CERN Yellow Reports: School Proceedings
Volume 1/2017

CERN-2017-004-SP

The CERN Accelerator School

Accelerators for Medical Applications

Vösendorf, Austria
26 May–5 June 2015

Editor: R. Bailey

CERN Yellow Reports: School Proceedings
Published by CERN, CH-1211 Geneva 23, Switzerland

ISBN 978–92–9083–455–7 (paperback)
ISBN 978–92–9083–456–4 (PDF)
ISSN 2519-8041 (Print)
ISSN 2519-805X (Online)
DOI https://doi.org/10.23730/CYRSP-2017-001

Accepted for publication by the CERN Report Editorial Board (CREB) on 19 June 2017

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

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

This volume should be cited as:

Proceedings of the CAS–CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol.1/2017, CERN-2017-004-SP (CERN, Geneva, 2017). https://doi.org/10.23730/CYRSP-2017-001

A contribution in this volume should be cited as:

[Author name(s)], in Proceedings of the CAS–CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017), pp. [first page] – [lastpage], https://doi.org/10.23730/CYRSP-2017-001. [first page]

Abstract

These proceedings collate lectures given at the course on Accelerators for Medical Applications, organised by the CERN Accelerator School (CAS). The course was held at the Eventhotel Pyramide, Vösendorf, Austria from 26 May to 5 June, in collaboration with MedAustron.

Following introductory lectures on radiobiological and oncological issues, the basic requirements on accelerators and beam delivery are reviewed. The medical applications of linear accelerators, cyclotrons and synchrotrons are then be treated in some detail, followed by lectures on the production and use of radioisotopes and a look at some of the acceleration techniques for the future.

Preface

The aim of the CERN Accelerator School (CAS) is to collect, preserve and disseminate the knowledge accumulated in the world's accelerator laboratories over the years. This applies not only to general accelerator physics, but also to related sub-systems and associated technologies, and to how these are adapted to particular requirements. This wider aim is achieved by means of specialized courses currently held twice per year. The topic of the first 2015 specialized course was Accelerators for Medical Applications and was held at the Eventhotel Pyramide, Vösendorf, Austria from 26 May to 5 June 2015.

The course was made possible through the fruitful collaboration with the MedAustron centre in Wiener Neustadt, in particular through the efforts of Thomas Schreiner and Ursula Schindler. The backing of the CERN management and the guidance of the CAS Advisory and Programme Committees enabled the course to take place, while the attention to detail of the Local Organising Committee and the management and staff of the Eventhotel Pyramide ensured that the school was held under optimum conditions.

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 76 participants of 29 nationalities, from institutes in many 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 Barbara Strasser and to the CERN Publishing Service, especially Valeria Brancolini for her very positive and efficient collaboration.

These proceedings have been published in paper (black and white) and electronic form. The electronic version, with full colour figures, can be found at https://e-publishing.cern.ch/index.php/CYRSP/issue/view/33.

Roger Bailey,
Head of the CERN Accelerator School

PROGRAMME

Accelerators for Medical Applications, 26 May – 5 June, Vösendorf, Austria, 2015

Time	**Tuesday 26 May**	**Wednesday 27 May**	**Thursday 28 May**	**Friday 29 May**	**Saturday 30 May**	**Sunday 31 May**	**Monday 1 June**	**Tuesday 2 June**	**Wednesday 3 June**	**Thursday 4 June**	**Friday 5 June**
08:30 09:30		Opening Talks	Overview of Particle Accelerators R. Bailey	Overview of Linacs A. Lombardi	Cyclotrons for Particle Therapy M. Schippers		Beam Dynamics in Synchrotrons I B. Holzer		Therapy Control and Patient Safety M. Grossman	FFAGs S. Sheehy	
09:30 10:30	**A**	Interaction of Particles with Matter A. Ferrari	Ion Sources for Medical Applications S. Gammino	Accelerating Structures A. Degiovanni	Magnetic Design and Beam Dynamics I W. Kleeven	**E** **X**	Beam Dynamics in Synchrotrons II B. Holzer		Applications of Radioisotopes U. Koester	PWA M. Roth	**D** **E**
		COFFEE	**COFFEE**	**COFFEE**	**COFFEE**		**Coffee**		**Coffee**	**Coffee**	**P**
11:00 12:00	**R** **R** **I**	Radiobiology of Particle Beams I P. Scalliet	Beam Instrumentation A. Peters	Beam Dynamics and Layout A. Lombardi	Magnetic Design and Beam Dynamics II W. Kleeven	**C** **U**	Extraction Methods K. Noda		Production of Radioisotopes for Medical Applications I T. Stora	Dielectric Laser Acceleration P. Hommelhoff	**A** **R**
12:00 13:00	**V** **A**	Radiobiology of Particle Beams II P. Scalliet	Gantries M. Pullia	Powering E. Montesinos	RF For Cyclotrons S. Brandenburg	**R** **S**	Beam Lines and Matching to Gantries M. Pullia	**Full Day Visit to MedAustron**	Production of Radioisotopes for Medical Applications II T. Stora	Case Study Presentations	**T** **U** **R**
	L	**LUNCH**	**LUNCH**	**LUNCH**	**LUNCH**		**Lunch**		**Lunch**	**LUNCH**	
14:30 15:30	**D**	Dose Delivery Concepts M. Donetti	Dose Delivery Instrumentation S. Giordanengo	Industrial Design T. Wilson	Transport and Energy Adjustment of Cyclotron Beams M. Schippers	**I** **O**	Medical Physics Commissioning D. Meer		Case Study Work	Case Study Presentations	**E**
15:30 16:30	**A** **Y**	Dose Delivery Verification S. Safai	Patient Workflow S. Delacroix	Case Study Work	Case Study Work	**N**	Case Study Work		Case Study Work	Case Study Presentations	**D** **A**
		TEA	**TEA**	**TEA**	**TEA**		**TEA**		**Tea**	**TEA**	
17:00 18:00	**Registra tion**	Case Studies Introduction M. Pullia	Imaging K. Parodi	Future Trends in Linacs A. Degiovanni	Future Trends in Cyclotrons T. Antaya		Future Trends in Synchrotrons J. Flanz		Case Study Work	Closing Talk Closing Reception	**Y**
19:30	**Dinner**	**Dinner**	**Dinner**	**Dinner**	**Dinner**	**Special Dinner**	**Dinner**	**Dinner**	**Dinner**	**Dinner**	

Contents

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015,
edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Radiobiological Characterization of Clinical Proton and Carbon-Ion Beams

P. Scalliet and J. Gueulette
Université Catholique de Louvain, Cliniques Universitaires Saint Luc, 10, avenue Hippocrate, 1200 Brussels, Belgium

Abstract

Electromagnetic radiation (photons) or particle beam (protons or heavy ions) have similar biological effects, i.e. damage to human cell DNA that eventually leads to cell death if not correctly repaired. The biological effects at the level of organs or organisms are explained by a progressive depletion of constitutive cells; below a given threshold, cell division is no longer sufficient to compensate for cell loss, up to a point where the entire organism (or organ) breaks down. The quantitative aspects of the biological effects are modulated by the microscopic distribution of energy deposits along the beam or particle tracks. In particular, the ionization density, i.e. the amount of energy deposited by unit path length (measured in keV/µm), has an influence on the biological effectiveness, i.e. the amount of damage per energy unit deposited (measured in gray or Gy, equivalent to 1 joule/kg). The ionization density is usually represented by the Linear Energy Transfer or LET, also expressed in keV/µm. Photon beams (X-rays, g-rays) are low-LET radiation, with a sparsely ionising characteristic. Particle beams have a higher LET, with a more dense distribution of energy deposits along the particle tracks. Protons are intermediary, with a LET larger than the photon one, but still belong to the 'radiobiological' group of low LET. The higher the ionization density, the higher the biological effectiveness per unit of dose. When comparing various radiation qualities, it appears that the ionization density is relatively homogeneous along photon tracks, whereas it strongly varies along particular tracks (protons, heavy ions). In the first instance, the biological effectiveness is proportional to the TEL, itself dependant on the particle beam energy. So, when the LET of a particle beam is increased, its biological effectiveness increases in proportion. Secondly, a low-energy beam (f.i. 4 MeV a rays) has a higher LET than a high-energy beam (f.i. 200 MeV a rays). As particle beams continuously loose their energy through their successive interactions with the irradiated medium, it ensues that the LET slowly increases along the beam path, down to a point where all energy has been imparted and the beam stops. Therefore, the biological effectiveness is not homogeneous along the beam path (like with low-LET radiation), with a strong reinforcement at the end of the particle tracks (in the Bragg peak). The modelization of the clinical effects of particle beams is therefore very challenging, as a variable biological weighting function needs to be incorporated in the planning process to account for the increase in biological effectiveness with the progressive loss of beam energy.

Keywords

Proton beam; carbon ion beam; radiobiology.

https://doi.org/10.23730/CYRSP-2017-001.1

1 Introduction

The performance of radiotherapy can be improved in two separate ways: (a) improvement of the ballistic selectivity (increasing the dose to the tumour while reducing the exposure of normal tissue) and (b) improvement of the biological effectiveness of the radiation (using radiation with a higher relative biological effectiveness, RBE). Clinical proton beams are ballistically superior but biologically equivalent to X-rays (and gamma rays), while carbon-ion beams are both ballistically and biologically more efficient than X-rays.

From the biological point of view, the effect of radiation on living material is mainly due to DNA damage and its consequences: DNA disruption, loss of genetic information, incapacitation of vital genes, and, eventually, cell death. Indeed, severe DNA damage rapidly induces cell apoptosis (a sophisticated mechanism of auto-destruction) if the DNA is not correctly repaired in time (a few hours). It may, however, happen that some cells manage to survive despite severe DNA damage, but they often do so with some amount of 'misrepair', i.e., incorrectly repaired DNA damage with a change in the information sequence (gene inactivation, gene promotion, etc.).

Cells that survive severe DNA damage are rare, but can be dangerous if they harbour gene alterations that may lead to cancer (loss of proliferation regulation, loss of apoptosis, and tissue invasion and colonization).

The sequence of events leading to cell death can be summarized in the following way:

- Energy is deposited in DNA (by primary and secondary electrons) along radiation tracks, in consecutive ionization events. In fact, what happens is an 'exchange' of energy between the radiation and peripheral electrons of the atoms constituting the DNA, causing electrons to break loose from atoms and collide further with other, neighbouring atoms. Often, a cluster of ionization arises from the accumulation of the primary energy exchange event and the interactions of secondary electrons.
- The appearance of positive charges in the DNA molecule causes a rearrangement or, worse, a complete disruption of the molecular structure if the energy imparted exceeds the binding energy of the atoms. The sequence of information coded in the DNA strands is therefore interrupted by a break.
- DNA rearrangement follows detection of damage. This process is amazingly fast: induction of repair enzymes starts within minutes after DNA damage.
- Repair induction (enzyme synthesis) occurs in proportion to the amount of damage, though only up to a certain point, as massive damage tends to saturate the repair mechanisms. Repair usually takes 4 to 6 hours. Slight damage can be repaired faster; severe damage takes longer. Owing to the dual structure of DNA, i.e., a structure of two long strands that mirror each other, a single-strand break is easily repaired, but a double-strand break is not. The repair mechanisms excise damaged DNA sequences and rebuild intact DNA by reading the 'mirror' strand. If both strands are damaged, the danger of a faulty repair (misrepair), i.e., a repair resulting in DNA with modified information, is larger.
- The cell cycle is arrested at the same time as repair induction, in order to 'lend' sufficient time for repair before the next cell division is triggered.
- Along with DNA repair, apoptosis is also triggered. This might look illogical but, in fact, it is a strong protective mechanism against misrepair. Apoptosis is a stepwise process, each step being reversible up to a certain 'no-return' point at which the process becomes unstoppable. This point is reached after a definite time period. If at that point the repair is not finished (because the damage is too dense), then apoptosis proceeds until cell death. Conversely, if the repair is finished before that point (because the damage is limited), then apoptosis is stopped and the cell

survives. How does this mechanism protect cells? In fact, it does not protect individual cells, but it protects the information conveyed by a cell population, by eliminating severely damaged cells that are at risk of misrepair and corruption of the information in the DNA.[1]

- The cell dies (as a result of apoptosis or misrepair) or survives (with either adequate or inadequate DNA repair).
- Tissue failure (in the case of normal tissue) occurs if enough cells have been destroyed, or cancer cure is achieved (in the case of cancer) if all cancer cells have been destroyed.

2 Density of ionization and microdosimetry

Whether the damage to DNA is light or severe depends in the first instance on the amount of energy dissipated in the molecule (the 'dose'), but it also depends on the density of ionizing tracks crossing the molecule. A few dense tracks are biologically more effective than several sparsely ionizing tracks. Therefore, smaller doses with 'dense' tracks are as effective in killing cells as larger doses with less dense tracks.

This density is related to the amount of energy per unit track length and to the distance between consecutive energy deposition events along the track of the particle (a photon in the case of X-rays, or a proton or carbon ion in hadron therapy); more specifically, it is related to the number of energy deposition or ionization events that occur in the short diameter of the DNA (a few nanometres). Densely ionizing radiation usually deposits enough energy to inactivate a cell in one single track, whereas sparsely ionizing radiation requires the cooperation of several tracks, each depositing a small amount of energy insufficient to kill a cell, to achieve the same result (Fig. 1).

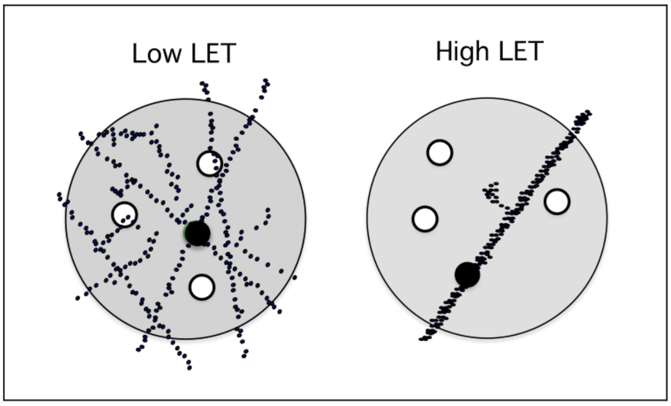


Fig. 1: Schematic representation of particle tracks for low-LET (left) and high-LET (right) radiation [1, 2]. For low-LET radiation, the inactivation of a radiosensitive target requires the conjunction of several tracks, whereas for high-LET radiation the impact of a single track is always fatal (closed circles).

The linear energy transfer (LET), a quantity expressed in keV/μm of particle track, measures the density of ionization per unit length along radiation tracks. Types of radiation can be sorted by their LET, with a customary distinction between low-LET (<10–20 keV/μm) and high-LET (>20 keV/μm) radiation.

Energy deposition in DNA is a quantized or random event, sometimes important, sometimes not. But the maximum energy that can be imparted in a single interaction depends directly on the LET. Low-LET radiation therefore very seldom kills cells with a 'single hit', whereas this is very common with high-LET radiation.

[1] This explains why embryos exposed to radiation most often do not survive. Indeed, inheritable DNA mutations have not been observed in the Hiroshima and Nagasaki survivors.

One particular dosimetric method allows individual energy deposition events to be measured in the eV range. This method is called 'microdosimetry', as its purpose is to describe energy exchange at the molecular level. In short, the overall concept is to miniaturize a dosimeter and to expose it to an extremely low particle fluence in order to register interactions separately. These interactions are measured by collecting electric charges created in a counter whose volume is artificially reduced by lowering the gas pressure in the measurement chamber. A very low pressure of a tissue-equivalent gas mimics a very small volume at normal atmospheric pressure, in the range of cubic micrometres. Nowadays, most microdosimetry is no longer done physically but done 'in silico' using Monte Carlo simulation methods.

Irradiating at low fluence and integrating all energy deposition events in a single graph yields a 'microdosimetric spectrum' specific to the radiation or particle type tested (Fig. 2). Small variations in the particle energy are reflected in small variations in the microdosimetric spectrum. In turn, variations in the microdosimetric spectrum illustrate differences in biological effectiveness, i.e., in the proportion of cells irreversibly damaged when it comes to cell kill.

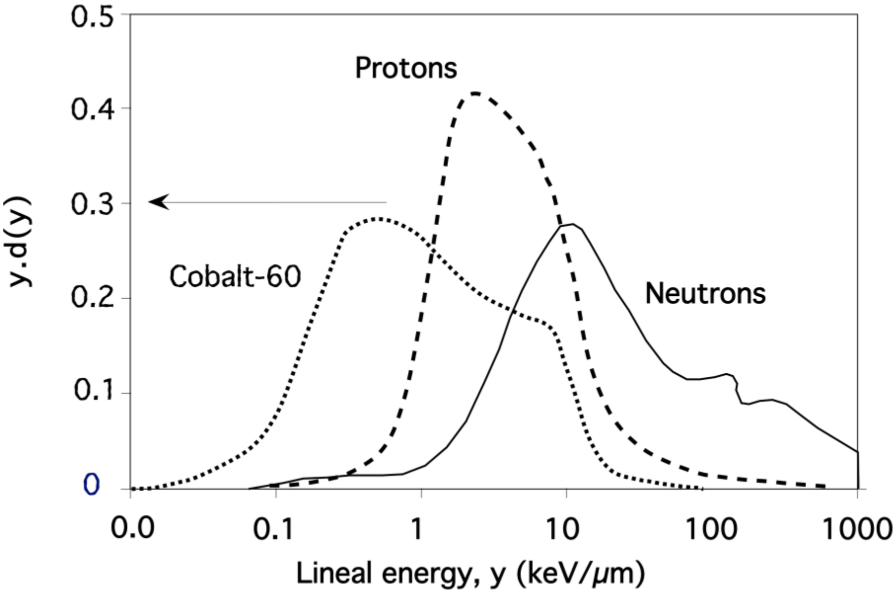


Fig. 2: Comparison of microdosimetric spectra of *y.d*(*y*) vs. *y* obtained for cobalt-60 gamma rays, 65 MeV protons, and p(65) + Be neutrons [3]; *y* is the lineal energy and *d*(*y*) is the probability density of the absorbed dose with respect to *y*. For cobalt-60 gamma rays, the maximum *y.d*(*y*) values occur at about 0.3 keV/µm. For protons and neutrons, the maxima are observed at about 3 keV/µm and 10 keV/µm, respectively.

3 Relative biological effectiveness

Again, minute energy deposition events are unable to damage DNA in a significant manner, whereas massive energy deposition events invariably kill the cell. When different types of radiation are compared (e.g., X-rays, neutrons, and alpha particles), the relationship between dose and cell survival shows that for the same amount of energy dissipated (i.e., the same radiation dose), the number of cells killed increases with the LET (Fig. 3).

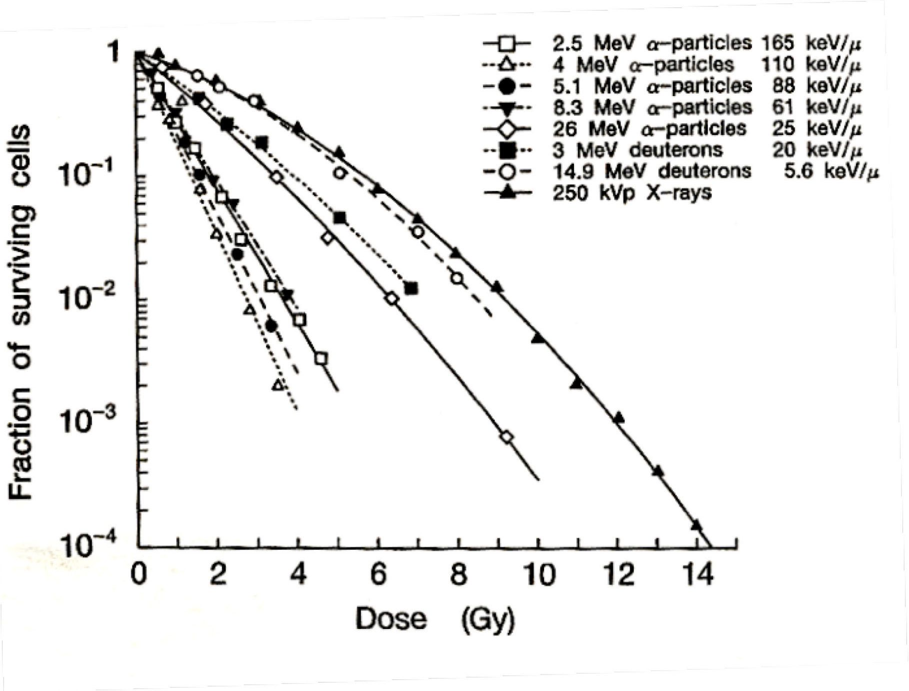


Fig. 3: Survival curves for cells exposed to radiation of different LET. The slopes of the curves become steeper as the LET of the radiation increases. The bending of the curves (i.e., the initial shoulder) is also progressively reduced. (Redrawn from Ref. [4].)

Conversely, for an identical cell kill level (known as an isoeffect), the dose required decreases as the LET increases. The ratio of the low-LET dose to the high-LET dose required for an isoeffect is called the RBE. It tends to increase with LET up to a maximum value, depending on the isoeffect considered (Fig. 5). Above this maximum value, the RBE decreases as the ionization density becomes very large, and most of the energy is wasted (an overkill phenomenon).

The RBE is a dimensionless quantity (as it is a ratio of doses) that compares the biological effectiveness of a given type of radiation with another type taken as a reference, usually cobalt-60 gamma rays. Thus, when cobalt-60 radiation is compared with itself, the RBE is 1. Up to an LET of around 20 keV/µm, the RBE remains stable at 1. Above this LET value, the RBE increases rapidly to a maximum value at around 100 keV/µm.

RBE values depend on the isoeffect level chosen for the comparison of radiation beams. Small radiation doses tend to increase the RBE, since at low doses, low-LET radiation is very ineffective in killing cells, whereas high-LET radiation is quite effective in doing so. At higher doses, low-LET radiation becomes more lethal and the difference in effect between low- and high-LET radiation becomes smaller. At very high doses, the RBE reaches a stable value that no longer depends on the dose (Fig. 4). The RBE also depends on the biological system under consideration.

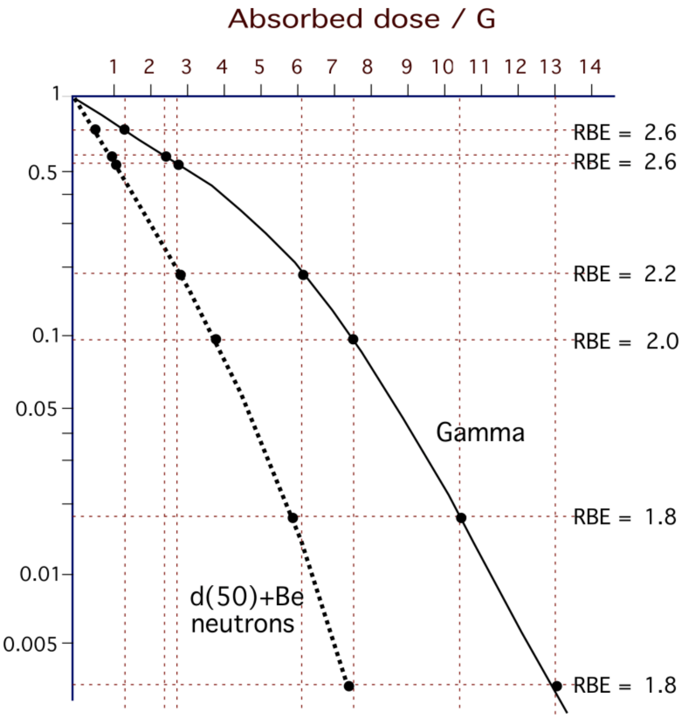


Fig. 4: Survival curves for intestinal crypt cells irradiated with neutrons or cobalt-60 gamma rays. The RBE is particularly variable in the initial part of the curves (i.e., for small doses), where it reaches its highest value. This variation is mainly due to the bending (i.e., the shoulder) of the gamma-ray curve. As the dose increases, the RBE stabilizes progressively, tending towards a minimum.

4 Time factor and repair

So, it is the spatial structure of the ionization events that characterizes the various types of radiation beams. If dense ionization crosses DNA, it will invariably destroy it beyond any possibility of repair. Too much information is lost in the event. Carbon-ion beams belong to the class of radiation that causes such dense ionization, i.e., high-LET radiation.

Low-LET radiation, conversely, only kills cells by the cooperation of several tracks that occur together spatially and in time: spatially to ‘build up’ damage at a particular molecular site, and in time because consecutive ‘hits’ must occur before the previous ‘hit’ has been repaired. For this reason, lowering the dose rate of the irradiation sharply decreases the biological effectiveness, as more time is available for the repair of damage before the next ionization takes place. Conversely, lowering the dose rate with high-LET radiation does not alter the biological effectiveness much, since a single hit is sufficient to inactivate a cell.

Another way to decrease the biological effectiveness of low-LET radiation is to deliver the dose in several small fractions, separated by sufficient time for DNA repair. In this case, spatial cooperation does not work, since small amounts of DNA damage are repaired between consecutive fractions. The longer the time between fractions, the more thorough the repair of the DNA. Again, this does not influence the biological effectiveness of high-LET radiation, as one single hit in the DNA is usually sufficient to kill the cell, without the need for spatial cooperation.

5 The effect of oxygen

It was observed early in the history of radiobiology that the radiosensitivity of cells depends on the partial pressure of oxygen in the immediate environment. When oxygen is present at normal atmospheric concentration, the radiosensitivity is at its highest. Lowering the partial pressure of oxygen progressively decreases the radiosensitivity, by a factor of that reaches 3 when oxygen is absent (or nearly absent) (Fig. 5).

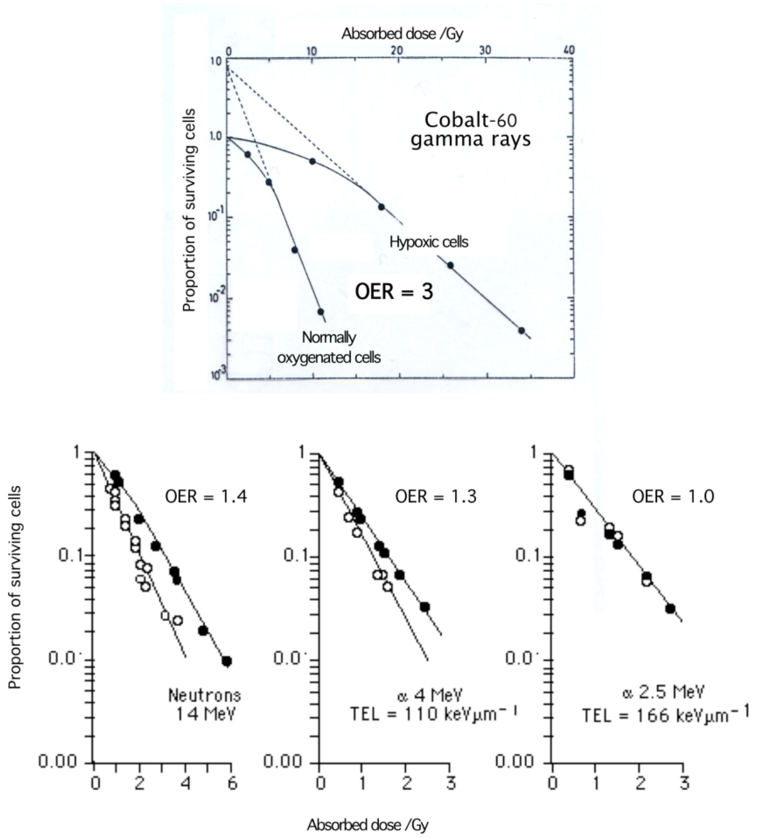


Fig. 5: Comparison of the influence of partial pressure of oxygen for radiation with different LETs. OER (Oxygen Enhancement Ratio) equals 3 for low-LET radiation; this value seems to be independent of the dose level and, to a certain extent, of the biological system. The OER value is thus commonly interpreted as a ‘scale factor’. The OER value decreases as the LET increases, down to 1 for very high-LET particles. Redrawn from Refs. [4, 5].

Obviously, no aerobic cell (as are human cells f.i.) is viable for more than a few minutes without any oxygen supply. But at a low partial pressure, some survival, often in a quiescent state, remains possible for hypoxic cells, which can then escape the effect of radiation doses that would otherwise be lethal.

In oncology, the blood supply of cancerous masses is commonly deficient, and large regions of hypoxia exist in virtually all tumours. This has proven to be a problem of importance in radiotherapy, and ways to overcome hypoxic radioresistance have been developed (see Section 7).

Oxygen can thus be considered as a ‘natural’ radiosensitizer. Without entering too much into details, we can say that the presence of oxygen ‘fixes’ damage in the DNA. What actually happens is that with low-LET radiation, most of the damage is absorbed by water, in the close vicinity of the DNA molecule. In physics terms, the DNA itself has a very small ‘cross-section’ for X-rays, and most of the damage to it is created by radiolysis of water, which creates free radicals, which in turn interact secondarily with the DNA molecule. The life-span of these free radicals and their ability to migrate some distance is influenced by the presence or absence of oxygen. It is said that oxygen is needed to ‘fix’ the damage (‘fix’ in the sense of fixation, not of repair). In its absence, the free radicals are less toxic to the DNA.

This influence of the partial pressure of oxygen is at its highest with low-LET radiation. As the LET of the radiation is progressively increased above 20 keV/μm, the sensitizing effect of oxygen progressively disappears because DNA damage is now usually the consequence of a direct, very dense hit on the molecule, whose fixation no longer requires the presence of oxygen. This is what makes carbon-ion beams clinically so attractive, since their LET is in the range where no oxygen is required to 'fix' the lethal DNA damage. Indeed, carbon ion therapy is advocated for the treatment of cancer types in which a large hypoxic component is suspected.

6 Cell cycle and cell division

Cell division is a lengthy and subtle process in which the cell duplicates its DNA before starting to physically divide. The structure of DNA allows this process to be precise and fail-safe (indeed, this is a condition for life). By duplicating each of the two strands constituting the DNA molecule, the cell doubles its set of chromosomes. The two sets then migrate in opposite directions, and the cell is cleaved between them.

DNA synthesis relies on a set of specific enzymes that 'gently' separate the two DNA strands and synthesize a new copy on each of them. The end result is two identical DNA molecules. The same enzymes are mobilized in the case of accidental DNA damage during the lifetime of a cell; they excise the damaged DNA section and then resynthesize the missing part of the DNA, using the other strand as a template for the exact restitution of the coded information.

The synthesis and repair enzymes normally have a very low concentration in the nucleus, at times not close to cell division. But any damage will trigger enzyme synthesis (in a matter of minutes) in order for repair to proceed efficiently.

Conversely, during cell division, the DNA synthesis enzymes are at their maximum concentration in the nucleus. If radiation damage is inflicted during cell division, the cell tends to be more resistant as the nucleus is already saturated with all of the enzymes needed for repair, especially at the end of DNA synthesis, when the enzymes are no longer required for duplication and therefore are free for binding to any new substrate (damaged DNA, for instance). When DNA synthesis is finished, the cell remains quiescent for some time (called the G2 phase), just before physically dividing. At that point the cell is at its most vulnerable to radiation damage, since repair is less effective in the presence of all the rearrangement needed for cell division.

The variation in the sensitivity of cells with the division cycle is more pronounced with low-LET radiation, quite logically, since repair plays an important role in the end result of radiation exposure. In contrast, the sensitivity to high-LET radiation damage is independent of the cell cycle, since repair plays no or only a very minor role in the end result (Fig. 6).

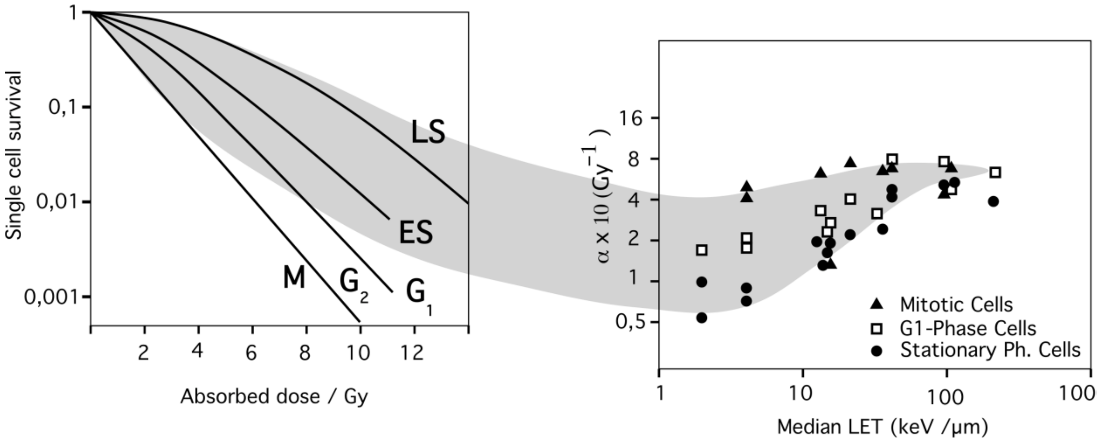


Fig. 6: Left panel: survival curves for cells irradiated in different phases of their mitotic cycle. Right panel: variation of the slope of the curves (the parameter α) as a function of LET. The gap between the slopes of the various curves (i.e., the difference in radiosensitivity) gradually becomes smaller as the LET increases.

7 Biological weighting function

The preceding considerations about the variations of radiosensitivity and the biological effectiveness of low- and high-LET radiation find quantitative expression in the so-called Biological Weighting Function (BWF), which is obtained by plotting the RBE against the LET (Fig. 7) [6, 7]. BWFs are specific to a given biological effect and given irradiation conditions, so that the peak value (the maximum of the RBE) and its place in the LET range may vary substantially. The main sources of variation are the dose and the oxygenation status of the biological material: the RBE of high-LET radiation with respect to cobalt-60 gamma rays increases as the dose decreases and when the partial pressure of oxygen decreases. Recall here that the higher RBE of carbon ions for hypoxic cells with respect to normally oxygenated cells is one of the main justifications for using these particles in radiotherapy.

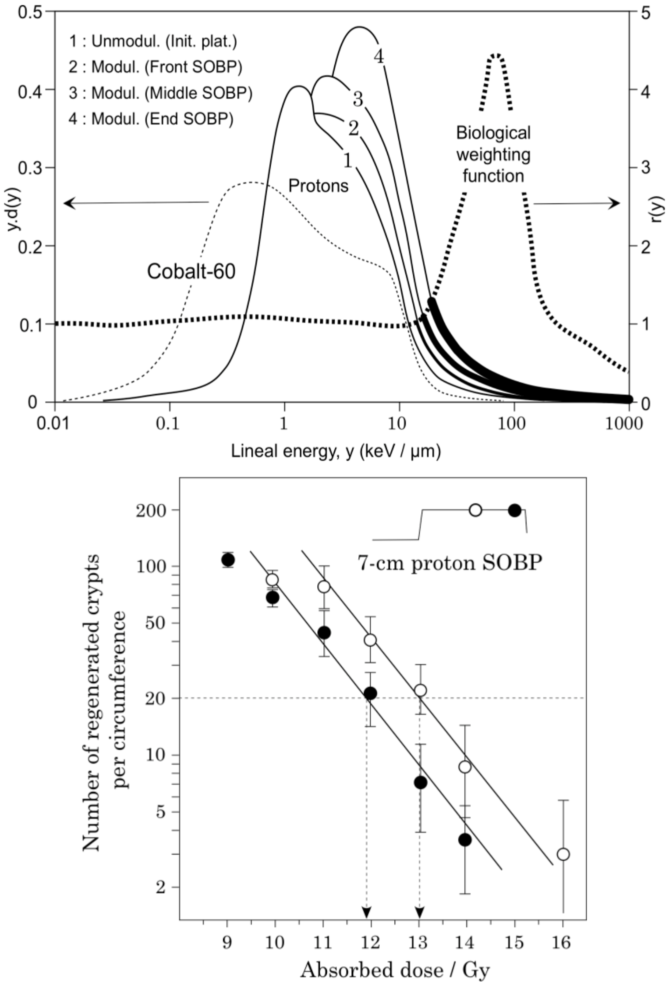


Fig. 7: Top: microdosimetric spectra for a 90 MeV energy-modulated proton beam. Measurements at four positions are shown (solid lines). These are compared with cobalt-60 gamma rays (dotted line). The BWF for different LET values has been superimposed (bold dotted line). The proton microdosimetric spectra are shifted towards higher LET values when measured more distally in the SOBP (Spread Out Bragg Peak - bold solid lines). This suggests that the proton RBE would also increase with depth. Bottom: dose–effect relationships for crypt regeneration in mice after irradiation in a single fraction with a 200 MeV energy-modulated proton beam at iThemba Labs, South Africa. The beam was modulated to produce a 7 cm SOBP. The open and closed circles correspond to irradiation in the middle and at the end of the SOBP, respectively (see the sketch of the dose profile at the top of the panel). Each point is the average of the readings for four mice. Parallel exponential regression curves were fitted through the points by a weighted least squares method. The error bars correspond to the 95% confidence intervals. In the case shown here, the RBE increases by 9% on moving from the middle to the end of the SOBP.

8 Proton beams

The RBE of protons is just above unity, usually around 1.1–1.15, indicating that the fractionation effect still matters with this type of radiation. Proton beams are a characteristic low-LET radiation.

A more precise examination of their microdosimetric spectrum, however, shows that small variations in LET can be observed along the particle paths. At the entrance to the irradiated medium, high-energy protons are sparsely ionizing, and thus typically low-LET. When they are close to the end of their path, at the place where all the remaining kinetic energy will be released (the Bragg peak), the LET increases sharply, though by only a modest amount. But this small change is sufficient to modify the RBE. Therefore, the RBE is not constant over the entire proton path.

The explanation is quite simple: as the protons enter the medium and penetrate deeper and deeper, they progressively release some of their kinetic energy and slow down, until they reach the end of their path. As the speed decreases, the distance between consecutive energy deposition events also decreases; hence, the LET increases, and the RBE increases in turn. Proton beams thus do not have a constant biological effectiveness along their path. But, again, the variation in RBE remains modest.

Precise radiobiological experiments with mice, using a model of intestinal toxicity, have been able to measure these RBE variations at the end of the proton path, by repeating measurements across the small distance covering the end of an extended Bragg peak. Radiobiological data demonstrate that the RBE indeed varies, and in the proportions predicted by the microdosimetric shift in LET.

9 Carbon-ion beams

The physics of carbon-ion beams is similar to that of proton beams (a plateau at the entrance point of the beam, followed by a sharp rise in dose at the end of the path, at the Bragg peak), but the mass of carbon ions is much greater. Therefore, carbon-ion beams are high-LET along their entire path, though with a similar pattern to that of protons: the LET is lower near the entrance point of the beam and at its highest at the Bragg peak.

The range of the RBE is close to 2–3 at the Bragg peak, and closer to 1.5–2 at the entrance point. The variation of the RBE along the path of a carbon ion is therefore much larger than for protons, which calls for some adjustment when irradiating under clinical conditions.

Usually, when a tumour is irradiated with X-rays or gamma rays (low-LET radiation), a homogeneous level of dose is delivered to the entire volume, with the intention of delivering an ‘isoeffect’ to the tumour. No specific attention to the size of the tumour is needed, as the biological effectiveness of low-LET radiation remains the same across the entire irradiated volume (RBE = 1).

This is not the case with carbon-ion beams, since the biological effectiveness increases significantly within the irradiated volume. If a homogeneous ‘biological’ dose is planned throughout the tumour volume, then some alteration to the ‘physical’ dose needs to be made, to compensate for the variation in RBE [8]. Precise measurements that illustrate this variation in RBE have been done with the clinical carbon-ion beam at HIMAC (Fig. 8).

