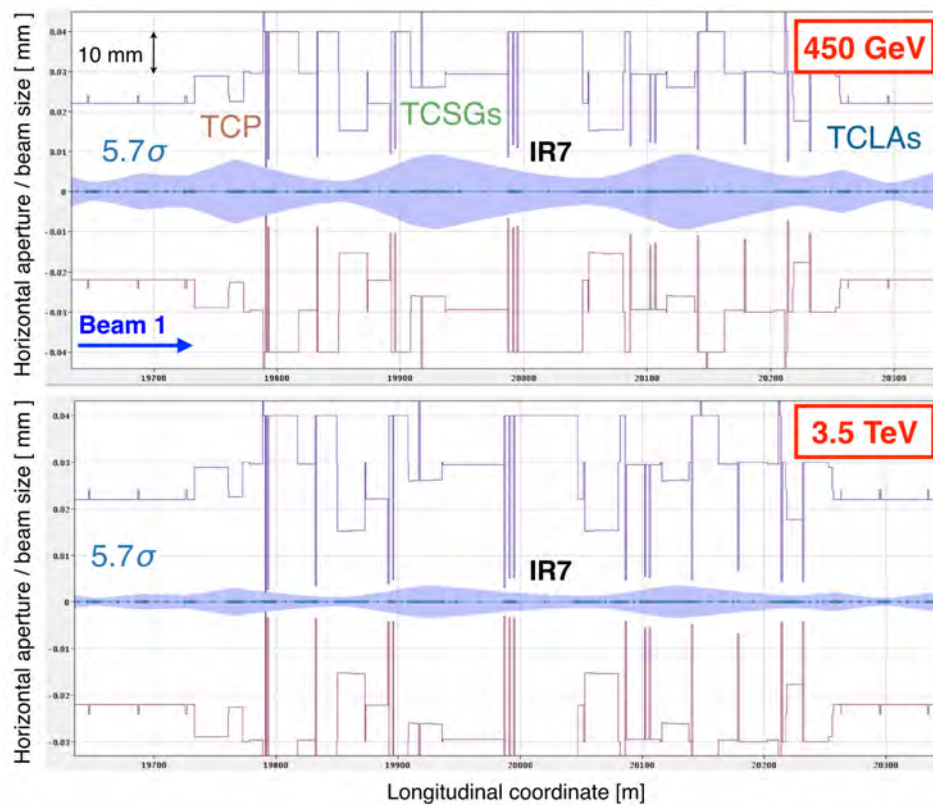
Figure 21 shows an example of collimator settings at injection (top) and 3.5 TeV (bottom), taken from the operation configuration of the LHC 2010 run [30]. The horizontal beam envelope at  $5.7\sigma_x$ , as defined by the TCP gaps, is shown, together with the values of the collimator half gap projected on the horizontal plane at each collimator (magenta bars). The TCPs were kept at a normalized aperture of  $5.7\sigma_z$  at all energies. The TCSGs were moved from  $6.7\sigma_z$  to  $8.5\sigma_z$ , and the TCLAs were moved from  $10.0\sigma_z$  to  $17.7\sigma_z$ . These relaxed top-energy settings were conceived to reduce the operational tolerances in the first year of the run [30] and were then subsequently tightened to improve the cleaning performance [31], reaching  $4.3\sigma_z$  in 2012. The collimator gap values in millimetres, as used for the 4 TeV operation at  $\beta^* = 60$  cm are shown in Fig. 22, where the transverse clearance left by the IR7 primary collimators and the distribution of gaps are shown. The smallest gap is 2.1 mm.

It is clear from Fig. 21 that a basic requirement for the LHC collimator design is that the jaws must be movable, as the gaps required at top energy to ensure optimum performance are not compatible with the larger beam sizes at injection. The need for small gaps at top energy also has important effects on the operational strategy of the collimation system because it necessitates dedicated beam-based alignment procedures, as collimators cannot be set deterministically to such small gaps without direct measurements to ‘find’ the local beam position and size.

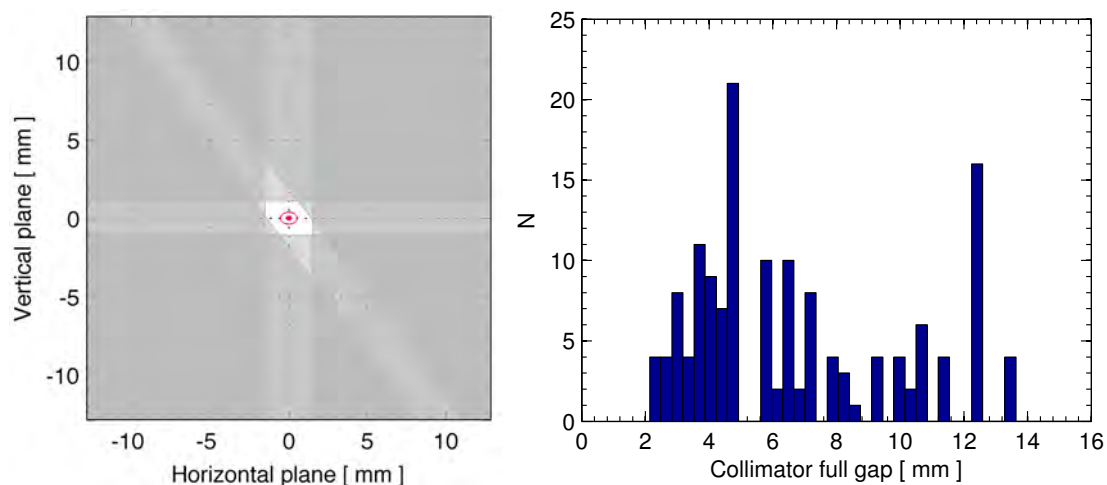
### 5.3.3 Beam-based set-up of LHC collimators

The LHC collimation system performance relies on respecting the well-defined hierarchy between collimator families. In practice, this involves knowing the beam orbit and beam size at each collimator, as shown in Fig. 23. With beam sizes as small as  $200 \mu\text{m}$  and orbit offsets of up to 2–3 mm, and in the presence of collimator alignment errors of up to a few hundred micrometres, the determination of optimum jaw positions can only be achieved through a series of measurements aimed at measuring the required parameters, which are referred to as *beam-based collimator alignments*.

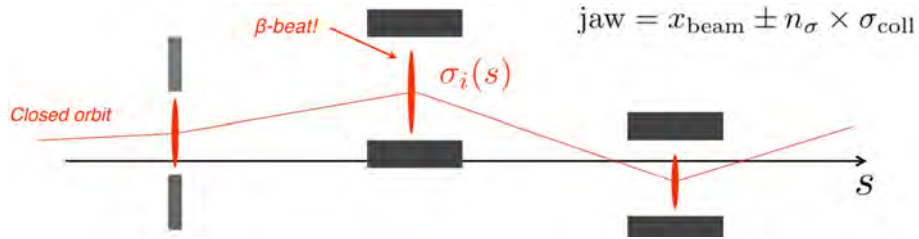
The procedure for collimation set-up at the LHC (Fig. 24) was established [30,32] based on experience gained with a prototype LHC collimator installed for beam tests in the Super Proton Synchrotron (SPS) [33]. The beam halo is shaped with a reference collimator (1), typically a primary collimator, which is closed to a known half gap of  $3 - 5\sigma$ . This reference halo is used to cross-align other collimators, by moving their jaws towards the beam in small steps of 5–20  $\mu\text{m}$  until the halo is *touched*, with symmetrical beam loss responses from either jaw (2). This gives the local orbit position. The reference collimator is then closed further (3) until it touches the halo again: this enables the gaps of the two collimators to be cross-calibrated. The average of the initial and final gaps of the reference collimator in



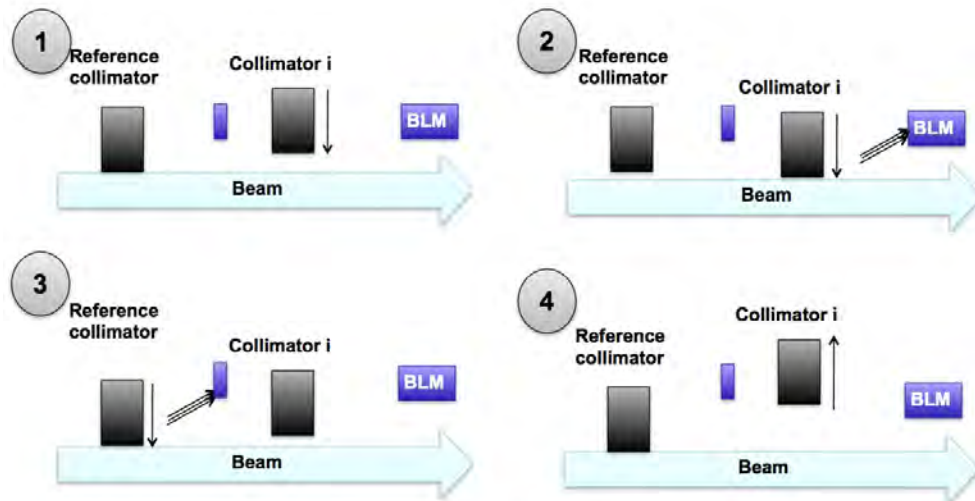
**Fig. 21:** Horizontal aperture, collimator jaw positions (vertical bars) and  $5.7\sigma$  beam envelope at (top) injection and (bottom) 3.5 TeV in betatron cleaning (IR7) from the LHC on-line model application [30]. IR, insertion region; TCG, target collimator (graphite); TCLA, target collimator (long absorber); TCP, target collimator (primary).