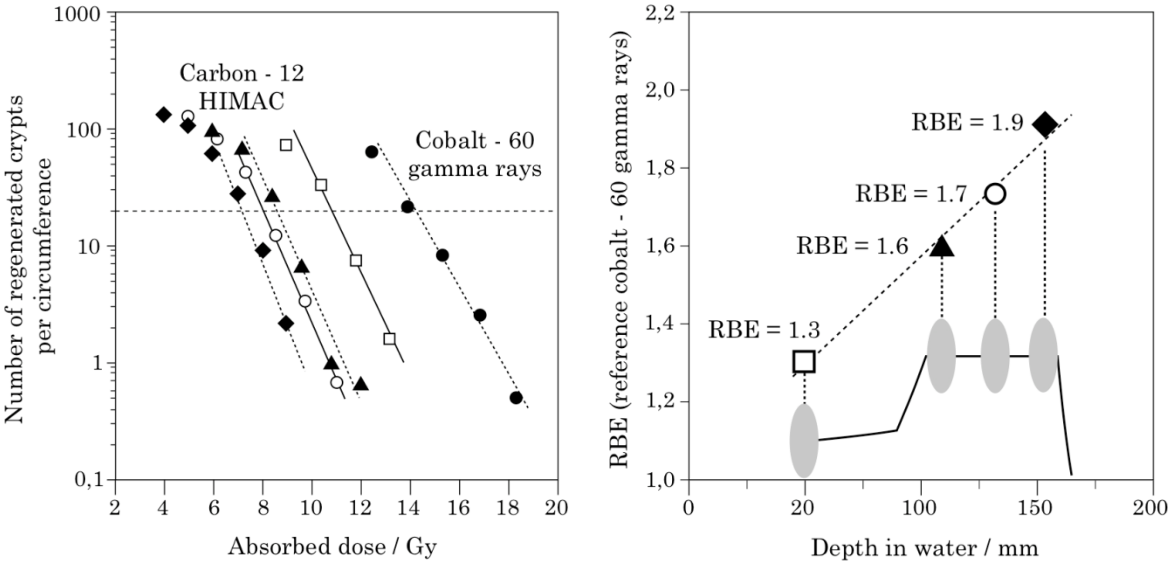


Fig. 8: Left panel: dose–effect relationships for intestinal crypt regeneration in mice after irradiation with cobalt-60 gamma rays or carbon-12 ions at the entrance plateau and at different positions in a 6 cm SOBP (the positions are shown in the sketch in the right panel). Right panel: the corresponding RBEs (reference cobalt-60 gamma rays), plotted against the depth, indicate a substantial increase in the RBE. As the irradiations were performed with single high doses, these RBEs are much lower than those for fractionated irradiations, which reach a value of approximately 3 at the end of the SOBP.

References

[1] D.T. Goodhead, *Can. J. Phys.* **68** (1990) 872. http://dx.doi.org/10.1139/p90-125

[2] M. Tubiana, *Radiobiologie, Radiothérapie et Radioprotection* (Mermann/Médecine, Paris, 2008).

[3] J. Gueulette, H.G. Menzel, P. Pihet and A. Wambersie, in *Recent Results in Cancer Research*, Eds. R. Engenhart-Cabillic and A. Wambersie (Springer, Berlin, 1998), pp. 31–53. http://dx.doi.org/10.1007/978-3-642-78774-4_2

[4] J.J. Broerse, G.W. Barensen, and G.R. van Kersen, *Int. J. Radiat. Biol.* **13** (1967) 559.

[5] G.W. Barendsen, C.J. Koot, G.R. van Kersen, C.K. Bewley, S.B. Field and C.J. Parnell, *Int. J. Radiat Biol.* **10** (1966) 317. http://dx.doi.org/10.1080/09553006614550421

[6] P. Pihet, H.G. Menzel, R. Schmidt, M. Beauduin and A. Wambersie, *Radiat. Prot. Dosim.* **31** (1990) 437.

[7] T. Loncol *et al.*, *Radiat. Prot. Dosim.* **52** (1994) 347.

[8] J. Gueulette and A. Wambersie, *J. Radiat. Res.* **48** Suppl. A (2007) A97. http://dx.doi.org/10.1269/jrr.48.a97

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Dose Delivery Concept and Instrumentation

S. Giordanengo[1] *and M. Donetti*[2]
[1]Istituto Nazionale di Fisica Nucleare (INFN), Torino, Italy
[2]Centro Nazionale di Adroterapia Oncologica (CNAO), Pavia, Italy

Abstract

Radiation therapy aims to deliver the prescribed amount of dose to a tumour at the same time as sparing the surrounding tissues as much as possible. In charged particle therapy, delivering the prescribed dose is equivalent to delivering the prescribed number of ions of a given energy at each position of the irradiation field. The accurate delivery is committed to a dose delivery (DD) system that shapes, guides and controls the beam before the patient entrance. Most of the early DD systems provided uniform lateral dose profiles by using different devices, mainly patient-specific, placed in the beam line to shape the three-dimensional final target dose. More recently, systems that provide highly conformal dose distributions using thousands of narrow beams at well-defined energy were developed which feature advanced scanning magnets and real-time beam monitors, without patient-specific hardware. This lecture will cover the general dose delivery concept as well as the different DD instrumentations depending mainly on the beam delivery technique and on the particle and accelerator types. Some characteristic worldwide DD and beam monitor systems will be mentioned.

Keywords

Dose delivery; beam shaping; beam scanning; beam monitoring.

1 Introduction

In charged particle therapy the presence of a dose delivery (DD) system is mandatory to deliver the dose as prescribed. In general this translates into controlling the beam characteristics, such as the number of particles, in the expected position and with a defined spatial distribution. The DD system connects the accelerator to the patient (see Fig. 1(b)) and operates on the charged particle beams to provide the 'patient-specific' beam, based on clinical requirements. In fact, at the highest level, the overall goal of radiotherapy is to deliver the specific amount of dose to the target and in an acceptable time.

The delivery technique together with the beam characteristics provided by the accelerator are the essential information and constraints used by the treatment planning system (TPS) to compute the optimum treatment (i.e. the list of parameters to set up the accelerator and the dose delivery instrumentation) in order to meet the prescriptions defined by radiation oncologists and medical physicists. The TPS uses computer algorithms and radiobiological models to simulate the interactions between the charged particle beams and the patient's anatomy and to determine the spatial distribution of the radiation dose provided by known beam characteristics.

The charged particle beam, accelerated by either a cyclotron or a synchrotron, when it reaches the treatment room, is quasi-monoenergetic and the transverse profile can be approximated to a Gaussian with a width of the order of a centimetre or less. With such a beam, the dose delivered to a tissue would be initially fairly constant, with a sharp peak toward the end of travel of the ions (Bragg peak). Since the depth of the Bragg peak increases with the particle energy, the beam has to be spread transversely and modulated in energy to become clinically useful. Thus, specific dose delivery instrumentation is mandatory to perform this task.

https://doi.org/10.23730/CYRSP-2017-001.13

For deep-seated tumours the treatment is usually composed of two or more irradiation fields to minimize the dose in surrounding organs, especially in the tissues passed through by the beam to reach the tumour, i.e. in the entrance volume. Radiation oncologists and medical physicists work together to define the desired treatment outcome, i.e. doses and constraints, which is then converted into beam parameters required by the therapy machine to deliver the dose. From the DD system and accelerator point of view, each irradiation field is a new and independent delivery, so, in the following, we will refer to treatment as a single-field irradiation.

As sketched in Fig. 1(a), the pristine particle beams, provided by any accelerator, are too small and narrow compared with the tumour dimension. The beam has to be modelled or deviated in order to conform the dose to the target volume (Fig. 1(a)). Some DD components could be in the accelerator vacuum pipe while others are in air, hosted in the very last part of the beam transport line called the nozzle.

The DD instrumentation depends on the nominal beam characteristics available at the DD system entrance (i.e. accelerator type and performance) and on the required beam specifications, which depend on the dose delivery technique and on the specific patient and target. Two general dose delivery techniques have been developed and used to deliver charged particle radiation therapy: the passive scattering and the scanning (or dynamic) beam widening [1–3]. We will see that specific DD systems exist for each combination of accelerator, particle and beam delivery technique.

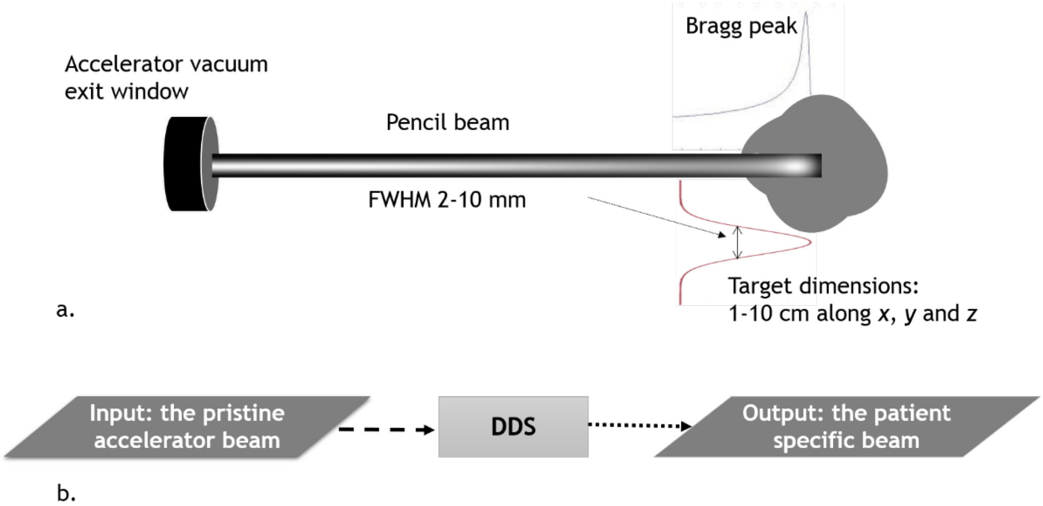

Bragg peak
Accelerator vacuum exit window
Pencil beam
FWHM 2-10 mm
Target dimensions: 1-10 cm along x, y and z
a.
Input: the pristine accelerator beam
DDS
Output: the patient specific beam
b.

Fig. 1: (a) Sketch of a charged particle beam from the accelerator vacuum exit window to the target; (b) scheme of principles for the role of the dose delivery system (DDS).

One fundamental role of any DD system is the continuous check of beam parameters by means of dedicated beam monitors in order to interrupt the irradiation in case of values out of ranges. Real-time feedbacks to DD instrumentation can come from targets or from independent detectors as well as from the accelerator control system in order to guarantee the treatment accuracy and safety.

1.1 From the clinical requirements to treatment delivery

Patient treatment is a process that involves a large number of systems developed and interfaced with each other in order to achieve desired clinical requirements. Therefore, the latter are the starting specification that guides the construction of a particle therapy centre. The DD system allows measuring the clinical beam parameters, such as dose, dose rate, range, distal falloff, penumbra and degree of dose conformity, through the physical beam parameters such as beam current, beam energy, beam shape and size and beam position. A single clinical parameter can depend on many beam parameters and vice versa. For example, a change in the beam energy or range, position or shape could influence the dose uniformity or the unwanted dose outside the target.

The beam features are the result of a combination of the design of the accelerator, the beam transport and the beam-spreading mechanisms; tolerances associated with each beam parameter are important and must be derived from the maximum tolerable uncertainties of the overall treatment process.

The DD system is one among several other systems, listed in the following, mandatory to perform any kind of radiotherapy treatment:

- accelerator;
- beam transport lines;
- dose delivery system;
- patient positioning system;
- patient position verification system;
- treatment planning system;
- oncological information system;
- safety system.

By selecting the particle type, the accelerator technology and the dose delivery technique, different outcomes can be achieved in term of good coverage of the target volume, treatment time, treatment cost and patient throughput. Strengths and weaknesses exist for each system.

The basic clinical requirements in the design of a facility are the maximum tumour depth (typically about 30 cm) and dimension (more than 40 cm for cranio-spinal irradiation), the treatment duration (i.e. 2 Gy/min/l) and the desired dose uniformity in the target (typically better than 1–2%). As an example for carbon and proton beams, these requests lead to the parameters and field characteristics listed in Table 1.

Table 1: Main beam parameters and target characteristics in charged particle therapy

Ion type	**Energy range (MeV/u)**	**Flux (N_{part}/s)**	**Transversal beam position resolution (mm)**	**Transversal field size (cm min–max)**	**FWHM (cm E_{min}–E_{max})**
Proton	70–250	10^9–10^{10}	±1	1–40	2.3–0.7
Carbon	70–400	10^7–10^8	±1	1–20	1–0.4

The bigger field size for proton beams is available with passive scattering techniques while for carbon ions 20 × 20 cm^2 is the recommended field dimension.

2 Dose delivery techniques

The delivery techniques that have been developed and used worldwide can be grouped under two broad categories: passive scattering and pencil beam scanning. Obviously each of these two categories has been implemented following different specific applications, but for sake of simplicity we will group them under the above classification: (1) 3D dose modulation with passive scattering techniques and (2) 3D dose modulation with pencil beam scanning techniques.

2.1 3D dose modulation with passive scattering techniques

The first charged particle delivery technique implemented has been the passive scattering. This method makes use of passive devices placed along the beam line that introduce scattering effects and energy degradation in order to spread out the Bragg peak transversally and along the beam direction; see Refs. [1–4].

This technique was taking advantage of the experience achieved with photon beams, known as conventional radiotherapy. In those applications, the photons (typically accelerated up to 18 MeV) are spread and collimated in the transverse plane by dose delivery devices placed in the accelerator nozzle. These devices are set to conform the irradiation field to the patient-specific case.

In the following sections we describe the use of the passive devices used in hadron therapy, such as: scatterers, range shifters, ridge filters, collimators, absorbers, apertures and range compensators.

2.1.1 Transversal beam modulation through beam spread and beam absorption

To increase the transverse dimensions of the pristine beam, one or two thin layers of high-*Z* materials, like lead and tantalum, are placed in the beam line. Materials with large values of *Z* are preferable because for a given energy loss the resulting multiple scattering angle is larger. Techniques using single or double scattering layers have been developed as shown in Fig. 2, reproduced from [5].

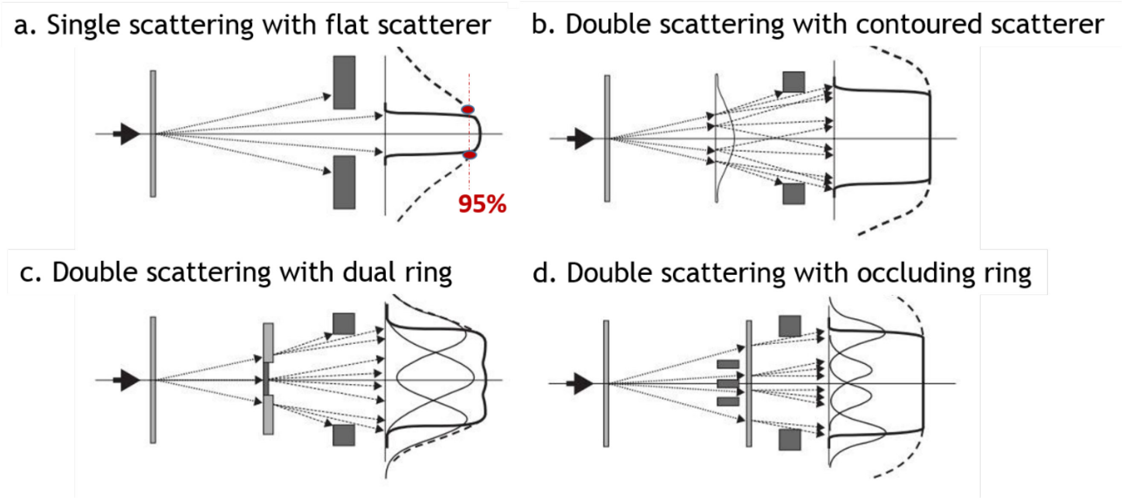

a. Single scattering with flat scatterer
95%
b. Double scattering with contoured scatterer
c. Double scattering with dual ring
d. Double scattering with occluding ring

Fig. 2: (Reproduced from [5]) Schematic representation of the single-scattering technique using a flat scatterer (a) and double-scattering techniques using a contoured scatterer (b), dual ring (c) and occluding ring (d). Dashed lines, lateral profile without aperture; solid lines, with aperture.

To keep the dose uniformity within ±2.5%, a collimator (aperture) blocks the beam outside the 95% dose level so the useful part of the beam is only the 5%. Because of this low efficiency, requiring relatively large beam currents and generating a high production of secondary neutrons, a single flat scatterer is limited to small fields with a diameter typically not exceeding ~7 cm (mainly ocular tumours). To achieve a larger and uniform dose distribution, a second scatterer has to be placed in the beam line. This type of system is called a ‘double-scattering system’ and can be implemented with three different types of second scatterers as shown in Fig. 2: (b) the contoured, (c) the dual ring and (d) the occluding ring. The shape of a contoured scatterer, thick in the centre and thin on the outside, increases the efficiency because the central particles scatter more to the outside and the flat profile increases (Fig. 2(b)). Typically, a first flat scatterer spreads the beam onto the contoured scatterer (second scatterer) that flattens out the profile at some distance. The characteristics of the contoured scatterer are optimized in combination with the scattering power of the first scatterer to obtain the desired dose distribution. Efficiencies of up to 45% can be obtained, significantly larger than in single scattering.

Charged particles hitting the centre of the contoured scatterer lose more energy than those going through the thinner parts at the periphery. To avoid a concave distortion of the distal isodose plane, with the range increasing away from the beam axis, energy compensation is applied to the contoured scatterer. A high-*Z* scattering material (lead, brass) is combined with a low-*Z* compensation material (plastic). The thickness of the two materials is designed to provide constant energy loss, while maintaining the appropriate scattering power variation. The thickness of the high-*Z* material decreases with distance from the axis, whereas the thickness of the compensating low-*Z* material increases. Note that energy compensation will increase the total water equivalent thickness of the scatterer and the energy of the particles entering the nozzle needs to be increased to achieve the same range in the patient as with a single scatterer.

An alternative to the contoured scatterer is the dual-ring scatterer shown in Fig. 2(c) and also described in Ref. [5]. It consists of a central disc made of a high-*Z* material (lead, tungsten) and a surrounding ring of a lower-*Z* material (aluminium, Lucite). The physical thickness of the outer ring is chosen such that the energy loss is equal (or close) to the energy loss in the central disc. The first flat scatterer spreads the beam onto the dual-ring scatterer. The central disc produces a Gaussian-like profile, and the ring produces an annulus-shaped profile, and both combine to produce a uniform profile at the isocentre (Fig. 2(c)). Because of the interface between the two materials, the dose distribution is not perfectly flat.

The third type of second scatterer is the dual ring, which blocks instead of scattering the central protons outward (Fig. 2(d)). The 'hole' created in the fluence distribution is filled in by scattering through a flat, second scatterer. Larger field sizes can be obtained by not just blocking the centre but by adding one or more occluding rings. Optimization of the ring diameters and the first scatterer power results in a flat dose distribution. Because the protons are not redistributed but blocked, the efficiency of an occluding ring system is significantly lower than for the contoured scatterers. The energy loss is smaller though, because a relatively thin second scatterer foil is needed to spread out the beam.

All the double scattering systems are sensitive to beam alignment while the contoured and dual-ring scatterers also depend on the beam phase space (see Section 3).

Once the lateral beam spread is achieved, a field-specific aperture (Fig. 3(a)), called a collimator, is used to conform it laterally such that the dose matches the maximum beams-eye-view extent of the target volume. The 2D projection of the target along the beam direction is presented in Fig. 3(b). Alternatively, a multileaf collimator (MLC), made of tungsten leaves [5] can be used (Fig. 3(c)) as a dynamic aperture.

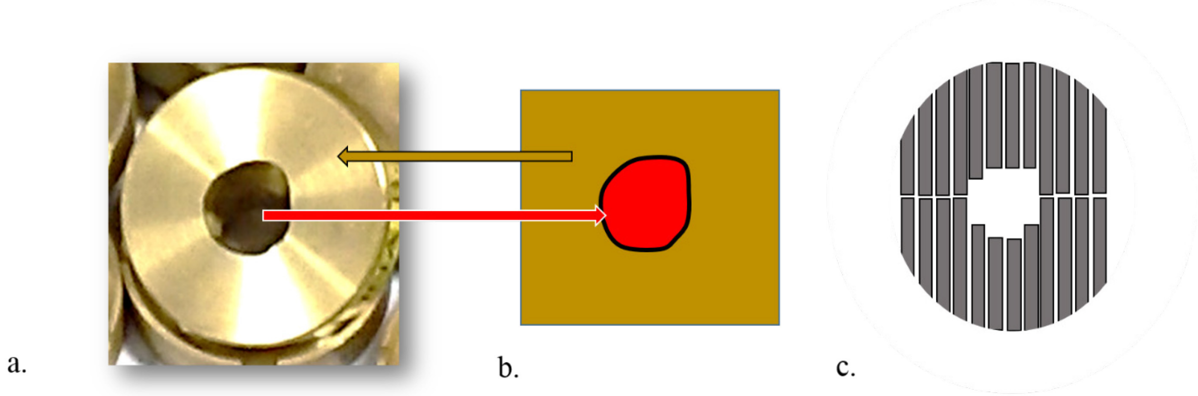


Fig. 3: (a) Example of brass field-specific aperture; (b) its 2D projection of the target; (c) scheme of a multileaf collimator.

Each delivered field requires apertures (or collimators) and range compensators to conform the dose to the target. The aperture absorbs the beam outside the target and conforms the beam laterally, while the range compensator is a patient-specific range shifter that conforms the dose delivered to the distal end of the target. It is also used to correct for patient surface irregularities and density heterogeneities in the beam path. Examples of lucite, wax and acrylic range compensators are shown in Fig. 4a, 4b and 5c respectively.

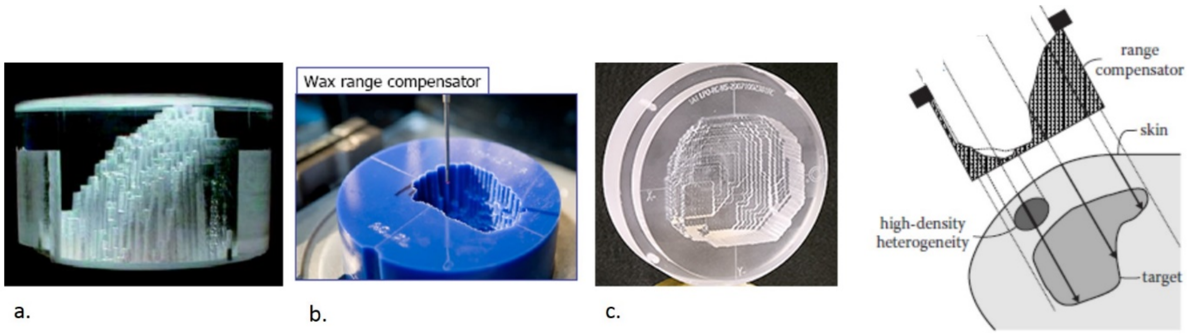


Fig. 4: Examples of range compensators made of lucite a); wax b) (photos by courtesy of MGH/LLUMC); acrylic c) (reproduced from http://www.oncolink.org/treatment/article.cfm?c=186&id=438) and a schematic representation of the application of a range compensator that compensates for the shape of the body entrance, the distal target shape, and inhomogeneities [5].

2.1.2 Longitudinal beam spread by beam range modulation

A uniform dose over the longitudinal extension of the tumour is obtained by modulating the beam range. The incident beam forms a flat dose region called the spread out Bragg peak (SOBP) by sequentially penetrating absorbers of variable thickness, e.g. via a range modulator. Each absorber contributes an individual pristine Bragg peak curve to the composite SOBP. A set of pristine peaks is delivered with decreasing depth and with reduced dose until the desired modulation is achieved. The result is the convolution of several Bragg peaks shifted in depth and with an appropriate weight, i.e. number of particles. Figure 5 shows a series of weighted pristine peaks as well as the resulting SOBP when these are superimposed.

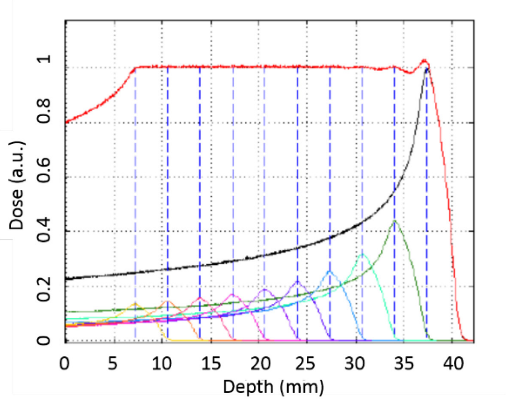


Fig. 5: Example of spread out Bragg peak (SOBP) to cover a 3 cm target extension in depth

The shape of the Bragg peak depends on the energy spread and scattering properties of the delivery system, so measured Bragg curves are used by TPS algorithms to determine the weights and the number of peaks necessary for each specific target volume treated in a specific facility.

The size of the SOBP is chosen to cover the largest extent in depth of the target volume. Since the SOBP size is constant over the entire target volume, in general, there is some pull back of the high-dose region into normal tissues proximal to the target volume.

2.1.2.1 Energy degraders: range shifters, range modulator wheels and compensators

Several different degraders have been developed and used either to change the fixed energy provided by the cyclotrons or to adapt the energy step and range provided by a given synchrotron. In Fig. 6, from the C. Ma and T. Lomax book [6], the following types are shown: (a) two or one adjustable wedges; (b) insertable slabs of graphite or Plexiglas, (c) rolled-up wedge; (d) insertable blocks with different thicknesses; (e) rotatable Plexiglas curved wedge; (f) adjustable multiwedge design.

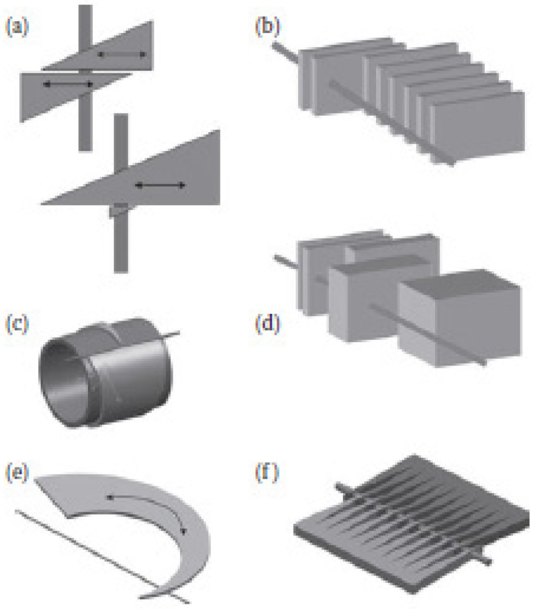


Fig. 6: Examples of energy degraders (reproduced from Ref. [6])

Different range modulator wheels, made of steps of varying thickness, are shown in Fig. 7. Each step corresponds to a pristine peak in the SOBP and the step thickness determines the range shift of that peak. When the wheel rotates in the beam, the steps are sequentially irradiated. The angular width of the step determines the number of protons hitting the step, and thus the weight of the pristine peak. By progressively increasing the step thickness while making the angular width smaller, a flat SOBP can be constructed. Like the range shifters used in energy stacking, modulator wheels are preferentially made of low-Z materials to limit scattering. Plastics (Plexiglas, Lexan) are often used, but for wheels that need to provide large range shifts and that are mounted in nozzles where space is limited, carbon and aluminium have been applied.

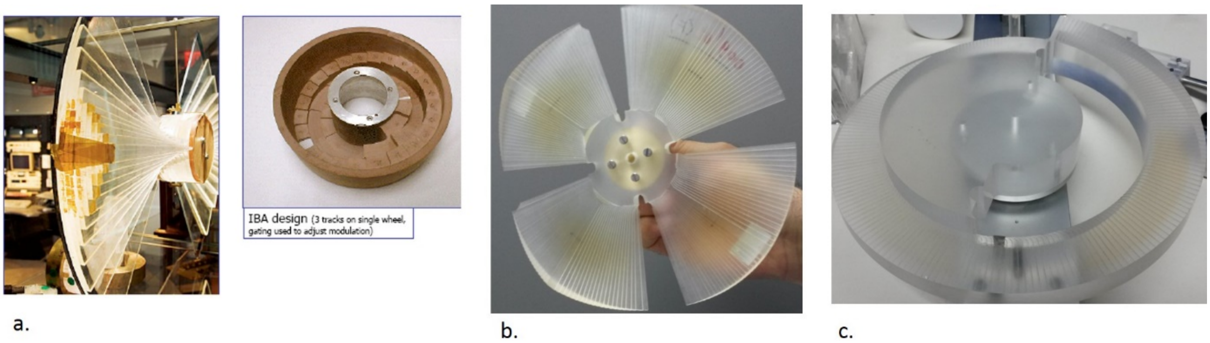

IBA design (3 tracks on single wheel, gating used to adjust modulation)
a.
b.
c.

Fig. 7: Examples of range modulator wheels (a. and b. photos by courtesy of IBA)

This technique is not optimal in terms of the conformity of the dose deposition and the use of a gantry is recommended. Particles heavier than protons produce fragments and an undesired dose is delivered after the Bragg peaks. Moreover, it results in an additional dose to the patient due to the neutrons produced by the beam passing through the various components.

2.2 Pencil beam scanning techniques: 3D modulated scanning ion therapy

The pencil beam scanning technique exploits the physical proprieties of charged particles: a thin ‘pencil’ beam (typically 3–10 mm FWHM (full width at half maximum) at isocentre) coming directly from the beam line is transported to the patient to achieve small depositions of dose. The tumour is irradiated by the superposition of a sequence of beamlets, hereafter called spots, each delivering a defined number of particles at a defined position with a monoenergetic beam (Fig. 8(b)). The target volume is segmented in several layers, called iso-energy slices (Fig. 8(c)), orthogonal to the beam direction, each corresponding to a different water equivalent beam penetration depth obtained by modifying the energy of the beam (directly through the accelerator for synchrotrons or through an energy-selection system for cyclotrons). The resulting dose delivery technique is a kind of 3D scanning of the target volume.

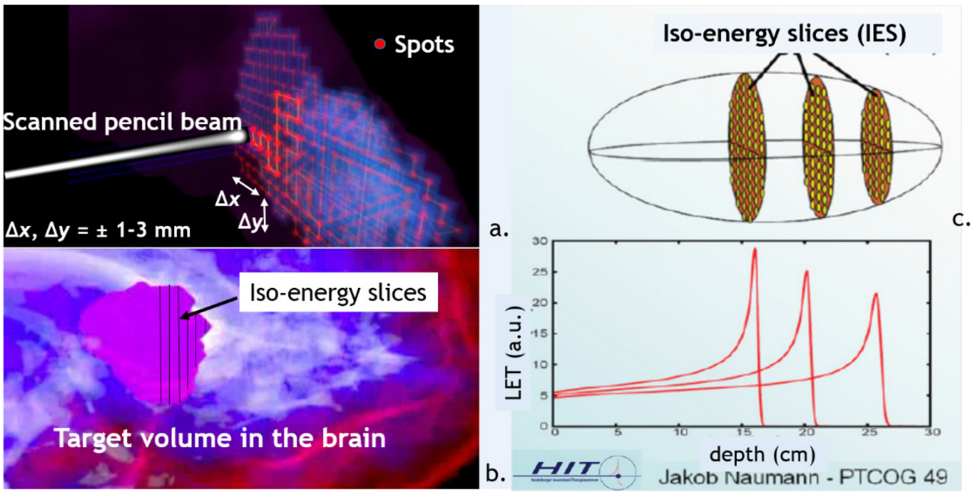


Fig. 8: Example of spot distribution (a) for a simulated target volume in the brain (b). (c) Iso-energy slices treated with different Bragg peaks (reproduced from Jakob Naumann presentation at PTCOG49).

The sweeping across the transverse plane is achieved by means of two independent scanning dipoles for horizontal and vertical beam deflections located at the end of the extraction line and several metres upstream of the patient. By changing the energy step and performing the irradiation slice by slice, a tumour of arbitrary shape can be successively irradiated from its most proximal to its most distal part or vice versa. This irradiation technique, only possible with charged particles, allows reaching very high dose conformation in radiotherapy, minimizing the radiation burden on healthy tissues. It reduces the neutron dose to the patient and removes the need for patient-specific devices. In addition, the dose modulation inside the target volume facilitates the application of individualized treatments with local boosts or other desired non-homogeneous dose distributions.

A careful treatment planning assigns the energy, number of particles and transverse position to each spot in order to deliver the optimum dose distribution to the patient. Multiple radiation fields, each characterized by a fixed beam entrance direction, are aimed by using a rotating gantry or, when no gantry is available, by rotating the patient. An isocentric plane is defined for each field, perpendicular to the beam entrance direction, which contains the axis of rotation, the isocentre being the origin of its reference system. Spot transverse positions are typically projected to the isocentric plane and specified relative to its reference system.

To obtain a range of more than 300 mm, the maximum beam energy is about 250 and 430 MeV/u for protons and carbon ions, respectively. The required field transverse size typically ranges from 200 × 200 mm^2 to 300 × 300 mm^2, with a 150 mm length for the longitudinal direction. These cover most of the targets that are usually treated. The required overall precision of the spot position is about 0.5 mm.

2.2.1 Flavours of scanning

As shown in Fig. 8(c), in pencil beam scanning the dose is generally painted with the pencil beams grouped in discrete energy layers (all beams of the same energy are sequentially applied and only then is the beam energy changed). There are several options to deliver the dose of an energy layer. The choice among them can influence the precision and the treatment dead time and provide different repainting capabilities (by repainting capability, we mean the possible number of repaintings used to reduce the error caused by target motion) [7-9].

Three methods of painting energy layers are available: (a) discrete spot scanning, (b) raster or quasi-discrete pencil beam scanning and (c) continuous line scanning.

With *discrete spot scanning* each pencil beam is delivered statically: after irradiation of a spot the beam is switched off and the position of the beam is changed to the next spot application. Once the dipoles have reached the new designated values, the beam is turned on again. This technique requires a fast beam shut-off system that can be operated at high frequencies. The dead time between two spots depends on the distance between two spots, on the average scanning magnet velocity and on the set-up time including all electronic delays. It is used at PSI [10], at MD Anderson [11] provided by Hitachi Co. and at the Rinecker Proton Therapy Center provided by Varian Co. At PSI Gantry 1, instead of moving the beam by scanning magnets in both transversal directions, it is also possible to move the patient table. As table motion is slow compared to scanning the beam transversally or even to changing the beam energy using the layer stacking method, the table movement is chosen as the slowest varying scan direction [10]. This is in contrast to most other facilities, where changing the beam energy takes most time and where the longitudinal scan direction is the slowest.

The pencil beam *raster scanning* (or *quasi-discrete scanning*) technique is a type of spot scanning that does not turn off the beam between two spots if these are close enough. This dose delivery method requires fast scanning magnets to minimize the dose between two points. This method was developed at GSI [12] and then implemented in the new synchrotron-based facilities like Heidelberg Ion-Beam Therapy Center (HIT) [13], Centro Nazionale di Adroterapia Oncologica (CNAO) [14], Austrian Hadron Therapy Centre (MedAustron) [15] and National Institute of Radiological Sciences (NIRS) [16, 17].

The *continuous line scanning* method has been implemented on Gantry 2 at PSI [18]. The delivery of each spot is replaced by the corresponding dose segment, a line piece that connects spots together over ±half of the grid distance between spots. A line segment is delivered with constant (highest possible) velocity and the required dose rate (varying spot dose) is controlled by adjusting the beam intensity at the ion source. By keeping the beam moving continuously at highest speed on a continuous path, the repainting capability is maximized and the errors due to organ motion are reduced. The required dose delivered along a line piece can be adjusted by changing either the beam intensity or the sweeper velocity.

The instrumentation required to perform such precise dose deliveries are: (a) two dedicated fast scanning magnets, one steering the beam along *Y* and the other along *X*. These magnets have to be operated with advanced power supplies. (b) The online beam monitors and (c) a dedicated real-time control system. In the following we will discuss in general the specifications giving more detailed descriptions of the solutions adopted at CNAO [14].

2.2.2 The scanning system

The matching of the dose profile to the tumour volume is obtained with the superposition of spots aimed to a specific position by changing the current circulating in the scanning dipoles. This process is supervised by a delivery control system which uses beam monitors for the online measurement of the beam position and fluence (see Section 3); when the prescribed number of particles in a spot is achieved the system steers the beam to the next spot.

The scanning system must be as fast as possible especially when the beam is not switched off during the movement between spots belonging to the same slice. For example, at CNAO, the fluence delivered during the transition from a spot to the next is ascribed to the destination spot [19]. The transition time, not accounted for by the treatment planning computation, influences the distribution of the delivered dose and has to be minimal [7, 8]. Additionally, with fast scanning capabilities, the overall treatment time is reduced compared to slower systems and the repainting option to treat moving targets can be implemented [9]. In the case of high repainting, one has to consider that the transient dose between spots becomes very significant compared to the static part of the spots.

In order to limit the dose inhomogeneity caused by the beam movement between spots within a few percent (±2.5% as clinically required), it has been estimated that, for typical beam fluxes of ~10^9 p/s and ~10^8 C^{6+}/s, a beam scanning velocity in excess of 20 mm/ms should be achieved. Considering a spot spacing of 1 to 3 mm, this implies a transient time lower than 200 μs for all the beam rigidities. For example, the Heavy-Ion Medical Accelerator in Chiba (HIMAC) is equipped with a fast scanning system [20], which provides beam-scanning velocities of 100 and 50 mm/ms for horizontal and vertical beam movements when measured at the isocentre. It allows covering a uniform 2D field having a 100 × 100 mm^2 size and spot spacing of 3 mm in a time as short as 40 ms. To fulfil these requirements, strong efforts are in general devoted to develop the fast scanning magnet and its power supply, the high-speed control system and the beam monitoring [20].

The scanning elements, dipole magnets and power supplies, designed for both protons and carbon ion beams, are the most demanding because of the wide beam rigidity ($B\rho$) range, between 1.1 T m for protons at 60 MeV/u and 6.3 T m for carbon ions at 400 MeV/u. For such elements, high ramp speed, low hysteresis and good accuracy are key points in the design.

As an example, a fixed horizontal beam line is sketched in Fig. 9, where the two dipoles are at distances of 5 and 6 m upstream of the isocentre. To cover with pencil beams a surface of 200 × 200 mm^2, a maximum bending angle of 16 mrad is necessary.

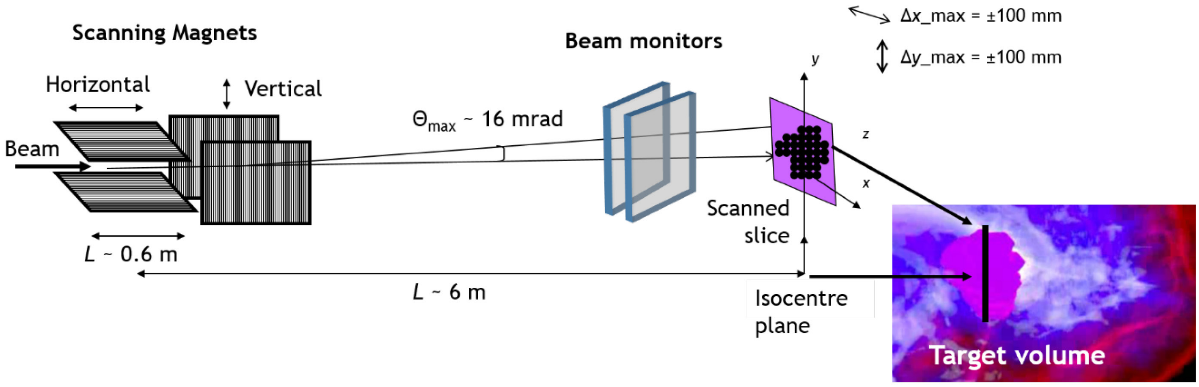


Fig. 9: Example of a fixed horizontal beamline for modulated spot scanning delivery

For the vertical line at CNAO, larger bending angles of 21.5 mrad are required because the scanning dipoles are closer to the isocentre due to the additional 90° bending magnet. At the boundaries the operating points of the scanning system of the CNAO treatment lines are listed in Table 2.

Table 2: Extreme operating points of the scanning system of the CNAO horizontal (above) and vertical (below) treatment lines. For each beam rigidity the maximum magnetic field and the maximum intensity of the power supplies for covering the 20 × 20 cm^2 field are reported.

H line						
Particles	**Energy [MeV/u]**	**(*B*ρ) [T m]**	***B*$_x$ [mT]**	***B*$_y$ [mT]**	**±*I*$_x$ [*A*] or [mA/0.1 mm]**	**±*I*$_y$ [A] or [mA/0.1 mm]**
Protons	60	1.14	30.47	33.99	59.05	65.87
Protons	250	2.43	64.95	72.46	125.87	140.43
C ions	120	3.26	87.13	97.21	168.86	188.39
C ions	400	6.36	169.98	189.64	329.42	367.52
V line						
Protons	60	1.14	35.04	44.32	67.91	85.89
Protons	250	2.43	74.70	94.48	144.77	183.10
C ions	120	3.26	100.22	126.74	194.23	245.62
C ions	400	6.36	195.51	247.27	378.9	479.21

At the isocentre plane the beam displacements Δx and Δy are given by $\Delta x = \alpha_x D_x$ and $\Delta y = \alpha_y D_y$ [m], where D_x and D_y are the distances between the scanning magnets (X and Y) and the isocentre. Assuming small beam deflection angles, α_x and α_y, the relation between α and B is the following:

$$\alpha = \frac{\int B\mathrm{d}l}{B\rho} \quad [\mathrm{rad}] \tag{1}$$

where $B\rho$ is the magnetic rigidity and is equal to p/q, p being the beam momentum and q the charge.

In a first approximation the relationship between the variation of magnetic field and beam displacement is given by

$$\Delta B = \frac{\Delta x\left(B\rho\right)}{D_x l_\mathrm{m}} \tag{2}$$

when assuming a constant field in the magnet length l_m.

Finally, the relationship between magnet current (I) and magnetic induction (B) in the dipole gap of height h_gap is given in first approximation by (3)

$$B = \frac{NI\mu_0}{h_\mathrm{gap}}. \tag{3}$$

For constant B, if $\int B\mathrm{d}l$ is the magnetic length (l_m) of the magnet, h_gap the dipole height and D_x and D_y the distances from the isocentre, the beam steps Δx and Δy are given by Eqs. (4) and (5):

$$\Delta x = \Delta I\left[\left(N\mu_0 / h_\mathrm{gap}\right)/\left(B\rho\right)\right] D_x l_\mathrm{m} \ , \tag{4}$$

$$\Delta y = \Delta I\left[\left(N\mu_0 / h_\mathrm{gap}\right)/\left(B\rho\right)\right] D_y l_\mathrm{m}. \tag{5}$$

In summary, several parameters influence the design and the performance of the scanning systems:

- Beam rigidity (*Bρ* characterizes the field strength required to bend the beam);
- Scanning speed (sets the maximum transit time between spots);
- Distance between spots (minimum and maximum beam shifts);
- Maximum field dimension (characterizes the maximum bending angles required);

- Scanning dipole positions (with $B\rho$ and field dimension determine the maximum bending angles);
- Magnetic length (l_m) of the dipole;
- Current ramp rate (indicating the required speed dI/dt of the power supplies);
- Beam intensity (determines the minimum spot duration);
- Dose modulation (fluence per spot proportional to the spot duration, which also sets the time between two different current settings).

As examples, the NIRS scanning system at HIMAC is described in Section 2.2.2 while the CNAO scanning dipoles and power supplies are reported in Sections 2.2.3 and 2.2.4.

2.2.3 The NIRS scanning system at HIMAC [20]

The new treatment facility at HIMAC is equipped with a 3D irradiation system for pencil beam scanning. The basic parameters of the scanning system are described as follows. To obtain the range of more than 300 mm, the maximum energy is chosen as 430 MeV/u. The required field size is 220 × 220 mm^2 for the transverse directions with a 150 mm length for the longitudinal direction. This covers most of the targets. Under these conditions, the new system must be as fast as possible to treat the moving target with rescanning. Based on the conceptual design study, the system was designed to provide a modulated dose delivery with beam scanning velocities of 100 and 50 mm/ms at the isocentre. These scanning velocities enable us to achieve the fastest irradiation time of around 40 ms for an example uniform 2D field having a 102 × 102 mm^2 size with spot spacing of 3 mm. The fast scanning magnet and its power supply, the high-speed control system and the beam monitoring were developed to fulfil these requirements.

The beam line, shown in Fig. 10, consists of the two scanning magnets (SMX and SMY), two screen monitors (SCN1 and SCN2), main and subflux monitors (DSNM and DSNS), position monitor (PSN), mini ridge filter (RGF) and range shifter (RSF). To achieve the fast beam scanning at the isocentre, the distances from SMX and SMY to the isocentre are designed to be 8.4 and 7.6 m, respectively. The vacuum beam exit window is made of 0.1 mm thick Kapton and located 1.3 m upstream from the isocentre. Beam monitors, RGF and RSF, are installed downstream of the vacuum window. The primary beam shutter (FST) and the neutron shutter (NST) will be placed just downstream from the SCN2 indicated by the dotted arrow in Fig. 10.

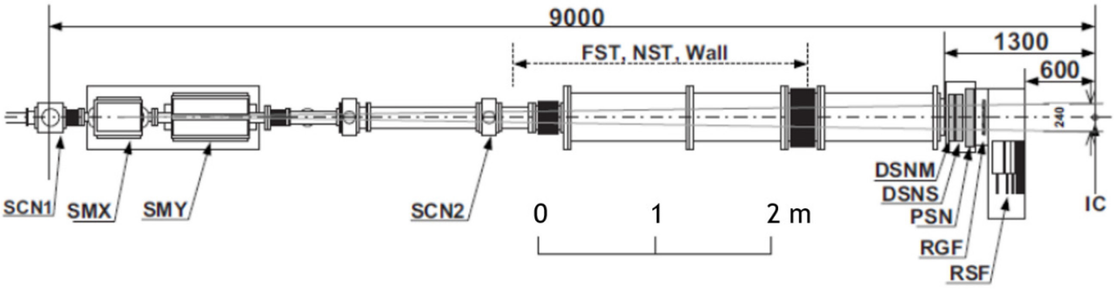


Fig. 10: Layout of the HIMAC scanning-irradiation system. SMX, horizontal scanning magnet; SMY, vertical scanning magnet; SCN1 and SCN2, screen monitors; DSNM and DSNS, main and subflux monitors; RGF, bar ridge filter; PSN, beam position monitor; RSF, range shifter; (FST), primary beam shutter; NST (neutron shutter). Unit: mm.

2.2.4 The CNAO scanning dipoles

The design of the scanning system had to consider several important and somewhat conflicting features: a good field uniformity, a limited hysteresis, a good linearity up to the largest magnetic fields and a good sensitivity, the latter being required for achieving a good scanning precision when low beam rigidities are used. In addition, a large field ramp was required in order to reach the desired ramp speed of 62 T/s to guarantee a scanning speed larger than 20 m/s for the all beam rigidities [21].

These features were met at CNAO by using for the two identical magnets lamination material with high saturation induction, low and uniform coercivity, and steel with a minimum amount of impurities and large grain size. Very thin (0.35 mm) yoke laminations were glued together to improve the torsion stiffness during the magnet assembly. The maximum magnetic field of 0.31 T is reached with a current of 606 A circulating in the coils. The magnet length is 553 mm and the gap size is 130 × 140 mm^2 with an inner good-field region of 120 × 120 mm^2. The coil is divided into three subcoils to optimize the field homogeneity, which was measured to be better than 0.2% in the good-field region. Several aspects affect the behaviour of the magnets when operated at high frequency.

First, the eddy currents induced in the magnet components (the conductor, the iron lamination and the mechanical structure) cause the field in the gap to be attenuated by a quantity $\Delta B(t)$ when a field ramp is applied. The value of ΔB depends on the imposed field ramp and on the geometrical and electrical characteristics of the magnet components. A careful design of the CNAO magnets limits the deviation to 1 G, decaying with a time constant of 300 µs due to eddy currents in the iron lamination with a field ramp of 62 T/s. A similar effect due to the eddy currents in the magnet conductor induces decays with a longer time constant (3 ms) but with much lower deviation (0.3 G).

Second, because the magnet inductance is a rate-dependent impedance, it has to be considered. For example, the CNAO magnet has an inductance of 4.4 mH and the electrical resistance is 26 mΩ. The load of the dipole magnet can be approximated at first order to a *R*–*L* circuit with a time constant of 170 ms. This value is three orders of magnitude larger compared to the average time interval required to move the beam between two spots; a dedicated and advanced power supply (PS), described in the next section, was developed to overcome this limit.

A picture of the CNAO scanning dipoles placed on one of the three horizontal beam lines is shown in Fig. 11.

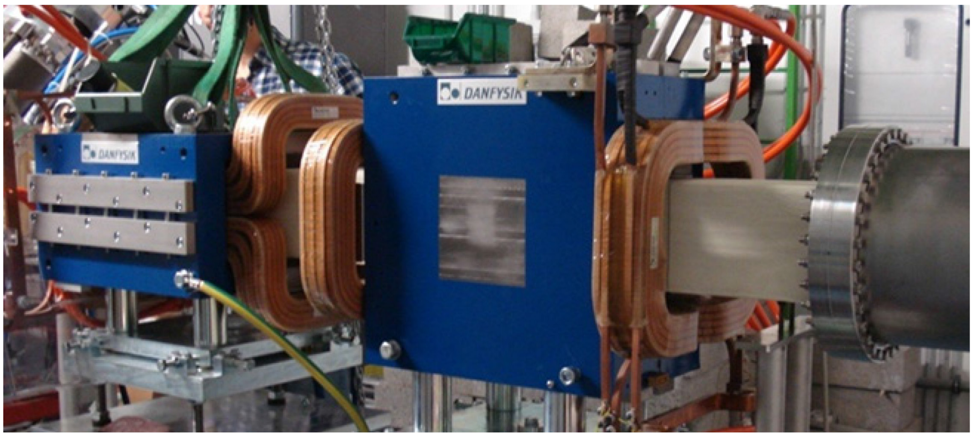


Fig. 11: Picture of the CNAO scanning dipoles for the horizontal beamline

2.2.5 The CNAO power supplies for scanning magnets

The time constant of scanning dipoles (higher than 150 ms) has to be shortened by three orders of magnitude to reach the required scanning speed of 20 mm/ms. The mechanism implemented for the CNAO PS is based on the pre-emphasis concept by delivering a large voltage step which is then aborted when the current is close to the required value. The precise adjustment is achieved then via smaller voltage steps.

The power supply is composed of three main modules, sketched in Fig. 12: a booster (BO) and two active filters (AFs). The BO is a high-voltage–high-current insulated gate bipolar transistor (IGBT) H-Bridge whereas the AFs are IGBT H-Bridges used in interleaved modulation.

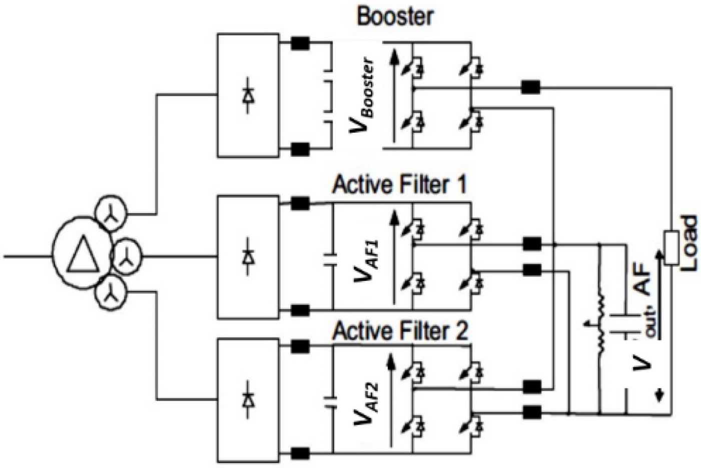


Fig. 12: Schematics of the CNAO power supply for the scanning dipoles

The BO provides a large voltage step (±660 V) with a ramp speed exceeding 100 kA/s when a current step larger than 2.5 A is demanded. The BO is switched off by a comparator when the current reaches a value close to the set value within ±0.6 A. The current is then driven to the final value by the control loops of the AFs. For current steps below 2.5 A only the AFs are used and the BO is short circuited.

The AFs provide both precise voltage adjustments during the transient and control for steady-state current. The precision of the delivered current is 55 mA, corresponding to 100 ppm for $I_{\max} = 550$ A. The AF control loop sensitivity is 36.4 mA.

The connection between the power supply and the magnet is achieved through a 10 m long shielded cable.

2.3 Dose delivery constraints for different particles and accelerators

Depending on the particle species used in particle therapy, different constraints apply to the dose delivery system and instrumentation. We will limit the discussion to protons and carbon ions, which are presently the only species used in clinics, and to the two types of accelerators available for hadron therapy, the cyclotron and the synchrotron. We will see how different beam characteristics lead to some basic differences in the dose delivery concept and instrumentation.

Briefly, the cyclotrons provide to the DD system only proton beams with a single fixed energy and with continuous and steady current. Therefore, to achieve a longitudinal dose distribution requires a fast degrader in the accelerator or range shifter plates for energy degradation in the dose delivery. Because of the presence of the absorbers, the beam flux produced by the accelerator must be higher compared with the clinical flux and the activation of the material along the beam line and in the dose delivery instrumentation must be carefully considered.

The synchrotrons accelerate both protons and heavier ions at fixed energies which can be selected within a large set of values. The beam energy range is usually chosen to avoid the use of range shifters. The synchrotrons work in pulsed mode, i.e. the extracted beam is not continuous and, during the extraction, called spill time, the beam current is generally not constant. The beam energy can be changed spill by spill through the real-time dose delivery control.

The transversal beam spread does not depend on the accelerator, so scattering and scanning modalities can be used both with cyclotrons and with synchrotrons.

The choice of spreading the beam with the passive scattering techniques allows delivering a dose with protons as was done with photons: through large and uniform radiation fields entering in the accelerator nozzle where patient-specific static devices conform the dose distribution. In these systems, the goal of the dose delivery is first to spread the protons in order to create a large square or circular field and then to degrade and absorb particles to conform to the required field shape. The robust and safe approach of passive scattering systems led to a large number of proton facilities. The technology for proton medical accelerators and for the dose delivery instrumentation has been widely improved in the last decade resulting in solutions available on the market with affordable costs. Additionally, in the treatment room gantries are widely adopted allowing patient irradiation from any beam direction.

The carbon ions show better physical properties providing a better conformation to the target when the modulated scanning technique is used and are particularly effective for some radio-resistant tumours. However, the low scattering angle and the projectile fragmentation is a constraint to use these particles with the passive scattering beam delivery techniques. As a consequence, the treatments with carbon ions have started later, in Europe at GSI first in the 1990s, where the active scanning technique to spread laterally the beam has been first designed and developed [12]. The main disadvantage of carbon ions is that a cyclotron accelerator for therapeutic heavy ions is still not available and presently only large synchrotrons can be used, increasing the system complexity and cost. Moreover, the beam particles fragment in the patient body, creating lighter particles with larger penetration depth, generate a tail beyond the Bragg peak, giving an unwanted dose after the treatment volume. Beam particles and fragments have different and variable relative biological effectiveness (RBE); this is why the radio-biological dose evaluation is complex and many details are still under investigation. Moreover, for carbon ion beams the gantry system is very large and expensive, so, up to now, a single system is clinically used [22], while the other carbon ion facilities have only fixed beam lines.

3 The beam monitoring system

3.1 Introduction

The delivery of the dose has to be performed to ensure the safety of the patient. Therefore, accurately monitoring the delivered dose in particle therapy is mandatory and is equivalent to accurately monitoring the correct delivery of the number of ions of the primary beam.

Additional checks may be required, which consist in measurements to control the beam parameters, listed in Table 1.

The treatment prescription is provided as a DICOM RT file by a TPS and consists of a set of data to set up the accelerator, the beam line and the beam shaper or scanning devices before and during the treatment delivery.

Passive scattering systems have different constraints and requirements for the beam monitors compared to systems for modulated pencil beam scanning DD systems. The latter must change the beam settings during the treatment; so, faster and real-time controls are mandatory.

For the scattering delivery technique, the type and position of scatterers, range shifters and compensators are provided by the TPS and verified before the irradiation starts. The total amount of monitor units, each corresponding to a given amount of dose, is also provided among prescriptions and used by the intensity beam monitor to stop the treatment when the total desired dose has been achieved (i.e. at the end of the field irradiated).

For pencil beam scanning deliveries, the prescriptions consist of data to define pre-treatment accelerator and beam line settings followed by a list of spots defined in terms of number of particles (or monitor units), energy and position. These quantities are used online by the dose delivery instrumentation (i.e. beam monitors and scanning devices) to guide the treatment. Note that additional beam monitoring requirements exist for pencil beam scanning because of the need to drive the delivery progress by acting on scanning magnets and on beam stopper devices.

The sensitivity of the monitoring system should match the maximum tolerable uncertainties that, for the pencil beam scanning, are listed in Table 3.

Table 3: Maximum uncertainties for the beam parameter. Note that the beam energy is not measured online by beam monitors but verified before the treatment starts by checks of the accelerator (for synchrotron) or beam line settings (for cyclotron).

Beam property	Maximum uncertainty
Beam flux	1–2% of the integral flux
Transversal position	0.5 mm
Transversal shape (FWHM)	1 mm
Energy	1–2%

3.2 Beam monitoring for scattering systems

For the scattering dose delivery technique the required beam characteristics are set before the irradiation and remain the same for the whole irradiation field. The beam fluence, position, shape and symmetry are continuously checked for safety purposes at a low frequency (Hz). However, beam parameter fluctuations within a small range are tolerable because the beam is then scattered over the total target volume. Thus, if the mechanical parts like scatterers, range shifters and compensators are in the correct position and the primary beam has the right energy, the correct beam delivery to the target volume is safe and ensured. The measured beam fluence is usually used by the DD system to stop the treatment when the total desired dose has been achieved.

The beam monitoring in passive beam delivery is generally performed by two independent dose monitoring devices located before the final collimator. One detector works as the main monitor (master) and the second works as the auxiliary (or sub or slave) monitor for redundancy. Additional beam profile and reference dose monitors measure the beam shape and intensity at the exit of the vacuum chamber to verify the particle distribution before the scattering and modulation.

The MD Anderson (Houston, Texas) passive scattering nozzle (Fig. 13) is taken as an example of a dose delivery system with gantry for a proton beam line [23].

After the beam exits the beam transport system, it passes through a vacuum window into the treatment delivery nozzle. As the beam traverses the nozzle it intercepts the following devices: (1) beam profile monitor, (2) reference dose monitor, (3) first scatterer with a range modulation wheel (which form a single, integrated unit), (4) second scatterer, (5) binary range shifter, (6) secondary dose monitor, (7) primary dose monitor, (8) multilayer Faraday cup (MLFC), (9) treatment field aperture, range compensator. The beam profile monitor, reference dose monitor, secondary dose monitor, primary dose monitor and MLFC are used to monitor various aspects of the beam while the remaining devices are used to shape or modify the beam.

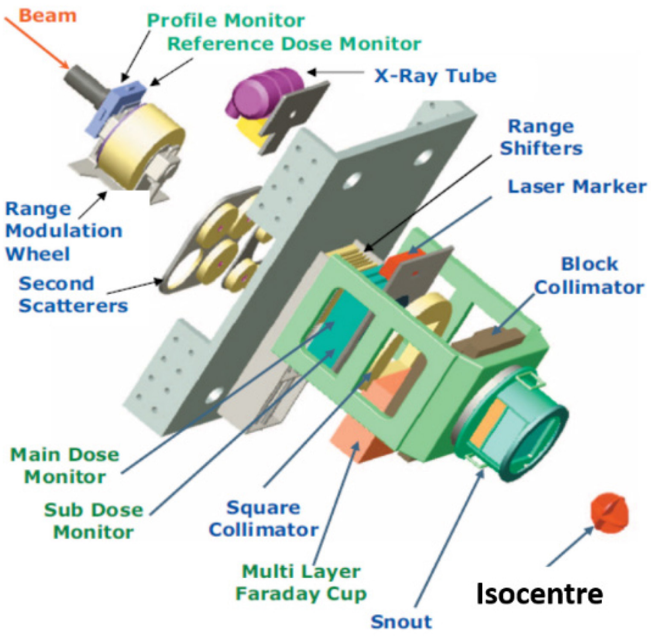


Fig. 13: Three-dimensional rendering of the MD Anderson passive scattering nozzle

Another example is the Catana (Centro di AdroTerapia e Applicazioni Nucleari Avanzate) fixed horizontal proton beam line for ocular treatments shown in Fig. 14, where two transmission ionization chambers for fluence measurement are followed by a strip chamber [24] for beam position and shape check.

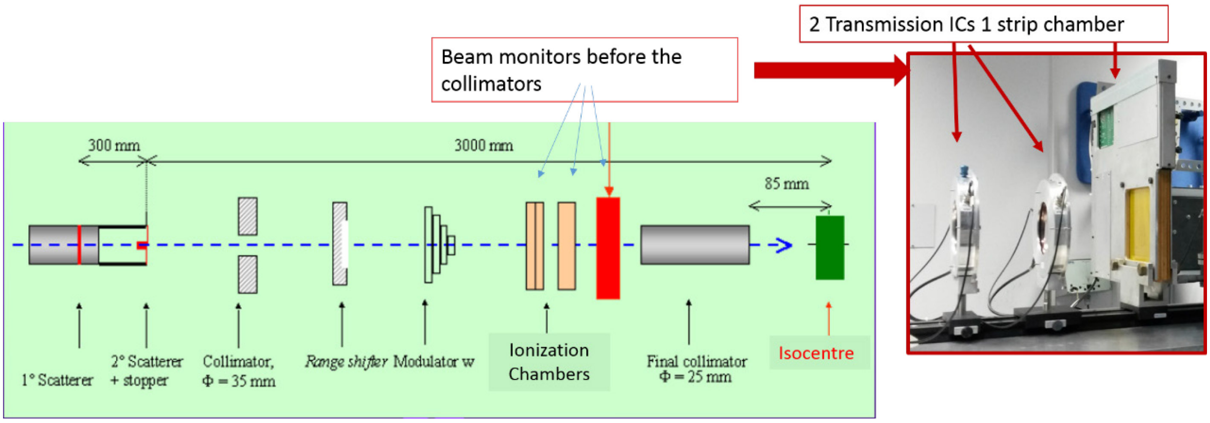


Fig. 14: Scheme of the Catana beam line with a picture of the two transmission ionization chambers and the strip chamber (in red).

3.3 Beam monitoring for pencil beam scanning techniques

Active scanning techniques require a precise, quick, stable and reliable beam monitoring system: the monitoring has to be repeated for each spot to meet the prescriptions, ensuring in addition the correct position of the spots. When the beam monitor has collected the prescribed amount of charge for a spot, the beam is steered to the next spot position by the DD control through the scanning magnets by changing the currents of the power supplies, see Fig. 15. This procedure is repeated until the last spot of the slice, i.e. a sequence of spots with the same energy, has been irradiated.

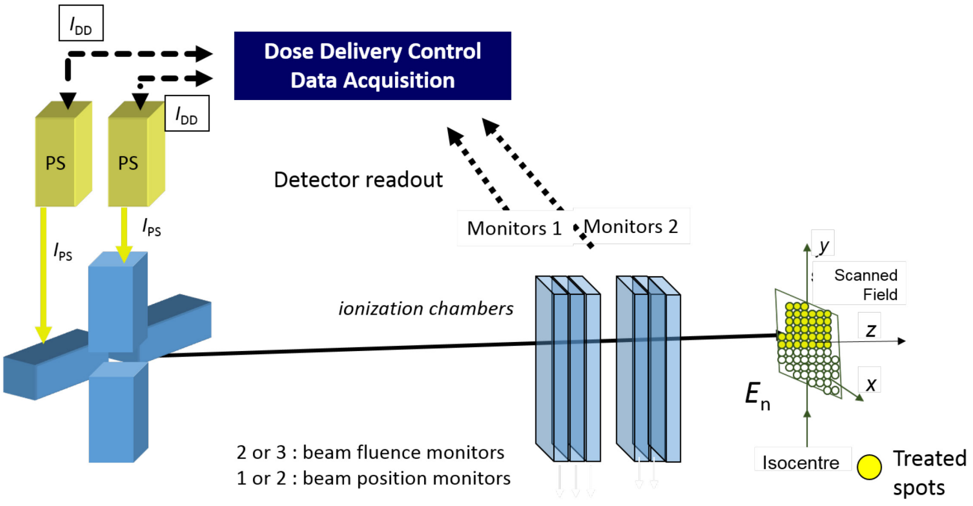


Fig. 15: Scheme of a primary beam monitor system for pencil beam scanning

Real-time and fast control systems are required to react to any condition leading to a potential hazard. Moreover, online feedback corrections (for example small corrections of beam position deviations based on beam monitor data) are recommended.

As beam monitors, thin detectors, fast and reliable for measuring beam fluence and position, are located just in front of the patient (Fig. 15). Arrays of parallel-plate ionization chambers with either a single large electrode or electrodes segmented in strips or pixels (see Section 3.1) as well as multiwire chambers (see Section 3.5) are well suited for this purpose. Special care is required in the detector design for high-intensity and pulsed beams. More details will be discussed in the following sections.

Secondary emission monitors (SEMs) may be used only with high dose rates because of their lower sensitivity (see Section 3.6).

In summary, the beam monitors, positioned as close as possible to the patient, have the tasks of measuring:

- the number of particles (i.e. integrated beam flux, often expressed in terms of monitor units);
- the transversal beam positions;
- the transversal beam shape (i.e. FWHM and symmetry).

The beam energy is typically not measured by the online beam monitors but guaranteed through proper checks of the accelerator and beam line settings (for synchrotrons) or by pre-treatment range measurements (for cyclotrons and synchrocyclotrons).

In pencil beam scanning systems that do not stop the beam during the transit between spots, a tiny fraction of the dose is delivered along the path. Since the transit time is typically very short, hundreds of microseconds, the beam monitoring system is unable to measure where the particles were actually delivered. The easiest choice is either to consider these particles as delivered entirely to the spot whose irradiation is terminated or to assign it to the following spot which is being irradiated. The latter is the choice implemented by the DD of CNAO. This choice overestimates the fluence as the beam approaches a given spot and, of course, it underestimates the dose as the beam moves away from the same spot. The two effects partially compensate, the net effect being sizeable for mainly the first and the last spots of a sequence. Moreover, to limit the transient dose when the distance between spot centres is larger than a pre-defined threshold, the irradiation can be paused. Optimization algorithms were also proposed to optimize the scanning path in order to minimize the dose delivered during the transit [7, 8].

3.3.1 Ionization chambers

The most widely used type of beam monitor for charged particle therapy is the parallel-plate ionization chamber. Compared to other detectors, it offers several advantages in terms of robustness, uniformity, easiness of operation and minimal perturbation of the beam. A sketch of an ionization chamber is shown in Fig. 16. Two parallel plates, made of metallized foils kept at a constant high voltage difference, generate a uniform electric field. Particles traversing the chamber ionize the gas in the volume bounded by the plates. The thickness of the plates and the gas are chosen to minimize the induced scattering and fragmentation of the beam particles.

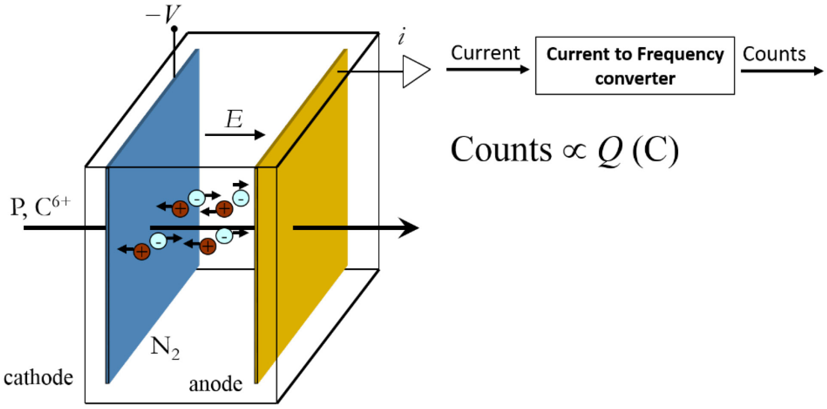


Fig. 16: Parallel-plate ionization chamber filled with nitrogen

Ionization chambers work efficiently if a large number of charge pairs are created and collected at the electrodes before they recombine to form neutral molecules. Therefore, recombination of the ionization charges at the highest expected dose rates must be taken into account in choosing

- the type of gas (usually air or nitrogen);
- the plate separation and the operational high voltage (i.e. the electric field, typically higher than 500 V/mm to ensure rapid charge collection with minimum recombination effects).

Also, multiplication of the produced charge should be avoided because it can lead to unacceptable errors in the beam intensity measurement.

There are different mechanisms through which these interactions can take place. Some of these interaction mechanisms can be predicted with a good level of accuracy by using statistical quantities such as cross-section and stopping power. A quantity that is extremely useful for radiation detection is the average energy needed to create an electron – ion pair in a gas. Ions can be formed either by direct interaction with the incident particle or through a secondary process in which some of the particle energy is first transferred to an energetic electron or delta ray. This quantity has to be considered as an effective energy, larger than the gas molecule ionization energy, to account for all the energy loss mechanisms not leading to ionization in the gas.

When the change in particle energy during the detector traversal can be neglected, the average charge produced per particle within the gas gap is given by Eq. (6):

$$\frac{Q}{\int_0^{Emax} f(E)\,\mathrm{d}E} = \frac{e\int_0^{Emax} \frac{\left(\frac{S(E)}{\rho}\right)_{\mathrm{g}} .\rho.x}{W(E)} f(E)\;\mathrm{d}E}{\int_0^{Emax} f(E)\,\mathrm{d}E}, \tag{6}$$

where $(S(E)/\rho)_g$ is the mass electronic stopping power for the gas g at particle energy E, ρ is the density, x is the mean free path in the gas, $W(E)$ is the mean energy needed to produce an ion pair for a charged particle of energy E losing $(S(E_p)_g \cdot\rho\cdot x)$ energy in the gas and $f(E)$ is the number of beam particles crossing the chamber with energies between E and E + dE. The formula can be simplified if a monoenergetic beam of energy E_p is considered:

$$\frac{Q}{N_p} = \frac{\left(\frac{S(E_p)}{\rho}\right)_g .\rho.x}{W(E)} \quad , \tag{7}$$

where N_p is the total number of particles crossing the chamber. For a monoenergetic beam of uniform intensity the formula is typically written as follows:

$$i = \varepsilon.\Phi.\frac{\left(\frac{S(E_p)}{\rho}\right)_g .\rho.x}{W(E)} \quad , \tag{8}$$

where i is the intensity of the current collected at the electrodes of the chamber, $\Phi = N_p/\Delta t$ is the beam flux and ε represents the collection efficiency. The charge Q measured at the electrodes is thus directly proportional to the number of particles crossing the gas gap N_p.

The practical quantity of interest on radiation detection is the total number of ion pairs created along the beam particle tracks. There is a minimum energy transfer from the incoming particle to the electron necessary so that the ionization of the atom occurs. In most of the gases used for radiation detectors (air, nitrogen, argon etc), the minimum ionization energy is between 10 and 25 eV. However, other mechanisms by which the incident particle may lose energy within the gas do not create ions. Examples are excitation processes in which an electron may be excited to a higher bound state in the atom without being completely removed. Therefore, the average energy loss by the incident particle per ion pair, defined as the W-value, is always substantially larger than the minimum ionization energy. The W-value is found to depend weakly on the particle type and energy and lies within 25–45 eV per charge pair for most gases and types of radiation (Table 4).

Table 4: W-value in several gases for electrons (data from ICRU 1979)

Gas	W-value (eV/i.p.)
Air	34.2
He	41.3
Ne	35.4
Ar	26.4
H_2	36.5
N_2	34.8
O_2	30.8
CO_2	33.0

Air- or nitrogen-filled chambers are typical choices and automated corrections for atmospheric pressure and temperature are recommended when the collected charge is transformed to number of beam particles delivered.

The charges created by the incident radiation are called primary charges to distinguish them from the ones produced by ionization caused by primary charge pairs. The W-value represents all such ionizations that occur in the active volume. For a particle that deposits energy ΔE inside a detector, the W-value can be used to determine the total number of ion pairs produced by

$$N = \frac{\Delta E}{W} \ . \tag{9}$$

In the unlikely case that the incident particle deposits all of its energy inside the detector gas, then ΔE would simply be the energy E of the particle. However, in a usual case only a negligible part of the total energy is lost in the gas; then the number of ion pairs is a function of the stopping power, as

$$N = \frac{1}{W}\frac{\mathrm{d}E}{\mathrm{d}x}\,\Delta x, \tag{10}$$

where Δx is the path covered by the particle. Sometimes it is more convenient, at least for comparison purposes, to calculate the number of ion pairs produced per unit length of the particle track as follows:

$$n = \frac{1}{W}\frac{\mathrm{d}E}{\mathrm{d}x}. \tag{11}$$

Drift velocity

In a gaseous detector, free electrons behave quite differently compared to the ions in the presence of the electric field and therefore the two types of charges should be studied separately. The ions are positively charged and much heavier than electrons and therefore move around quite sluggishly compared to the electrons. In the ionization chambers, the output signal can be measured from the positive or from the negative electrode. In both cases, however, what is measured is actually the charge induced by the change in the electric field inside the active volume. Hence, the drifts of electrons and ions both contribute to the overall output pulse. This implies that understanding the drift of positive charges is as important in a chamber as the electrons.

In the presence of an externally applied electric field, ions move toward the negative electrode with a drift velocity that is much lower than that of electrons. Assuming that N ions are produced in $x = 0$ at $t = 0$, the distribution at time t of these ions can be fairly accurately characterized by a Gaussian distribution of the form

$$\mathrm{d}N = \frac{N}{\sqrt{4\pi Dt}}\,\mathrm{e}^{-(x-tv_\mathrm{d})^2/4Dt}\,\mathrm{d}x, \tag{12}$$

where v_d is the drift velocity of ions, i.e. the average velocity of the cloud of ions moving along the electric field lines, and D is a temperature-dependent diffusion coefficient per unit time. The drift velocity is much lower than the instantaneous velocity of ions. Drift velocity is an important parameter, since it indicates how quickly the ions reach the cathode and get collected. It can be fairly accurately predicted from the relation

$$v_\mathrm{d} = \mu^+ \frac{E}{P}. \tag{13}$$

Here E is the applied electric field, P is the pressure of the gas and μ^+ is the mobility of ions in the gas. Mobility is related to the mean free path of the ion in the gas, the energy loss per impact and the energy distribution.

If the chamber is filled with electronegative gases like air, electrons are rapidly attached to the neutral gas molecules giving rise to negative ions drifting with a velocity similar to the positive ions, though in the opposite direction. On the contrary, if a non-electronegative gas like nitrogen is used, free electrons will drift to the anode with a velocity two to three orders of magnitude larger than the positive ions.

Free electrons, owing to their small mass, are rapidly accelerated between collisions and thus gain energy. Along the electric field lines, the electrons drift with velocity v_d, which is usually an order of magnitude smaller than the velocity of thermal motion, v_e. However, the magnitude of the drift velocity (Fig. 17) depends on the applied electric field E as follows:

$$v_\mathrm{d} = \frac{2eEl_\mathrm{mt}}{3m_\mathrm{e}v_\mathrm{e}}, \tag{14}$$

where l_mt is the mean momentum transfer path of electrons.

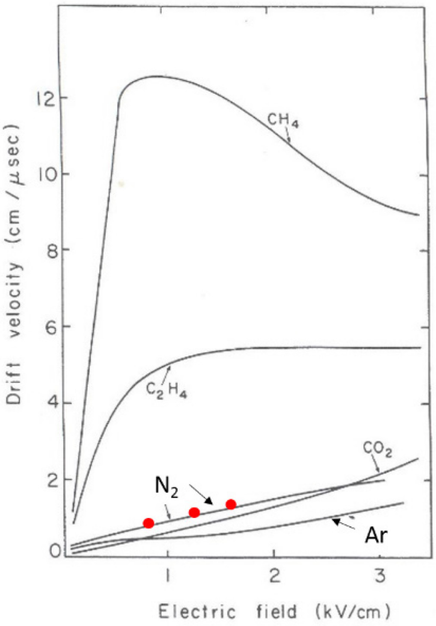


Fig. 17: Electron drift velocity as a function of the electric field (*V*/m) for different gases (reproduced from [25])

Table 5 shows the time required to an electron and an ion to cross a 5 mm gap of a nitrogen-filled ionization chamber with the electrodes polarized at 500 V.

Table 5: Example of ion and electron collection times for 5 mm gap of a nitrogen-filled ionization chamber

Charge	Gas	E (V/cm)	v (cm/s)	Gap (mm)	Time (μs)
Ions	N_2	1000	3×10^3	5	150
Electrons	N_2	1000	10^6	5	0.5

The current measured by this chamber, given a total charge Q released uniformly in the chamber by the beam passing through, is provided by both ions and electrons collected at the electrodes at different rates as shown in Fig. 18. If the integral of this curve is used to measure the charge Q, and hence the number of particles N_p that crossed the chamber, a measurement time long enough to collect all the signal is required or a systematic underestimation will occur. This is particularly relevant for a pulsed beam structure if the number of particles delivered in each pulse needs to be determined with high accuracy.

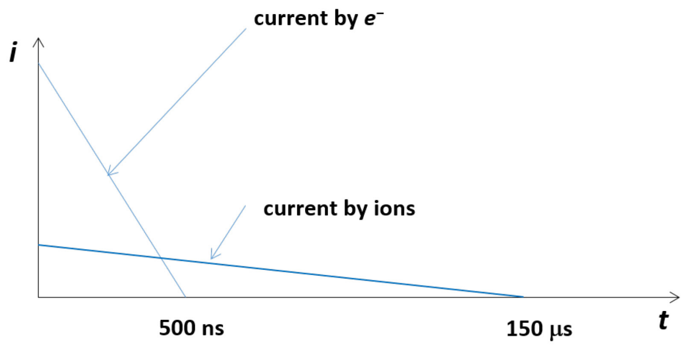


Fig. 18: Different contributions to the measured current in a beam monitor characterized by a 5 mm gap filled with nitrogen and supplied with 500 V (E = 1000 V/cm).

The charge collection efficiency ε is mainly determined by the recombination of charges in the gas volume. As the voltage difference between the electrodes increases, the resulting electric field separates the ion pairs with increasing drift velocity, reducing the equilibrium concentration of ions within the gas. The ionization current thus increases due to fewer charges lost to recombination, at first almost linearly with voltage, then more slowly and finally asymptotically approaches the saturation current for the given radiation intensity. At this level, the electric field is large enough to effectively suppress the recombination to a negligible level, and all the original charges created through the ionization process contribute to the ionization current. Increasing the voltage further cannot increase the current because all charges are already collected and their rate of formation is constant. Therefore, the saturation current is the measured current if all the ions formed in the chamber by the radiation are able to reach the electrodes. Different detector designs (i.e. gap and gas type) and settings (i.e. high voltage between electrodes) lead to different saturation curves. In Fig. 19, we show the charge collected as a function of the applied voltage for transmission ionization chambers filled with air with different gaps that have been irradiated with 3 × 10^7 carbon ions per spill. For these data, the spill length was about one second.

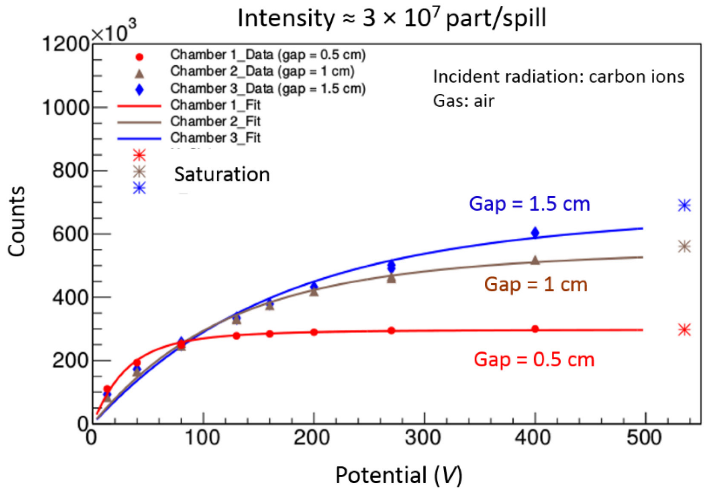


Fig. 19: Detector counts (i.e. charge collected) from parallel-plate ionization chambers with different gaps, irradiated with 3 × 10^7 carbon ions per spill (i.e. per second), as a function of the voltage.