**Fig. 22:** Left: Beam clearance for the LHC beams, as defined by the primary collimator gaps. Right: Distribution of collimator gaps, as adopted for operation at 4 TeV and  $\beta^* = 60$  cm in 2012. In 2015, the same IR7 settings in millimetres are used for the 6.5 TeV operation.



**Fig. 23:** Collimator jaw positions at various locations in the ring, where the closed orbit and beam size are different. Proper collimator set-up requires direct measurements of beam position and size, to ensure that the collimator hierarchy is respected.



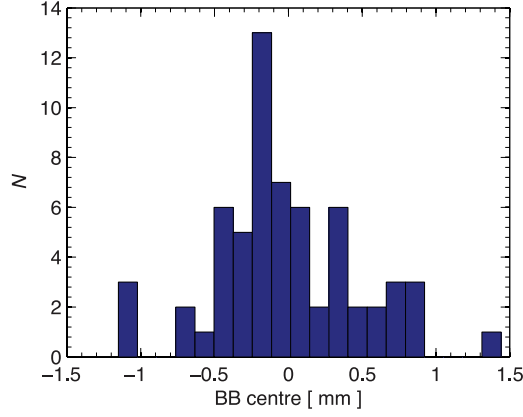
**Fig. 24:** The collimator set-up procedure used to determine beam orbit and relative beam size to that of a reference collimator, for operational settings generation [32]. BLM, beam loss monitor.

units of  $n_\sigma$  gives the normalized gap of the other collimator. Finally, the latter collimator is opened to its nominal settings (4). This ensures that the relative retraction with respect to the reference collimator is respected, even in the presence of different beta-beating at the two locations. An example of *beam-based collimator centres* measured in 2012 is shown in Fig. 25. This reinforces the previous assertion that beam-based alignment is mandatory for a proper collimator set-up at the LHC.

This set-up procedure is precise but time consuming. During the initial commissioning in 2010, it was carried out manually for each collimator. An automated feedback system between collimator movements and the beam loss monitor signal has been developed, enabling the set-up time to be improved significantly and dramatically reducing the number of spurious beam aborts from human error. A detailed treatment of this optimization of beam-based alignment is beyond the scope of this lecture but can be found in Ref. [34].

### 5.3.4 Collimator setting generation for operation

Beam-based alignment must be done for each collimator in the ring, for every relevant machine configuration (injection, top energy before and after squeeze, collision). To minimize the risk of damaging the collimators while approaching them to the beams, the alignment is carried out with the minimum intensity that allows reliable orbit measurements, i.e., with a few bunches of nominal bunch intensity. Let us now assume that the local orbit,  $x_{\text{beam}}$ , and beam size,  $\sigma_{\text{coll}}$ , are calculated at every collimator in each discrete point of the operational cycle.



**Fig. 25:** Distribution of beam-based centres of LHC collimators as a result of the alignment campaign of 2012. Shifts of up to more than 1.5 mm are found from the cumulative effects of orbit misalignments, electronics offsets of the beam position monitor system, alignment error of the collimators with respect to the reference orbit, etc.

While collimators are installed in a variety of azimuthal orientations (see Fig. 26), the jaw movement is in one dimension, along the collimator plane. For arbitrary collimator angles  $\theta_{\text{coll}}$ , the *effective* beam size in the collimation plane,  $\sigma_{\text{coll}}$  is computed from the horizontal and vertical sizes as

$$\sigma_{\text{coll}} = \sqrt{\sigma_x^2 \cos(\theta_{\text{coll}})^2 + \sigma_y^2 \sin(\theta_{\text{coll}})^2}, \quad (14)$$

where  $\sigma_z$ ,  $z \equiv (x, y)$ , is calculated as in Eq. (3). The collimator half gap is calculated as  $h = n_\sigma \times \sigma_{\text{coll}}$  and the jaw positions around the beam position,  $x_{\text{beam}}$ , are given by

$$\text{jaw} = x_{\text{beam}} \pm n_\sigma \times \sigma_{\text{coll}}. \quad (15)$$

Note that each jaw has two motors, which allow the tilt angle to be adjusted with respect to the beam envelope. In the following, the tilt angle is assumed to be zero. Stepping motors can be driven through arbitrary functions of time. The motion of collimators around the ring can be synchronized through timing events at the microsecond level [35, 36]. This is necessary to ensure optimum collimator settings during critical machine phases, such as the energy ramp and the betatron squeeze. To this end, continuous setting functions must be generated from the beam-based parameters through scaling rules versus beam energy and optics.

Let us calculate, for example, the ramp functions, starting from settings values at injection ('0') and flat-top ('1'). The half gap during the energy ramp is expressed as a function of the energy:

$$h(\gamma) = n_\sigma(\gamma) \times \sigma_{\text{coll}}(\gamma), \quad (16)$$

where  $\gamma = \gamma(t)$  is the relativistic gamma function. For the LHC, it is sufficient to use linear functions in  $\gamma$  for  $n_\sigma$  and  $\sigma_{\text{coll}}$ . A linear interpolation between the beam-based parameters at injection and flat-top yields:

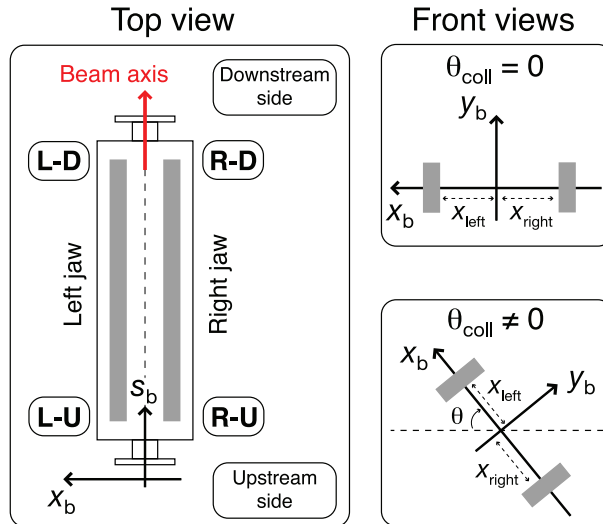
$$h(\gamma) = \left[ n_{\sigma,0} + \frac{n_{\sigma,1} - n_{\sigma,0}}{\gamma_1 - \gamma_0} (\gamma - \gamma_0) \right] \times \frac{1}{\sqrt{\gamma}} \left[ \frac{\sqrt{\epsilon_1 \beta_1} - \sqrt{\epsilon_0 \beta_0}}{\gamma_1 - \gamma_0} (\gamma - \gamma_0) \right]. \quad (17)$$

The beam centre is also expressed as a linear function of  $\gamma$  to give the jaw position as

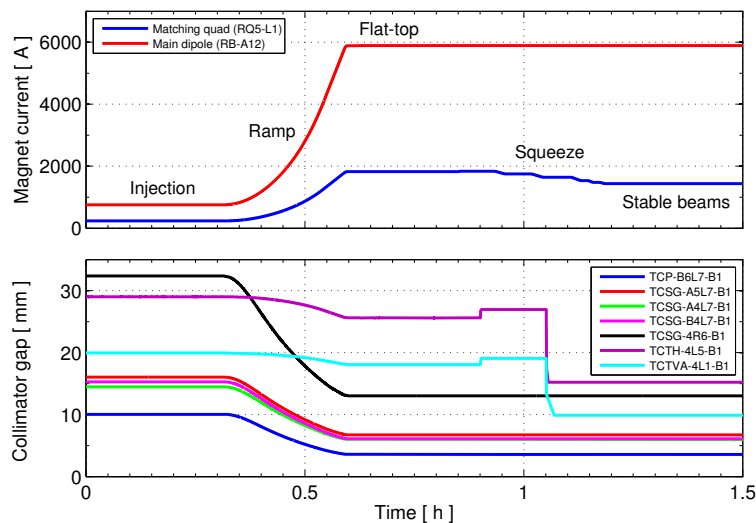
$$\text{jaw}(\gamma) = \left[ x_{\text{beam},0} + \frac{x_{\text{beam},1} - x_{\text{beam},0}}{\gamma_1 - \gamma_0} (\gamma - \gamma_0) \right] \pm h(\gamma). \quad (18)$$

Note that the beam size  $\sigma_{\text{coll}} = \sigma_{\text{coll}}(\gamma)$  is also a function of the optics and therefore might change, typically for the tertiary collimators in the experimental regions, during the betatron squeeze [37]. This





**Fig. 26:** Top and front views of a collimator, with labels and naming conventions. Each jaw has two motors that move the jaws in the collimation plane: horizontal ( $\theta_{coll} = 0$ ), vertical ( $\theta_{coll} = \pi/2$ ) or skew planes. D, downstream; L, left; R, right; U, upstream.