Parallel-plate geometry simplifies the electrode design generating a uniform electric field. The electrodes of the beam monitors are perpendicular to the particle beam axis and a single large-area electrode is used to measure at high rate, typically larger than 100 kHz, the beam flux. Parallel-plate geometry can also be used efficiently for measuring the beam shape, beam position and beam fluence using segmented electrodes.

In the segmented type, each element measures the collected charge independently of the other elements. Such chambers have a variety of uses such as measuring the beam profile and position or measuring the two-dimensional beam profile distribution. The layouts of such devices depend on the required spatial resolution and on the total area to be covered. As is shown in Fig. 20, strip chambers can provide high resolution (~100 μm) and faster (10 kHz) beam position evaluation compared with pixel chambers (~200 μm and 5 kHz). In Fig. 20, different anodes with different segmentations are shown.

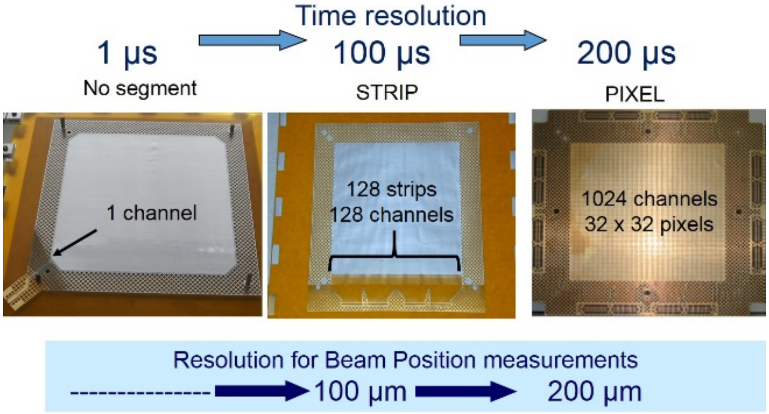


Fig. 20: Picture of anodes for ionization chambers segmented in different shapes

The practical limit on the reduction of the size of the collecting elements is determined by the maximum size of the irradiated fields, by the maximum density of channels on the anode (or cathode) surface and by the detector readout capability.

The detector response should be position-independent, especially for pencil beam scanning dose deliveries, because the beam spot is scanned across the whole detector. The beam-induced ionization in the gas detector should be identical at any position in the detector while the beam direction has a normal incidence with the detector. The main parameter affecting the uniformity is the flatness of the electrodes that may determine variation of the gas gap.

The charge collected in the ionization chambers depends on the pressure and temperature of the gas, as well as on the voltage across the gap. These quantities have to be continuously measured and checked; the read values are used to correct the gain of the chambers. Appropriate interlock procedures are required whenever any of the values is outside the expected range.

For passive scattering dose delivery, beam-centring systems (including segmented ionization chambers), which are capable of detecting misalignments between the central axes of the beam and the scattering devices, are required. Dose-monitoring detectors can either intercept the entire beam area or just the central portion. The former requires larger detectors and measurements are more reliable as they are less dependent on beam alignment variations. Beam steering is much more sensitive in double- than in single-scatterer DD systems.

Beam position and width

The beam position is usually evaluated online through the measurements of ionization chambers segmented in strips or of multiwire detectors (see Section 3.5). The distributions of the charges during a typical clinical application are shown in Fig. 21 as a function of the position; they represent the projections of the beam phase space (or beam size) respectively on the horizontal (X) and vertical (Y) directions.

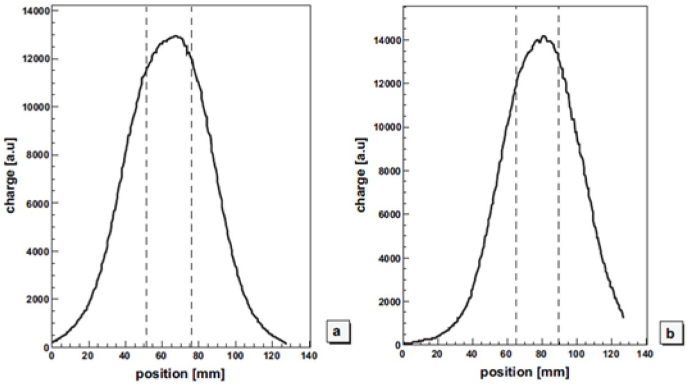


Fig. 21: Transversal beam profiles as measured from strip-segmented ionization chamber

An example of a charge distribution as measured with a strip chamber is shown in Fig. 22.

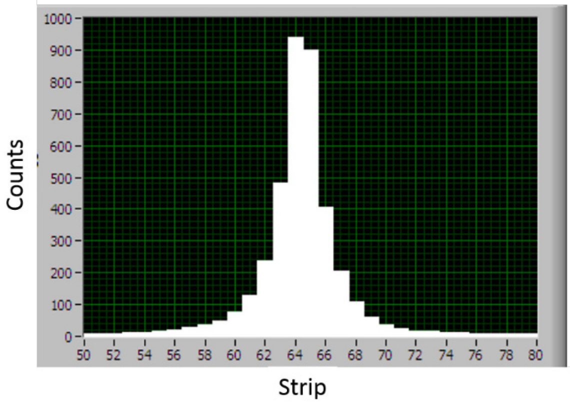


Fig. 22: Charge distribution over strips for a proton beam

The beam position can be evaluated as a weighted average through the following equations:

$$X = \frac{\sum x_i^{\text{strip}} \times S_i}{\sum S_i} \; ; \qquad Y = \frac{\sum y_i^{\text{strip}} \times S_i}{\sum S_i}, \tag{15}$$

where x_i and y_i are the coordinates of the strip centre, S_i is the charge signal measured in the strip i and the sum extends over all the strips of the chamber.

To avoid wrong position estimation because of noise or background currents, the values measured by strips far from the beam should be neglected. For example, at CNAO, the strips which measure a single count are not considered in Eq. (15) because this corresponds to the average expected background current for each readout channel.

The FWHM or the σ of the charge distribution is used to estimate the transversal beam width:

$$\text{FWHM}_x = 2.35\sqrt{\frac{1}{S}\Sigma_i\, S_i \left(x_i - X_{\text{beam}}\right)^2}, \tag{16}$$

where $S = \Sigma_i\, S_i$ and X_{beam} is the beam position along x as from Eq. (15).

Beam shape (flatness and symmetry)

For passive scattering systems, where a uniform transverse beam distribution is covering the irradiation fields, beam monitors can be used to check the beam flatness and symmetry before and during the treatment. The flatness *Fl* is used to estimate the maximum percentage of deviation from the average dose delivered in a reference region; following the conventional definition used in radiotherapy, it is given by Eq. (17):

$$Fl = \frac{\left(D_{max} - D_{min}\right)}{\left(D_{max} + D_{min}\right)} \times 100, \tag{17}$$

where D_{max} and D_{min} are the maximum and minimum doses measured in the selected uniform region; a negative sign is conventionally assigned to Fl if D_{max} occurs on the left-half part of the detector.

The symmetry S_y is defined by considering the integrated absorbed doses D_{l} and D_{r} in each half of the field about the centre of the selected region, where the quantity $\left|1 - \frac{D_r}{D_l}\right|$ reaches its maximum:

$$S_y = \left(\frac{D_r}{D_l}\right) \times 100. \tag{18}$$

The clinical tolerances imposed on these parameters are typically $Fl < 2$–3% and $|S_{\mathrm{y}} - 100\%| < 3\%$.

A symmetry in the beam size is essential, especially at the entrance of passive scattering and gantry systems. In the latter, the shape and position of the beam at the patient and the transmission through the gantry should not depend on the gantry angle.

The beam symmetry can also be evaluated by using the skewnesses γ_x and γ_y calculated from the transversal beam projections, as follows:

$$\gamma_x = \frac{\sum_i S_i \cdot (x_i - \mu_x)^3}{\sum_i S_i \cdot \sigma_x^3} \tag{19}$$

and similarly for y. In Eq. (19), S_i is the charge measured in the strip i, x_i is the coordinate of the strip centre and μ_x and σ_x are the mean and the r.m.s. of the distribution. For a symmetric beam projection the expected value of the skewness is zero, while positive and negative values indicate asymmetric beam particle distributions as shown in Fig. 23.

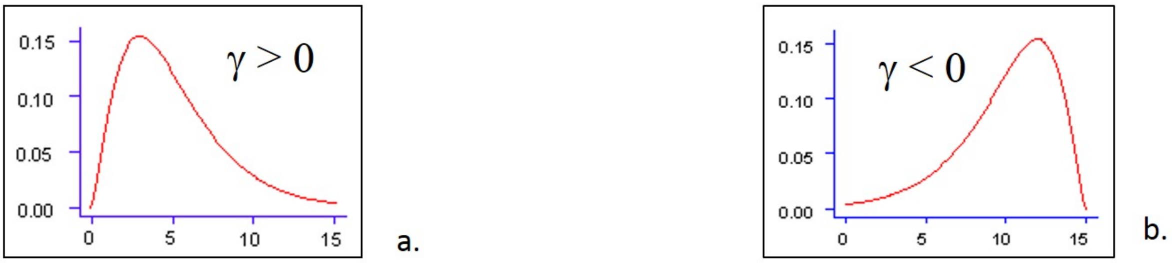


Fig. 23: Examples of asymmetric particle beam distributions with positive (a) and negative (b) skewness values

3.4 Detector readout

The current collected by each anode of the ionization chambers (i.e. the single channel of the integral beam monitors and each strip or pixel of segmented anodes) is converted into a digital frequency of increments or decrements of a counter, proportional to the current itself. The time integral of the frequency corresponds to a number of counts that is proportional to the collected charge and thus to the energy released in the detector active area. The resulting counts are stored in a register. In the more advanced applications ASICs (application-specific integrated circuits) are specifically developed equipped with ADCs (analogue to digital converters) and registers. Real-time and safe data operations, required for control in clinical applications, are often demanded by field programmable gate arrays (FPGAs), which, besides a high flexibility, guarantee fast and deterministic data processing. A direct connection of the FPGA output with the interlock collector is recommended to perform prompt and safe actions. A scheme with the sequence of signals for a typical detector readout is shown in Fig. 24.

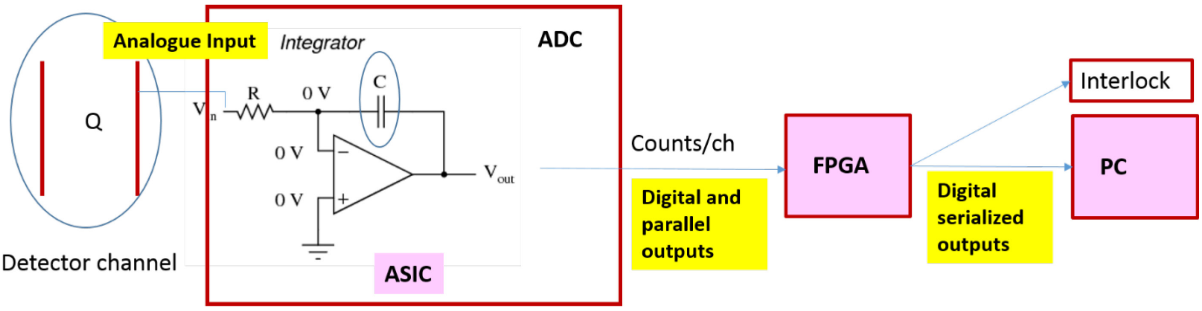


Fig. 24: Sequence of signals for a typical detector readout

3.5 Multiwire ionization chambers

A multiwire ionization chamber [25], in its simplest form, is a grid of parallel thin anode wires between two cathode planes as shown in Fig. 25(a). Under application of a symmetric voltage difference, ionization electrons released in the upper and lower gas volumes drift with a constant speed to the anodes, where, because of the high field gradient, they are amplified in an avalanche. The backdrift of ions produced in the avalanche away from the anode induces a negative charge on the wire and positive charges in all surrounding electrodes (adjacent anodes and cathode planes).

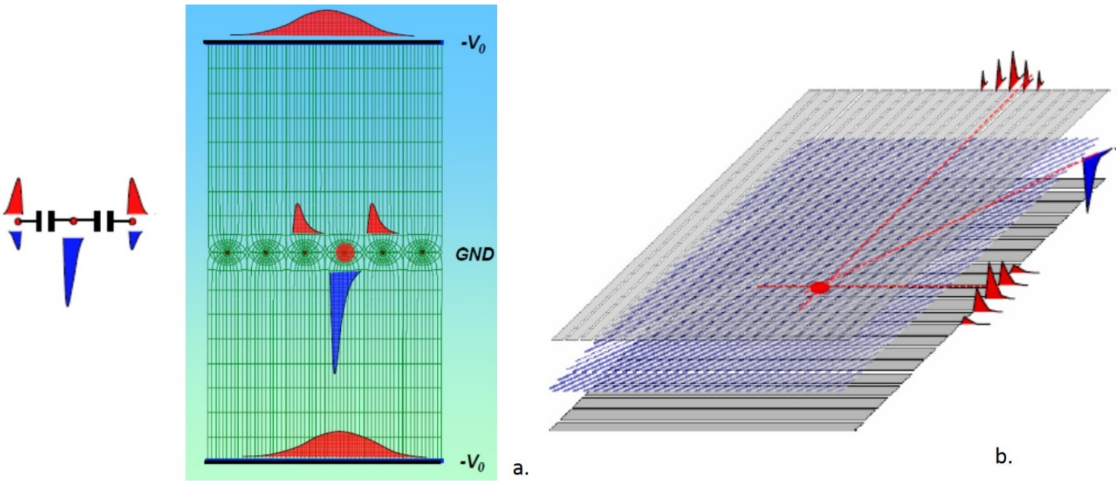


Fig. 25: (a) Scheme of a MWIC; (b) perpendicular cathode strips

Signal planes are alternated with high-voltage planes. The high-voltage plane can also be a solid foil conductor instead of parallel wires. While the common application of wire chambers in particle physics is to measure individual particles, in medical applications wire chambers are often used as integrating devices: the signal collected on a wire during a given time is proportional to the number of particles crossing the gas volume around the full length of the wire. In Fig. 25(b), an example of perpendicular cathode strips (printed circuit or wire groups) is shown. The centre of gravity of the induced charge distribution on cathode strips provides the *X* and *Y* projections of the avalanche position. The gain of such devices is dependent on the gas used, its pressure and the wire spacing, and is an exponential function of the high voltage. Clinically, with these devices a submillimetre resolution with a wire pitch of approximately 1 mm can be reached. Wire chambers are typically operated in a proportional mode and called multiwire proportional chambers (MWPCs), where a multiplication of the initial ionization occurs. However, the required high electrical fields can easily lead to saturation for high-intensity beams and only the beam tails can be monitored. This problem can be overcome by operating the wire chambers in an ionization mode, thus improving their performance. Multiwire chambers are primarily used in many centres for beam monitoring during beam line tuning and for on line beam control, for example at the first hadron therapy medical centre (Loma Linda University Medical Center [26]), at the GSI experimental carbon ion therapy beam line and at HIT [13]. The fact that good spatial resolution is obtained with a small number of signals and variable sensitivity makes these devices extremely practical for this purpose.

3.6 Secondary emission monitor

A secondary emission monitor (SEM) [27] is particularly well suited for large proton fluences. It consists of one or more thin metallic foils mounted in an ultra-high-vacuum enclosure. As the foil is traversed by protons, electrons are released, resulting in a net current flow that provides the signal. A thin Al foil is placed in the hadron beam path at 45° with respect to the beam direction. The energy lost by the beam in the foil is transferred to the electrons of the medium. Then, γ-rays are produced together with electrons of the energy below 50 eV that are called secondary electrons (SEs). The number of SEs ejected from the foil is proportional to the local beam intensity. The SEs are accelerated and focused by an electrostatic field towards an imaging device or a position-sensitive sensor, which provides the beam intensity and its position [28]. The secondary emission monitors are precise and reliable but relatively large and expensive. Moreover, frequent checks of these monitors are necessary because they may suffer from performance changes due to deposits or damages of the electrode surface.

3.7 Beam monitor calibration for scanned proton and carbon ion beams

The beam monitor (BM) for scanned charged particle beams measures the charge (C) collected during a given beam delivery and needs to be calibrated in units of particles per monitor unit (MU) (see Eq. (20)). The calibration has to be performed in a region with constant energy loss and usually the plateau region (i.e., the entrance region) is selected. The calibration is beam-energy dependent and must be validated (or determined) for all the available beam energies. The determination of the number of particles is based on a reference measurement of absorbed dose to water for a fixed set of energies. For example, a Farmer chamber having a calibration factor in terms of absorbed dose to water (reference beam quality is a Co^{60} beam) is placed in a phantom in the centre of a homogeneous monoenergetic square field and irradiated with a regular grid of proton or carbon ion spots.

The calibration factor at energy E, $K(E)$, is defined as the number of particles N per monitor unit MU, and is given by

$$K\left(E\right)=\frac{N}{\mathrm{MU}}=\frac{D_{\mathrm{meas}}}{\mathrm{MU}\times S_{E(x)}}\Delta x\Delta y\,, \tag{20}$$

where D_{meas} is the absorbed dose measured in the phantom, $S_{E(x)}$ is the mass stopping power of protons or carbon ions with the initial energy E at the depth of measurement x, Δx and Δy are the spacings between two consecutive spots in the transversal direction (for example, 3 and 2 mm for protons and carbon ions, respectively). A pre-requisite for the application of Eq. (20) is that the scanned field delivers a homogeneous dose in the transverse plane. The measurement has to be performed at some representative energies E_i (six at HIT [29] and nine at CNAO [30]) ranging from the minimum to the maximum, which corresponds to minimum and maximum depths of penetration in water. The delivered number of particles has to be set to a constant for each spot, corresponding to the selected number of MUs. The collected data, as shown in Fig. 26, can be fitted with a third-order polynomial curve, $K(E)$, that represents the BM calibration curve, one for each beam line and particle type and used for treatments by the software managing the dose delivery.

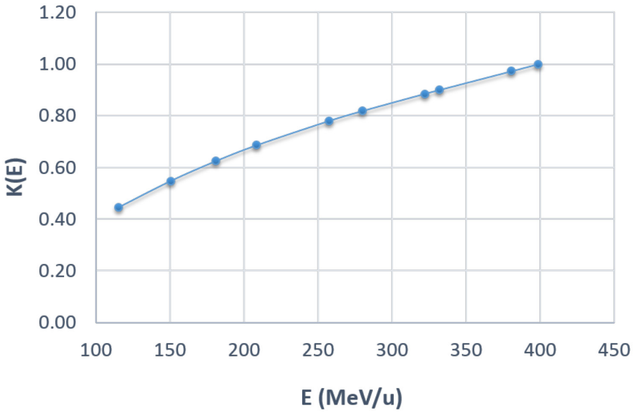


Fig. 26: Monitor calibration factor $K(E)$ as measured according to Eq. (20) for six different energies of the carbon ion beam.

3.8 Worldwide beam monitoring systems

3.8.1 The CNAO beam monitors

The CNAO dose delivery system [19] measures at a fixed frequency the number of delivered particles, the beam transversal position and dimension by means of five parallel-plate ionization chambers [31] filled with nitrogen. The monitor chambers are enclosed in two independent aluminium boxes: BOX1 and BOX2 (Fig. 27). The BOX1 contains an integral chamber (INT1) with a large-area anode for the measurement of the beam flux, followed by two chambers with the anodes segmented in 128 strips, 1.65 mm wide, respectively with vertical (StripX) and horizontal (StripY) orientations, which provide the measurement of the beam position and beam width projected along two orthogonal directions. The BOX2 contains a second integral chamber (INT2) followed by a chamber with the anode segmented in 32 × 32 pixels, each 6.6 mm wide (PIX). The measurements accomplished by the BOX2 detectors are beam flux, position and width.

The total water equivalent thickness of these chambers has been measured to be approximately 0.9 mm. The material budget interposed by the beam detectors is spread along approximately 20 cm and it is one of the main contributions to the beam lateral dispersion. To minimize the effect, the boxes are installed close to the patient, at approximately 70 cm from the isocentre.

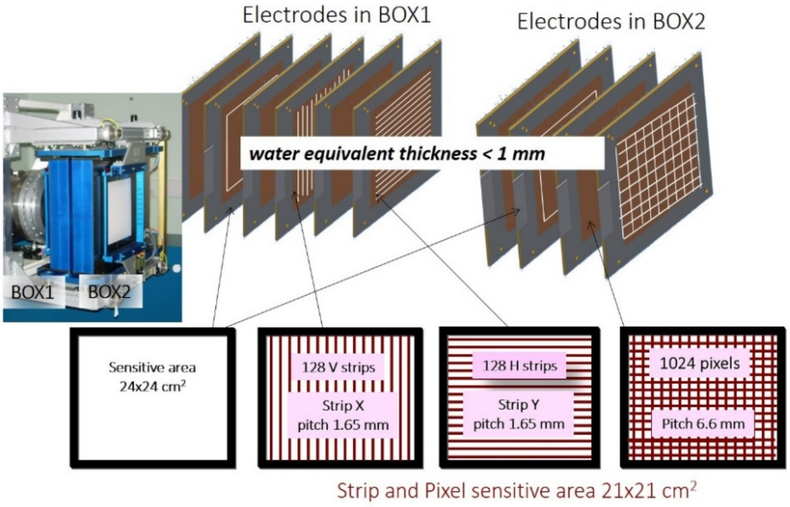


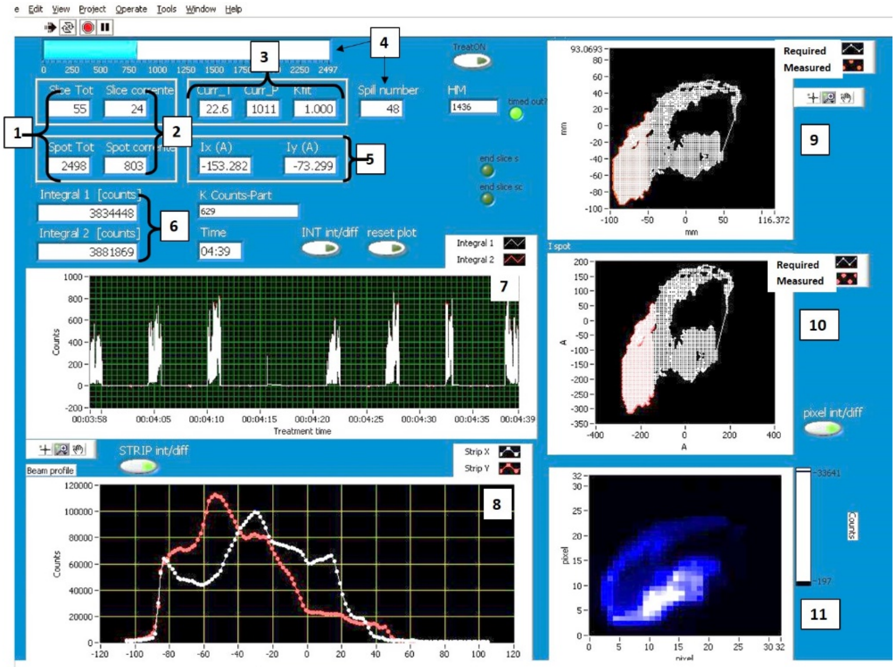
Fig. 27: Picture and sketch of the CNAO beam monitor detectors (BOX1 and BOX2)

The detector front-end readout is based on custom-designed boards, which host ASICs custom designed for this purpose. The chosen architecture and technology allow good sensitivity, with a minimum measured charge of 200 fC corresponding to a number of protons ranging from 7.2 × 10^3 at the energy of 62 MeV to 1.9 × 10^4 at the energy of 226 MeV. Similarly, the range of the number of carbon ions extends from 3.4 × 10^2 at the energy of 115 MeV/u to 7.7 × 10^2 at 399 MeV/u. The background current is limited to 200 fA; this ensures negligible error on the dose delivered to each spot.

The customized electronics is based on large-scale integration to provide many readout channels with uniform channel-to-channel behaviour, and allows a large segmentation of the active area.

The charge collected in the ionization chambers depends on the pressure and temperature of the gas, as well as on the high voltage across the gap. These quantities are monitored by a set of transducers installed in each box and controlled by a peripheral interface controller (PIC). Values are periodically read and checked by the PIC against pre-set values. The measured deviations are used to correct the gain of the chambers. Appropriate interlock procedures are activated whenever any of the values is outside the expected range.

The data, provided by the beam monitors in the CNAO local control room during the dose delivery, are shown in Fig. 28.



Slice Tot 55
Slice corrente 24
Spot Tot 2498
Spot corrente 803
Curr_T 22.6
Curr_P 1011
Kfit 1.000
Spill number 48
HM 1436
Ix (A) -153.282
Iy (A) -73.299
Integral 1 [counts] 3834448
Integral 2 [counts] 3881869
K Counts-Part 629
Time 04:39

Fig. 28: Snapshot of the data published continuously on a monitor display in the CNAO local control room: (1) total number of slices and spots of this field; (2) number of the slice and the spot under treatment; (3) measured temperature, pressure and the flux correction factor; (4) number of the spill and delivery progress bar; (5) PS current set-points; (6) INT1 and INT2 total counts; (7) flux measured by INT1 and INT2; (8) StripX and StripY total counts; (9) spot positions in millimetres; (10) PS currents in amperes: measured (full dots covering partial field) and required (small dots covering the overall field); (11) 2D flux measured by PIX chamber.

3.8.2 The PSI Gantry 1 beam monitors [32]

The Paul Scherrer Institute Gantry 1 is equipped with four ionization beam chambers in the gantry nozzle to monitor the incident proton beam. The beam flux monitor 1 (Mon1) integrates the output ionization charge and determines the spot dose. When the integrated ionization charge from Mon1 reaches the expected value, the fast magnetic kicker switches off the beam. The main beam flux monitors, Mon1 and Mon2, are parallel-plate transmission ionization chambers (Fig. 29(a)) used to check the applied spot dose, while a strip ionization chamber (Fig. 29(b)) verifies the position and the width of the applied beam against the prescribed values.

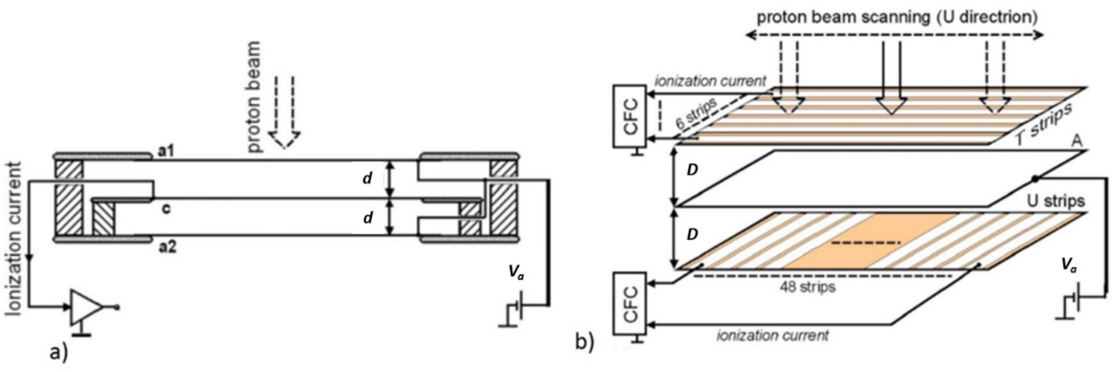


Fig. 29: (Reproduced from [32]) (a) Schematic representation of the PSI G1 parallel-plate ionization chamber (Mon1 and Mon2); V_a (2000 V) is the applied voltage to the anodes (a1 and a2) and d is the spacing between the anode and the cathode (c). Mon1 has $d = 0.5$ cm and Mon2 has $d = 1$ cm; (b) schematic representation of the PSI G1 strip ionization chamber.

The Mon1 and Mon2 chambers are filled with ambient air because these satisfy all the technical requirements and simplifies operation and maintenance compared to a closed gas system. The high-voltage planes (a1 and a2) are made of 20 μm thick Mylar foils single coated with aluminium with a thickness of less than 0.1 μm. The applied voltage (V_a) is 2000 V. The signal plane consists of a 20 μm thick aluminium foil. The chamber window covers the swept beam area (±9 cm) and is 23 × 3 cm^2 in size. Because the average duration of one spot is about 10 ms, the ionization charge collection time should be of the order of 0.1 ms in order to reach a precision of 1%. With the chamber gaps, Mon1 has an ion collection time t_c of about 90 μs, and Mon2 has a t_c of about 350 μs, having derived the collection time with the usual relationship:

$$t_c = \frac{d^2}{\mu . V_a}, \tag{21}$$

where V_a is the applied voltage, d is the spacing between the high-voltage plane and the signal plane and μ is the mobility of charged particles in the electric field [in dry air, the average mobility μ is about 1.4 cm^2/(s V)]. The slower collection time of Mon2 is acceptable because it is used only as a back-up element and does not affect the precision of the dose delivery.

The strip chamber has a voltage plane made of 20 μm thick Mylar foil coated with aluminium layers (less than 0.1 μm thick). The signal planes consist of 20 μm thick Kapton foils coated with 4 mm wide aluminium strips. The spacing between two strips is 0.4 mm. The strip aluminium coating is about 1 μm thick. The 48 strips U and six strips T provide readings of the beam profiles in the horizontal and vertical directions. The counting gas is ambient air. By calculating the centre of gravity of the outputs of the strips hit by the beam, its position can be determined to approximately a tenth of a millimetre. The sums of the strip monitor outputs are also used as additional dose checks.

3.8.3 The Loma Linda University Medical Center beam monitors

At Loma Linda University Medical Center, a synchrotron produces accelerated protons to final energies between 70 and 250 MeV. The double scattering and the pencil beam scanning deliveries are available. A transmission ionization chamber (TIC) and a secondary electron emission monitor measure the integrated number of protons per spill or per treatment. Helium has been chosen as the filling gas because of helium's fast positive-ion mobility. As previously described, half of the intensity signal is due to the collection of electrons and half of the signal is due to the collection of positive ions. With field strength of 1 kV cm^{-1} and an anode to cathode distance of 0.5 cm, the electron signal is detected quite fast, less than 10 ns, and the positive ions are collected linearly over time from 0 to 50 µs after the passage of the primary proton. The intensity signal is sampled at 50 kHz (every 20 µs) and processed (requiring another 10 µs) for determining if the beam position should be moved again [33].

The beam position and beam profile are monitored using three retractable multiwire ion chambers (MWICs). The wire chamber resolution is 2 mm. A 25 × 25 cm^2 ion chamber segmented into 400 square pads was placed after the range modulator to monitor the dose delivered to the target volume. The detector consists of a 20 × 20 array of pads (each 1.25 × 1.25 cm^2) from a thin sheet of gold-plated Kapton. Any of the pads in the central region of the pad plane can be used to monitor and control the dose delivered to the target and the remaining pads are also used to study the transverse dose distribution [26].

3.8.4 The Ion Beam Application (IBA) beam monitors for Pencil Beam Scanning (PBS) [34]

A 320 × 320 mm^2 parallel-plate ionization chamber composed of 15 Mylar foils separated by 5 mm air gaps has been developed to be installed in the IBA nozzles for PBS. The detector is composed of two identical units IC2 and IC3 with independent power supply and electronic acquisition set-up, for redundancy requirements. As sketched in Fig. 30, each unit is composed of five 2.5 µm Mylar electrodes coated on both sides with aluminium or gold. Three are connected to the high voltage while the two others are measurement electrodes at ground potential, one being used for dose measurement (uniform film) and the second one for beam position measurement (striped film) along one axis (horizontal for IC2 and vertical for IC3). Apart from the two units, three other films are connected to the ground to ensure the electrostatic pressure equilibrium. In addition, two thicker (25 µm) Mylar films are used to cover both entrance and exit windows. The whole chamber is 6.86 cm thick for a total water equivalent thickness of 187 µm.

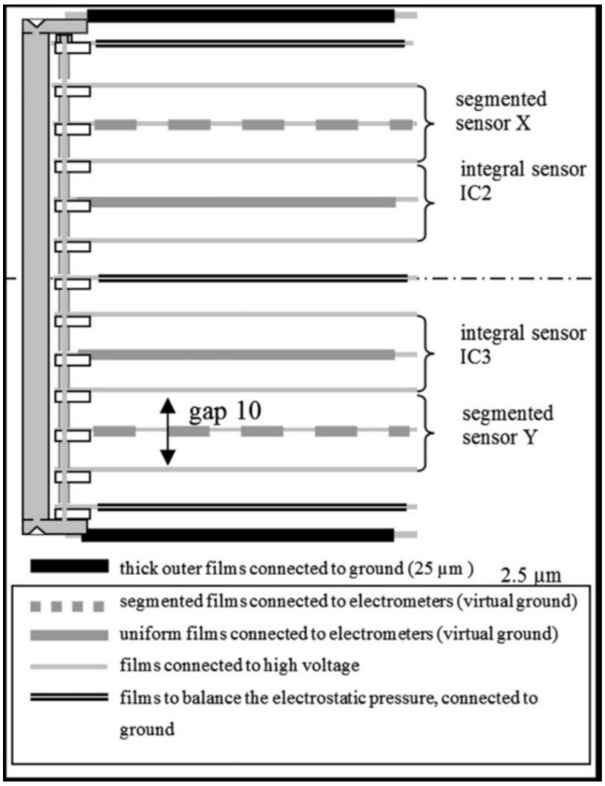

segmented sensor X
integral sensor IC2
integral sensor IC3
gap 10
segmented sensor Y
thick outer films connected to ground (25 µm)
2.5 µm
segmented films connected to electrometers (virtual ground)
uniform films connected to electrometers (virtual ground)
films connected to high voltage
films to balance the electrostatic pressure, connected to ground

Fig. 30: (Reproduced from [34]) Vertical section of the IBA beam monitors for PBS

The position measurement layers are plain Mylar films covered by gold strips on both sides deposited by vacuum evaporation. There are 64 4.8 mm strips with a 5 mm step, oriented to give the *x* and *y* positions for IC2 and IC3, respectively. Dose measurement layers are single-layer aluminized Mylar foils covered in the plain side by a thin gold layer (200 nm). The dose is measured by integrating the signal collected on each film at a frequency of 2 kHz, allowing a time resolution of 500 µs.

3.8.5 The MD Anderson beam monitor system

The MD Anderson passive scattering nozzle [23], shown in Fig. 13, houses the following devices to monitor various aspects of the beam: the beam profile monitor, reference dose monitor, secondary dose monitor, primary dose monitor and multilayer Faraday cup.

The profile monitor measures the beam profile, the position of the beam centre and the beam width. The profile monitor sends a beam interlock signal that stops the beam if an out-of-tolerance condition is detected, i.e. if any measured value is unacceptably different from the pre-set value.

User-adjustable tolerance tables contain tolerances for a warning signal when the maximum tolerance value is approached and a pause signal when the maximum tolerance value has been reached. The primary and secondary dose monitors are segmented ionization chambers whose primary functions are to measure the monitor units (MUs) delivered and to terminate the treatment when the prescribed MUs have been delivered. For any specific treatment, the secondary dose monitor is set a few percent higher than the primary dose monitor so that the treatment is normally terminated by the primary dose monitor.

These two segmented ionization chambers also serve as beam flatness and symmetry monitors. They consist of a quadrant fan-shaped electrode and eight concentric circular electrodes. The quadrant fan-shaped electrode measures the beam centre, while the eight concentric electrodes measure the lateral beam distribution.

4 The dose delivery control concept

The dose delivery instrumentation has a crucial role within the complex architecture of a facility for charged particle radiotherapy. Their advanced control software and IT systems, mainly in the context of high-speed scanning techniques, can have a significant level of risk as a side effect.

Therefore, DD instrumentation needs to be integrated in control (or interlock) systems, which immediately stop the irradiation as a reaction to any condition leading to a potential hazard.

At CNAO, for example, the safety of the treatment mainly relies on two interlock systems: the patient interlock system (PIS) and the safety interlock system (SIS). These systems collect any error conditions and either force the immediate interruption of the beam delivery or inhibit the operations as long as the conditions persist. The PIS is dedicated to the patient safety by acting on the beam chopper to interrupt the treatment when an interlock occurs. It manages both short interruptions (a few seconds) and treatment termination and recovery. Additionally, for recovery purposes, one battery back-up device, called a dose delivered recovery system, continuously receives, stores and displays the last treated slice and the spot of each slice during irradiation [19]. For passive scattering systems, the backed-up storage of measured monitor units is also recommended.

At PSI, a rigorous separation has been implemented between the tasks and responsibilities of the machine as a beam delivery system and a user who asks for a beam with specified characteristics and decides whether the beam is accepted for treatment. A machine control system (MCS) controls the accelerator's and beam line's performance. Then, each treatment area has its own therapy control system (TCS) [35].

Acknowledgements

We wish to thank F. Marchetto and R. Sacchi for useful comments and a careful review of this document.

References

[1] H. Blattmann, *Radiat. Environ. Biophys.* **31**(3) (1992) 219. http://dx.doi.org/10.1007/BF01214829

[2] Nuclear Physics European Collaboration Committee (NuPECC), *Nuclear Physics for Medicine* (2014), Chap. 1, www.nupecc.org

[3] W.D. Newhauser and R. Zhang, *Phys. Med. Biol.* **60**(8) (2015) R155. http://dx.doi.org/10.1088/0031-9155/60/8/R155

[4] I.C.R.U., Clinical proton dosimetry Part 1: Beam production, beam delivery and measurements of absorbed dose, Technical Report, International Commission on Radiation Units and Measurements (1998).

[5] H. Paganetti, *Proton Therapy Physics* (Series in Medical Physics and Biomedical Engineering), Eds. J.G. Webster *et al.* (CRC Press, 2012) - ISBN-13: 978-1-4398-3645-3.

[6] C. Ma and T. Lomax, *Proton and Carbon Ion Therapy, Ed.* W.R. Hendee (CRC Press, 2012).

[7] J.H. Kang *et al.*, *Med. Phys.* **34**(9) (2007) 3457. http://dx.doi.org/10.1118/1.2760025

[8] J. Pardo Montero *et al.*, *Med. Phys.* **36**(6) (2009) 2043. http://dx.doi.org/10.1118/1.3121506

[9] A.-C. Knopf *et al.*, *Phys. Med. Biol.* **56**(22) (2011) 7257. http://dx.doi.org/10.1088/0031-9155/56/22/016

[10] E. Pedroni *et al.*, *Med. Phys.* **22**(1) (1995) 37. http://dx.doi.org/10.1118/1.597522

[11] A. Smith *et al.*, *Med. Phys.* **36**(9) (2009) 4068. http://dx.doi.org/10.1118/1.3187229

[12] T. Haberer *et al.*, *Nucl. Instrum. Methods Phys. Res. Sect. A* **330**(1–2) (1993) 296. http://dx.doi.org/10.1016/0168-9002(93)91335-K

[13] S.E. Combs *et al.*, *Radiother. Oncol.* **95**(1) (2010) 41. http://dx.doi.org/10.1016/j.radonc.2010.02.016

[14] S. Rossi, *Phys. Med.* **31**(4) (2015) 333. http://dx.doi.org/10.1016/j.ejmp.2015.03.001

[15] M. Stock *et al.*, Development of clinical programs for carbon ion beam therapy at MedAustron (2015), *Int. J. Particle Ther.*

[16] T. Furukawa *et al.*, *Med. Phys.* **34**(3) (2007) 1085. http://dx.doi.org/10.1118/1.2558213

[17] K. Noda *et al.*, *Nucl. Instrum. Methods Phys. Res. Sect. B* **266**(10) (2008) 2182. http://dx.doi.org/10.1016/j.nimb.2008.02.075

[18] E. Pedroni *et al.*, *Phys. Med. Biol.* **50**(3) (2005) 541. http://dx.doi.org/10.1088/0031-9155/50/3/011

[19] S. Giordanengo *et al.*, *Med. Phys.* **42**(1) (2015) 263. http://dx.doi.org/10.1118/1.4903276

[20] T. Furukawa *et al.*, *Med. Phys.* **37**(11) (2010) 5672. http://dx.doi.org/10.1118/1.3501313

[21] S. Giordanengo *et al.*, *Nucl. Instrum. Methods Phys. Res. Sect. A* **613**(2) (2010) 317. http://dx.doi.org/10.1016/j.nima.2009.11.068

[22] M. Galonska *et al.*, The hit gantry: from commissioning to operation, IPAC 2013: Proc. 4th Int. Particle Accelerator Conf., Shanghai, 2013, p. 3636.

[23] A. Smith *et al.*, *Med. Phys.* **36**(9) (2009) 4068. http://dx.doi.org/10.1118/1.3187229

[24] N. Givehchi *et al.*, *Phys. Med.* **27**(4) (2011) 233. http://dx.doi.org/10.1016/j.ejmp.2010.10.004

[25] F. Sauli, Principles of operation of multiwire proportional and drift chambers, CERN-1977-009 (CERN, Geneva, 1977). http://dx.doi.org/10.5170/CERN-1977-009

[26] M. Coutrakon *et al.*, *Med. Phys.* **18**(6) (1991) 1093. http://dx.doi.org/10.1118/1.596617

[27] E.J. Sternglass, *Phys. Rev.* **108**(1) (1957) 1. http://dx.doi.org/10.1103/PhysRev.108.1

[28] L. Badano *et al.*, *IEEE Trans. Nucl. Sci.* **52**(4) (2005) 830. http://dx.doi.org/10.1109/TNS.2005.852620

[29] O. Jakel *et al.*, *Med. Phys.* **31**(5) (2004) 1009. http://dx.doi.org/10.1118/1.1689011

[30] A. Mirandola *et al.*, *Med. Phys.* **42**(9) (2015) 5287. http://dx.doi.org/10.1118/1.4928397

[31] S. Giordanengo *et al.*, *Nucl. Instrum. Methods A* **698**(11) (2013) 202. http://dx.doi.org/10.1016/j.nima.2012.10.004

[32] S. Lin *et al.*, *Med. Phys.* **36**(11) (2009) 5331. http://dx.doi.org/10.1118/1.3244034

[33] G. Coutrakon *et al.*, *Phys. Med. Biol.* **55**(23) (2010) 7081. http://dx.doi.org/10.1088/0031-9155/55/23/S09

[34] C. Courtois *et al.*, *Nucl. Instrum. Methods A* **736** (2014) 112. http://dx.doi.org/10.1016/j.nima.2013.10.014

[35] H. Tsujii *et al.*, *Carbon-ion Radiotherapy* (Springer, 2014)

Bibliography

U. Linz, *Ion Beam Therapy (Fundamentals, Technology, Clinical Applications)* (Springer, 2012).

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Dose Delivery Verification

S. Safai
Paul Scherrer Institut, Villigen, Switzerland

Abstract
This paper focuses on some dosimetry aspects of proton therapy and pencil beam scanning based on the experience accumulated at Paul Scherrer Institute (PSI). The basic formalism for absolute dosimetry in proton therapy is outlined and the two main techniques and equipment to perform the primary beam monitor chamber calibration are presented. Depth–dose curve and lateral beam width measurements are exposed and discussed in detail, with particular attention to the size of the ionization chamber and the characteristic of scintillating–CCD dosimetry systems, respectively. It is also explained how the angular–spatial distribution of individual pencil beams can be determined in practice. The equipment and the techniques for performing regular machine-specific quality checks are focused on (i) output constancy checks, (ii) pencil beam position and size checks and (iii) beam energy checks. Finally, patient-specific verification is addressed.

Keywords
Proton therapy; pencil beam scanning; reference dosimetry; angular–spatial distribution; quality assurance.

1 Introduction

Dosimetry plays an essential role in the clinical activities of any centre that offers radiation therapy as a modality to treat patients afflicted by tumours. After the installation and tuning of any treatment unit, dosimetry is necessary first to accept and then to characterize such a unit in what are typically referred to as *acceptance* and *clinical commissioning*, respectively. After a successful conclusion of the commissioning phase, dosimetry is then required on a regular basis as part of the quality assurance programme. As part of that programme, dosimetric quality and consistency checks are typically repeated on a daily, weekly, monthly and yearly basis to provide the confidence that the system is behaving as expected. Particle-therapy centres are not an exception and adhere strictly to this well-established practice.

In our overview on dosimetry in particle therapy we will follow the sequence outlined above by first looking into the dosimetry equipment and techniques for clinical commissioning (Sections 2 and 3) and then into those for periodic checks (Section 4) with particular attention to those aspects relevant to proton therapy and pencil beam scanning (PBS).

2 Absolute dosimetry

The most relevant task of absolute dosimetry at commissioning is the calibration of the primary beam monitor chamber (BMC). The primary beam monitor chamber—usually a large parallel-plate ionization chamber situated in the nozzle of a gantry—is typically calibrated in terms of monitor units (MUs) per dose–area product (MU/$D_{\mathrm{w}}A$) or alternatively in terms of MUs per proton (MU/p) [1]. The preferred choice is, to a large extent, dictated by the requirements of the treatment planning system (TPS), which, to produce a desired dose distribution, could predict either the dose per pencil beam (most commercially

https://doi.org/10.23730/CYRSP-2017-001.49

available TPSs) or the number of protons per pencil beam (e.g. the in-house TPS PSIPLAN of PSI). The former would require a MU/$D_\text{w}A$ calibration that could be derived with ionization chamber measurements based on IAEA TRS-398 [2], the latter a MU/p calibration that could be derived with Faraday cup measurements. In what follows we will first review the basics of the TRS-398 Code of Practice for protons and then explain the two major techniques to calibrate the BMC.

2.1 Code of practice: the basic formalism

ICRU 78 [3] adopted the IAEA TRS-398 Code of Practice for reference dosimetry in proton therapy. The reader is therefore referred to the latter for a comprehensive overview of reference dosimetry in particle therapy. What follows is a short overview from that report.

The basic formalism according to TRS-398 described the absorbed dose to water $D_{\text{w},Q}$ for a beam quality Q in the following way:

$$D_{\text{w},Q} = M_Q N_{D,\text{w},Q_0} k_{Q,Q_0}. \tag{1}$$

Here M_Q is the instrument reading at user beam quality Q, corrected for all influence quantities, N_{D,w,Q_0} is the absorbed dose to water calibration coefficient for calibration beam quality Q_0 (usually ^{60}Co) and k_{Q,Q_0} is the beam quality factor to correct for effects of differences between calibration beam quality Q_0 and user beam quality Q.

As primary standard laboratories have no, or very limited, access to proton beams, the reference beam quality remains ^{60}Co and the values for k_{Q,Q_0} tabulated in TRS-398 for protons are derived by calculation rather than experimentally, which introduces additional uncertainties in reference dosimetry with protons.

Typically reference dosimetry in a proton beam is performed by measuring the dose with cylindrical ionization chambers placed in the plateau region of a spread-out Bragg peak (SOBP). The beam quality index for the proton beam is defined as the residual range R_res in g/cm^2 at a measurements depth z, with

$$R_\text{res} = R_\text{p} - z. \tag{2}$$

Here R_p is the practical range, i.e. the depth at which the absorbed dose beyond the Bragg peak falls to 10% of its maximum value.

2.2 Beam monitor chamber calibration

2.2.1 With ionization chambers

To calibrate the BMC, the use of SOPBs, even though it is the standard for reference dosimetry, may not be advisable for proton pencil beam scanning under certain conditions, as first pointed out by Jäkel *et al.* [4] for heavy ions. In fact, its applicability depends on the location of the energy modulation system along the beam line with respect to the BMC, as explained below.

(a) If the energy modulation is performed downstream from the BMC, for instance with the use of range shifters in the nozzle, then the BMC will always ‘see’ the same energy for all the energy layers delivered in a given SOBP. As such, only one calibration value has to be determined for a given energy tune. In this case BMC calibration with reference dosimetry in the middle of SOBPs is applicable and should be performed for every energy tune.

(b) If, on the other hand, the energy modulation is performed upstream from the BMC, for instance with the use of a degrader just after the accelerator, then the BMC will ‘see’ different energies, a different one for every energy layer in a given SOBP. Since, in this case, each energy layer corresponds to a different energy tune, the calibration has to be performed individually for every deliverable energy or at least for a subset of deliverable energies.

Hence, the BMC calibration with reference dosimetry in the middle of SOBPs is not advisable in this case. A much more practical approach is the calibration performed with use of small parallel-plate ionization chambers (e.g. Markus chambers) placed at shallow depth in water, e.g. at z equal to 2 g/cm^2, in the middle of a 10 × 10 cm^2 monoenergetic energy layer, to be repeated for all, or a subset of, deliverable energies. Equation (2) is then used to compute the beam quality index for a given energy layer in order to extract the corresponding k_{Q,Q_0}. With this method parallel-plate ionization chambers are preferred to cylindrical chambers since the measurements are performed in a gradient region. As such, particular attention should be given in positioning the chambers at the desired depth z.

2.2.2 *With Faraday cups*

The use of Faraday cups is the direct way to calibrate the BMC in terms of MU/p.

Typically the Faraday cup is placed at the exit of the nozzle in air. All energy tunes, or a subset of deliverable energy tunes, are individually delivered. After delivery, the number of protons is determined directly by the measured charge in the Faraday cup; this number is recorded together with the number of MUs registered by the BMC for a given energy. The ratio between the two numbers will then provide the calibration.

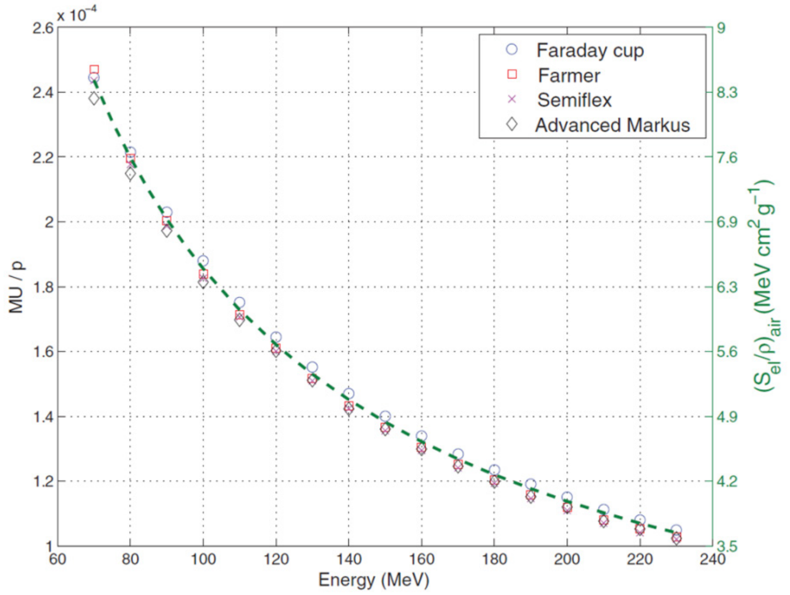


Fig. 1: Comparison between a BMC calibration performed with a Faraday cup and the calibration performed with ionization chambers [1].

At Paul Scherrer Institut (PSI) we compared the calibration performed with a Faraday cup with the calibration derived with different ionization chambers (Fig. 1). The calibrations agree within 3%. Details of this comparison have been published by Gomà *et al.* [1].

2.2.3 *Validation*

Regardless of the method used to calibrate the BMC, reference dosimetry in SOBPs following TRS-398 remains a valuable way to validate the calibration determined with that particular method and to introduce additional correction factors if a significant discrepancy is observed. As an example, at PSI we opted for the Faraday cup method in Gantry 2, which has an upstream energy modulation system as described in Section 2.2.1(b). After the BMC calibration with the Faraday cup, the dose was verified following TRS-398 in the middle of several SOPBs located at different depths. The difference between the measured dose and the expected one was then used to introduce an energy-dependent correction factor in the calibration.

3 Relative dosimetry

At clinical commissioning *relative* dosimetry plays an essential role in the collection of the reference dosimetric beam data required by the TPS for dose calculations. The type and quantity of the data to be collected depend on the requirement of a specific TPS. In general, a comprehensive set of pencil beam integral depth–dose curves and lateral beam widths (i.e. spot sizes) is needed, typically measured for a subset of the deliverable energies. We will illustrate the type of equipment and techniques to perform such measurements.

3.1 Depth–dose curve measurements

Integral depth–dose curves for monoenergetic pencil beams are usually measured with large circular parallel-plate ionization chambers with a diameter (∅) of at least 8 cm. These chambers have an excellent resolution in depth and are large enough to collect the dose deposited by both primary and secondary particles. The chamber is placed in a water tank and while a pencil beam is delivered, either with constant beam ON or on a spot-by-spot basis, the chamber is moved along the beam axis to measure the entire Bragg peak curve (Fig. 2). At each position in depth the integral signal measured by the chamber is recorded and later normalized with either the MU measured by the BMC or the signal from a large reference chamber positioned at the entrance of the tank.

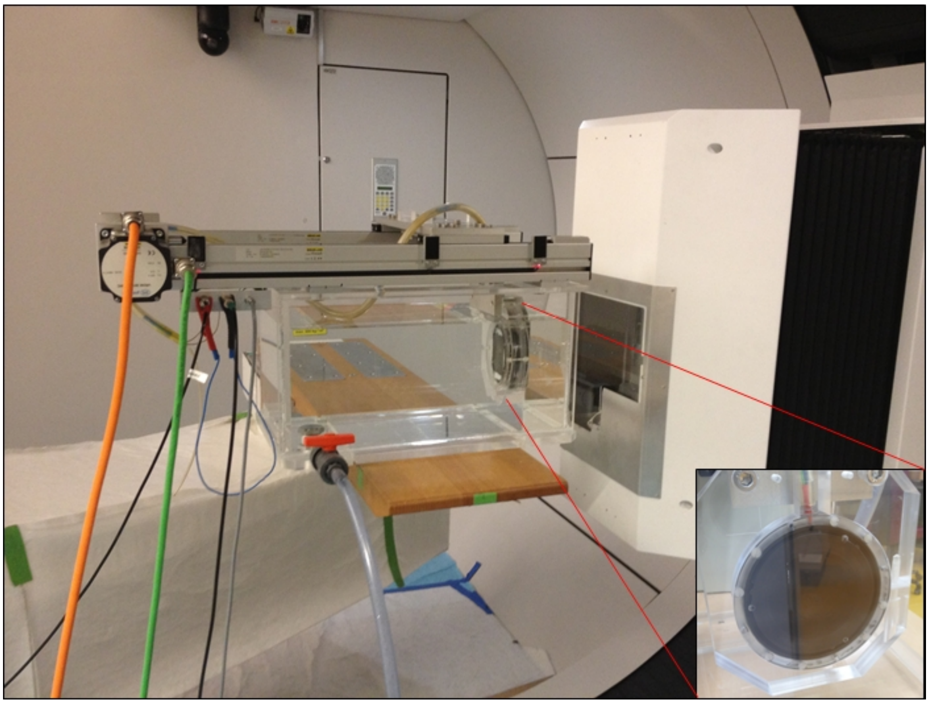


Fig. 2: Experimental set-up at PSI Gantry 2 for range measurements in water

A question that is often asked is what should be the ideal size of the chamber for such measurements. Figure 3 shows in a logarithmic scale the lateral profile of a pencil beam measured at mid range in water for a 150 MeV pencil beam. The primary protons contribute mainly to the central envelope of the distribution and can be well described with a Gaussian function (the first Gaussian in the plot). The sigma of this function is often used to characterize the lateral pencil beam width (also referred to as ‘spot size’). The secondary particles, as a result of nuclear interactions of the primary beam in the medium, deposit dose not only in the central envelope but also outside that region and create a so-called *halo* of deposited dose around the primary beam. The halo is mainly produced by large scattered secondary protons. In its most simplistic form the halo can also be described with a Gaussian function but with a significantly larger sigma [5]. The sigma of this second Gaussian could be up to 2-cm large [5]. Figure 3 shows that the second Gaussian can well describe the beam halo up to 4 cm from the central axis. Beyond that point there could be additional dose deposited that would need a better mathematical description. From this we learn that an 8-cm (∅) circular ionization chamber could miss some of the dose deposited at larger radius and that a wider chamber could be advisable. On the other hand, if a treatment planning system is unable to properly describe the additional dose deposited outside that boundary, e.g. if only a two-Gaussian model is implemented, then an 8-cm chamber could be sufficient and the use of a larger one could be counterproductive. Hence, the answer to the original question is not so straightforward and depends also on how well the TPS is able to describe the halo. At PSI we decided to use an 8-cm chamber. Figure 4 shows the comparison between 8-cm and a 12-cm chambers for five different energies measured at PSI. A small but significant difference is observed only at mid range for high-energy beams.

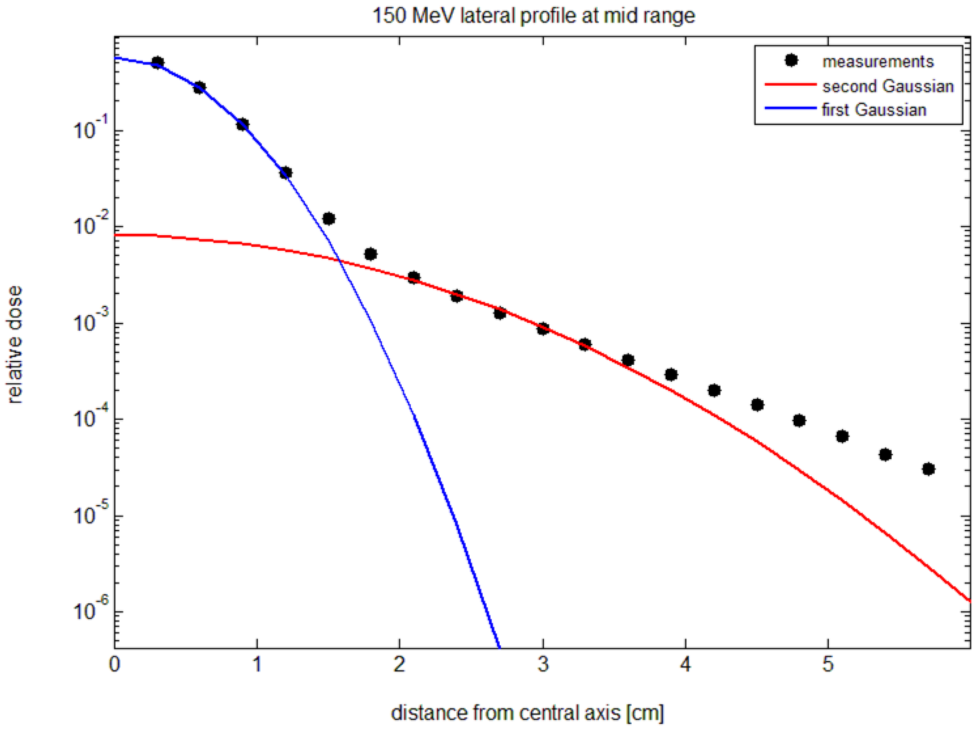


Fig. 3: Transversal profile of a 150 MeV pencil beam measured in water at mid range

3.2 Lateral beam width measurements

In this section first a brief exposure of the fundamentals of pencil beam propagation in air is given followed by how to determine in practice the spot size and angular–spatial distribution of individual pencil beams.

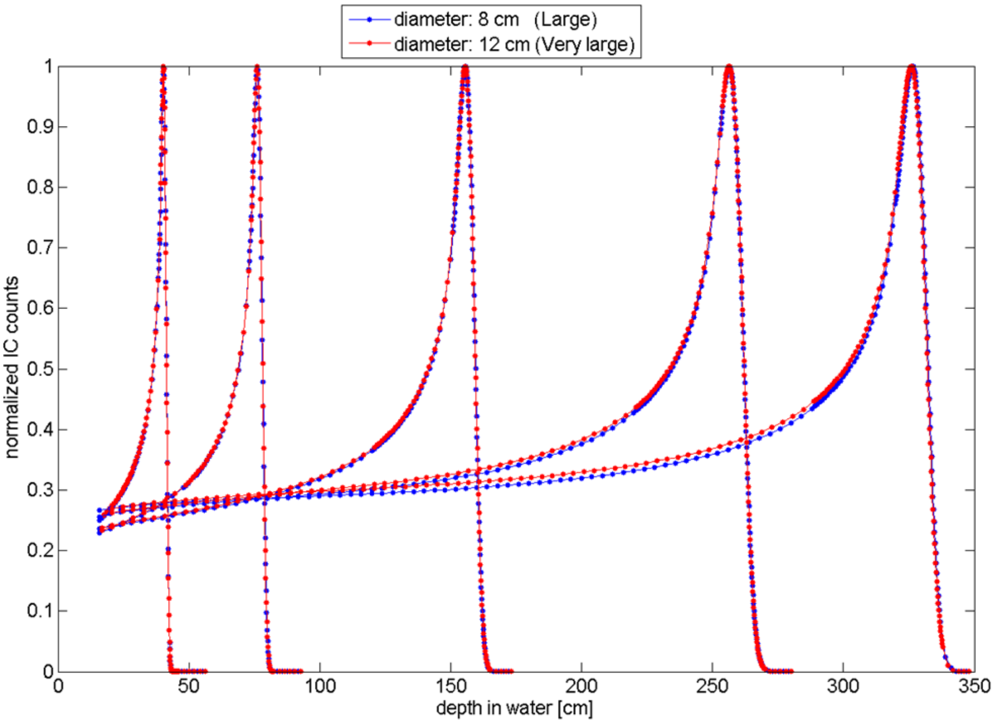


Fig. 4: Depth–dose curves measured in water with 8-cm and 12-cm-large parallel-plate ionization chambers

3.2.1 Pencil beam propagation in air: generalized Fermi–Eyges theory

Generally the beam optic of a PBS gantry is designed and tuned to bring the waist of the beam as close as possible to the gantry isocentre, i.e. the spot is smallest at isocentre. In reality multiple Coulomb scattering (MCS) in the nozzle and air moves the waist further upstream from the isocentre, in particular for low-energy beams, for which MCS is more pronounced. As the lateral size of an individual pencil beam changes as the beam propagates in air, a unique value to describe the size of such a beam is, in general, not sufficient. Generalized Fermi–Eyges theory can well describe the propagation of a pencil beam in air by parametrizing the angular–spatial distribution of the pencil beam with three parameters, the so-called *moments* of the distribution, A_0, A_1 and A_2 [6, 7]. The angular–spatial distribution, ASD, represents the Eyges solution to Fermi's diffusion equation and can be described as follows (for the x coordinate):

$$\mathrm{ASD}_x(x,\theta,z) = \frac{1}{\pi\sqrt{A_0A_2-A_1^2}}\mathrm{e}^{-\frac{(A_0x^2-2A_1x\theta+A_2\theta^2)}{(A_0A_2-A_1^2)}} . \tag{3}$$

Here ASD_x describes the distribution of the protons within a spot both in angle and in position projected onto the x–z plane at a given longitudinal position z. In general, A_0, A_1 and A_2 are a function of z and they represent the doubled angular variance, doubled covariance and doubled spatial variance, respectively. Hence, we can write

$$A_0(z) = 2\sigma_\theta^2(z), \tag{4}$$

$$A_1(z) = 2\mathrm{Cov}(x,\theta,z), \tag{5}$$

$$A_2(z) = 2\sigma_x^2(z), \tag{6}$$

where σ_x and σ_θ is the spatial and angular spreads, respectively. As in general the multiple Coulomb scattering in air is small, the propagation of the beam at the exit of the nozzle and in the proximity of the isocentre can be expressed as it would propagate in vacuum, i.e. with

$$A_2(z) = A_{0,0}z^2 + 2A_{1,0}z + A_{2,0,} \tag{7}$$

where $A_{0,0}$, $A_{1,0}$ and $A_{2,0}$ represent the initial moments of the distribution at the reference position $z = 0$. When $A_2(z)$ is known, then the spot size (expressed with σ_x) at a given position z can be computed with Eq. (6), i.e.

$$\sigma_x = \sqrt{A_2(z)/2}. \tag{8}$$

3.2.2 Angular–spatial distribution determination in practice

Nowadays treatment planning systems can describe the beam propagation in air with Eq. (7) or with similar equations. From Eq. (7), we learn that the doubled spatial variance A_2 is a quadratic function of z and that the initial moments are the coefficients of this function.

Experimentally the coefficients are determined by measuring A_2 at, at least, three different planes in z. A quadratic fit to the measured A_2 as a function of z will provide the coefficients, i.e. the initial moments of the angular distribution, which will be used by the TPS to predict the spot size at any other plane in air via Eqs. (7) and (8) (see Fig. 5). The values of A_2 for the fit are typically determined experimentally by measuring high-resolution 2D lateral profiles of individual pencil beams at those planes. A 2D Gaussian fit is then performed on the 2D lateral profiles to obtain σ_x and σ_y. Equation (6) and the analogous equation for y are then used to compute A_2 at those planes.

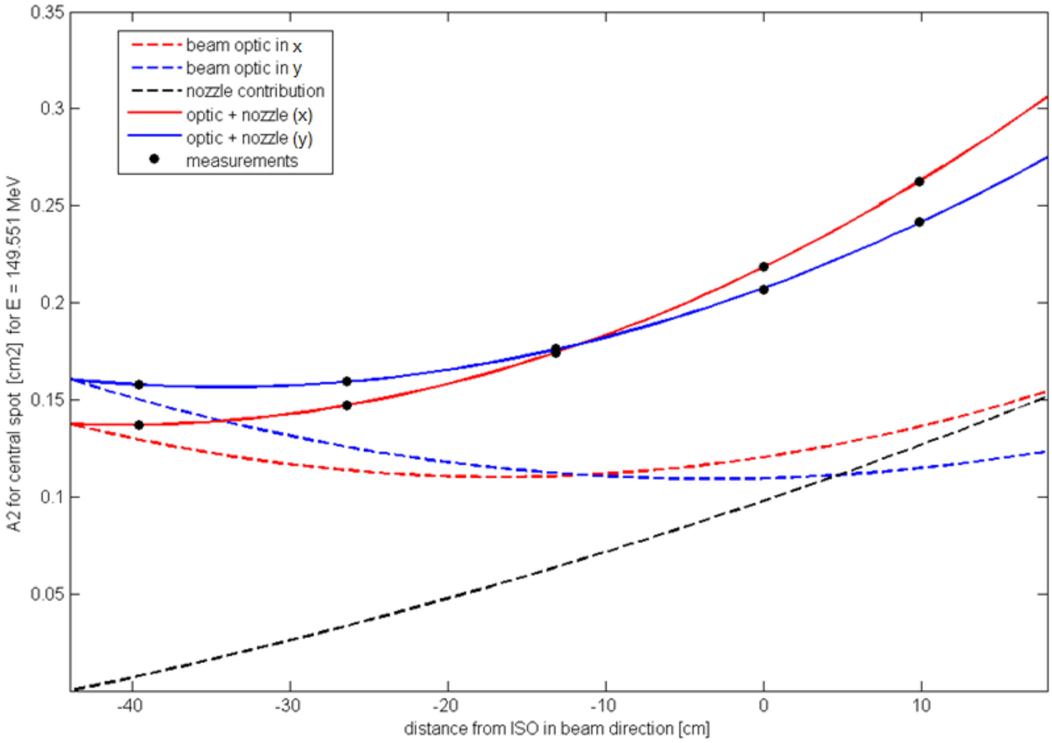


Fig. 5: Double spatial variance A_2 for a 150 MeV pencil beam measured in air at PSI Gantry 2 (dots), and the quadratic fit through the solid lines. The black dashed line shows the nozzle contribution due to the MCS of the nozzle alone. The dashed coloured lines are derived from the difference between the fit and the nozzle contribution and can be interpreted as the propagation of A_2 due to the beam optic alone.

For this kind of measurement it is advisable to use high-resolution detectors, such as scintillating screens or Gafchromic films (e.g. EBT3). Scintillating screens are usually used in combination with a mirror and a CCD; therefore hereafter we will refer to this equipment as *scintillating–CCD dosimetry systems* (Fig. 6).

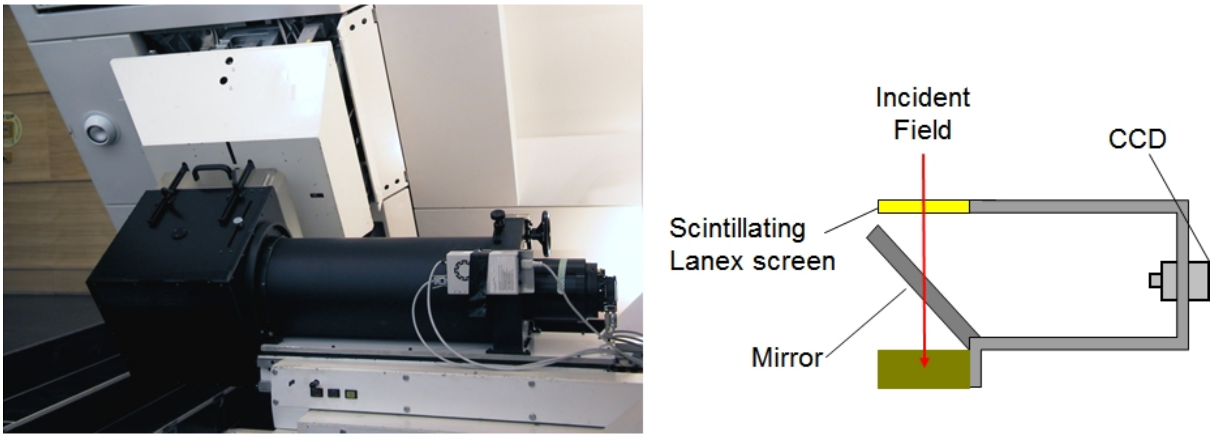


Fig. 6: Scintillating–CCD dosimetry system at PSI Gantry 1 (left) and a schematic representation (right)

It should be noted that the response of both Gafchromic films and scintillating screens depends on the energy and the linear energy transfer (LET – the energy transferred to the medium per unit of path length) [8, 9]. The response decreases with decreasing energy and increasing LET. Therefore, a pencil beam depth– dose curve measured with such systems shows a supressed Bragg peak compared to a curve measured with ionization chambers. This phenomenon is documented in the literature and has been often called the *quenching effect*. The quenching effect is irrelevant for transversal dose distribution measurements when the energy spectrum across the measurement plane can be assumed constant, which is generally the case for spot-size determination in air [10].

Scintillating–CCD systems have the advantage compared to Gafchromic films to be linear with dose and to have an electronic readout providing faster and more accurate results. With these systems a large quantity of data can be recorded and analysed almost simultaneously, representing one of the most efficient ways of collecting the necessary dosimetric data at commissioning.

4 Periodic checks

4.1 Machine-specific dosimetry

For PBS there are three main important dosimetry aspects that have to be verified constantly, on a daily, monthly and yearly basis, i.e.:

i. absolute dose (output constancy check);

ii. pencil beam position and size (including alignment at isocentre);

iii. beam energy (range measurements).

The rationale for these checks is to identify as early as possible problems with: (i) the monitor calibration and/or in general with the system, (ii) the scanning system and/or beam line optic and (iii) the energy selection system and/or beam line.

4.1.1 Output constancy checks

Output constancy checks are typically performed with reference ionization chambers, either in a phantom (daily checks) or in water (yearly checks), in the middle of flat dose distributions (i.e. SOBPs).

4.1.2 Pencil beam position and size checks

Scintillating–CCD systems, strip chambers, Gafchromic films and amorphous-Si detectors are among the devices that can be used to perform constancy checks of the beam size and of the accuracy of beam positioning. Compact systems like strip chambers could be used on a daily basis, while larger systems like the scintillating–CCD systems could be employed on a monthly or yearly basis, even though it is difficult to generalize. As a matter of fact, there are now new commercial products coming to the market designed to use scintillating–CCD systems for daily quality assurance, which are quite promising.

The performance of the scanning system, i.e. the accuracy of pencil beam positioning, is probably the one more prone to changes over time than any other system (e.g. energy-selection system); at least, this is our experience at PSI. The accuracy of beam positioning is typically gantry angle and energy dependent and could fluctuate from day to day. It could also depend on the ramping scheme of the beam line and the sequence in which the energy layers are delivered. Therefore, particular attention should be given to the frequency and comprehensiveness of the consistency checks for this system. Both the alignment of the beam at isocentre should be verified (absolute beam position) as well as the relative position between neighbouring spots. The homogeneity of large energy layers is quite sensitive to the precision of the placement of the individual pencil beams within the layers. Hence, the delivery of such layers in air, covering the lateral extent of the scanning region, on a 2D high-resolution detector (e.g. scintillating–CCD system), is an effective way to verify the performance of the scanning system.

4.1.3 Beam energy checks

The energy is indirectly checked by verifying the range and/or the shape of the Bragg curve. Range measurements in water as described in Section 3.1 are time consuming and are usually repeated only on a yearly basis. Multilayer ionization chambers (MLICs), on the other hand, can record a full Bragg curve extremely fast in a single measurement [11]. If, on top of that, the MLIC can be synchronized with the beam delivery then hundreds of energies could be measured in a few minutes. MLICs are cross-calibrated against measurements in water and the reproducibility of the measured range is very high, well below the typical tolerance of 1 mm for range checks. If well integrated, a MLIC could be a powerful tool to perform daily range measurements or it can be employed for comprehensive weekly or monthly checks.

The use of wedges in combination with 2D detectors (e.g. scintillating–CCD systems) is another alternative to perform energy checks, even though the number of energies that can be verified in the same session is small. Another method is to look at the ratio of the signal measured by at least two small detectors (ionization chambers or diodes) placed at different depths. The ratio is than compared to a reference value, which is characteristic for a given energy.

4.2 Patient-specific dosimetry

Patient-specific verifications are typically performed by measuring 2D dose profiles at different planes with 2D arrays of ionization chambers in phantoms. For comparison the planned dose is recalculated for that particular phantom and geometry used during verification. Sometimes 2D arrays are only used to verify the relative dose distribution. In this case, the absolute dose is additionally verified with calibrated ionization chambers (e.g. pin-point chambers) in a water phantom for one of a few selected reference points. The field under consideration would have to be applied at least twice, one for relative and one for absolute dose verification. Ideally the dose should be verified under the proper treatment gantry angle for that field, but this would require a rotatable phantom. When a rotatable phantom is not available, then, on top of the field verification under a reference angle, it is advisable to run the field also under the treatment angle but without performing a dose measurement, just to verify that the field is deliverable. This is for the reasons pointed out in Section 4.1.2; the dedicated monitor in the nozzle could identify potential problems of spot positioning for that field.

Each individual field is typically verified and the output corrected if a discrepancy is observed between measured and planned doses. At PSI, only after several years of experience and improvements in the in-house dose calculation engine did we reach the confidence to drop the verification of every individual field planned for Gantry 1. On the other hand, Gantry 2 has been in operation only for a few years and, therefore, as of now, all planned fields for this gantry are being verified.

4.2.1 Log file analysis

Dose calculations based on the parameters registered in log files during beam delivery, such as spot position and delivered MUs, could become a relevant tool in the verification of the delivered dose and could reduce the amount of measurements to be performed for each planned field prior to the treatment [12]. This approach could also verify on a daily basis the delivered dose and could therefore be used to adapt the treatment if necessary.

References

[1] C. Gomà *et al.*, *Phys. Med. Biol.* **59** (2014) 4961. http://dx.doi.org/10.1088/0031-9155/59/17/4961

[2] P. Andreo, Absorbed dose determination in external beam radiotherapy, IAEA Technical Reports Series 398 (2000).

[3] International Commission on Radiation Units and Measurements, Prescribing, recording, and reporting proton-beam therapy, ICRU Report 78 (2007).

[4] O. Jäkel *et al.*, *Med. Phys.* **31** (2004) 1009. http://dx.doi.org/10.1118/1.1689011

[5] E. Pedroni *et al.*, *Phys. Med. Biol.* **50** (2005) 541. http://dx.doi.org/10.1088/0031-9155/50/3/011

[6] S. Safai *et al.*, *Phys. Med. Biol.* **53** (2008) 1729. http://dx.doi.org/10.1088/0031-9155/53/6/016

[7] L. Eyges, *Phys. Rev.* **74** (1948) 1534. http://dx.doi.org/10.1103/physrev.74.1534

[8] S. Reinhardt *et al.*, *Med. Phys.* **39** (2012) 5257. http://dx.doi.org/10.1118/1.4737890

[9] S.N. Boon, *Dosimetry and Quality Control of Scanning Proton Beams* (Ponsen & Looijen, Wageningen, 1998).

[10] C. Gomà *et al.*, *Phys. Med. Biol.* **58** (2013) 2509. http://dx.doi.org/10.1088/0031-9155/58/8/2509

[11] S. Lin *et al.*, *Med. Phys.* **36** (2009) 5331. http://dx.doi.org/10.1118/1.3244034

[12] G. Meier *et al.*, *Phys. Med. Biol.* **60** (2015) 2819. tp://dx.doi.org/10.1088/0031-9155/60/7/2819

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Ion Sources for Medical Applications

S. Gammino
Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali del Sud, Catania, Italy

Abstract
Ion sources are key components of accelerators devoted to different types of medical applications: hadron-therapy facilities (accelerating protons or carbon ions), high-intensity accelerators for boron-neutron capture therapy (using intense proton beams), and facilities for isotope production (using different ion species). The three types of application present different requirements in terms of ion beam quality, reproducibility, and beam availability. Different characteristics of ion sources will be described, along with the reasons why they are particularly interesting or largely used.

Keywords
Ion sources; plasma; hadron therapy; accelerators.

1 Introduction

For medical facilities, ion-source design must be oriented according to the needs of three demanding applications:

- hadron-therapy facilities (accelerating protons or carbon ions);
- BNCT (boron-neutron capture therapy, accelerating protons);
- isotope production (producing different ion species).

The requirements for an ion source devoted to medical applications are not unequivocally determined. High currents are required (between hundreds of µA and tens of mA), but there are more stringent constraints than the one on beam intensity. The characteristics for ion sources devoted to these purposes can be summarized as follows:

a) ability to produce the necessary beam current for the treatment (with a contingency margin of 20% or more), i.e. 200–400 eµA for carbon ions and three-times more for protons (up to tens of mA of protons in the case of BNCT facilities);

b) beam emittance lower than accelerator acceptance, typically 0.5-0.7 π mm mrad normalized;

c) high stability, low beam ripple, and high reproducibility;

d) user friendly;

e) high MTBF (mean time between failure);

f) low maintenance.

Requirements d)–f) are very important for all types of application, since hospital facilities need versatile devices with a short tuning time.

Some sources may be adapted for different applications, i.e. a source producing multiply charged ions may work for hadron therapy and for isotope production, while intense beams of protons are used for BNCT and for isotope production. Often, in order to minimize the spare parts and the overall complexity, the same type of ion source is made available in two copies, in order to be interchangeable. Uptime is essential for medical applications, more than for ion sources devoted to nuclear-physics

https://doi.org/10.23730/CYRSP-2017-001.59

accelerators. The above-mentioned characteristics restrict the scope of possible ion sources to those high-current ion sources based on the formation of a dense and hot plasma. The ions are generated inside the plasma, and then extracted by beams of electrostatic optical elements. The production of the plasma, the extraction of ions from the plasma, and the beam management coincide to fulfil the requirements, and particular care is given to each stage to avoid the emittance growth (and poor transport of the beam to the users) and the beam halo formation.

1.1 General technical characteristics

The definition of ion-source properties is given by the mechanism of plasma formation. In a few words, a good knowledge of the plasma parameters simplifies the fulfilment of the requirements. Laboratory plasmas—including the ones generated in compact ion sources—can be produced in several manners, and they differ from each other in the electron temperature, electron density, and plasma lifetime. Electrical discharges under vacuum, electron beams passing through neutral gases or, in a more complicated manner, electromagnetic waves interacting with gases or vapours, in the presence of a well-shaped magnetic field, are the mechanisms used for inducing discharges in gaseous systems, and producing plasmas with suitable characteristics as sources of singly charged or multi-charged ions. Some common features of any ion source used in research or medical laboratories are summarized here:

- a cylindrically shaped under-vacuum cavity with metallic walls, with vapour or gas fluxed before turning on the plasma;
- some flanges for gas feeding, ovens, electron beams, and other tools for plasma discharge tuning;
- a system feeding the energy needed to the discharge (electrostatic energy, electromagnetic waves, etc.);
- a system for plasma confinement (e.g. magnetic trap);
- an ion extraction system able to optically manipulate the beam, ensuring good collimation, low energy spread, and low emittances.