**Fig. 27:** Operational cycle for selected collimators for a typical LHC fill. Top: Measured magnet currents versus time. Bottom: Collimator gaps versus time.

notation can be generalized in a straightforward way by considering functions of  $\beta^*$  instead of  $\gamma$  for the parameters involved. An example of collimator gaps versus time during a full LHC cycle is given in Fig. 27 (bottom graph), together with the LHC dipole and matching quadrupole currents, to indicate the times of the ramp and squeeze phases (top graph).

The operation of the collimation system is automated by sequences that are run at every fill, enabling operation crews to run smoothly through the different sets of the cycle settings. The operation mode can only work thanks to the excellent stability of the LHC orbit and optics and of the collimator hardware itself. So far, only one beam-based alignment per year has been required [38].

**Table 3:** Minimal horizontal and vertical apertures at injection (450 GeV) for warm and cold elements

| Parameter                         | Value                                |
|-----------------------------------|--------------------------------------|
| High stored beam energy           | 360 MJ/beam                          |
| Large transverse energy density   | 1 GJ/mm <sup>2</sup>                 |
| Activation of collimation inserts | 1–15 mSv/h                           |
| Small spot sizes at high energy   | ≈200 μm                              |
| Collimation close to beam         | 6–7σ                                 |
| Small collimator gaps             | 2.1 mm (at 7 TeV)                    |
| Big and distributed system        | 110 devices, ≈500 degrees of freedom |

## 6 Collimator design for high-power accelerators

The key parameters for the design of the LHC collimator are summarized in Table 3. The list emphasizes the challenges in terms of quenching of superconducting magnets, damage, heating of components and radiation doses, which must be addressed by a optimized design. It is important to note that the design must ensure adequate mechanical stability during jaw position changes and in the presence of important heat loads. Other aspects related to materials choice to ensure robustness and limited impedance are addressed in a companion paper [18]. Details of the final collimator design deployed for the LHC can be found in Refs. [14, 39]. Here, only the main design features are given.

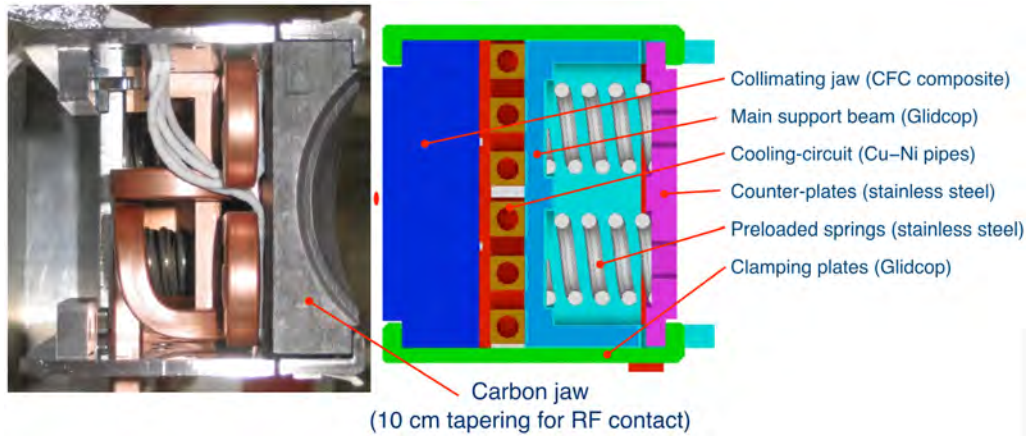
The LHC collimators are high-precision devices that ensure the correct hierarchy along the 27 km long ring with beam sizes as small as 200 μm. Each collimator has two jaws, of different lengths and materials, depending on functionality (Table 2). Each jaw can be independently moved by two stepping motors. Key features of the design are: (1) a jaw flatness of about 40 μm along the 1 m long active jaw surface; (2) a surface roughness less than 2 μm; (3) a 5 μm positioning resolution; (4) an overall setting reproducibility below 20 μm [35]; (5) a minimal gap of 0.5 mm; (6) evacuated heat loads of up to 7 kW in a steady-state regime and of up to 30 kW in transient conditions.

Primary and secondary collimators are made of a robust CFC that is designed to withstand beam impacts without significant permanent damage for the worst failure cases, such as impacts of a full injection batch of  $288 \times 1.15 \times 10^{11}$  protons at 450 GeV and of up to  $8 \times 1.15 \times 10^{11}$  protons at 7 TeV [39]. Other collimators made of heavy tungsten alloy or copper, obviously, do not have the same robustness and are only utilized at larger distances from the circulating beams, where maximum absorption is needed.

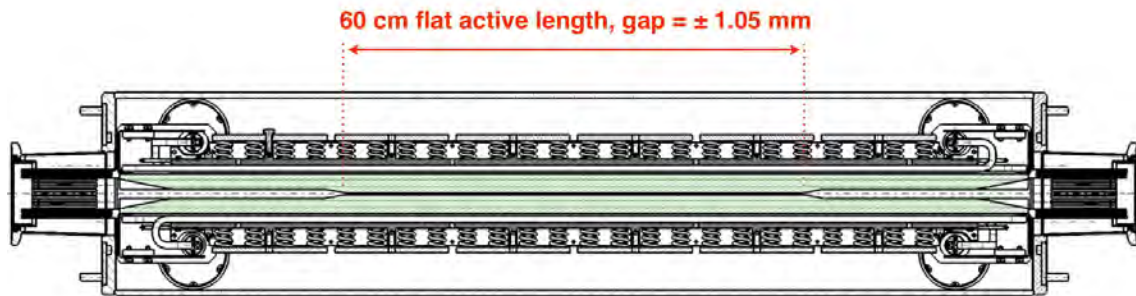
The cross-section of the primary and secondary collimator jaws, with a 2.5 cm thick active CFC part and a cooling system underneath, is shown in Fig. 28. The design drawing on the right side of the picture is compared with a real jaw prototype on the left, built during the initial production phases to verify the manufacturing quality. Two parallel jaws are mounted in the vacuum tank, as shown in Fig. 29 for a primary collimator. In Fig. 29, the jaws are actually shown set to the operational position for the vertical collimator with the tightest gaps, as in Fig. 22.

Figure 30 shows a horizontal and a 45° tilted LHC collimator. Their vacuum tank is still open to show the CFC jaws inside. An example of the tunnel installation layout for a IR7 collimator is given in Fig. 31. This is a horizontal TCLA collimator. Notice, next to the collimator, a yellow support that supports a vacuum pump that is installed next to each collimator. A beam loss monitor, not visible in the photograph, is also connected to this support, to record losses generated locally when the beam is intercepted by the collimator jaws.

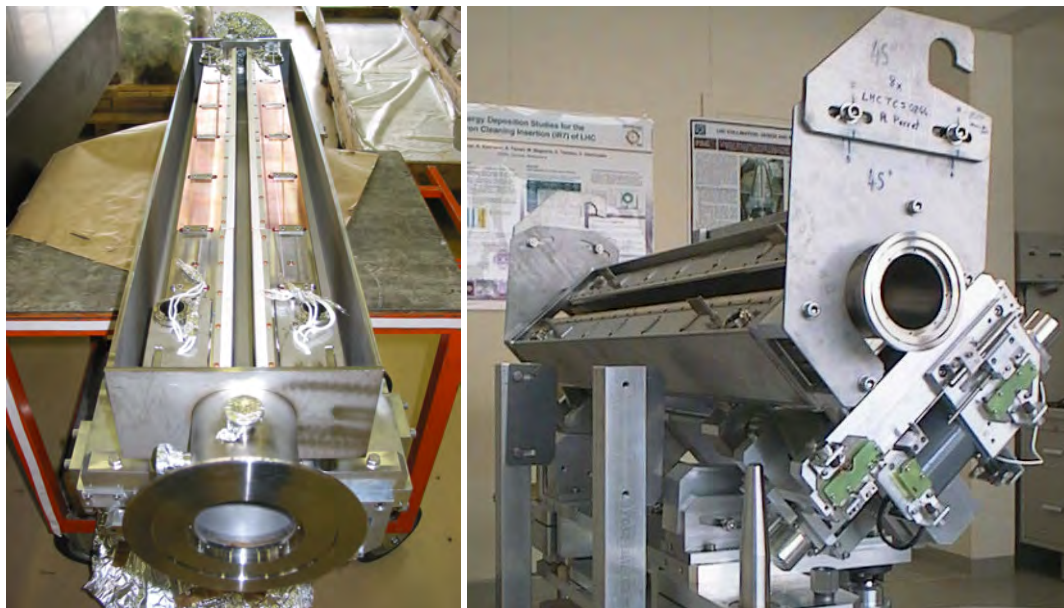
The collimator design has been recently improved by adding two beam position monitors on either extremity of each jaw [27]. An example of a CFC jaw prototype with this new design is shown in Fig. 32. This feature allows faster collimator alignment as well as constant monitoring of the beam orbit at the



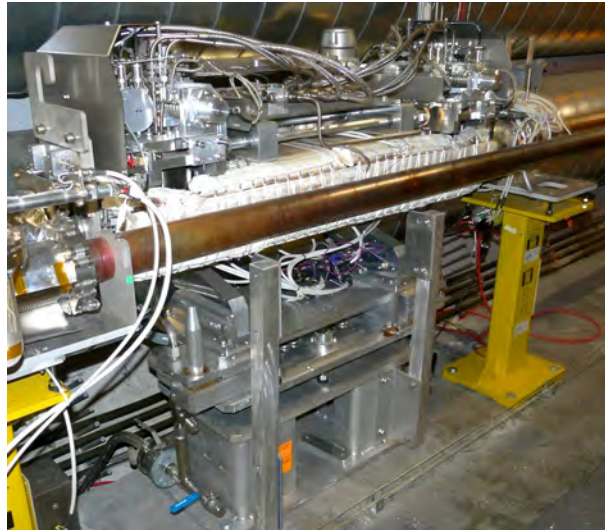
**Fig. 28:** Cross-section of the LHC collimator jaws. Left: real prototype. Right: design drawing. The position of the beam is shown by the red ellipse, as if the two jaws were those of a horizontal collimator. A sandwich structure, with cooling circuits clamped on the CFC plate of the active part, is optimized to minimize deformation of the structure during steady loss conditions [39].



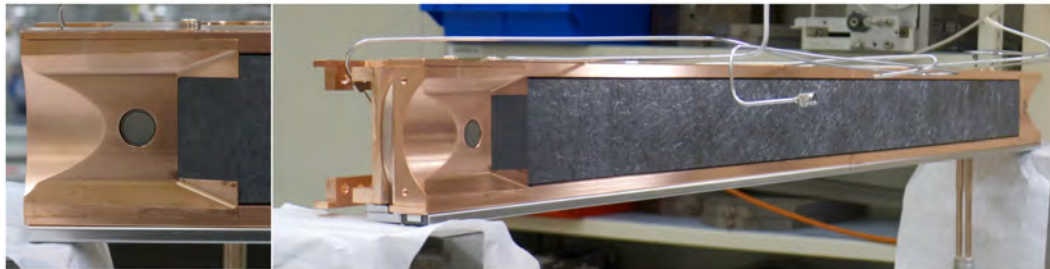
**Fig. 29:** Design of the LHC primary collimator. The two jaws can move independently, thanks to four stepping motors enabling position and angular adjustment with respect to the beam. This design is essentially identical to that of the secondary collimator except that the jaws are tapered to an effective length of 60 cm instead of 100 cm.



**Fig. 30:** Horizontal (left) and skew (right) LHC collimators with open tank, showing movable jaws. The support allows assembly in the same collimator tank of all the required orientations.



**Fig. 31:** Active absorber TCLA.B6R7.B1 as installed in the betatron cleaning insert. The stepping motors that control jaw position and angle are visible on top of the vacuum tank. The pipe of the opposing beam is also shown.

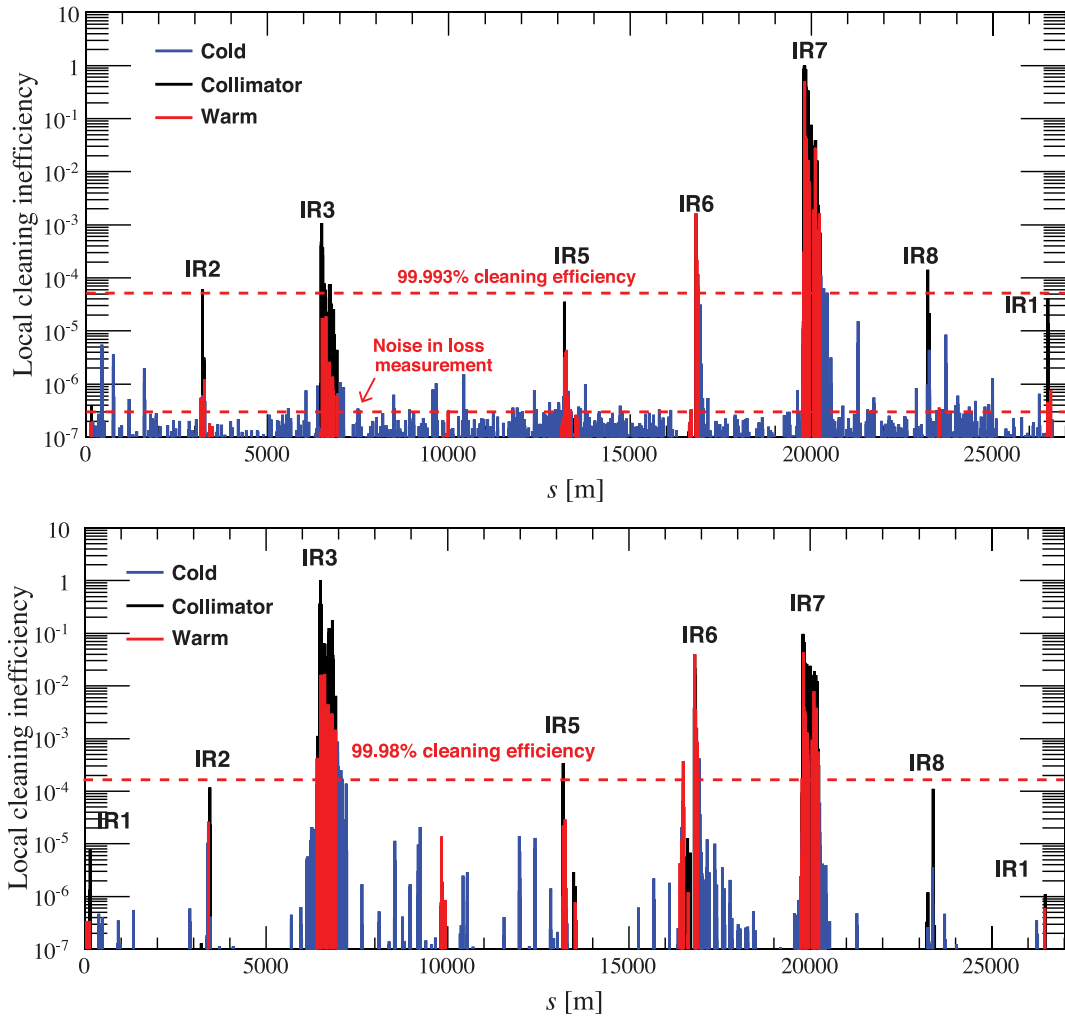


**Fig. 32:** New CFC jaw with integrated beam position monitors at each extremity for installation in IR6 (see Fig. 17). A variant of this design, made with a Glidcop support and tungsten heavy alloy inserts on the active jaw part, is used for the new TCTP tertiary collimators in all IRs.

collimator, as opposed to the beam-loss-monitor-based alignment that can currently only be performed during dedicated low-intensity commissioning fills. The beam position monitor buttons will improve collimation performance significantly in terms of operational efficiency and flexibility, by reducing the machine time spent on aligning collimators and the  $\beta^*$  reach [39]. The beam-position-monitor-embedded design is considered the baseline for future upgraded collimator design.

## 7 Cleaning performance of the LHC beam collimation

The cleaning performance of the LHC collimation system is measured by intentionally generating transverse and off-momentum beam losses while measuring losses around the ring. This is done with low intensities circulating in the machine. A few bunches are excited by driving the betatron tune close to resonance or by adding transverse noise with the transverse damper. The latter method is preferred, as it can act in a bunch-by-bunch mode so one fill can be used for several loss maps. Large losses of the momentum cleaning can instead be generated by changing the radio frequency. These so-called *loss maps* are used to validate, empirically, the response of the collimation system in the presence of high loss rates. This is an essential part of the validation of the LHC machine protection functionality, as discussed in Ref. [9]. In particular, loss maps are used to verify: (1) that the hierarchy is respected, by checking that the relative loss rates at the different collimators are in agreement with predictions or within tolerable levels; (2) that the leakage of losses to the other machine equipment, in particular

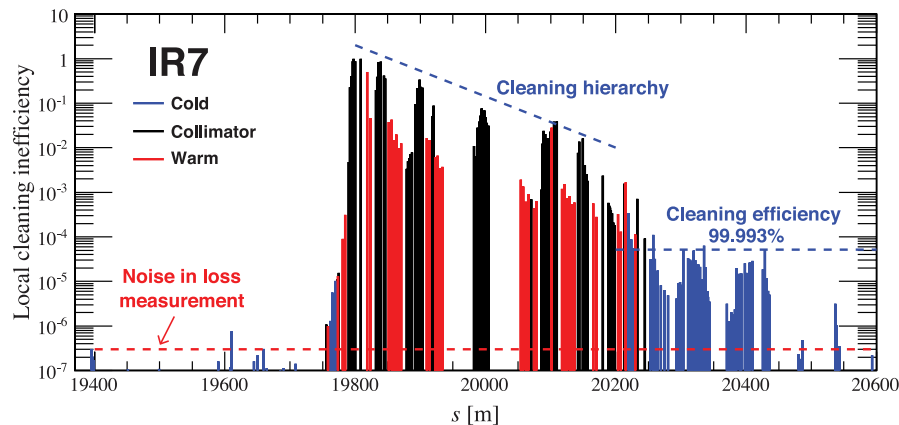


**Fig. 33:** Betatron (top) and off-momentum (bottom) loss maps obtained at the LHC at 4 TeV with beams squeezed to 60 cm in IR1 (ATLAS) and IR5 (CMS), showing the beam losses recorded at about 4000 beam loss monitors around the ring, normalized to the highest measured signal. Betatron losses are generated in IR7 by adding noise to the kickers of the transverse damper of clockwise beam 1. IR3 losses are generated by changing the radio frequency until the full beam is intercepted by the IR3 TCP. Both beams are excited at the same time as their frequencies are synchronized. Courtesy of B. Salvachua [38].