Other ancillary systems include pumps, control systems, and several kinds of diagnostics (for both the plasma and the produced beam). More information is available in [1–3].

Ion sources may be chosen from among a large variety, e.g.:

1) PIG: i.e. Penning ion sources [4];
2) EBISs: i.e. electron beam ion sources;
3) LISs: i.e. laser ion sources;
4) ECRISs: i.e. electron cyclotron ion sources.
5) H-source ‘family’:
   a) helicon sources;
   b) surface plasma sources;
   c) volume sources;
   d) RF sources.

Sources 1)–3) in the above list are not convenient for the use in medical facilities for several reasons. PIG sources are unfit for medical applications due to their short lifetime. EBISs are more appropriate for nuclear-physics facilities requiring low currents but extremely high charge states. LIS produced beams are affected by high-energy spreads. Sources 4) and 5) (including all the different kinds of sources belonging to this ‘family’) are therefore the most used in worldwide medical facilities. The motivation of their use and the specifications that are particularly relevant will be hereinafter described, while for the operation principle of each type of ion source, we refer to [1–3].

2 Hadron therapy facilities

Because of the increasing interest in different ion species and beam features, and because of the larger social impact of hadron therapy, a large part of this note is devoted to such ion sources.

Hadron therapy, up to now, has used very few types of particles, thus limiting the requirements for ion sources to proton (even for neutron generation) and carbon beams, in most cases. Rarely, deuteron, helium, oxygen, neon and argon beams have been required.

2.1 Ion sources for proton therapy

The current rate requested for proton therapy is always in the range of 1 mA or even less, so the types of proton sources [1–3] used as injectors for proton-therapy accelerators are:

- multicusp volume or surface plasma sources for H^-;
- 2.45 GHz microwave discharge ion source;
- duoplasmatron;
- PIG ion sources.

Several types of H^- sources are used, based on different mechanisms of plasma generation and beam formation. Surface plasma sources, with magnetrons, Penning ion gauge (PIG) ion sources with and without convertors, as well as magnetic-multipole volume sources, with and without caesium, are used. The methods of igniting and maintaining magnetically confined plasmas may be quite different: hot and cold cathodes, radio frequency and microwave power, etc. The extraction systems are specialized and use magnetic and electric fields to guide the beam towards the low-energy beam transport line, providing cw (continuous wave) beams or pulsed beams with ad hoc time structures.

Simple commercial multicusp sources are often used to produce 1–2 mA of protons [5] while more complex devices are developed for larger currents. Several improvements have been implemented to the H^- sources in the last decades; e.g. at Lawrence Berkeley National Laboratories (LBNL, USA) many efforts were made in developing innovative RF antennas for the optimization of RF matching to the plasma. In 1990, Leung *et al.* [6] reported the use of inductively generated plasma for producing H^- beams ‘with almost no lifetime limitation’. The efficiency was shown to be higher than the case of a source with a filament. In 1993, the same group reported [7] a three-fold gain in H^- beam using a collar with a SAES company Cs dispenser. In 1996, Saadatmand *et al.* [8] increased the current to 70–100 mA running at 10 Hz 0.1 ms with the SSC (Superconducting Super Collider) source modelled after the LBNL source. The H^- beam appeared to be stable for up to 8 hrs.

Additional advancements were achieved at LBNL with the SNS (Spallation Neutron Source) H^- ion source: a caesium-enhanced, multicusp ion source [9]. For the ion extraction and transport, an advanced Low-Energy Beam Transport (LEBT) line was designed in order to properly manipulate the extracted beam. The very compact LEBT, shown in Fig. 1, was made of a two-lenses, electro-static system, only 12 cm long. Lens 2 was split into four quadrants to steer, chop, and blank the beam.

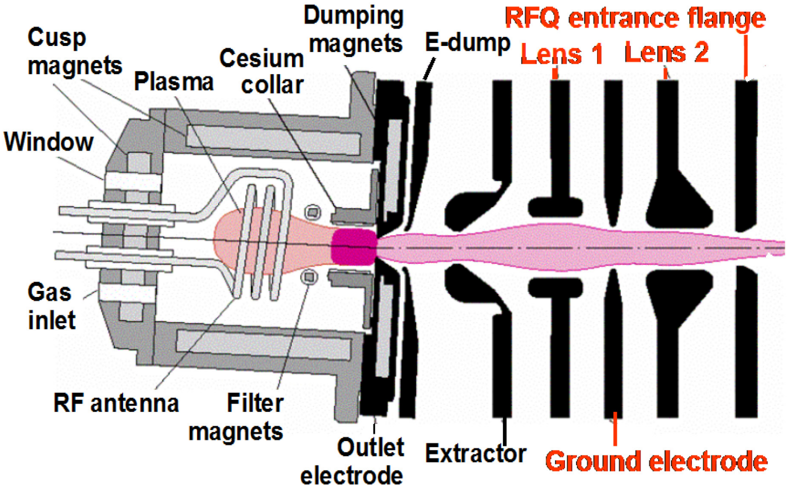


Fig. 1: Sketch of the H⁻ source with a compact LEBT [10]

The results are good and the source is suitable with some limitations for medical accelerators: a 30 mA peak current is produced with horizontal r.m.s. normalized emittance of 0.115 mm mrad, and, even when running Cs free, the H⁻ volume source produces 10–15 mA of H⁻.

Another widely used H⁻ source is the so-called Helicon Discharge Surface Plasma Source [10]. The versatility of this source, along with its ability to produce different ion species, the fairly long MTBF, and relatively good stability, might be applicable to different medical facilities. Figure 2 shows the main subsystems of these devices, including: 1–gas valve; 2–discharge volume; 3–discharge vessel; 4–helicon saddle-like antenna; 5–magnetic coil; 6–ion/atom converter; 7–electron flux; 8–emission aperture (slit); 9–extraction electrode; 10–suppression /steering electrode; and 11–ground.

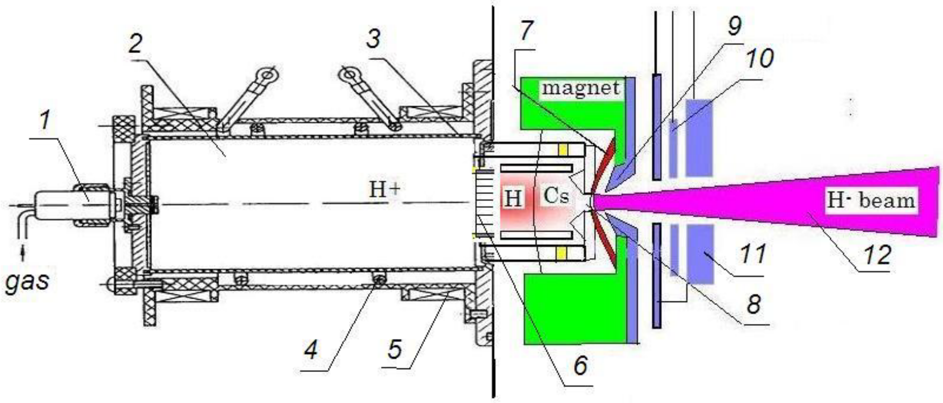


Fig. 2: Scheme of helicon discharge surface plasma source

The H⁺ sources which may be suitable for medical facilities are the 2.45GHz off-resonance discharge microwave ion sources (also known as microwave discharge ion sources—MDIS) [11–12]. They present many advantages in terms of compactness, high reliability, ability to operate in cw mode or in pulsed mode, reproducibility, and low maintenance. High-current proton beams may be delivered, up to 100 mA**,** with low transversal r.m.s. normalized emittance, of the order of 0.20 to 0.30 π mm mrad. The major advantage of MDIS comes from the absence of antennas, which makes this equipment reliable for long duration operations (even months) with good reproducibility. This is a condition for computer-controlled operations, without intervention of operators. The scheme of the source is simple and is shown in Fig. 3.

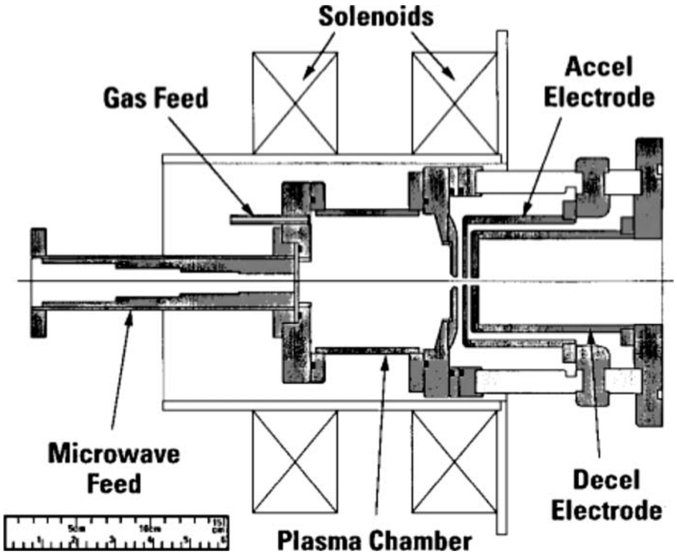


Fig. 3: Original MDIS design at Chalk River National Laboratory [11]

2.2 Ion sources for carbon therapy

ECRISs, EBISs, and laser ion sources may be able to produce multiply charged ions, but only ECRISs are suitable for carbon therapy (and proton therapy as well), as they can provide beam currents of hundreds of microamperes with good long-term stability and reproducibility.

The operating principle and additional information may be found in [1, 13] along with the scaling laws that determine the beam characteristics. Some conditions have been explored in the 1990s that permit a long enough plasma confinement time to be obtained and multiply charged ion beams to be produced. For the sake of brevity, let us say that the conditions needed to produce intense beams of C^{4+} and C^{6+} used in hadron therapy may be summarized by the high-B mode concept, which states that for a high-frequency Electron Cyclotron Resonance (ECR) source, the magnetic confinement should obey the rule

$$B/B_{\mathrm{ECR}} > 2.$$

According to this concept, a higher B/B_{ECR} increases the electron temperature and the ion lifetime and fully stripped heavy ions are obtained. It does not explain the evolution of the electron temperature T_e and its dependence on the power density in the plasma chamber (microwave coupling to plasma is limited by other instabilities in addition to the magnetohydrodynamical ones, which instead limit the ion lifetime). Moreover, the charge exchange process becomes crucial for inner-shell electrons and, even in presence of high production rates of highly charged ions (HCI) in the plasma, few HCI appear in the extracted beam if the pressure is not further improved, in the range of 10^{-7} mbar or better. The operating frequency also plays a role and the ECRISs for hadron therapy use microwave generators with frequencies equal to, or higher than, 10 GHz.

The description of the ECRIS operating principles in [1, 13] does not complete the scope of specific requirements of ECRISs for hadron-therapy facilities. In fact, though the conditions of magnetic confinement, vacuum, and microwave frequency are satisfied, the result may be not sufficient for the full exploitation of an accelerator facility for treatment, and additional conditions must be satisfied in order to ensure patient safety and successful treatment. Additional interlocks on the RF system, or high voltage power supplies, permit the medical operator to ensure the patient's safety, while for successful treatment, it is necessary to have a stable emittance figure and a beam ripple as low as possible.

It should be also considered that ECRISs must be adapted to make them suitable for hospital facilities:

- a hospital facility is not adapted to the ‘difficult case’ of multi-parameter ECR sources that can be managed in a laboratory environment, so the most advanced third generation ECRIS must be discarded [14];
- a ‘high-performance conventional ECRIS’ has large operating costs for electricity, which is not the case for a source with permanent magnets, even if its performance is worse;
- minimizing the number of components subject to failure makes the source reliability better.

Standard ECR ion sources have been specialized for carbon therapy. The two series most commonly known worldwide are the Supernanogan type and the so-called KEI series. Their main features are listed in Table 1. Figure 4 shows a picture of the two ion sources [15–16].

Table 1: Main features of commercial ion sources

Source	**Supernanogan**	**KEI series**
Type	ECR	ECR
Magnets	permanent	Permanent
Ion	Carbon	Carbon
Charge/current	4+/200–250 µA	4+/240–430 µA
Extraction voltage	24 kV	30 kV
Frequency	14.25–14.75 GHz	9.75–10.25 GHz
Operation	CW	Pulse
Gas	CO_2	CH_4

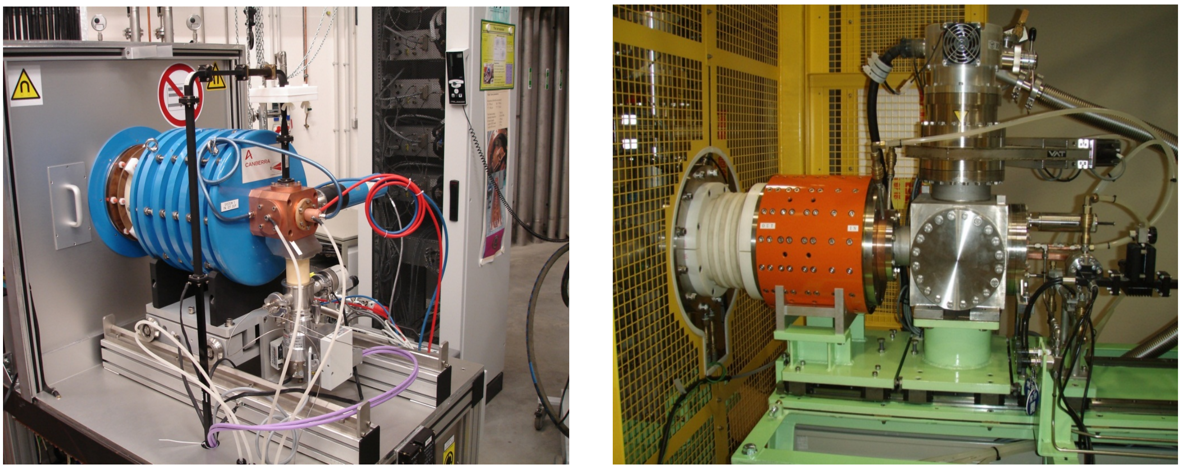


Fig. 4: (left) A Supernanogan-type ion source [15], and (right) a KEI-type source [16]

The KEI sources are used in Japanese facilities, whilst the Supernanogan type are largely used in Europe, and they became a standard solution as beam injectors of hadron-therapy facilities, in spite of their limited beam brightness. The Supernanogan source is equipped with a double-wall, water-cooled plasma chamber, with a 7 mm diameter aperture for beam extraction. The permanent-magnet system provides the axial and radial confinement (axial field from 0.4 to 1.2 T, radial field 1.1 T), and the extraction system and RF injection have been optimized through the years for the maximum reliability.

The RF injection consists in the so-called, copper-made 'magic cube', made of a waveguide to coaxial converter with a tuner to minimize the reflected power. An RF window is used for the junction between the magic cube at high vacuum and the waveguide at atmospheric pressure.

The injection flange hosts also a DC bias system to add electrons to the plasma and decrease the plasma potential. An RF generator, of about 400 W at 14.5 GHz, is used for feeding the plasma (the effective power used in operation is below 300 W). In order to make a further optimization of the beam properties (both current and emittance) possible, flexible frequency variable travelling-wave-tube amplifiers (TWTAs) are used in order to exploit the so-called frequency tuning effect (FTE) [17]. The main beam parameters, and the improvements proposed by INFN-LNS scientists, are listed in Table 2 (implementation of a frequency tuning device for the microwave injection; design of a new extraction system for further improvement of the beam emittance and beam stability; and changes to the gas input system in order to achieve a much better stability).

The FTE was applied since significant improvements were observed at INFN-LNS when a source was fed by a klystron or a Travelling-Wave-Tube (TWT) based generator. It was then understood that FTE does not only change the plasma density, but it also strongly affects the beam formation dynamics. In Fig. 5, it is shown that a wide fluctuation of the output beam current is obtained by varying the feeding frequency for a C^{4+} ion extracted by the Supernanogan source of CNAO, Pavia.

The impact of the wave frequency on the formation and spatial distribution of the warm electrons, including the effects on ion dynamics, was investigated in [18] showing that: i) the energy absorption is influenced by the electromagnetic field modal distribution inside the cavity, affecting the heating rapidity; ii) the frequency also impacts on the density distribution—warm electrons are mostly formed where a high field intensity exists; and iii) the RF heating near resonance induces the formation of a non-homogeneously distributed plasma. All this information can be exploited for our purpose, because the variation of the electron energy distribution fraction may optimize the production of some charge states while the electromagnetic mode distribution may improve the beam emittance.

Table 2: Main beam parameters and improvement steps of the Supernanogan ion source (made with the aid of INFN-LNS ion-sources team).

Ions	Current (requested) [μA]	Current (available) [μA]	After improvements by INFN-LNS [μA]	Emittance (requested) [π mm mrad]	Emittance (measured) [π mm mrad]	Stability [99.8%]
C^{4+}	200	200	250	0.75	0.56	36 h
H_2^+	1000	1000		0.75	0.42	2 h
H_3^+	700	600	1000	0.75	0.67	8 h
He^+	500	500		0.75	0.60	2 h

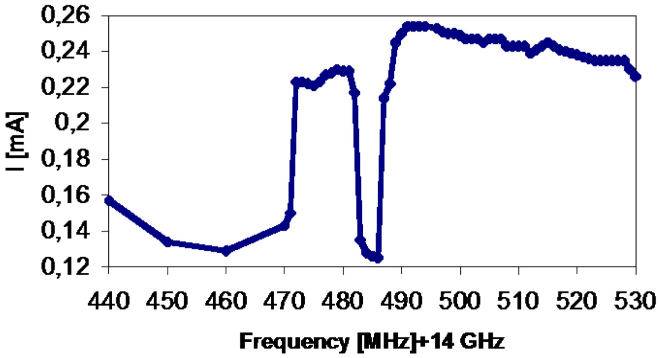


Fig. 5: Experimental evidence of frequency tuning effect with the Supernanogan-type ECRIS of CNAO, around 14 GHz [17].

Additional ‘alternative heating schemes’—like two-frequency heating (TFH)—have proved very helpful in improving the beam stability, which indeed represents a key challenge for cancer treatment machines. TFH consists of simultaneously launching two electromagnetic waves into the plasma, at different frequencies, in order to improve the plasma confinement and the overall stability. This, in turn, decreases the extracted beam ripple, as can be seen in Fig. 6.

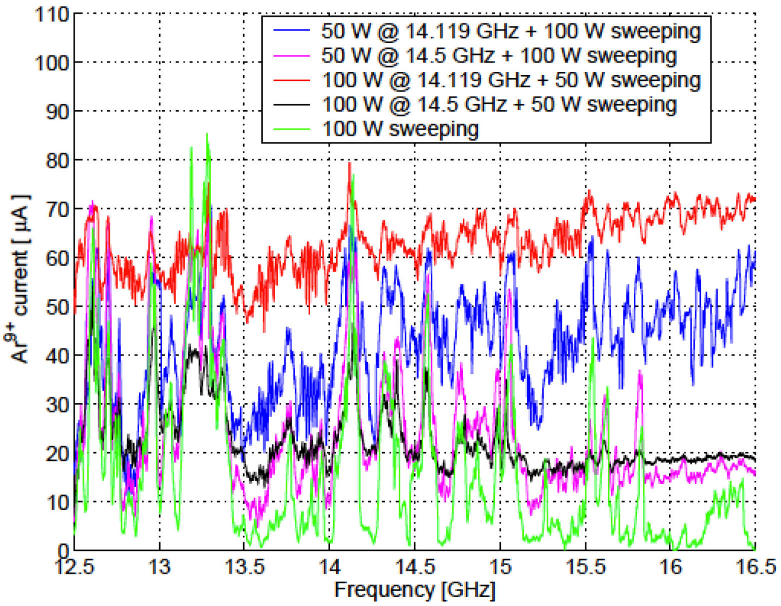


Fig. 6: Impact of TFH—two frequency heating in terms of current and beam ripple damping [19]

Future perspectives in hadron therapy—like the production of higher carbon-ion currents and metallic ions as lithium or beryllium—will require more performing ECRISs.

The AISHa (Advanced Ion Source for Hadrontherapy) project [20] will try to overcome such challenges. AISHa is a hybrid ECRIS: the radial confining field is obtained by means of a permanent magnet hexapole, while the axial field is obtained with a set of four superconducting coils. The superconducting system will be helium free at 4.2 K, by using two cryocoolers. The magnetic field values are following the scaling laws (R. Geller) and the high-*B*-mode concept [13–14]. The operating frequency of 18 GHz has been chosen to maximize the plasma density, taking into account the availability of commercial microwave tubes and the specificity of the installation in hospital environments. The electric insulation is chosen to be 40 kV, for daily operation above 30 kV. A sketch of the AISHa ion source is shown in Fig. 7, while the main features are listed in Table 3.

The set of four superconducting coils independently energized will permit a flexible magnetic trap to be realized and the electron energy distribution function (EEDF) to be adapted to the users’ needs. The use of a broadband microwave generator able to provide signal with complex spectrum content, will permit the frequency to be tuned efficiently, increasing the electron density, and therefore, the performance, in terms of current and average charge state produced.

The chamber dimensions and the injection system have been designed to optimize the microwave coupling to the plasma chamber, taking into account the need for space to house the oven for metallic ion beam production. This feature may represent a significant step above the current technology, with the goal of keeping a very low beam ripple, even in the presence of larger plasma density and high oven temperature.

The ability of the AISHa source to operate with larger RF power may be applied to the production of C^{6+} for compact cyclotrons. In fact, the use of an accelerator chain based on a synchrotron increases the cost of a heavy-ion-therapy facility, but, up to now, the success of compact cyclotrons accelerating C^{6+} has been limited by the insufficient stability, reliability, and reproducibility of high-performance ECRISs. For example, the HEC (high-voltage extraction configuration) source developed at NIRS-Chiba (National Institute of Radiological Sciences) is a powerful ECRIS operating at 18 GHz, and its performances are more than double the KEI sources, but conversely, its application in a hospital has not been judged as viable by the research group that developed it, because of its complexity and reproducibility.

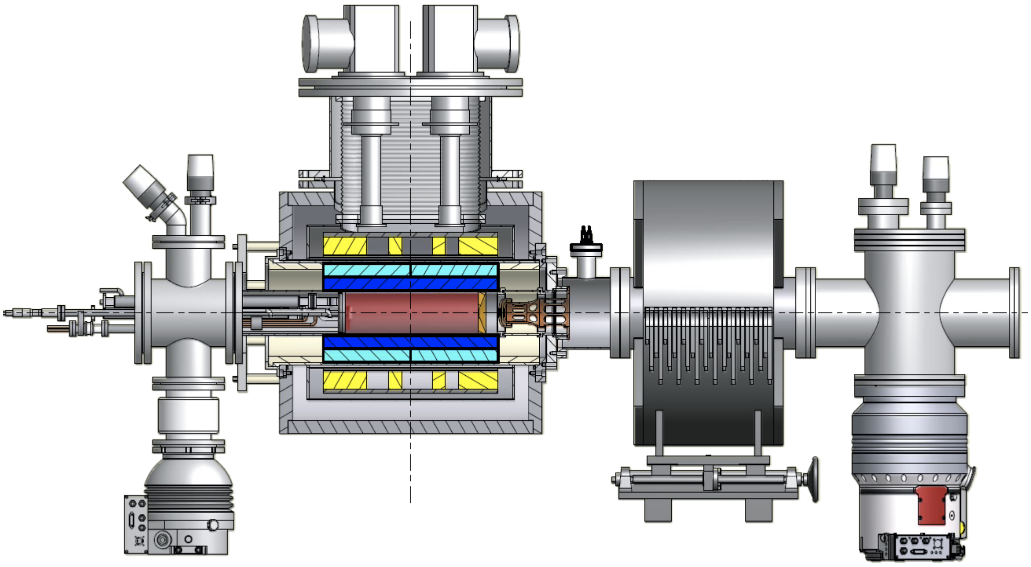


Fig. 7: Sketch of the AISHa ion source [20]

Table 3: Main operating parameters of the AISHa source

Parameter	Value
Radial field	1.3 T
Axial field	2.6 T–0.4 T–1.5 T
Operating frequencies	18 GHz (TFH)
Operating power	1 kW
Extraction voltage	40 kV
Chamber diameter / length	*Φ* 92 mm / 300 mm
LHe	Free
Iron yoke diameter / length	42 cm / 60 cm
Source weight estimation	480 kg

2.3 Ion sources for BNCT

Ideal ion sources for BNCT should fulfil the requirements of neutron-flux production via reactions induced by intense proton beams. The beam current (with a margin of 20% or more) should lie in the range of many mA of protons. Even in this case, high stability and high reproducibility are mandatory, along with a very high MTBF.

Two types of ion sources are particularly suitable for BNCT facilities: multicusp sources of H^- beams and MDIS for H^+.

A typical example is the multicusp source of Kyoto University. The source is able to provide 15 mA of H$^-$ beams with 0.66 π mm mrad 4 r.m.s. normalized emittance [21]. The source was developed in 1990 for injection in a TR70 cyclotron. The first version was able to produce up to 7 mA H$^-$ beams, but the upgraded version (1994) allowed the source performances to improve by up to 15 mA cw at 5 kW of arc power.

MDIS ion sources can be useful sources for BNCT since they generate beams characterized by low emittance (< 0.2 π mm mrad) and beam ripple. They are able to guarantee very high MTBF due to their reliability. Figure 8 shows the Trasco Intense Proton Source (TRIPS), a high-intensity microwave source, whose goal is the injection of a maximum proton current of 35 mA in a radio-frequency quadrupole (RFQ), with an r.m.s.-normalized emittance lower than 0.2 π mm mrad at the operating voltage of 80 kV [22]. TRIPS was originally developed as a proton source for driving a subcritical reactor to transmute nuclear waste, but it is considered for BNCT facilities at INFN. It has been tested for reliability, which reached an interesting value of 99.8% over 142 h.

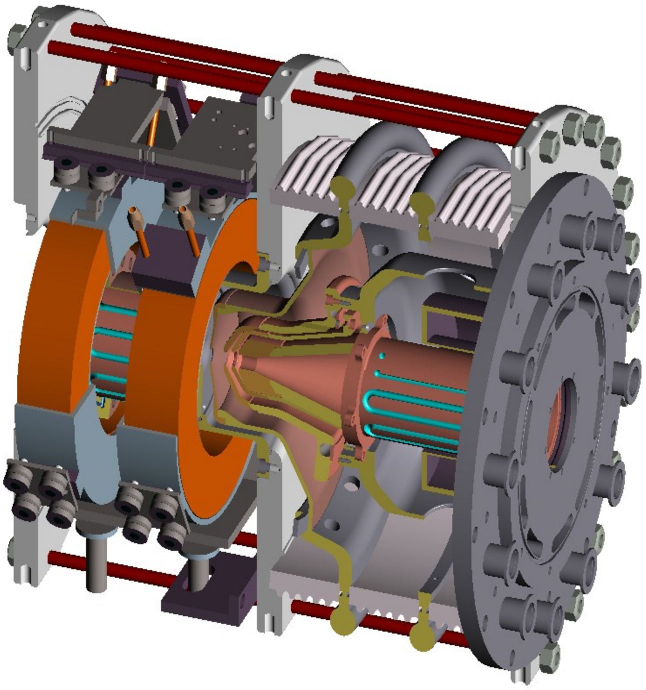


Fig. 8: Layout of the TRIPS proton source designed and tested at INFN-LNS [22]

2.4 Ion sources for isotope production

Ion sources for isotope production must be characterized by beam currents as large as possible (depending on the limits on target reliability) and emittance lower than the accelerator acceptance. Also, in this case, a high MTBF is required. Commercial cyclotrons for the in-situ production of protons or deuterons at energies in the range 10–100 MeV, with beam currents up to 2 mA, are available.

H$^-$ ion sources are often used, such as a multicusp ion source, or RF-driven ion sources and helicon sources. Other options, such as ECR-driven ion sources, have been considered but they have a narrow application because of the larger cost and complexity.

It can be seen that the development of sources for isotope production is not subject to breakthrough but is rather the outcome of technological steps. For example, the development of the filament-driven surface conversion ion source regarded the filament properties. In fact, the filament strength and longevity depend on the material grain size, impurities, and processing. For example, adding 3% of rhenium into tungsten filaments has been proven to enhance the material properties in other applications.

The improvement of the filament material would allow longer lifetimes at the same performance level or higher currents at the same lifetime. The improvement of the source temperature control permits higher performance, provided that the lifetime does not change significantly. Furthermore, the improvements in beam current allow the size of the extraction aperture to be reduced, leading to the decrease of the emittance.

The feasibility of RF-driven H^- ion sources has been attractive because of their longer lifetimes, obtained by means of an external antenna, and their larger plasma density. Furthermore, surface production is more effective and technologically simpler [23]. The neutral pressure seems to be the limiting parameter in the case of the helicon source, whilst the 2.45 GHz MDIS could be a useful kind of source for overcoming any limitations in future years.

3 Conclusions

In this presentation, I have outlined some peculiar aspects of the science and technology of ion sources, rather than the working principles of the different types of ion sources which cannot be described in a short paper; they are available in the CERN Accelerator School textbook and in other relevant textbooks [1–3]. More relevance was given to the comprehension of the requirements set by the different medical applications along with the ability to provide the needed specifications of the source over a long time period, with decent reproducibility in order to guarantee the satisfaction of the users, particularly important in the case of therapy. There is plenty of room for future improvements coming from a better knowledge of the basic processes and improved ability to simulate them and to forecast the produced beam properties. The development of better ion sources may permit the success of treatment to be increased, the cost to be decreased, and finally the application of accelerators to be extended to medical facilities as the result of increased social awareness. The optimization of performance of ion sources and their matching to the accelerators may improve the budget of the isotope production facilities and may extend the variety of isotopes with great advantages for medical diagnostics.

Acknowledgement

I wish to thank my colleague David Mascali for the editing and final revision of this paper.

References

[1] Proceedings of the CAS-CERN Accelerator School: Ion sources, Senec, Slovakia, 29 May—8 June 2012, CERN–2013–007 (CERN, Geneva, 2013). http://dx.doi.org/10.5170/CERN-2013-007

[2] B. Wolf, *Handbook of Ion Sources* (CRC Press, Boca Raton, FL, 1995).

[3] I.G. Brown, *The Physics and Technology of Ion Sources* (New York, Wiley, 1989).

[4] P. Cohilis and Y. Jongen, Proc. Eur. Part. Acc. Conference, Sitges, 1996.
http://accelconf.web.cern.ch/AccelConf/e96/PAPERS/THPG/THP075G.PDF

[5] D. Wutte *et al.*, *Nucl. Instr. Meth. Phys. Res.* B **142**(3) (1998) 409.
http://dx.doi.org/10.1016/S0168-583X(98)00221-3

[6] K.N. Leung et al., *Rev. Sci. Instr.* **61** (1990), 1110;

[7] K.N. Leung, D.A. Bachman and D.S. McDonald, *Rev. Sci. Instrum.* 64, 970 (1993).
http://dx.doi.org/10.1063/1.1144152

[8] K. Saadatmand *et al. Rev. Sci. Instr.* **66**(6) (1995) 1348. http://dx.doi.org/10.1063/1.1145519

[9] M. P. Stockli, *Rev. Sci. Instr.* **85**(2) (2014) 02B137. http://dx.doi.org/10.1063/1.4862205

[10] F.F. Chen, in *High Density Plasma Sources*, Ed. O.A. Popov (Noyes Publications, Norwich, 1996), Chap. 1, pp. 1-75. http://www.seas.ucla.edu/~ffchen/Publs/Chen155R.pdf

[11] T. Taylor and J.Wills, *Nucl. Instr. Meth. Phys. Res.* A **309**(1–2) (1991) 37. http://dx.doi.org/10.1016/0168-9002(91)90090-D

[12] S. Gammino *et al.*, *Rev. Sci. Instr.* **81** (2010) 02B313. http://dx.doi.org/10.1063/1.3266145

[13] R. Geller, *Electron Cyclotron Resonance Ion Sources and ECR Plasmas* (Institute of Physics, Philadelphia, PA, 1996).

[14] G. Ciavola *et al.*, Proc. Italian Vacuum Association Conference (AIV) (2009), unpublished.

[15] C. Bieth, S. Kantas, P. Sortais, D. Kanjilal and G. Rodrigues, Recent developments in ECR sources, Proc. of the XXXIII European Cyclotron Progress Meeting, Warszawa-Kraków*,* 2002 [Nukleonika **48** (Supplement 2) (2003) S93].

[16] M. Muramatsu *et al.*, *Rev. Sci. Instr.* **81** (2010) 02A327. http://dx.doi.org/10.1063/1.3273055

[17] S. Gammino, *High Energy Phys. Nucl. Phys*. **31**(S1) (2007) 137.

[18] D. Mascali *et al.*, *Rev. Sci. Instr.* **85** (2014) 02A956. http://dx.doi.org/10.1063/1.4858115

[19] F. Maimone *et al.*, *Rev. Sci. Instr.* **82**(12) (2011) 123302. http://dx.doi.org/10.1063/1.3665673

[20] G. Ciavola *et al.*, Proc. of Workshop on ECR Ion Sources, Sydney, 2012. http://accelconf.web.cern.ch/AccelConf/ECRIS2012/papers/tupp08.pdf

[21] T. Kuo *et al.*, *Rev. Sci. Instr.* **67**(3) (1996) 1314. http://dx.doi.org/10.1063/1.1146704

[22] L. Celona, *et al.*, *Rev. Sci. Instr.* **71**(2) (2000) 771. http://dx.doi.org/10.1063/1.1150289

[23] J. Peters, *Rev. Sci. Instr.* **69**(2) (1998) 992. http://dx.doi.org/10.1063/1.1148620

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Imaging in Radiotherapy

K. Parodi
Ludwig-Maximilians-Universität München, Munich, Germany

Abstract
With the continued evolution of modern radiation therapy towards high-precision delivery of high therapeutic doses to a tumour while optimally sparing surrounding healthy tissue, imaging is becoming a crucial component for identifying the intended target, positioning it properly at the treatment site, and, in more advanced research applications, visualizing the treatment delivery. This contribution reviews the main roles of imaging in modern external beam radiation therapy, with special emphasis on emerging ion beam therapy techniques aimed at exploiting the favourable properties of the interaction of ions with matter to achieve unprecedented ballistic accuracy in dose delivery.

Keywords
Imaging; radiation therapy; ion beam therapy; treatment planning; treatment delivery; treatment verification.

1 Introduction

Over the last two decades, modern radiation therapy with external photon beams has evolved considerably, with the introduction of new delivery techniques such as the use of intensity modulation [1] and rotational therapy [2] for spatio-temporal variation of the radiation dose. In addition to photons, other types of ionizing radiation have been explored with the goal of increased dose conformity for better cure and/or reduced toxicity. In particular, ion beam therapy, especially with proton beams, is rapidly emerging as a promising radiation therapy technique owing to its superior ability to concentrate the beam energy in the tumour while better sparing normal tissue and critical organs compared with photons (Fig. 1). In this context, state-of-the-art technologies that can exploit the charged nature of ions by magnetically steering narrow pencil beams of selected energy over the tumour are being introduced into clinical practice to achieve even better conformity of dose delivery [4].

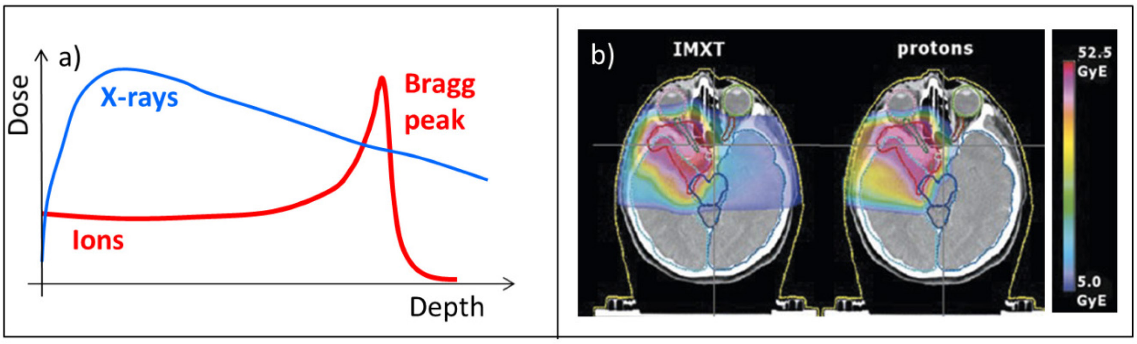


Fig. 1: (a) Illustrative representation of depth–dose profiles of photons and ions in water. (b) Treatment plans for a skull-base tumour (adapted from Ref. [3]), comparing the use of photons, delivered with advanced intensity modulation radiation therapy (IMXT), and state-of-the-art treatment with scanned protons, to illustrate the increased tumour–dose conformity of ion therapy due to the characteristic Bragg peak shown in part (a).

https://doi.org/10.23730/CYRSP-2017-001.71

The growing ability of modern radiation therapy techniques (using both photon and ion beams) to sculpt the radiation dose tightly to arbitrarily complex tumour shapes (Fig. 1(b)) has tightened the demands on imaging technologies, which play a fundamental role in all three of the stages of treatment planning, treatment delivery, and, in more recent research work, *in vivo* treatment verification aimed at potential plan adaptation, as illustrated in Fig. 2. In the following, the main features of the evolution of imaging technologies for photon and ion radiation therapy will be reviewed, as well as ongoing research, with emphasis on *in vivo* verification of ion beam therapy.

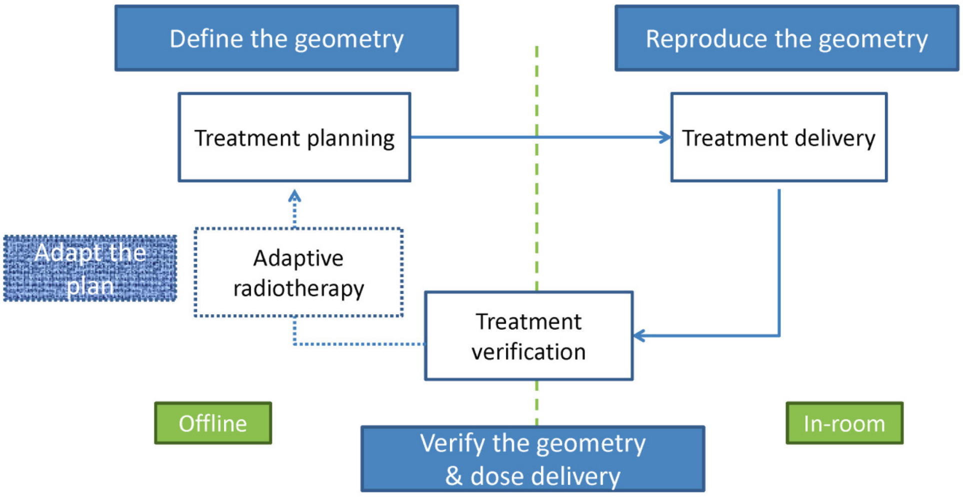


Fig. 2: Schematic representation of the role of imaging in the radiation therapy chain, from treatment planning to delivery and verification, with the potential for adaptation. Workflows that must be performed in the treatment room (in-room) are separated by the dashed line from those which can be performed either inside or outside the treatment room (offline) (adapted by courtesy of C. Gianoli, LMU Munich).

2 Imaging in radiation therapy

2.1 Imaging for treatment planning

The recent physics- and engineering-driven advances in the conformity of dose delivery in radiation therapy were largely motivated by an increased ability to visualize the internal anatomy of the patient with millimetre or submillimetre resolution, as was first made possible by the introduction of volumetric X-ray-based Computed Tomography (CT) in the 1970s [5] and, a few years later, Magnetic Resonance Imaging (MRI) [6]. While X-ray CT still provides the basic patient-specific information for calculating the interaction of radiation with matter for treatment planning, its limited soft-tissue contrast often prevents correct identification of the macroscopic shape of the tumour, called the Gross Target Volume (GTV), for the purpose of delineation. Hence, complementary morphological information from CT and MRI imaging, including physiological motion captured using dedicated time-resolved (4D) acquisition strategies such as respiratory correlated 4D-CT for specific anatomical regions, is typically used to identify the Internal Target Volume (ITV) and Planning Target Volume (PTV). The latter is defined in such a way as to encompass the microscopic extent of the tumour (called the Clinical Target Volume, CTV) expanded by appropriate safety margins, to guarantee coverage of the tumour in the presence of various sources of uncertainty in the treatment.