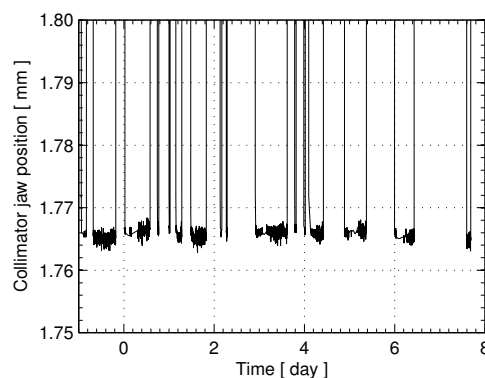
superconducting magnets, are as expected; (3) that the system performance remain stable during long periods when beam-based alignment is not repeated.

Examples of betatron and off-momentum loss maps are shown in Fig. 33. These maps were recorded in 2012 at 4 TeV, with beams squeezed to 60 cm in IR1 and IR5. At the LHC, beam losses are recorded by about 3600 beam loss monitors around the ring [40]. To estimate the cleaning inefficiency, losses at each monitor are normalized to the highest measured signal, i.e., next to the primary collimators. This is shown in Fig. 33 as a function of the longitudinal coordinate  $s$ . It is seen that inefficiencies less than  $\sim 10^{-4}$  were achieved. In all IRs, the largest losses are recorded at the collimators (black bars), as expected. The cold locations with the highest losses are the dispersion suppressors downstream of the cleaning insert, as predicted in simulations (see Fig. 15).

The IR7 losses are given in Fig. 34. The limiting locations with the worst cleaning are the dispersion suppressors on either side of IR7 (the right side for beam 1). A cleaning efficiency above 99.993% was achieved. Note that, with the exception of a few isolated peaks in the dispersion suppressor, the



**Fig. 34:** Enlargement of the top graph of Fig. 33, showing details of losses in IR7. The limiting location for betatron cleaning is given by the losses on the cold magnets in the dispersion suppressor immediately downstream of IR7.



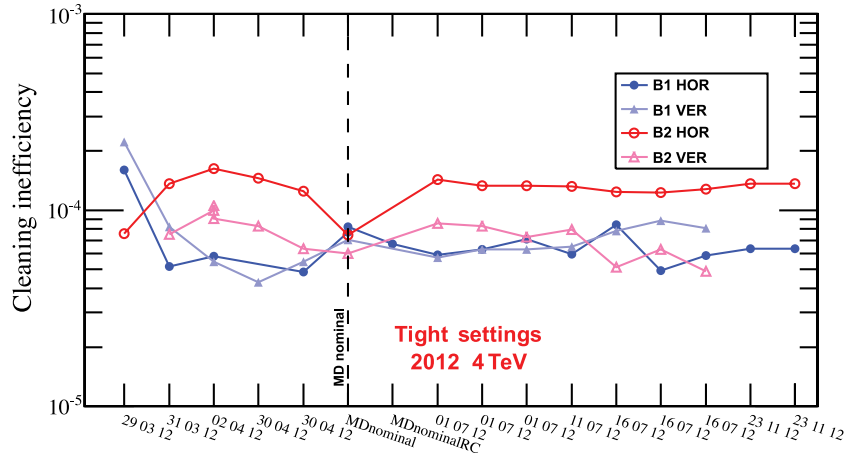
**Fig. 35:** End-of-ramp settings for one TCP jaw as a function of time over 9 days, showing micrometre reproducibility [30].

rest of the cold machine experiences losses that are more than one order of magnitude smaller, i.e., close to the noise of the beam loss monitor system. In simulations, losses are sampled using 10 cm bins, counting the number of beam particles hitting the aperture. In measurements, losses are measured at the discrete locations of beam loss monitors that record the flux of ionizing particles in the beam loss monitor volume. Clearly, these two quantities cannot be directly compared without additional simulations of energy deposition, starting from the multiturn loss pattern. Detailed discussion of this aspect is beyond the scope of this lecture. It suffices to say that the agreement between simulations and measurements is good [41].

The fill-to-fill reproducibility of the collimator positions is of the order of a few micrometres. A typical example for one jaw of a TCP collimator is given in Fig. 35. This is a key ingredient for the system performance because the collimator settings are not realigned. This stability of the hardware, together with the outstanding fill-to-fill reproducibility of optics and orbit at the LHC, makes it possible to maintain excellent collimation performance with one single beam-based alignment per year in IR3/6/7. As an example, in Fig. 36 the cleaning inefficiencies at the worst locations in the rings are shown for each beam and loss plane. It can be seen that the stability of the measured cleaning is indeed remarkable.

## 8 Advanced collimation concepts for enhanced beam collimation

Other advanced collimation concepts have been under study in the last year, as possible methods of improving the performance of the LHC multistage system. In this section, the main topics presently under



**Fig. 36:** Collimation cleaning inefficiency at the worst location in the dispersion suppressors at either side of IR7 for both beams and planes, as measured throughout the 2012 operation with protons (4 TeV,  $\beta^* = 60$  cm). B, beam; HOR, horizontal; VER, vertical. Courtesy of B. Salvachua [38].

investigation are introduced. Possible immediate applications of such advanced concepts are already under consideration for the high-luminosity upgrade of the LHC.

### 8.1 Local dispersion suppressor cleaning

Protons and ions interacting with the collimators in IR7 emerge from the IR with a changed magnetic rigidity. This represents a source of local heat deposition in the cold dispersion suppressor magnets downstream of IR7, where the dispersion starts to increase: these losses are the limiting locations for collimation cleaning, i.e., they are the highest cold losses around the ring. This may pose a certain risk for inducing magnet quenches, in particular, in view of the higher intensities expected for HL-LHC. This problem arises for halo collimation of both proton and heavy-ion beams.

A possible solution to this problem is to add local collimators in the dispersion suppressors, which is only feasible with a major change of the cold layout at the locations where the dispersion starts to increase. Indeed, the existing system's multistage cleaning is not efficient at catching these dispersive losses. Clearly, the need for local collimation depends on the absolute level of losses achieved in operation and the quench limit of superconducting magnets. In view of the uncertainties in the scaling of the current system performance for operation at 7 TeV, it is important to take appropriate margins, to minimize the risk of limitation in the future.

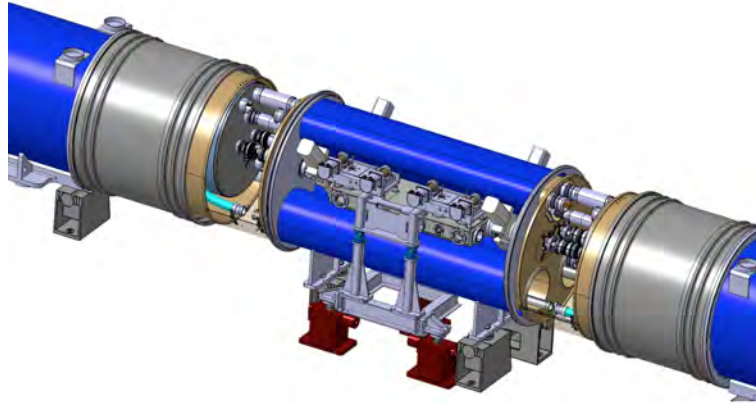
A solution with minimum impact on the cold section layout is to replace the existing 15 m long dipoles with two shorter, higher-field magnets, by freeing enough space to install a warm collimator. This solution is illustrated in Fig. 37. It requires an 11 T dipole field to free sufficient space for a warm collimator to be installed in a dedicated cryogenics by-pass system, as shown in Fig. 38. Even in this tight space limitation, an adequate solution can be found. New dipoles and collimators are being prototyped at CERN, providing a viable solution for IR7 cleaning upgrades that might already be available for a long LHC stop planned for 2019. Note that this solution is modular and was designed to be implemented easily in any existing dipole location. It can therefore also be used to improve cleaning around collision points, if necessary, as is foreseen for the ALICE ion experiment [42].

### 8.2 Status on research and development on novel collimator materials

The LHC impedance budget is largely dominated by the contribution of the LHC collimators. For this reason, the current collimation system was conceived in such a way that it can be easily upgraded to reduce the impedance [14]: every secondary collimator slot in IR3 and IR7 features a companion slot



**Fig. 37:** Longitudinal integration of a TCLD collimator between two short 11 T dipoles. Courtesy of D. Ramos



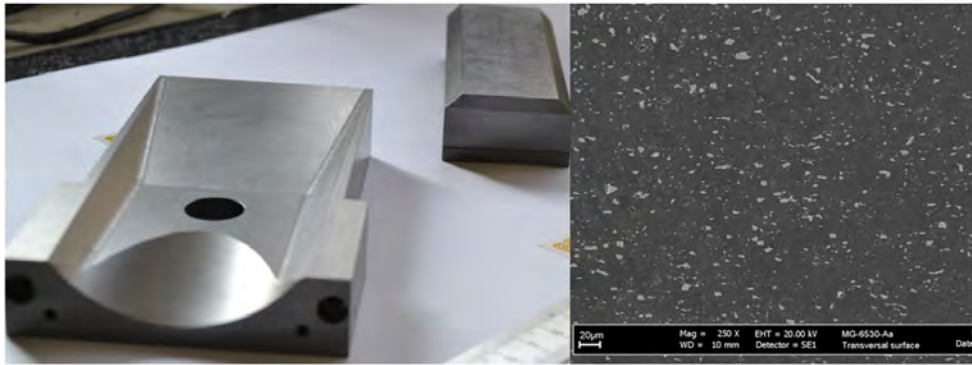
**Fig. 38:** Three-dimensional view of the TCLD installation in the cryogenic by-pass region between two 11 T dipoles. Courtesy of D. Ramos and L. Gentini.

for the future installation of a low-impedance secondary collimator. A total of 22 slots in IR7 and 8 slots in IR3 are already cabled for a quick installation of new collimators—referred to as TCSPMs—which can either replace or supplement the existing TCSG collimators. The TCSPMs will include pick-ups for orbit measurements ('P') and will be based on metal composites ('M'). In addition, limited robustness against beam losses of the present tungsten collimator has already limited the LHC performance of run I in terms of  $\beta^*$  reach, because adequate margins had to be taken in the collimation hierarchy to shield the existing tertiary collimators properly [28].

A rich programme of research and development was initiated, to find novel material with optimum response to thermomechanical stress and with reduced impedance, to improve various limitations of current LHC collimator materials. More details are given in a companion paper [18]. Simulations predict that beam stability can be re-established for all HL-LHC scenarios if the CFC of the existing secondary collimators is replaced, at least in the betatron cleaning insertion (IR7), with a jaw material having an electrical conductivity a factor of 50 to 100 higher than CFC [43]. This improvement could easily be achieved if the jaw material were made of highly conductive metals, such as copper or molybdenum. However, secondary collimators in IR7 also play a crucial role in LHC machine protection and might be exposed to large beam losses. Therefore, collimator materials and designs must also be robust against beam failure. The driving requirements for the development of new materials are thus: (i) low resistive-wall impedance, to avoid beam instabilities; (ii) high cleaning efficiency; (iii) high geometrical stability, to maintain the precision of the collimator jaw during operation despite temperature changes; and (iv) high structural robustness, in case of accidental events, such as single-turn losses.

The current baseline for the upgraded secondary collimators relies on novel carbon-based materials, such as molybdenum carbide-graphite (MoGr), a ceramic composite jointly developed by CERN and Brevetti Bizz, in which the presence of carbides and carbon fibres strongly catalyses the graphitic ordering of carbon during high-temperature processing, enhancing its thermal and electrical properties (Fig. 39). To further improve their surface electrical conductivity, these materials could be coated with pure molybdenum or other lower- $Z$  refractory coatings. Replacing all existing CFC secondary collimators in both IR7 and IR3 with bulk MoGr or MoGr coated with 5  $\mu\text{m}$  thick pure molybdenum would reduce the total LHC impedance by 40% or 60%, respectively.





**Fig. 39:** Left: MoGr components recently produced by Brevetti Bizz (Italy) for a jaw prototype. Jaw extremity (dimensions of  $147 \times 88 \times 25 \text{ mm}^3$ ) and jaw absorbing block ( $125 \times 45 \times \text{mm}^3$ ) are shown. A jaw assembly includes two jaw extremities (taperings) and eight blocks. Right: Detail of microstructure; the graphite matrix is visible, together with molybdenum carbide grains of about  $5 \mu\text{m}$ . Courtesy of A. Bertarelli.

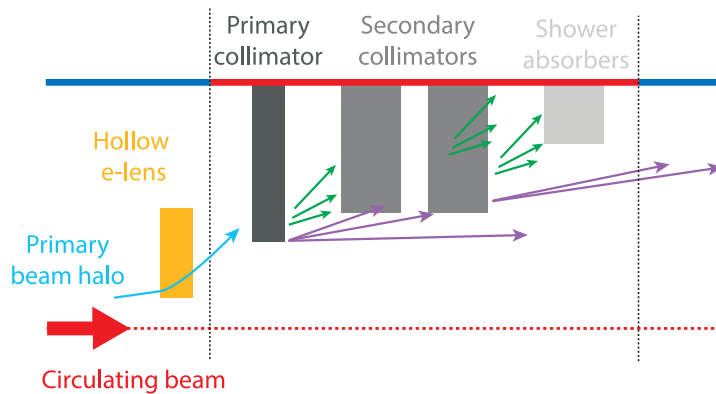
The new collimator design, along with novel materials and possible alternative coatings must be validated for operation in the LHC. For these purposes, a rich programme of validation is in progress, involving tests at the CERN HiRadMat facility [44], to address robustness against beam impact, mechanical engineering prototyping, beam tests at the LHC, and experimental verification of the material response under high radiation doses. It is anticipated that this test will be completed, and the production of new, low-impedance, highly robust collimators in the LHC started, by 2019.

### 8.3 Halo diffusion control techniques

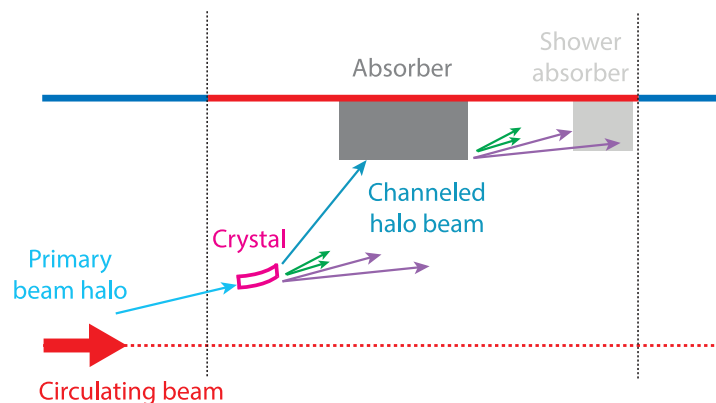
The 2012 operation experience indicates that the LHC collimation would profit from halo control mechanisms. The idea is that, by controlling the diffusion speed of halo particles in an aperture range between the core and the TCP opening ( $\approx 3\text{--}5\sigma_z$ ), one can act on the time profile of the losses. The main goal is to reduce loss rates that would otherwise take place in a short time, or simply to control the static population of halo particles in a certain aperture range. For example, it is expected that losses caused by orbit drifts [45] during the squeeze (see Figs. 5 and 6) can be strongly reduced by actively depleting the halo population.

One of the best candidate techniques for achieving active halo control at the LHC is to use the hollow e-lens collimation concept [46, 47]. A hollow electron beam, running coaxially to the proton or ion beam, is used to generate an annular beam in the transverse ( $x, y$ ) plane. This hollow beam induces an electromagnetic field, which affects halo particles above a certain transverse amplitude and can change their transverse speed. The working principle is illustrated in Fig. 40. A solid experimental basis achieved at the Tevatron indicates that this solution is very promising for the LHC. The design for an hollow e-lens for the LHC is ongoing (see [48] and references therein).

Conversely, in the case of loss spike limitations at the LHC during run II, the hollow e-lens solution would not be viable because it could only be implemented over a time-scale of a few years [49]. It is, therefore, crucial to work on alternatives that, if necessary, might be implemented on an appropriate time-scale. Two alternatives are currently being considered: tune modulation through noise in the current of lattice quadrupoles, as outlined in Ref. [50], and narrow-band excitation of halo particles, using the transverse damper system. Though very different from the hardware point of view, both these techniques rely on exciting tail particles through resonances induced in the tune space. This method works on the assumption that there is a correlation between halo particles with large amplitudes and corresponding tune shifts in tune space (de-tuning with amplitude). Clearly, both methods require solid experimental verification in a very low noise machine, like the LHC, in particular, to demonstrate that these types of excitation do not perturb the beam core emittance. Unlike hollow e-lenses, which act directly in the transverse plane by affecting particles at amplitudes above the inner radius of the hollow beam, resonance



**Fig. 40:** Integration of hollow e-lenses as halo diffusers in the present collimation system.



**Fig. 41:** Crystal collimation concept foreseen the use of a bent crystal to channel halo particle in one single passage to a dedicated absorber, reducing significantly the number of secondary collimators.

excitation methods require a good knowledge of the beam core and tail particle tunes, even in dynamic phases of the operational cycle. It is planned to test these techniques experimentally in LHC run II.

#### 8.4 Crystal collimation

Highly pure bent crystals can be used to steer high-energy particles that become trapped between the potential of parallel lattice planes [51]. Equivalent bending fields of up to hundreds of teslas can be achieved in crystals with a length of only 3–4 mm, enabling, in principle, halo particles to be steered to a well-defined point, with obvious potential applications to beam collimation. As opposed to standard primary collimators based on amorphous materials, which require several secondary collimators and absorbers to catch the products developed through the interaction with matter (Fig. 14), one single absorber per collimation plane is, in theory, sufficient in a crystal-based collimation system. This is shown in Fig. 41.