Beyond morphology, tumour identification increasingly relies also on metabolic and biological features, as assessed by functional imaging via visualization of the spatio-temporal accumulation of injected tracers. In nuclear medicine, Single-Photon Emission Computed Tomography (SPECT) and, more frequently, Positron Emission Tomography (PET) are employed to detect the decay of radionuclides that label selected molecules, resulting in the emission of energetic (a few hundred keV) single photons or annihilation gamma rays [7]. Additional functional information can be provided by special sequences of MRI images that are capable of highlighting physiological processes in the microenvironment of a tumour, as has recently been investigated [8]. Besides contributing to a reduction in inter-observer variability in the delineation of the relevant target volume for treatment planning (Fig. 3) [9], functional imaging opens up the possibility of considering an additional dimension via the definition of the Biological Target Volume (BTV), which describes the complex, heterogeneous microenvironment of the tumour, including subareas of different radiosensitivity [10].

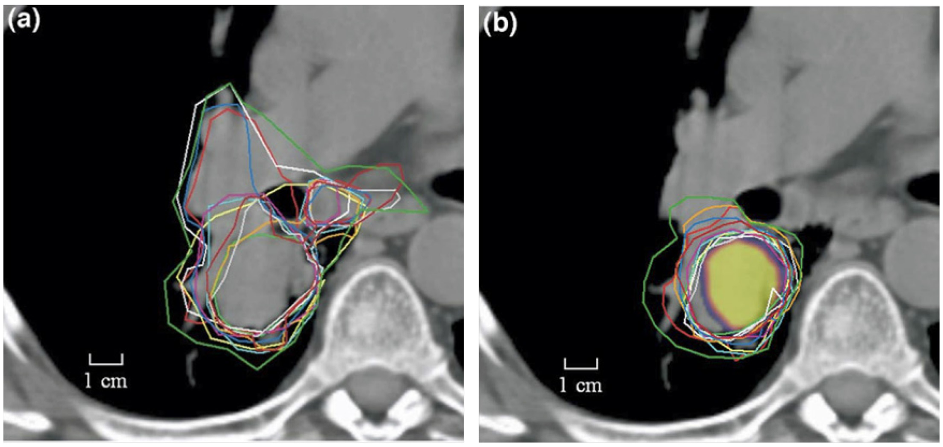


Fig. 3: PET data from an ^{18}F-labelled fluorodeoxyglucose (FDG) tracer, combined with CT imaging, can reduce inter-observer variability in tumour delineation. The coloured contours were defined independently by different radiation oncologists using CT imaging alone in (a) and using PET data in (b), where improved agreement can be seen (reproduced from Ref. [9] with permission.)

Hence, modern radiation therapy is making increasing use of images obtained from different modalities, co-registered using software or hardware fusion (in the case of state-of-the-art combined imaging devices such as PET/CT and PET/MRI scanners), to provide complementary information to the physician to help in the challenging task of segmentation of the tumour and organs at risk, as well as in identification of radioresistant subareas of a tumour which might require a locally enhanced radiation dose for more successful tumour eradication. A treatment plan is then generated using computational engines that optimize the intended dose, in terms of both prescription of the dose to the tumour and constraints on the dose to critical organs, according to the structures identified in the multimodal images used. In this process, the use of CT information is necessary to adapt the well-characterized properties of the interaction of the beam with water to a patient-specific scenario, using the relative electron density (or relative stopping power in the case of ions) extracted from the measured Hounsfield Units (HU) or CT numbers.

2.2 Imaging for treatment delivery

Successful delivery of the planned treatment requires correct replication of the patient's position and anatomy to reproduce properly the conditions used in the dose calculation, which is based on images acquired from the patient (as described in Section 2.1) immobilized in the treatment position typically days or weeks before the start of the fractionated therapeutic course of treatment. Although external alignment based on lasers and skin markers or other optical sensors can provide an initial approximate method of positioning that is easy to implement in practice and does not involve exposure to ionizing radiation, the correct positioning of deep-seated tumours requires imaging of the patient's internal anatomy at the treatment site. To this end, individual orthogonal X-ray radiographs have traditionally been used to identify the tumour position indirectly by the alignment of visible anatomical landmarks, such as bony structures or implanted radio-opaque fiducials. However, the restriction to planar images in which contrast from soft tissue is almost absent severely limits the ability to ensure correct positioning of the tumour volume at the treatment site. Hence, the major improvements which have been achieved in the precision of dose delivery have stimulated advances in dedicated instrumentation for *in situ* image guidance, from the use of simultaneous stereoscopic X-ray projections [11] to the widespread use of volumetric kilovoltage X-ray Cone-Beam Computed Tomography (CBCT) [12] and, very recently, MRI (still limited to only a few installations for photon therapy) [13]. In all cases, advanced image processing is used to compare the acquired images with the intended position in the planning CT images, to provide an appropriate position correction to be executed by a robotic positioning system with three (translational) or six (translational and rotational) degrees of freedom. In this way, it can be better ensured that the intended target is accurately positioned on each day of a fractionated treatment, so that the tumour is properly hit while optimally sparing critical structures.

In addition to providing information for position correction of the patient at the treatment site prior to dose delivery, in a technique typically referred to as Image-Guided RadioTherapy (IGRT), the new volumetric representation of the patient can also be used to assess potential anatomical changes, which may make an adaptation of the initial plan necessary, in a technique called Adaptive RadioTherapy (ART). Such adaptation can be done on the basis of a dosimetric evaluation of the initial treatment plan on the modified anatomy of the patient. However, none of the techniques proposed so far for anatomical image guidance in the treatment position at the exact dose delivery site (i.e., without the need to move the immobilized patient to an in-room CT scanner, as is done in some installations) provides accurate CT numbers for dosimetric calculations. This limitation is especially crucial for particle therapy, since the interaction of the ion beam is extremely sensitive to the stopping properties of the tissue, which influence the finite penetration depth of the beam where the maximum energy deposition occurs (the Bragg peak in Fig. 1(a)). Hence, several methods have recently been proposed to recover HU numbers of quality equivalent to the planning CT images for dose calculation, starting from noisy (mostly due to scattering) X-ray-based images produced in CBCT acquisitions (Fig. 4) (see [14–16] and the references therein), or from properly interpreted (e.g., via manual or threshold-based segmentation) MRI images (e.g., [17]). These methods are being extensively investigated. In this way, the new anatomical information can be used to evaluate the dosimetric consequences of anatomical variations prior to treatment delivery, in order to support a decision about whether a new treatment plan needs to be developed to counteract any relevant changes that may have occurred with respect to the initial planning situation. However, updated anatomical imaging cannot provide information about the actual dose application, as will be discussed in the next section.

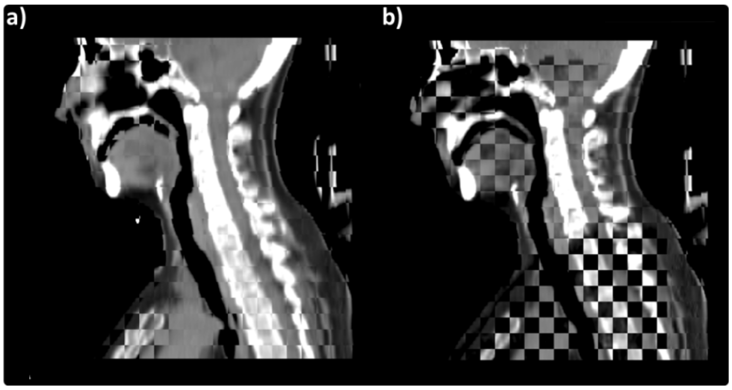


Fig. 4: Chequerboard comparison of two intensity-corrected CBCT images, obtained with (a) deformable image registration of the original planning CT image into the CBCT image and (b) a simple HU scaling of the CBCT image based on a population-based look-up table, against a high-quality CT image acquired a few days apart from the CBCT image but weeks after the original planning CT image, in order to capture non-negligible anatomical changes with respect to the planning scenario. Adapted from Ref. [15] with permission.

2.3 Imaging for treatment verification

In comparison to photons, ions pose the additional challenge of uncertainty in their range, i.e., inaccuracy in the knowledge of the *in vivo* stopping position of the beam, which determines the position of maximum dose delivery, or Bragg peak. Currently, an intrinsic range uncertainty of 1–3% or even more is associated with the conversion of X-ray CT images to the stopping power ratio of the ions relative to water for calculating the intended treatment [18]; this is in addition to possible set-up errors and anatomical changes (Section 2.2). Although morphological image guidance in clinical ion beam therapy is still predominantly being performed with orthogonal or oblique X-rays, and volumetric (cone-beam or on-rail) CT is just entering routine use in clinical practice, several new methods are being investigated to address the problem of range uncertainty in order to complement or even eventually replace X-ray guidance. Such methods are aimed either at improving the knowledge of the stopping properties of the tissue by using the beam itself for imaging (ion radiography/tomography [19, 20]) or at measuring the stopping position of the ion beam in the target by exploiting secondary emission generated by nuclear reactions (using PET and prompt gamma monitoring [21]), as illustrated in Fig. 5 and described in the following.

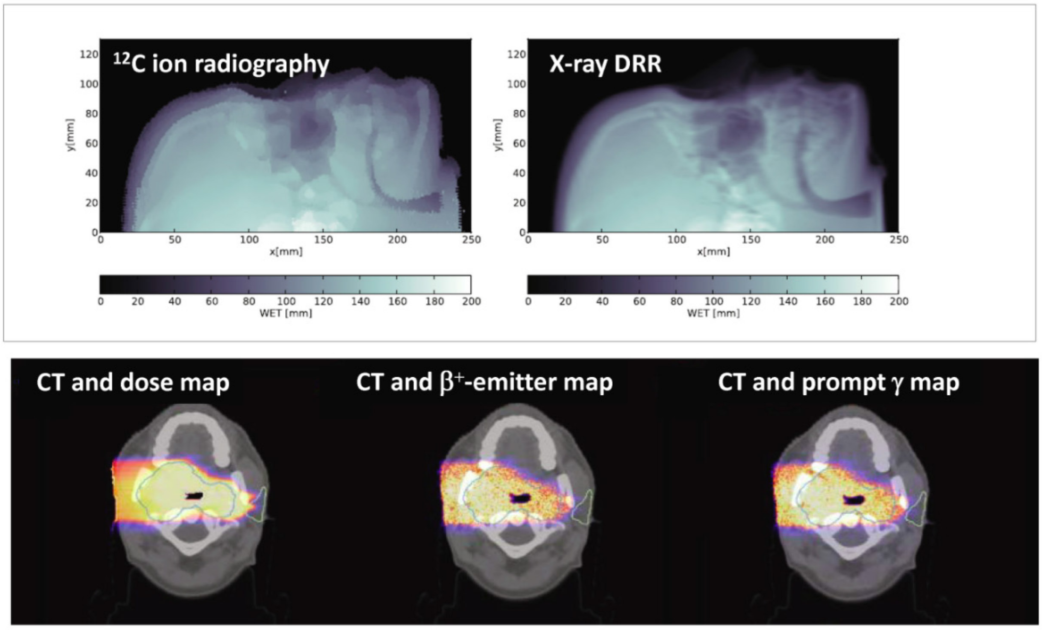


Fig. 5: Top: comparison of measured carbon-ion radiography (left) and digitally reconstructed radiography from an X-ray CT image (right) for an anthropomorphic Alderson head phantom, both calibrated to water-equivalent thickness. Adapted from Ref. [22]. (© Institute of Physics and Engineering in Medicine. Reproduced by permission of IOP Publishing. All rights reserved.) Bottom: comparison of CT-based Monte Carlo-calculated positron emitter maps (middle) and prompt gamma distributions (right) for the original plan of a proton treatment of a patient with a head-and-neck tumour (left), all recalculated from an intensity-corrected CBCT image based on the method of Fig. 4(a).

In facilities that are able to produce sufficiently high beam energies for transmission imaging, calibrated ion radiography can provide two-dimensional information about the water-equivalent thickness of the object traversed (Fig. 5, top), which can be used to verify the position of a subject at the treatment site using a dose that is typically lower than that for X-ray images [23], as well as to refine the knowledge of the integral stopping properties of the tissue in order to adjust the planned treatment prior to irradiation [24]. By rotation of the subject or the beam, several projections can be collected to provide tomographic data, yielding volumetric information about the stopping power ratio of the ions in tissue relative to water, which is approximately independent of the beam energy and the ion species. The set-ups proposed for this depend on whether monoenergetic or polyenergetic beams will be used, and include particle trackers for identification of the most likely ion path (especially in the case of the most strongly scattered protons) followed by residual-energy measurements in calorimeters or range telescopes [25], and simpler planar systems for fluence or energy loss measurements of suitably modulated beams [26]. To date, however, no commercial system exists and most research efforts are being devoted to the development of medium-size prototypes, with a focus on clinical application to cranial sites.

In addition to radiographic or tomographic verification of the stopping properties of tissue by ion-based transmission imaging prior to treatment (or even during treatment sessions), nuclear-based methods offer the opportunity to exploit emission signals correlated with the interaction of the beam with tissue during the therapeutic irradiation. To this end, PET imaging has been the method most often investigated clinically, owing to the more mature and readily available detector instrumentation for imaging the transient production of positron emitters in nuclear fragmentation reactions in tissue. Owing to the half-lives of the isotopes formed, ranging from milliseconds or seconds up to 20 minutes, the irradiation-induced PET signal can be measured during or shortly after therapeutic treatment, using instrumentation integrated into the irradiation unit (called in-beam instrumentation) or installed nearby, inside (in-room) or outside (offline) the treatment room [21, 27]. Although the detected PET signal does not have a straightforward correlation with the dose delivered, it can be compared either with predictions based on advanced CT-based computations, using the same HU-range calibration as for treatment planning, or with PET measurements from previous fractions. Clinical experience reported so far with different implementations and ion species indicates the feasibility of validating the HU-range calibration curve *in vivo* [28], as well as of detecting deviations between the planned and actual treatment delivery due to positioning inaccuracies or major anatomical changes [29–31], thus opening up the possibility of adaptation of the treatment before the next treatment delivery (and hence inter-fractional adaptation). However, it has also highlighted methodological challenges, associated mainly with the adaptation of instrumentation originally developed for different applications in diagnostic nuclear medicine or small-animal imaging. Hence, there are ongoing developments aimed especially at new-generation in-beam PET solutions, which rely either on limited-angle dual-head instrumentation exploiting ultrafast time-of-flight (TOF) detectors [32, 33] or on innovative full-ring designs such as the axially shifted ‘openPET’ [34], together with improvements in computational modelling and data analysis.

Despite the promising clinical results reported by some groups, PET-based verification is intrinsically limited by the delayed emission from the radioactive decay and by the physiological washout that occurs in the time that elapses between production of the positron emitter and detection [21, 27], unless extremely short-lived (millisecond) emitters are used with ultrafast in-beam TOF-PET scanners for quasi-real-time imaging, as has been proposed recently [35, 36]. Conversely, prompt gamma monitoring provides a real-time indirect assessment of the range of the beam from the depth distribution of energetic photons emitted in very fast (on an approximate scale of nanoseconds or less) de-excitation processes, following inelastic scattering and nuclear reactions induced by ion irradiation. Owing to the high photon energies of several MeV and the need for collimated detection for spatial observation of the prompt-gamma–ion-range correlation, several different detection schemes have been proposed, which range from mechanical collimators seen by single or multiple scintillation detectors to more complex Compton cameras, exploiting electronic collimation making use of Compton kinematics (see Refs. [21, 27] and the references therein). Some additional promising developments are aimed at exploiting spectroscopic information about the nucleus-specific, characteristic prompt gamma emission, together with spatial correlation information captured with mechanical collimation [37], or rely on simplified uncollimated detection with very fast scintillators to exploit only the prompt-gamma–proton-range correlation in the time domain [38]. So far, only a prototype system, featuring a single-slit collimator viewed by scintillator detectors, has been realized by a commercial vendor; this is at the stage of clinical evaluation at a few selected proton therapy facilities [39]. The system is limited to a one-dimensional projection of the detected signal and is thus unable to perform volumetric imaging. Nevertheless, in the first clinical applications to passively scattered proton therapy treatment, it has been reported that it is feasible to detect inter-fractional range variations of a few millimetres by comparing prompt gamma profiles measured at different treatment fractions [40]. These findings were corroborated by additional dosimetric recalculations of the treatment plan using new anatomical representations of the patient captured by an on-rail CT system installed in the treatment room [39], thus supporting the promise of the rapidly emerging modality of prompt gamma monitoring for new possibilities of online treatment verification and, potentially, intra-fractional adaptation.

3 Conclusion

With the increasing ability of modern photon and ion radiation therapy to provide highly conformal dose delivery, imaging is becoming of paramount importance in all of the main stages of the therapy chain, from treatment planning to treatment delivery and *in vivo* verification, and will be crucial in potential future inter- and intra-fractional adaptation. This contribution has reviewed the main features of the evolution of imaging in radiation therapy, from novel trends in functional imaging for tumour delineation and new treatment concepts, to integrated devices for *in situ* anatomical confirmation and, in the case of the rapidly emerging field of ion beam therapy, unconventional applications of transmission imaging and nuclear-based techniques for *in vivo* ion range verification.

Acknowledgements

The author would like to thank Drs Gianoli, Dedes, Kurz, and Landry at LMU Munich for providing material used in the graphical illustrations. In this context, support from DFG (under the MAP and carbon ion computed tomography projects) and BMBF (under the SPARTA project) is gratefully acknowledged.

References

[1] T. Bortfeld, *Phys. Med. Biol.* **51** (2006) R363. http://dx.doi.org/10.1088/0031-9155/51/13/R21

[2] K. Otto, *Med. Phys.* **35** (2008) 310. http://dx.doi.org/10.1118/1.2818738

[3] Nuclear Physics European Collaboration Committee (NuPECC), Nuclear physics for medicine, report (2014), Ch. I.9, p. 41.

[4] H. Owen *et al.*, *Int. J. Mod. Phys. A* **29** (2014) 1441002. http://dx.doi.org/10.1142/S0217751X14410024

[5] H. Ambrose and G. Hounsfield, *Br. J. Radiol.* **46** (1973) 148. http://dx.doi.org/10.1259/0007-1285-46-552-1023

[6] P.C. Lauterbur, *Nature* **242** (1973) 190. http://dx.doi.org/10.1038/242190a0

[7] T. Carlier and C. Bailly, *Front. Med.* **2** (2015) 18. http://dx.doi.org/10.3389/fmed.2015.00018

[8] M. Duarte Guimaraes *et al.*, *Radiologia Brasileira* **47** (2014) 101. http://dx.doi.org/10.1590/S0100-39842014000200013

[9] R.J. Steenbakkers *et al.*, *Int. J. Radiat. Oncol. Biol. Phys.* **64** (2006) 435. http://dx.doi.org/10.1016/j.ijrobp.2005.06.034

[10] C.C. Ling *et al.*, *Int. J. Radiat. Oncol. Biol. Phys*. **47** (2000) 551. http://dx.doi.org/10.1016/S0360-3016(00)00467-3

[11] E. Infusino *et al.*, *J. Appl. Clin. Med. Phys*. **16** (2015) 5102.

[12] D.A. Jaffray *et al.*, *Int. J. Radiat. Oncol. Biol. Phys.* **53** (2002) 1337. http://dx.doi.org/10.1016/S0360-3016(02)02884-5

[13] S. Mutic *et al.*, *Semin. Radiat. Oncol.* **24** (2014) 196. http://dx.doi.org/10.1016/j.semradonc.2014.02.008

[14] M. Peroni *et al.*, *Int. J. Radiat. Oncol. Biol. Phys.* **84** (2012) e427. http://dx.doi.org/10.1016/j.ijrobp.2012.04.003

[15] C. Kurz *et al.*, *Acta Oncol.* **54** (2015) 1651. http://dx.doi.org/10.3109/0284186X.2015.1061206

[16] Y.-K. Park *et al.*, *Med. Phys.* **42** (2015) 4449. http://dx.doi.org/10.1118/1.4923179

[17] S.J. Hoogcarspel *et al.*, *Phys. Med. Biol.* **59** (2014) 7383 http://dx.doi.org/10.1088/0031-9155/59/23/7383

[18] H. Paganetti, *Phys. Med. Biol.* **57** (2012) R99. http://dx.doi.org/10.1088/0031-9155/57/11/R99

[19] G. Poludniowski *et al.*, *Br. J. Radiol.* **88** (2015) 20150134. http://dx.doi.org/10.1259/bjr.20150134

[20] K. Parodi, *Phys. Med.* **30** (2014) 539. http://dx.doi.org/10.1016/j.ejmp.2014.02.004

[21] A.C. Knopf and A. Lomax, *Phys. Med. Biol.* **58** (2013) R131. http://dx.doi.org/10.1088/0031-9155/58/15/R131

[22] I. Rinaldi *et al.*, *Phys. Med. Biol.* **59** (2014) 3041. http://dx.doi.org/10.1088/0031-9155/59/12/3041

[23] U. Schneider *et al.*, *Med. Phys.* **31** (2004) 1046. http://dx.doi.org/10.1118/1.1690713

[24] P.J. Doolan *et al.*, *Phys. Med. Biol.* **60** (2015) 1901. http://dx.doi.org/10.1088/0031-9155/60/5/1901

[25] V.A. Bashkirov *et al.*, *Med. Phys.* **43** (2016) 664. http://dx.doi.org/10.1118/1.4939255

[26] H. Ryu *et al.*, *Phys. Med. Biol.* **53** (2008) 5461. http://dx.doi.org/10.1088/0031-9155/53/19/01

[27] K. Parodi, *Nucl. Instrum. Methods Phys. Res. Sect. A* **809** (2016) 113. http://dx.doi.org/10.1016/j.nima.2015.06.056

[28] W. Enghardt *et al.*, *Nucl. Instrum. Methods Phys. Res. Sect. A* **525** (2004) 284. http://dx.doi.org/10.1016/j.nima.2004.03.128

[29] W. Enghardt *et al.*, *Radiother. Oncol.* **73** Suppl. 2 (2004) S96. http://dx.doi.org/10.1016/S0167-8140(04)80024-0

[30] T. Nishio *et al.*, *Int. J. Radiat. Oncol. Biol. Phys.* **76** (2010) 277. http://dx.doi.org/10.1016/j.ijrobp.2009.05.065

[31] J. Bauer *et al.*, *Radiother. Oncol.* **107** (2013) 218. http://dx.doi.org/10.1016/j.radonc.2013.02.018

[32] P. Crespo *et al.*, *Phys. Med. Biol.* **52** (2007) 6795. http://dx.doi.org/10.1088/0031-9155/52/23/002

[33] G. Sportelli *et al.*, *Phys. Med. Biol.* **59** (2014) 43. http://dx.doi.org/10.1088/0031-9155/59/1/43

[34] H. Tashima *et al.*, *Phys. Med. Biol.* **61** (2016) 1795. http://dx.doi.org/10.1088/0031-9155/61/4/1795

[35] P. Crespo *et al.*, *IEEE Trans. Nucl. Sci.* **52** (2005) 980. http://dx.doi.org/10.1109/TNS.2005.852637

[36] P. Dendooven, *Phys. Med. Biol.* **60** (2015) 8923. http://dx.doi.org/10.1088/0031-9155/60/23/8923

[37] J.M. Verburg and J. Seco, *Phys. Med. Biol.* **59** (2014) 7089. http://dx.doi.org/10.1088/0031-9155/59/23/7089

[38] C. Golnik *et al.*, *Phys. Med. Biol.* **59** (2014) 5399. http://dx.doi.org/10.1088/0031-9155/59/18/5399

[39] J. Smeets *et al.*, *Phys. Med. Biol.* **57** (2012) 3371. http://dx.doi.org/10.1088/0031-9155/57/11/3371

[40] C. Richter *et al.*, *Radiother. Oncol.* **118** (2016) 232. http://dx.doi.org/10.1016/j.radonc.2016.01.004

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Overview of Linacs

A.M. Lombardi
CERN, Geneva, Switzerland

Abstract
In this paper, we give an overview of the different types of linac accelerators, with special emphasis on their use for a hadron-therapy facility.

Keywords
Linacs; medical.

1 Introduction

Linac stands for LINear ACcelerator: a single pass device that increases the energy of a charged particle by means of an (radio frequency, RF) electric field and is equipped with magnetic elements (quadrupoles, solenoids, bending magnets) to keep the charged particle on a given trajectory. The motion equation of a charged particle in an electromagnetic field can be written as

$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = q\cdot\left(\vec{E}+\vec{v}\times\vec{B}\right), \tag{1}$$

where

$\vec{p} = \text{momentum} = \gamma m_0 \vec{v}$,

$q, m_0 = $ charge, mass,

$\vec{E}, \vec{B} = $ electric field, magnetic field,

$t = $ time,

$\vec{x} = $ position vector, and

$\vec{v} = \dfrac{\mathrm{d}\vec{x}}{\mathrm{d}t} = $ velocity.

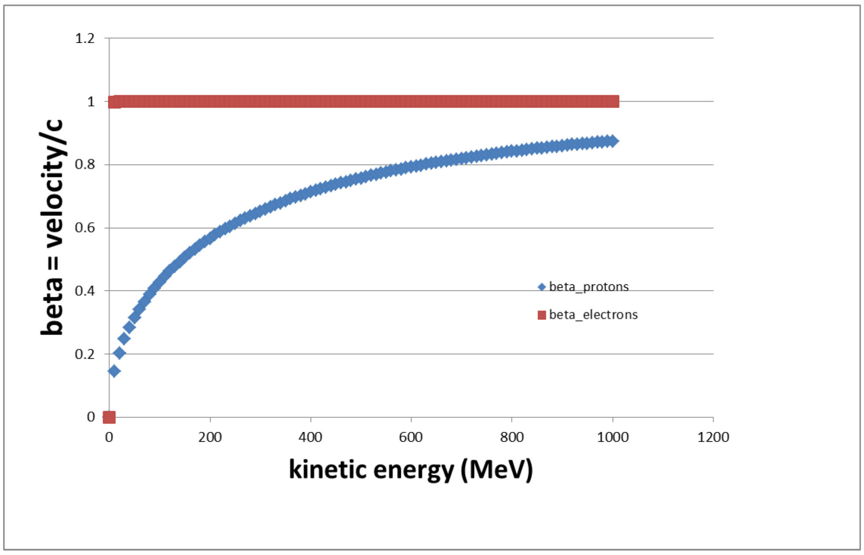
If we rewrite Eq. (1) in a slightly different form, we can identify the key parameters of a linear accelerator. The relativistic factor gamma on the left-hand side of Eq. (2) indicates whether we are in the non-relativistic, semi-relativistic, or fully relativistic regime. The factor q/m_0 on the right-hand side indicates what type of particle the linac is designed for, and the E term and B term on the right-hand side indicate the type of RF structure and the type of focusing, respectively.

$$\frac{d}{dt}\left(\gamma\frac{d\vec{x}}{dt}\right) = \frac{q}{m_0}\cdot\left(\vec{E}+\frac{d\vec{x}}{dt}\times\vec{B}\right). \tag{2}$$

It is interesting to recall the dependence of the relativistic beta for electrons and protons, both particles being used in medical linacs. From Fig.1, we can see that the electrons are relativistic at a few MeV of energy, whereas the protons (and the carbon ions) are never fully relativistic in the energy range interesting for hadron therapy (250 MeV for protons and 450 MeV/u for carbon ions).

Electron linacs for medical applications are very compact. They operate in the energy range between 4 and 25 MeV and they are, nowadays, commercially available. They are, therefore, not the subject of this lecture.

https://doi.org/10.23730/CYRSP-2017-001.79



Proton energy (MeV)	**Relativistic beta**	**Relativistic gamma**
7	0.12	1.01
250	0.6	1.26
450	0.74	1.48

Fig. 1: Relativistic beta for electrons (red) and protons (blue)

The predecessors of modern RF hadron linacs are electrostatic linacs, based on a static field to accelerate particles. This type of linac, used until not long ago in the first stage of acceleration before the dimensions of the system became impractical, is limited to energies of a few MeV. Acceleration by a time-varying electromagnetic field overcomes the limitation of static acceleration. The first experiment towards an RF linac was done by the Norwegian physicist, R. Wideroe [1], in 1928 based on a proposal by Ising dated 1925 [2]. A group {please leave bunch]of potassium ions was accelerated to 50 keV in a system of drift tubes in an evacuated glass cylinder. The available generator provided 25 keV at 1 MHz. It was not until 1931 that the first linac was developed by Sloan and Lawrence at Berkeley Laboratory [3]. The man who brought the linac from an experiment to a practical accelerator was Luis Alvarez. He realized that to proceed to higher energies it was necessary to increase, by an order of magnitude, the frequency and to enclose the drift tubes in an RF cavity (resonator). A 200 MHz 12 m-long Drift Tube Linac (DTL), built by Luis Alvarez at the University of California in 1955 [4], accelerated protons from 4 to 32 MeV. The development of the first linac was made possible by the availability of high-frequency power generators developed for radar application during World War II.

2 Linacs for medical applications

In this section, we will explore in more detail hadron linacs for medical applications. The parameter γ of Eq. (2) is in the range 1 to 1.48 and the q/m goes from 1 (protons) to 1/3 for carbon ions. In this framework, the particles are non-relativistic (or semi-relativistic towards the end of the acceleration), therefore, the choice of the type of RF structure is critical for the efficiency (and cost) of the accelerator complex. In Table 1, the different types of RF structures used in medical facilities (existing and in the design stage) are listed.

Table 1: RF structures used in existing and planned facilities

Type of structure	Used in
Radio Frequency Quadrupole	HIT,CNAO, MEDAUSTRON, ADAM,TULIP2.0
Interdigital-H structure	HIT CNAO MEDAUSTRON
Drift Tube Linac	IMPLART (ENEA FRASCATI) / TULIP2.0
Cell Coupled Linac also called Side Coupled Linac	ADAM, TULIP

2.1 Transverse electric or transverse magnetic modes, and cavity modes

Let us recall the Maxwell equation for *E* and *B* fields [5]:

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)\vec{E}=0. \tag{3}$$

In free space, the solution of the above equation shows that electromagnetic fields are of the transverse electromagnetic, TEM, type: the electric and magnetic field vectors are perpendicular to each other and to the direction of propagation. In a bounded medium, e.g. a cavity, the solution of the equation must satisfy the boundary conditions:

$$\vec{E}_{//}=\vec{0} \text{ and } \vec{B}_{\perp}=\vec{0}\,; \tag{4}$$

therefore, only either transverse electric (TE) or transverse magnetic (TM) modes are possible.

In TE modes, the electric field is perpendicular to the direction of propagation, whereas in TM modes, the magnetic field is perpendicular to the direction of propagation. In a cylindrical cavity, they are denoted as TE_{nml} and TM_{nml}, respectively, where the indices *n*, *m*, and *l* refer to the azimuthal, radial, and longitudinal components.

Figure 2 shows the first two TE modes in a cylindrical cavity. The cut is perpendicular to the direction of propagation of a beam, the lines represent the electric field, and the dots and the crosses are the points where the magnetic field, parallel to the direction of propagation, enters/exits from the paper.

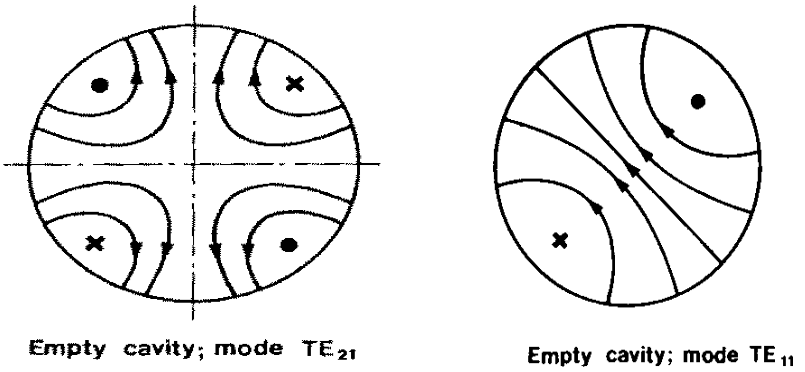


Fig. 2: The first two TE modes in a cylindrical cavity: dipole mode TE_{110} and quadrupole mode TE_{210}

Figure 3 shows the first TM mode in a cylindrical cavity. On the left-hand side, a perpendicular cut where the beam enters the cavity from left to right, and on the right-hand side, a transverse cut where the beam enters into the page.

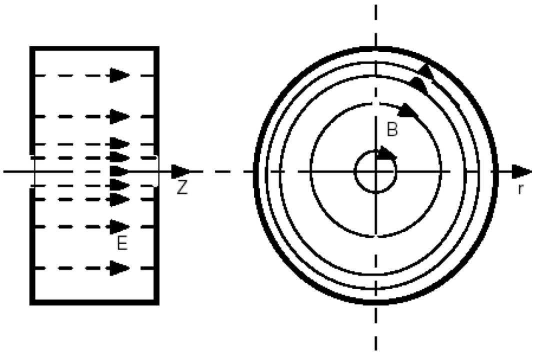


Fig. 3: The TM$_{010}$ mode: most commonly used accelerating mode

A linac is composed of a string of cavities, depending on the phase shift of the field in one cavity with respect to the adjacent cavity, we can identify three main cavity modes. These are:

- 0-mode: zero-degree phase shift from cell to cell, so fields in adjacent cells are in phase. The best example is a DTL;
- π-mode: 180-degree phase shift from cell to cell, so fields in adjacent cells are out of phase. The best example is multi-cell superconducting cavities;
- $\pi/2$ mode: 90-degree phase shift from cell to cell. In practice, these are bi-periodic structures with two kinds of cells, accelerating cavities, and coupling cavities. The CCL operates in a $\pi/2$ structure mode. This is the preferred mode for very long multi-cell cavities, because of very good field stability.

2.2 The radio frequency quadrupole

A radio frequency quadrupole (RFQ) [6, 7, 8] is composed of a cavity loaded with four electrodes and therefore, forced to resonate in the TE$_{210}$ mode. A sketch is shown in Fig. 4.

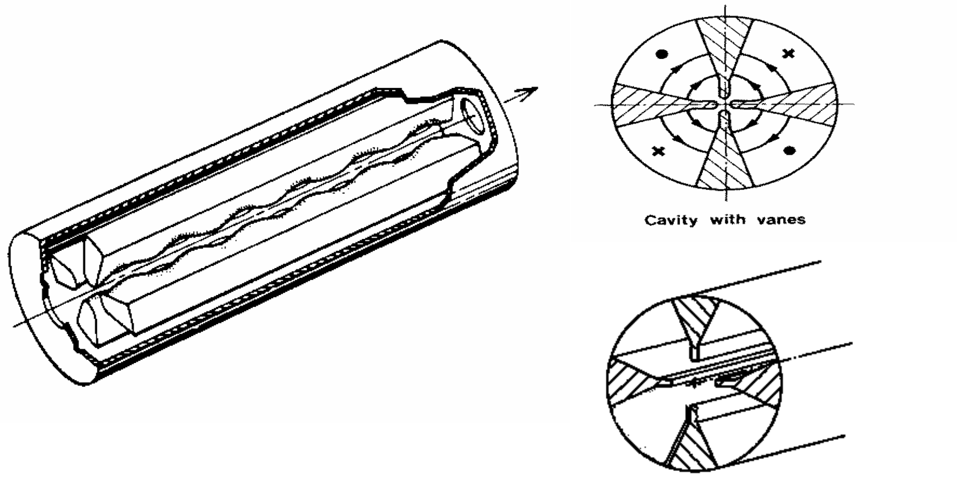


Fig. 4: Sketch of an RFQ cavity, courtesy of T. Wangler

The transverse field in an RFQ provides an alternating-gradient focusing structure with a period length equal to $\beta\lambda$, where β is the velocity of the particle and λ is the RF wavelength. This can be more easily understood by looking at the sketch in Fig.5 and 6. Let us assume that the top and bottom electrodes have positive polarity during the time DT1, DT3…, whereas the left and right electrodes have negative polarity during this time. A positively charged ion beam will pushed away from the top/bottom electrodes and pulled towards the left/right. During this time (DT1, DT3...), the RFQ behaves like a defocusing quadrupole. Conversely, during the time intervals DT2, DT4…, the RFQ behaves like a focusing quadrupole. No force is exerted on the beam at the zero crossing between positive and negative electric fields. During any of the time intervals DT, the beam has travelled a distance equal to $\beta\lambda$ /2 and therefore has seen a smooth focusing channel with period $\beta\lambda$.

Acceleration in an RFQ is achieved by periodically modulating the electrodes in the longitudinal direction, as shown in Fig. 7. The periodic longitudinal modulation, which is 180 degrees out of phase in the top/bottom electrodes with respect to the left/right pair, deforms the pattern of the pure TE mode by creating a longitudinal component proportional to the depth of the modulation. The synchronism between the increasing velocity of the particle and the longitudinal component can be controlled with the wavelength of the modulation.

The RFQ is the accelerator that has bridged the gap between a proton/hadron source and a conventional (TM mode) accelerator. Its strong points are the electric focusing which allows a low-energy beam to be accepted, and the adiabatic bunching which preserves beam quality and allows high capture. These two features combined have increased the efficiency of the very first phase of pre-acceleration from 50% to more than 90%, also in the presence of a strong space charge. Besides, as the transverse and longitudinal dynamics are machined in the electrode microstructure, the RFQ is very easy to operate in routine runs of an accelerator complex.

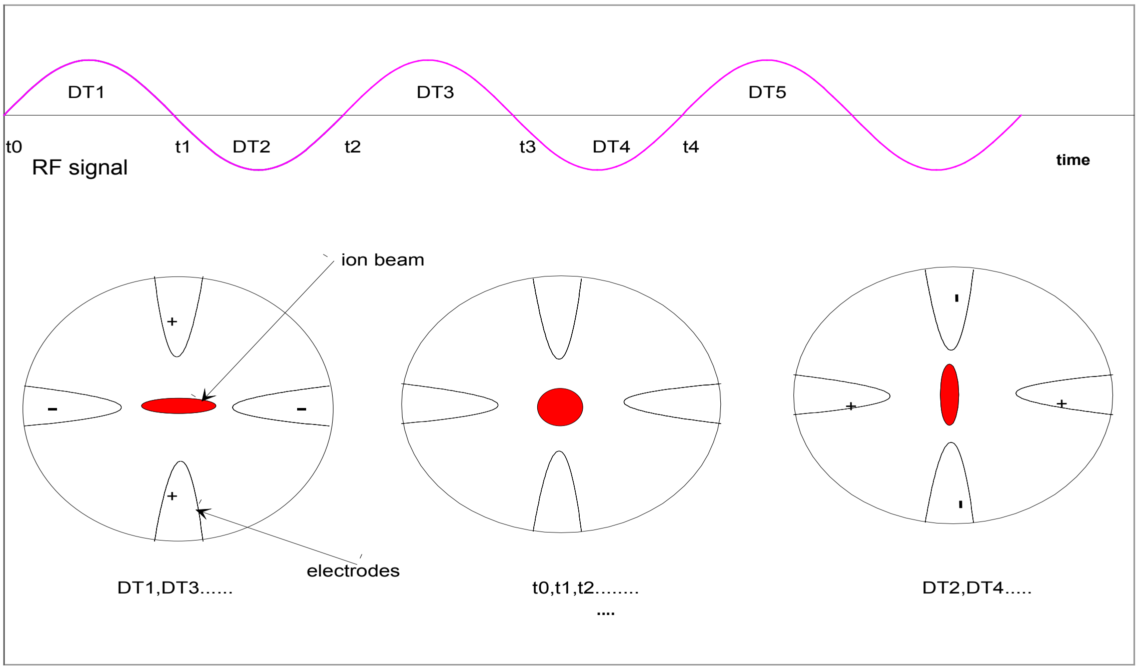


Fig. 5: Time-varying transverse focusing field in an RFQ

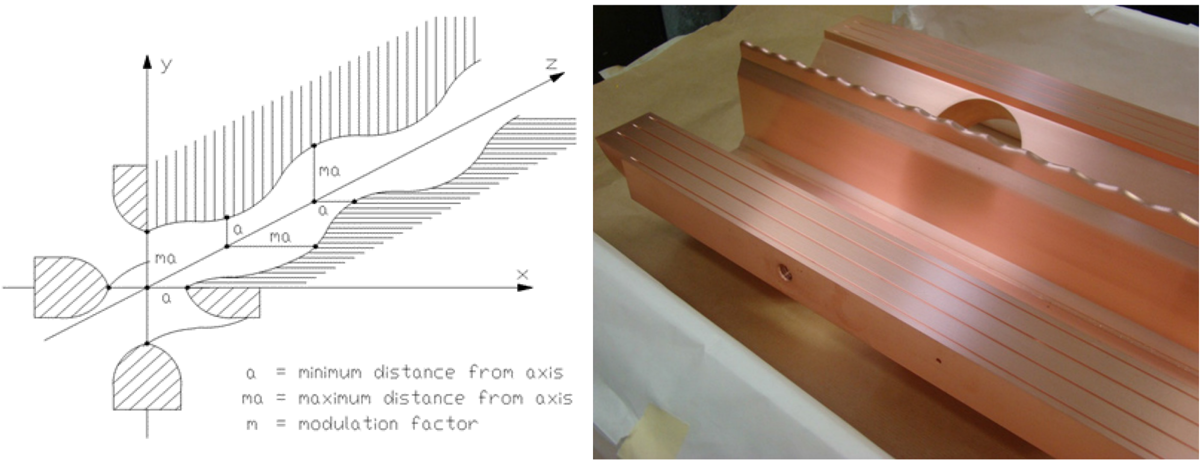

y
z
x
ma
a
ma
a
ma
a
a = minimum distance from axis
ma = maximum distance from axis
m = modulation factor

Fig. 6: Longitudinal modulation on the RFQ electrodes

2.3 Interdigital H structure

A structure used in all existing hadron-therapy facilities is the Interdigital H (IH) structure. A natural structure to follow the RFQ in low-current machines, it is composed of drift tubes alternately held by lateral stems, and sections including magnetic quadrupoles for the transverse focusing, as shown in Fig. 7.