In addition to the reduction of secondary collimators, nuclear interactions with well-aligned crystals are much reduced compared with a primary collimator, provided that high channelling efficiencies of halo particles can be achieved (particles impinging on the crystal are to be channelled within a few turns). This is expected to reduce dispersive beam losses in the dispersion suppressor of the betatron cleaning insertion significantly, compared with the existing system, which is limited by the leakage of particles from the primary collimators. Simulations indicate a possible gain of between 5 and 10 [52], even for a layout without an optimized absorber design. The crystal collimation option is particularly interesting for collimation of heavy-ion beams, thanks to the reduced probability of ion dissociation and fragmentation compared with current primary collimators.

Another potential of crystal collimation is a strong reduction of the machine impedance, since (1) only a small number of collimator absorbers is required and (2) the absorbers can be spaced much farther apart, owing to the large bending angle from the crystal (40–50  $\mu\text{rad}$  instead of a few microradians from multiple Coulomb scattering in the primary collimator). Conversely, an appropriate absorber design must be conceived to handle the design peak loss rates, of 1 MW during 10 s, expected for the LHC upgrade [2]. Other potential issues concern the machine protection aspects of this scheme, which has not yet been studied in detail, and operational aspects for crystals that require mechanical angular stability in the submicroradian range through the operational cycle. Note that the critical angle beyond which channelling is lost is  $\approx 2 \mu\text{rad}$  at 7 TeV.

Promising results were achieved in dedicated crystal collimation tests at the SPS, performed from 2009 within the UA9 experiment [53–55]. However, some outstanding issues on the feasibility of the crystal collimation concept for the LHC can only be addressed by beam tests at high energy in the LHC. For this purpose, a study at the LHC has been proposed, and will take place in the LHC run II [52, 56]. Tests at the LHC will address the feasibility of the crystal collimation concept with LHC beam conditions, in particular, to demonstrate that such a system can provide better cleaning than the present high-performance system throughout the operational cycle.

### Acknowledgements

The material presented here is the result of the work of many people and was presented on behalf of the LHC collimation project team. Past and present team members are gratefully acknowledged. In particular, R. Bruce, D. Mirarchi, B. Salvachua, and G. Valentino are sincerely thanked, for providing material for this document and for providing useful comments on the manuscript.

### References

- [1] LHC design report, Vol. 1. The LHC main ring, edited by O.S. Brüning *et al.*, CERN-2004-003-V-1 (CERN, Geneva, 2004), <http://dx.doi.org/10.5170/CERN-2004-003-V-1>
- [2] L. Rossi, LHC upgrade plans: options and strategy, Proc. 2nd Int. Particle Accelerator Conf., San Sebastian, 2011 [*Conf. Proc. C* **110904** (2011) 908].
- [3] L. Rossi *et al.*, HL-LHC preliminary design report, CERN-ACC-2014-0300 (2014). <https://cds.cern.ch/record/1972604/files/CERN-ACC-2014-0300.pdf>
- [4] N. Mokhov *et al.*, *J. Instrum.* **6** (2011) T08005. <http://dx.doi.org/10.1088/1748-0221/6/08/T08005>
- [5] H. Hahn *et al.*, *Nucl. Instrum. Methods Phys. Res. A* **499**(2–3) (2003) 245. [http://dx.doi.org/10.1016/S0168-9002\(02\)01938-1](http://dx.doi.org/10.1016/S0168-9002(02)01938-1)
- [6] M. Lamont, The first years of LHC operation for luminosity production, Proc. 4th Int. Particle Accelerator Conf. (IPAC 2013), 12–17 May 2013, Shanghai, China. C13-05-12 (2013) p. MOYAB101.
- [7] B. Auchmann *et al.*, *Phys. Rev. ST Accel. Beams* **18**(6) (2015) 061002. <http://dx.doi.org/10.1103/PhysRevSTAB.18.061002>
- [8] R. Schmidt, Introduction to accelerator protection course, these proceedings.
- [9] J. Wenninger, Machine protection and operation for LHC, these proceedings.
- [10] R. Bruce *et al.*, *Nucl. Instrum. Methods Phys. Res. A* **729** (2013) 825. <http://dx.doi.org/10.1016/j.nima.2013.08.058>
- [11] G. Valentino *et al.*, *Phys. Rev. ST Accel. Beams* **16**(2) (2013) 021003. <http://dx.doi.org/10.1103/PhysRevSTAB.16.021003>
- [12] K.H. Mess and M. Seidel, *Nucl. Instrum. Methods Phys. Res. A* **351**(2-3) (1994) 279. [http://dx.doi.org/10.1016/0168-9002\(94\)91354-4](http://dx.doi.org/10.1016/0168-9002(94)91354-4)

- [13] H. Burkhardt *et al.*, Collimation down to 2 sigmas in special physics runs in the LHC, Proc. 4th Int. Particle Accelerator Conf. (IPAC 2013), 12–17 May 2013, Shanghai, China. C13-05-12, CERN-ACC-2013-0144 (2013).
- [14] R. Aßmann *et al.*, The final collimation system for the LHC, Proc. 10th European Particle Accelerator Conf., Edinburgh, UK, 26–30 June 2006, CERN-LHC-Project-Report-919 (2006), p. 986.
- [15] R. Aßmann, Collimators and cleaning: could this limit the LHC performance? LHC Performance Workshop, Chamonix XII, Chamonix, France, 2003.
- [16] S. Redaelli *et al.*, LHC aperture and commissioning of the collimation system, Proc. Chamonix 2005 LHC Project Workshop, Chamonix, France, 2005, CERN-AB-2005-014 (2005), p. 268.
- [17] V. Kain, Beam dynamics and beam losses—circular machines, these proceedings.
- [18] A. Bertarelli, Beam-induced damage mechanisms and their calculation, these proceedings.
- [19] G. Robert-Demolaize *et al.*, A new version of SixTrack with collimation and aperture interface [*Conf. Proc. C* **0505161** (2005) 4084].
- [20] F. Cerutti, Beam material interaction, heating and activation, these proceedings.
- [21] S. Redaelli, M. C. Alabau-Pons, M. Giovannozzi, G. Muller, F. Schmidt, R. Tomas and J. Wenninger, “LHC Aperture Measurements,” *Conf. Proc. C* **100523**, MOPEC010 (2010).
- [22] R.W. Aßmann *et al.*, “Aperture Determination in the LHC Based on an Emittance Blowup Technique with Collimator Position Scan,” *Conf. Proc. C* **110904**, 1810 (2011).
- [23] J.B. Jeanneret and R. Ostojic, Geometrical acceptance in LHC ver. 5.0, CERN-LHC-PROJECT-NOTE-111 (1997).
- [24] S. Redaelli *et al.*, Aperture and optics—measurements and conclusions, Proc. 3rd Evian Workshop on LHC beam operation, Evian-les-bains, France, 12–14 Dec 2011, CERN-ATS-2012-083 (2012), p. 77.
- [25] M. Seidel, The proton collimation system of HERA, DESY-94-103 (1994).
- [26] J.B. Jeanneret, *Phys. Rev. ST Accel. Beams* **1**(8) (1998) 081001.  
<http://dx.doi.org/10.1103/PhysRevSTAB.1.081001>
- [27] F. Carra *et al.*, LHC collimators with embedded beam position monitors: a new advanced mechanical design, Proc. IPAC2011, San Sebastian, IPAC-2011-TUPS035 (2011).
- [28] R. Bruce *et al.*, *Phys. Rev. ST Accel. Beams* **18**(6), (2015) 061001.  
<http://dx.doi.org/10.1103/PhysRevSTAB.18.061001>
- [29] R. Bruce *et al.*, Baseline LHC machine parameters and configuration of the 2015 proton run, Chamonix2014: LHC Performance Workshop, Chamonix, France, 22–25 September 2014, p. 100.
- [30] S. Redaelli *et al.*, Operational performance of the LHC collimation, Proc. HB2010 workshop, Morschach, CH (2010). <http://epaper.kek.jp/HB2010/>
- [31] R. Bruce *et al.*, LHC  $\beta^*$  reach in 2012, LHC Operation Workshop, Evian, 2011.  
<http://indico.cern.ch/event/155520>
- [32] D. Wollmann *et al.*, First cleaning with LHC collimators [*Conf. Proc. C* **100523** (2010) TUOAMH01].
- [33] S. Redaelli *et al.*, Operational experience with a LHC collimator prototype in the CERN SPS, Proc. PAC09, Vancouver, Canada, 2009.
- [34] G. Valentino, Ph.D. thesis, University of Malta (2103). Also as CERN-THESIS-2013-208 (2013) and CERN-ACC-2014-0062 (2014).
- [35] A. Masi *et al.*, Measured performance of the LHC collimator low-level control system, Proc. ICALEPCS09, Kobe, Japan, 2009.
- [36] S. Redaelli *et al.*, Final implementation and performance of the LHC collimator control system, Proc. PAC09, Vancouver, Canada, 2009.

- [37] S. Redaelli *et al.*, Performance of ramp and squeeze at the Large Hadron Collider, Proc. HB2010 workshop, Morschach, CH (2010). <http://epaper.kek.jp/HB2010/>
- [38] B. Salvachua *et al.*, Cleaning performance of the LHC collimation system up to 4 TeV, Conf. C13-05-12 (2013), p. MOPWO048.
- [39] A. Bertarelli *et al.*, The mechanical design for the LHC collimators, Proc. EPAC2004, Lucerne (2004), p. 545. <http://accelconf.web.cern.ch/accelconf/e04/papers/mop1t008.pdf>
- [40] B. Dehning, Beam loss monitors at the LHC, these proceedings.
- [41] R. Bruce *et al.*, *Phys. Rev. ST Accel. Beams* **17**(8) (2014) 081004.  
<http://dx.doi.org/10.1103/PhysRevSTAB.17.081004>
- [42] R. Bruce *et al.*, Conceptual design of IR collimation, WP5 report of the FP7-HiLumi programme, CERN-ACC-2014-0293 (2014).  
<https://cds.cern.ch/record/1972595/files/CERN-ACC-2014-0293.pdf>
- [43] E. Métral *et al.*, Intensity limitations, WP5 report of the FP7-HiLumi programme, CERN-ACC-2014-0297 (2014).  
<https://cds.cern.ch/record/1972601/files/CERN-ACC-2014-0297.pdf>
- [44] H. Gaillard *et al.*, HiRadMat: a new irradiation facility for material testing at CERN [*Conf. Proc. C* **110904** (2011) 1665].
- [45] S. Redaelli, *et al.*, Experience with high-intensity beam scraping and tail populations at the Large Hadron Collider, Proc. 4th Int. Particle Accelerator Conf., Shanghai, China, 12–17 May 2013, C13-05-12 (2013), p. MOPWO039, CERN Report CERN-ACC-2013-0063 (2013).
- [46] V. Shiltsev, Proc. 3rd CARE-HHH-APD Workshop (LHC-LUMI-06), Valencia, Spain, CERN-2007-002 (2007), p. 92.
- [47] V. Shiltsev, Proc. CARE-HHH-APD Workshop (BEAM07), Geneva, Switzerland, CERN-2008-005 (2008), p. 46.
- [48] G. Stancari *et al.*, Conceptual design of hollow electron lenses for beam halo control in the Large Hadron Collider, CERN-ACC-2014-0248 (2014).
- [49] S. Redaelli *et al.*, Plans for deployment of hollow electron lenses at the LHC for enhanced beam collimation, Proc. 6th Int. Particle Accelerator Conf., Richmond, VA, USA, 3–8 May 2015.
- [50] O. Brüning and F. Willeke, *Phys. Rev. Lett.* **76**(20) (1996) 3719.  
<http://dx.doi.org/10.1103/PhysRevLett.76.3719>
- [51] W. Scandale, *Int. J. Mod. Phys. A* **25**(S1) (2010) 70.  
<http://dx.doi.org/10.1142/S0217751X1004992X>
- [52] D. Mirarchi, Ph.D. thesis, Imperial College, London, 2015.
- [53] W. Scandale *et al.*, *Phys. Lett. B* **703**(5) (2011) 547.  
<http://dx.doi.org/10.1016/j.physletb.2011.08.023>
- [54] W. Scandale *et al.*, *Phys. Lett. B* **726**(1-3) (2013) 182.  
<http://dx.doi.org/10.1016/j.physletb.2013.08.028>
- [55] W. Scandale *et al.*, *Phys. Lett. B* **714**(2-5) (2012) 231.  
<http://dx.doi.org/10.1016/j.physletb.2012.07.006>
- [56] D. Mirarchi *et al.*, Final layout and expected cleaning for the first crystal-assisted collimation test at the LHC, Proc. 5th Int. Particle Accelerator Conf., IPAC2014, Dresden, Germany, 16–20 June 2014, C14-06-16.



## Participants

ALTINBAS, Z. Brookhaven National Laboratory, Long Island, US  
ANDERSSON, R. ESS, Lund, SE  
APOLLONIO, A. La Sapienza University, Rome, IT  
AUCHMANN, B. CERN, Geneva, CH  
BAILEY, R. CERN, Geneva, CH  
BARLETTA, W. USPAS, Batavia, US  
BAUER, J. SLAC, Menlo Park, US  
BERTARELLI, A. CERN, Geneva, CH  
BESANA, M.I. CERN, Geneva, CH  
BLAHA, J. SLAC, Menlo Park, US  
BREEDING, E. Oak Ridge National Lab, Oak Ridge, US  
CARRONE, E. SLAC, Menlo Park, US  
CERUTTI, F. CERN, Geneva, CH  
CHITNIS, P. Stony Brook University, Stony Brook, US  
CHOLAKIAN, A.E. Harvard University, Cambridge, US  
CONLON, M. ESS, Lund, SE  
DEHNING, B. CERN, Geneva, CH  
DOOM, L. Brookhaven National Laboratory, Long Island, US  
GALAMBOS, J. Oak Ridge National Lab, Oak Ridge, US  
GEELHOED, M. Fermilab, Batavia, US  
GILLESPIE, E. Oak Ridge National Lab, Oak Ridge, US  
GORZAWSKI, A. EPFL and CERN, Lausanne and Geneva, CH  
HAKULINEN, T. CERN, Geneva, CH  
HARRISON, M. RadiaBeam Technologies, Santa Monica, US  
HUSCHAUER, A. Vienna University of Technology, Vienna, A  
HWANG, K. Indiana University, Bloomington, US  
JAROSZ, M. ESS, Lund, SE  
JI, Y. Illinois Institute of Technology, Chicago, US  
JOBE, K. SLAC, Menlo Park, US  
KAIN, V. CERN, Geneva, CH  
KALLIOKOSKI, M. CERN, Geneva, CH  
KASTRIOTOU, M. University of Liverpool and CERN, Liverpool and Geneva, UK and CH  
KIM, S-H. Oak Ridge National Lab, Oak Ridge, US  
KOWALSKA, M. Warsaw University of Technology, Warsaw, PL  
LARI, L. ESS, Lund, SE  
LEE, H-S. Pohang Accelerator Laboratory, Pohang, ROK  
LENSCH, T. DESY, Hamburg, DE  
LIANG, T. Georgia Tech and SLAC, Atlanta and Menlo Park, US  
LIU, A. Indiana University and Fermilab, Bloomington and Batavia, US  
MAGNIN, N. CERN, Geneva, CH  
MALYZHENKOV, A. Northern Illinois University, DeKalb, US  
MARKIEWICZ, T. SLAC, Menlo Park, US  
MASCIA, A. SCRIPPS Health, San Diego, US  
MOKHOV, N. Fermilab, Batavia, US  
MOMPO, R. CERN, Geneva, CH  
MONTANO, R. ESS, Lund, SE  
NORDT, A. ESS, Lund, SE  
PFEFFER, H. Fermilab, Batavia, US  
PHAM, A. Michigan State University, East Lansing, US

|                          |   |
|--------------------------|---|
| PLUM, M.                 | Oak Ridge National Lab, Oak Ridge, US                           |
| RAFIQUE, H.              | University of Huddersfield, West Yorkshire, UK                  |
| KWANKASEM, A.            | Synchrotron Light Research Institute, Nakhon Ratchasima, TH, CH |
| REDAELLI, S.             | CERN, Geneva, CH  |
| RODRIGUEZ MATEOS, F.     | CERN, Geneva, CH  |
| ROKNI, S.                | SLAC, Menlo Park, US  |
| ROSS, M.                 | SLAC, Menlo Park, US  |
| SALVACHUA FERRANDO, M.B. | CERN, Geneva, CH  |
| SANTAMARIA GARCIA, A.    | EPFL and CERN, Lausanne and Geneva, CH                          |
| SCHMIDT, R.              | CERN, Geneva, CH  |
| SCHULTZ, R.              | Fermilab, Batavia, US   |
| SHEA, T.                 | ESS, Lund, SE   |
| SKORDIS, E.              | CERN, Geneva, CH  |
| SOLFAROLI CAMILLOCCI, M. | CERN, Geneva, CH  |
| SONMEZ, O.               | Istanbul Technical University, Istanbul, TR                     |
| STECKERT, J.             | CERN, Geneva, CH  |
| STEIMEL, J.              | Fermilab, Batavia, US   |
| STEIN, O.                | University of Hamburg and CERN, Hamburg and Geneva, DE, CH      |
| SUETSUGU, Y.             | KEK, Ibaraki, J   |
| SZYMCZYK, K.             | National Centre for Nuclear Research, Warsaw, PL                |
| TAO, F.                  | SLAC, Menlo Park, US  |
| THOMAS, J.               | MIT, Cambridge, US  |
| TROPIN, I.               | Fermilab, Batavia, US   |
| TURNER, C.               | SLAC, Menlo Park, US  |
| VALENTINO, G.            | CERN, Geneva, CH  |
| VELOTTI, F.M.            | EPFL and CERN, Lausanne and Geneva, CH                          |
| WAMSAT, T.               | DESY, Hamburg, DE   |
| WARZYBOK, P.             | National Centre for Nuclear Research, Warsaw, PL                |
| WENNINGER, J.            | CERN, Geneva, CH  |
| WERNER, M.               | DESY, Hamburg, DE   |
| WILLEKE, F.              | Brookhaven National Laboratory, Long Island, US                 |
| WOLTER, M.               | Fermilab, Batavia, US   |
| XU, C.                   | SLAC, Menlo Park, US  |
| YEE-RENDON, B.           | Instituto Politecnico Nacional, Mexico City, MEX                |