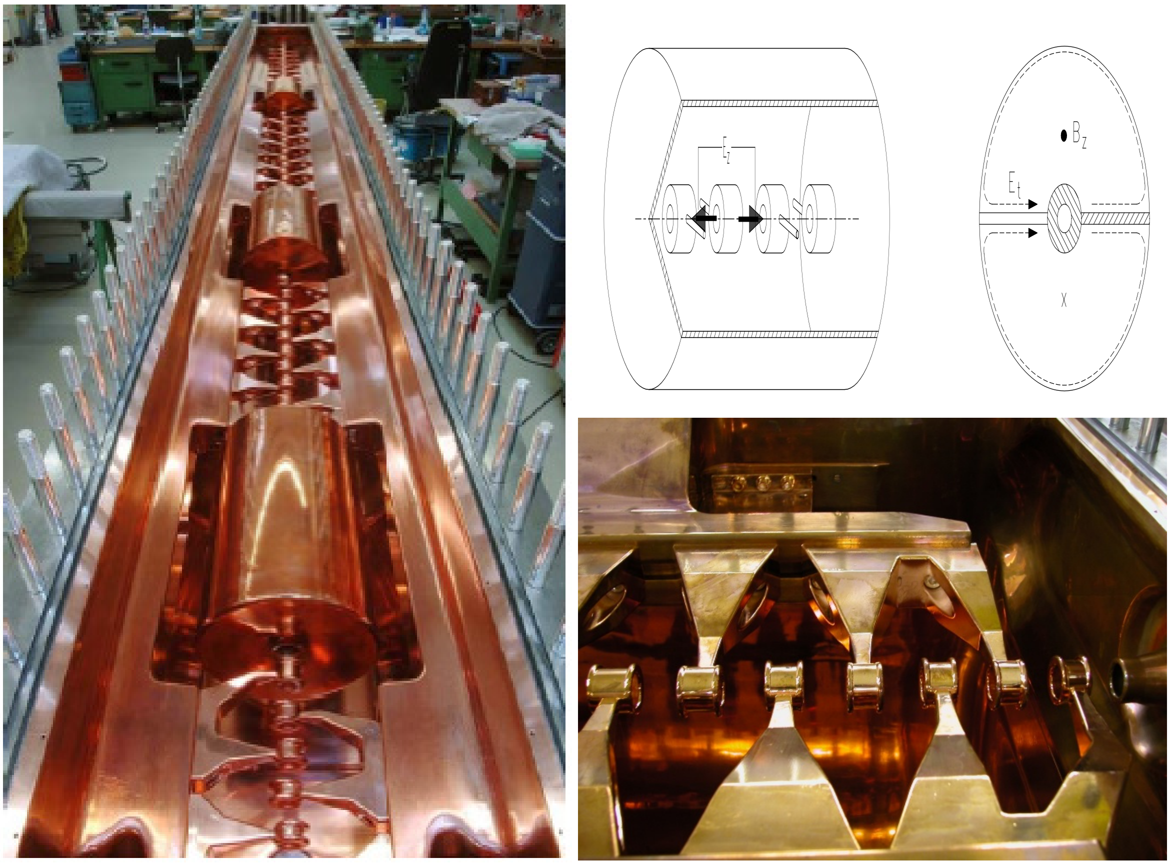


Fig. 7: The IH structure

The resonating mode of the cavity is a dipole mode, the TE_{110}. The cavity is equipped with thin drift tubes, and alternating the stems on each side of the drift tubes produces a field in the direction of propagation of the beam which accelerates the beam. The focusing is provided by quadrupole triplets located inside the tank in a dedicated section. The IH structure is very efficient in the low-beta region (β = 0.02 to 0.08) and at low frequency (up to 200 MHz). It is adapted for low-beta ion acceleration, ideal in a facility which accelerates both protons and carbon ions.

2.4 The Drift Tube Linac

The DTL or Alvarez Linac is a cavity resonating in the TM_{010} mode equipped with drift tubes, typically housing quadrupole lenses. A sketch is shown in Fig. 8. In Fig. 9, the field of the fundamental accelerating mode is shown. The length of each drift tube is adapted to the velocity of the beam as the particle has to spend half the RF period inside the drift and half inside the accelerating gap.

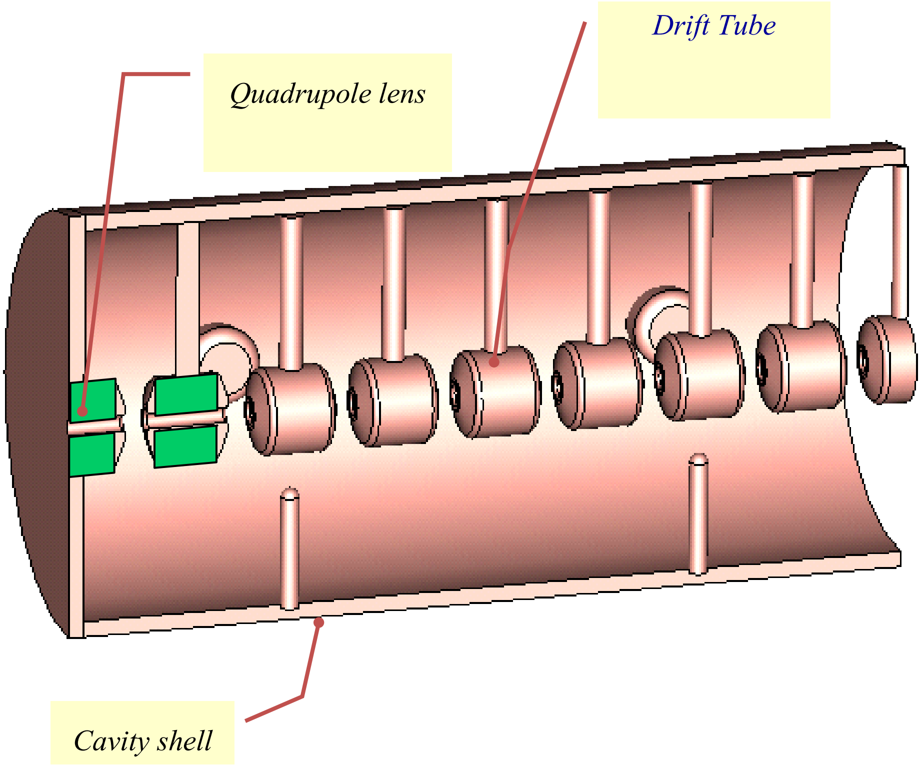


Fig. 8: Sketch of the DTL

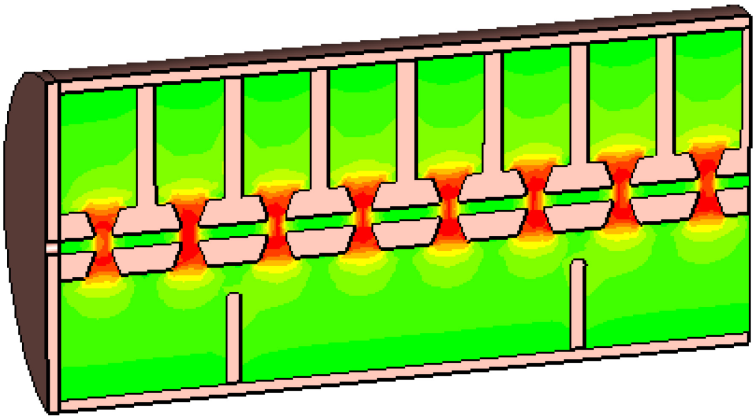


Fig. 9: Field distribution in a DTL

2.5 The Side-Coupled Linac

Side-Coupled or Cell-Coupled Linacs are strings of cavities operating in the in $\pi/2$ mode. Focusing elements are placed outside the cavities and they are completely independent. They are used when the beam is sufficiently energetic.

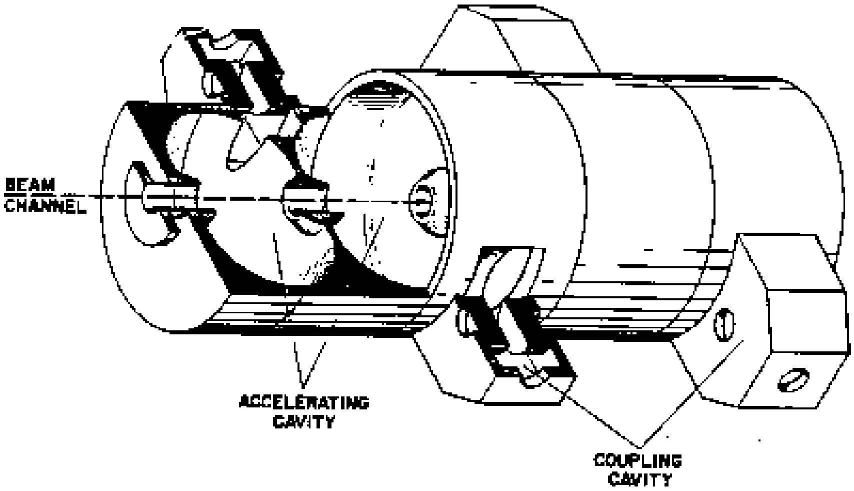


Fig. 10: Sketch of SCL

3 Quality factors

Each type of structure is adapted for a different energy range and for a different use. In the following, we describe a selection of quality factors to be considered when choosing a structure. A simple sketch of a cavity for illustration purpose s is shown in Fig. 11.

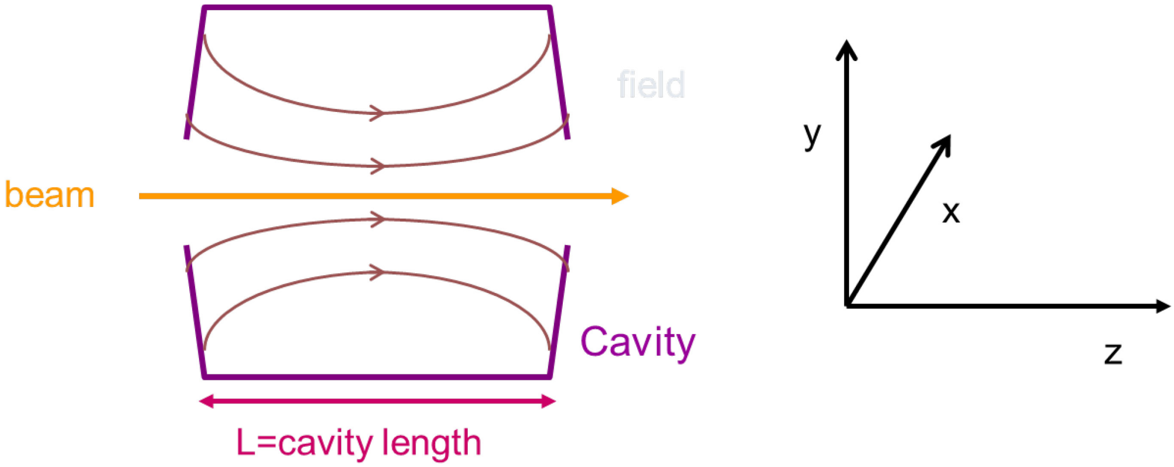


Fig. 11: Sketch of an RF cavity

3.1 Maximum field/average field

The average electric field (usually indicated with E_0 and measured in V/m) is the space average of the electric field along the direction of propagation of the beam in a given moment in time when the field is maximum.

If we write the electric field as

$$E(x,y,z,t) = E(x,y,z)\mathrm{e}^{-\mathrm{j}\omega t}, \tag{5}$$

and we take the space average

$$E_0 = \frac{1}{L}\int_0^L E_z(x=0, y=0, z)\mathrm{d}z \ , \tag{6}$$

we obtain a value that gives a measure of how much field is available for acceleration, which depends on the cavity shape, the resonating mode, and the frequency.

The limit to the field in a normal conducting cavity comes from sparking and a useful criterion was determined by W.D. Kilpatrick in 1950 [9], relating the maximum peak surface field to the frequency of a cavity, showing a dependence as in Fig. 12. The maximum surface peak field obtainable depends also on the surface quality and, nowadays, fields up to twice the Kilpatrick limit are obtained. Nevertheless, for a conservative design, a maximum surface field not exceeding 1.7 times the Kilpatrick field is generally advised.

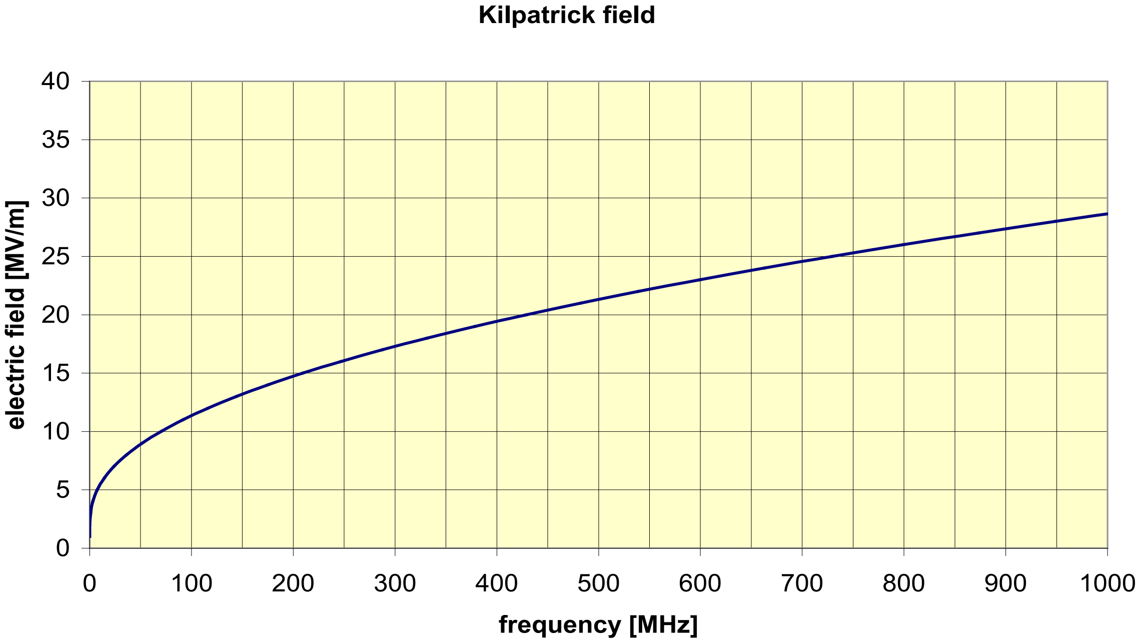


Fig. 12: The Kilpatrick criterion

3.2 Transit time factor

The transit time factor (indicated with *T* and dimensionless) is defined as the maximum energy gain per charge of a particle traversing a cavity over the average voltage of the cavity.

To better understand the meaning of this important parameter, let us write the field as

$$E_z\left(x, y, z, t\right) = E_z\left(x, y, z\right)\mathrm{e}^{-\mathrm{i}(\omega t)} \tag{7}$$

The energy gain of a particle entering the cavity on axis at phase φ is

$$\Delta W = \int_0^L qE_z(o, o, z)\mathrm{e}^{-\mathrm{i}(\omega t+\varphi)}\mathrm{d}z. \tag{8}$$

Assuming constant velocity through the cavity, we can relate position and time via

$$z = v \cdot t = \beta ct. \tag{9}$$

We can write the energy gain as

$$\Delta W = qE_0 LT\cos(\varphi) \tag{10}$$

and define the transit time factor as

$$T = \frac{\left| \int_0^L E_z(z) \mathrm{e}^{-\mathrm{j}\left(\frac{\omega z}{\beta c}\right)} \mathrm{d}z \right|}{\int_0^L E_z(z) \mathrm{d}z}. \tag{11}$$

T depends on the particle velocity and on the gap length. It does not depend on the absolute value of the field.

For a cylindrical pillbox resonating in the TM$_{010}$ mode, assuming a square-wave field distribution, the transit time factor turns out to be

$$T = \frac{\sin\left(\frac{\pi L}{\beta\lambda}\right)}{\left(\frac{\pi L}{\beta\lambda}\right)}. \tag{12}$$

This simple expression tells us that the length of the cavity must be adapted to the energy of the particle to be accelerated, otherwise the efficiency of acceleration can drop drastically.

3.3 Effective shunt impedance

The effective shunt impedance (indicated with ZTT and measured in Ω/m) is defined as the ratio of the average effective electric field squared (E_0T) to the power (P) per unit length (L) dissipated on the wall surface:

$$\mathrm{ZTT} = \frac{(E_0T)^2}{} \cdot \frac{L}{P}. \tag{13}$$

It is independent of the field level and cavity length. It depends on the cavity mode and geometry, and on the velocity of the particle to be accelerated. It is a measure of how much energy a charged particle can gain for 1 W of power when travelling over 1 m of structure. More on the shunt impedance can be found in these proceedings [10].

3.4 Comparison of structure

A summary table for the structures discussed in this lecture is reported below (Table 2).

Table 2: Comparison of structures discussed in this lecture

Type of structure	Ideal range of beta	Frequency	Effective gradient	
Radio frequency quadrupole	Low–0.05	40–400 MHz	1 MV/m (350 MHz)	Ions/protons
Interdigital H structure	0.02–0.08	40–400 MHz	4.5 MV/m (200 MHz)	Ions/protons
Drift Tube Linac	0.04–0.5	100–400 MHz	3.5 MV/m (350 MHz)	Ions/protons
Cell-Coupled Linac also called Side-Coupled Linac	Ideal Beta=1 But as low as beta 0.3	800–3000 MHz	20 MV/m (3000 MHz)	protons

References

[1] R. Wideröe, Über ein neues Prinzip zur Herstellung hoher Spannungen, *Archiv für Elektrotechnik* (in German) **21** (1928) 387. http://dx.doi.org/10.1007/BF01656341

[2] G. Ising, Prinzip Einer Methode Zur Herstellung Von Kanalstrahlen Hoher Voltzahl. *Arkiv för matematik, astronomi och fysik* (in German) **18** (1928) 1.

[3] E.O Lawrence and D. H. Sloan. Production of heavy high speed ions without the use of high voltages. *Phys. Rev.*, **38** (1931) 2021.

[4] L.W. Alvarez, Berkeley Proton Linear Accelerator, Radiation Laboratory, University of California, Berkeley (October 13, 1953).

[5] T. Wangler, *RF Linear Accelerators*, 2nd ed. (Wiley-VCH, Weinheim, 2008). http://dx.doi.org/10.1002/9783527623426

[6] R.H. Stokes and T.P. Wangler, Radio Frequency Quadrupoles and their applications, *Annual Review of Nuclear and Particle Science* **38** (1988) 97 1989. http://dx.doi.org/10.1146/annurev.ns.38.120188.000525

[7] K.R. Crandall, R.H. Stokes and T.P. Wangler, RF quadrupole beam dynamics design study, Proc. Linear Accelerator Conference, Montauk, 1979, p. 205.

[8] M. Weiss, Radio frequency quadrupole, in Proceedings of the CAS-CERN Accelerator School: Accelerator Physics, Aarhus, Denmark, 16—26 September 1986, edited by S. Turner, CERN-1987-010 (CERN, Geneva, 1987), pp. 196-230. http://dx.doi.org/10.5170/CERN-1987-010.196

[9] W.D. Kilpatrick, *Rev. Sci. Instrum.* **28**(10) (1957) 824. http://dx.doi.org/10.1063/1.1715731

[10] A. Degiovanni, these proceedings.

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Accelerating Structures

A. Degiovanni
CERN, Geneva, Switzerland

Abstract
In this lecture the basic concepts of electromagnetic waves in accelerating structures are discussed. After a short introduction on the propagation of electromagnetic waves and the concept of travelling-wave and standing-wave structures, the most important parameters and figures of merit used to describe the acceleration efficiency and the performance of accelerating structures in terms of power consumption are presented. The process of radio-frequency design optimization is also discussed. Finally, a review of several types of structure using different accelerating modes is given, with examples of interest for medical accelerators.

Keywords
Linac; hadron therapy; RF design; resonant mode; accelerating cavity.

1 Introduction

The aim of this lecture on ‘accelerating structures’ is to introduce the basic concepts and ideas related to the design and use of a radio-frequency (RF) accelerating system in linacs to an audience with backgrounds in several different fields. It does not cover the discussion of RF systems used for other types of machines (such as cyclotrons or synchrotrons, which are discussed in other lectures of these proceedings) and it does not pretend to be a detailed discussion of an extremely vast and complex field of accelerator physics and engineering. For the interested reader, a detailed introduction and description of this subject can be found in the Bibliography.

Accelerating structures are metallic resonant cavities used to accelerate beams of charged particles (i.e. to increase their energy). The acceleration of a charged beam is obtained by the interaction of the particles with the electromagnetic (EM) field confined in such void spaces delimited by metallic boundaries.

The force experienced by a particle of charge q passing through an electromagnetic field with a certain velocity v is described by the following equation:

$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = q\left(\vec{E} + \vec{v} \times \vec{B}\right). \tag{1}$$

The particle velocity $\boldsymbol{v} = v_x\,\mathbf{i} + v_y\,\mathbf{j} + v_z\,\mathbf{k}$ can be approximated—in the paraxial approximation (i.e. when $v_z >> v_x, v_y$)—by $\boldsymbol{v} = \beta c\,\mathbf{k}$. In this case, it is clear that only the electric field can change the particle momentum along the z-axis. So in order to accelerate a charged beam along the structure axis, a longitudinal component of the electric field E_z is needed.

2 Electromagnetic waves in RF structures

Electromagnetic waves are described by Maxwell’s equations. In free space, they propagate as spherical waves and the intensity of the electromagnetic field decays as $1/r$, where r is the distance from the point-like source to the measurement point. By using cylindrical and rectangular pipes, called waveguides, electromagnetic waves can propagate with very small losses.

https://doi.org/10.23730/CYRSP-2017-001.91

In a wave-guiding system, the electric and magnetic fields are solutions of Maxwell's equations propagating along the guiding direction (the *z* direction) and confined in the near vicinity of the guiding structure. They are mathematically described by the following equations:

$$E(x,y,z,t)=E(x,y)\mathrm{e}^{\mathrm{j}\omega t-\mathrm{j}k_z z}\,, \tag{2}$$

$$H(x,y,z,t)=H(x,y)\mathrm{e}^{\mathrm{j}\omega t-\mathrm{j}k_z z}\,. \tag{3}$$

They represent homogenous plane waves, characterized by the wave vector ***k***. The magnitude of the wave vector ***k*** is related to the EM-field angular frequency ω by the relation $k = \omega / c$. The wave vector ***k*** can be decomposed in its longitudinal and transverse components, respectively k_z and k_t.

The precise relationship between ω and k_z (generally called the *waveguide-propagation constant* or *wave number*) is called the *dispersion relation* and can be written as follows:

$$\omega=\sqrt{\omega_c^2+k_z^2c^2}\,. \tag{4}$$

The dispersion diagram (or Brillouin diagram) shows the frequencies of electromagnetic waves that can be transmitted through a hollow conductor and is obtained by plotting Eq. (4) in a ω–k_z scatter plot. For rectangular and circular waveguides the dispersion relation has a hyperbolic shape as shown in Fig. 1.

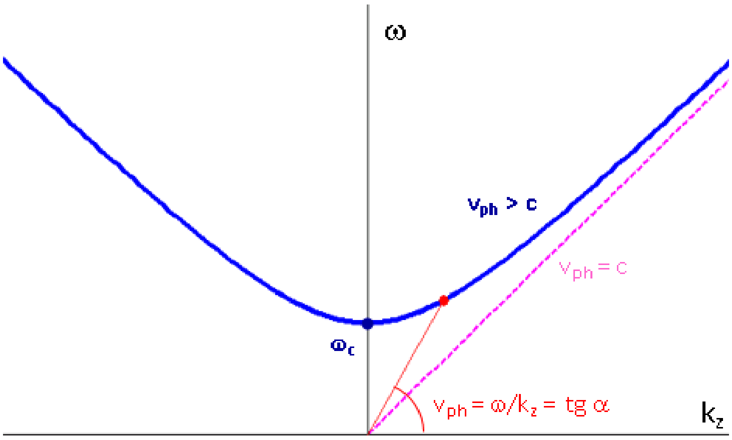


Fig. 1: Dispersion relation of a waveguide system

The term ω_c is the so-called *cut-off frequency*. Below the cut-off frequency, there is no propagation of the EM field. The boundary conditions for each waveguide type (i.e. the geometry of the inner section of the waveguide) force ω_c to take on only certain values. At each excitation frequency is an associated phase velocity, the velocity at which a certain phase travels in the waveguide.

The *phase velocity* is defined as the ratio between the excitation frequency ω (which is the angular frequency of the EM field once excited) and the waveguide-propagation constant k_z. This corresponds to the tangent of the angle between the line connecting the origin with the working point and the horizontal axis on a dispersion diagram:

$$v_{\mathrm{ph}}=\omega/k_z=\tan\alpha\,. \tag{5}$$

In this condition, to be synchronized all the time with an accelerating *E*-field, a particle travelling inside the waveguide has to travel at $v = v_{\mathrm{ph}} > c$. This is clearly not possible and it will be shown in Section 2.2 how the phase velocity of the EM field can be reduced.

Energy and information travel at the group velocity $v_\mathrm{g} = \mathrm{d}\omega / \mathrm{d}k$, which corresponds to the slope of the dispersion curve in the Brillouin diagram ($v_\mathrm{g} < c < v_\mathrm{ph}$).

2.1 Electromagnetic waves propagating modes

Solutions of Maxwell's equations can be classified in three families:

- TEM (Transverse Electric and Magnetic field) modes, where both electric and magnetic longitudinal components of the field are zero;
- TE (Transverse Electric) modes, where only the electric longitudinal component of the field is zero;
- TM (Transverse Magnetic) modes, where only the magnetic longitudinal component of the field is zero.

In bounded media (like a waveguide), the TEM modes are not allowed, because one of the field components must be in the direction of propagation to satisfy the boundary conditions. Only TM and TE modes are considered for propagation in bounded media. Subscripts are added to indicate different modes: TM_{mnp} and TE_{mnp}. The meaning of these subscripts is different for rectangular and circular cavities:

- in rectangular cavities, the subscripts m and n are the number of half waves in the x and y directions, respectively. The additional subscript p indicates the number of longitudinal half waves (i.e. in the z direction);
- in circular cavities, the subscript m indicates the number of full period variations of the field component in the azimuthal direction and n is the number of zeros of the axial field component in the radial direction. The subscript p indicates always the number of longitudinal half waves in the z direction.

As already noted, in order to accelerate particles, a mode with a *longitudinal E-field* component on an axis is needed. A TM mode can be used for such purpose. Therefore, the simplest propagating mode is the TM_{01}.

2.2 Disc-loaded accelerating structure

A way to reduce the phase velocity of the TM_{01} mode is to periodically add discs along the cylindrical waveguide, obtaining the so-called *disc-loaded waveguide*. The addition of discs inside the cylindrical waveguide, spaced by a distance l, induces *multiple reflections* between the discs. This action causes the dispersion curve to change (as shown in Fig. 2).

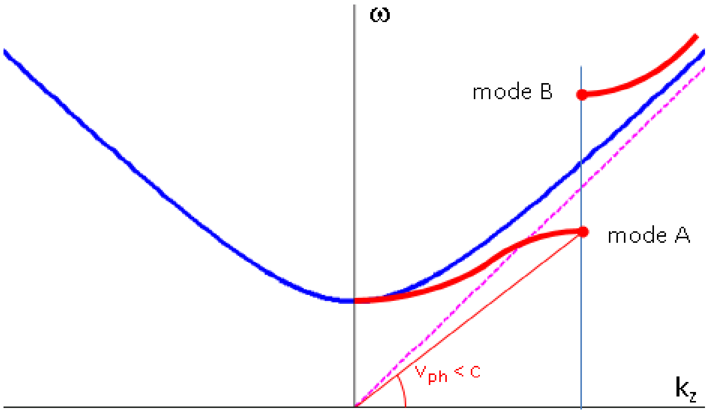


Fig. 2: Dispersion relation of a disc-loaded waveguide with stop band

In particular, for $k_z = 0$ or $k_z = \infty$ (corresponding to wavelength $\lambda_{ph} = \infty$ and $\lambda_{ph} = 0$), the wave does not see the effect of the discs. On the other hand, for waveguide wavelength $\lambda_{ph}= 2l$, the wave is confined between two discs and presents two polarizations (mode A and mode B in Fig. 2). These two modes have the same wavelength but different frequencies; the dispersion curve splits into two branches, separated by a *stop band*, where no propagation is allowed. Mode A can be used to accelerate particles since its phase velocity is smaller than the speed of light. On the other hand, mode B can propagate through the waveguide, but cannot be used to accelerate particles.

2.3 Travelling-wave and standing-wave structures

The disc-loaded waveguide is an example of a multi-cell structure. In such structures, a cell is defined by the space between two consecutive irises. The disc-loaded waveguide can be operated in two different conditions, i.e. as a travelling wave (TW) or a standing wave (SW).

In TW mode, the structure (or accelerating tube) has an input and output aperture (couplers) from which RF power can be, respectively, fed into the structure and extracted from it. The field propagates through each cell and the distance between the discs *l* determines the phase advance between each cell:

$$\Delta\varphi = \frac{l}{\beta\lambda}2\pi. \tag{6}$$

When designing a linac, Eq. (6) is used to fix the length of the cells based on the choice of the frequency of excitation, the speed of the particle that needs to be accelerated, and the mode of operation.

SW modes are generated by the sum of two waves travelling in opposite directions, adding up in the different cells (Fig. 3). The boundary conditions at both ends impose that the electric field must be perpendicular to the reflecting plane. This results in the fact that only some modes on the disc-loaded dispersion curve are allowed. The resonant modes are characterized by the phase advance between each cell and depend on the number of coupled cells.

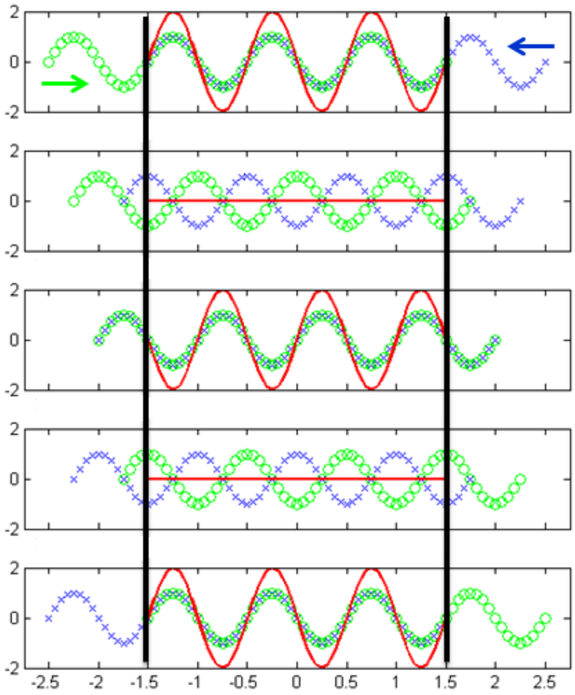


Fig. 3: Example of an SW pattern (red wave) obtained by superposition of two opposite travelling waves (blue and green curves). An SW cavity is obtained by inserting two metallic boundaries in correspondence of the two black lines.

Figure 4 shows the comparison of the dispersion curve of a multi-cell TW structure (left) and of a SW structure made of seven coupled resonators (right). As opposed to the case of TW structures, in SW structures there is no real power flow ($v_g = 0$) and, therefore, only one coupling port is needed for the excitation of the field inside the structure.

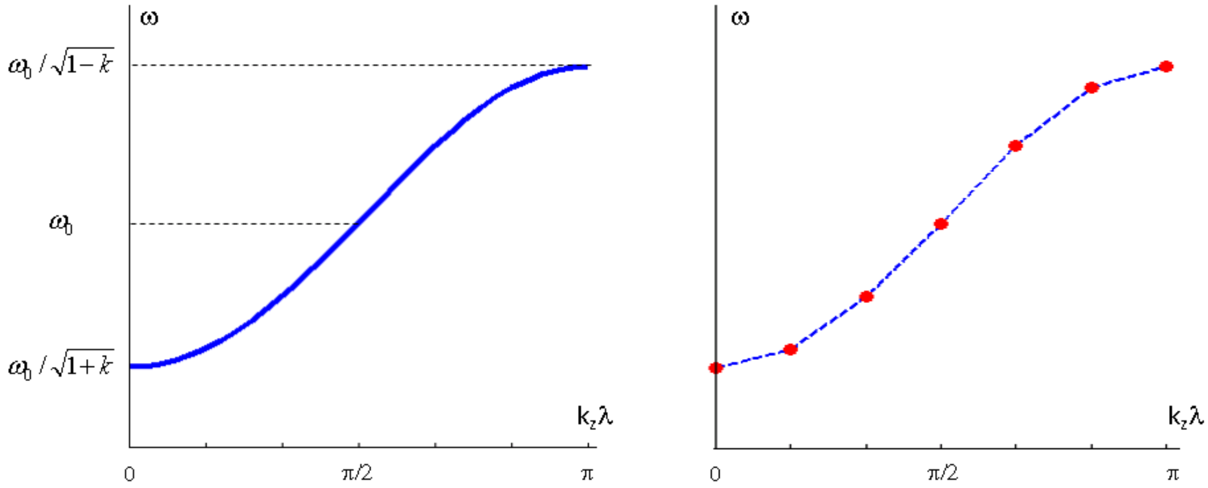


Fig. 4: Dispersion curve for coupled cavity oscillators with coupling factor *k*. In the case of TW mode (left), the dispersion curve is continuous (blue line), while for SW mode, only some modes are allowed (red dots).

The name of the resonant mode is typically given by the phase advance between consecutive cells. So, for example, a π-mode structure indicates a structure where the accelerating field is at maximum but in opposite directions in consecutive cells (it follows the scheme +1, –1, +1, –1, ...).

2.4 The pillbox cavity

A very important and instructive example of an accelerating structure is given by the so-called pillbox cavity, like the one shown in Fig. 5. It represents the simplest type of resonant electromagnetic structure. To use it as an accelerator cavity, one has to open two bore holes along the axis, introducing a small perturbation to the results obtained analytically [1].

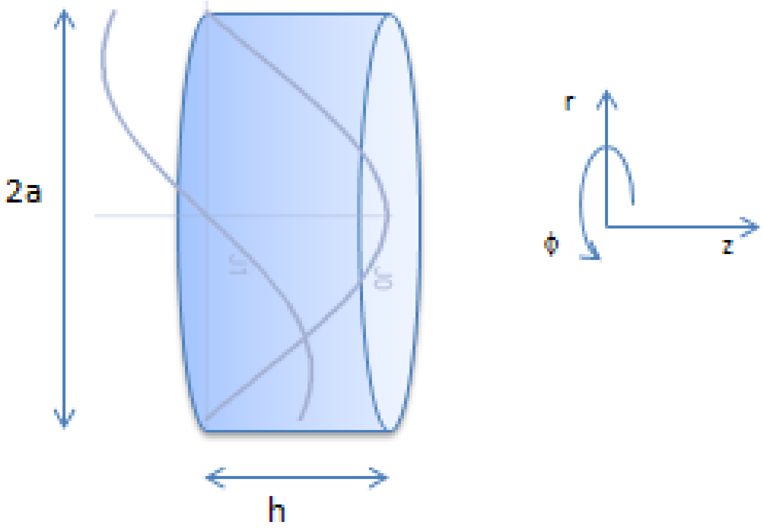


Fig. 5: Pillbox cavity, with longitudinal electric and azimuthal magnetic field components

For the simple geometry shown in Fig. 5, an analytical solution of the field components does exist and is given by the well-known Bessel's functions, J and J'. In particular, the longitudinal component of the electric field and the azimuthal component of the magnetic field are described by

$$E_z = E_0 J_m\left(k_{mn} r\right)\cos\left(m\varphi\right)\cos\left(\frac{p\pi z}{h}\right)\mathrm{e}^{\mathrm{j}\omega t}, \tag{7}$$

$$B_\varphi = -\mathrm{j}\omega\frac{a}{x_{mn}c^2}E_0 J_m^{'}\left(k_{mn} r\right)\cos\left(m\varphi\right)\cos\left(\frac{p\pi z}{h}\right)\mathrm{e}^{\mathrm{j}\omega t}. \tag{8}$$

From a physical point of view, one can think that the boundary conditions on the cavity walls ($E_\parallel = 0$) force the fields to exist only at certain quantized resonant frequencies. The resonant condition is fulfilled when the wavelength is such that an integer number of half-wavelengths fits along each direction. The resonant frequency is given by the dispersion relation

$$\omega^2 / c^2 = k_{mn}^2 + k_z^2 = \left(x_{mn} / a\right)^2 + (p\pi / h)^2. \tag{9}$$

For the TM$_{010}$ mode, one has m = 0, n = 1, and p = 0, so (7) and (8) can be written as

$$E_z = E_0 J_0\left(\frac{2.405}{a}r\right)\cos\left(\omega t\right), \tag{10}$$

$$B_\varphi = -\frac{E_0}{c}J_0^{'}\left(\frac{2.405}{a}r\right)\sin\left(\omega t\right). \tag{11}$$

3 Structure characteristics

Some of the fundamental properties of accelerating structures, which provide information about their performance, are described in the following. During the RF design phase, they are typically calculated using simulation tools that are able to solve Maxwell's equations inside complex geometrical structures. Some examples of such structures will be shown in Section 4. Depending on the application foreseen, one or more of such quantities are optimized by changing the geometry of the inner shape or by adding or removing features (for instance, by adding 'nose cones' close to the beam axis). It is important to underline that during the process of designing an accelerator, the RF optimization goes together with other considerations which can impose several constraints, such as beam dynamics, vacuum, beam instrumentations, and mechanical integration.

3.1 Energy gain in an accelerating gap

One of the most important quantities to evaluate for an accelerating cell of length L is the *axial accelerating voltage*, defined as

$$V_{acc} = \int_0^L E_z e^{j\frac{\omega}{\beta c}z}\, dz. \tag{12}$$

The exponential factor accounts for the variation of the field, while particles with velocity βc traverse the accelerating gap. The integral is taken over a distance L in which the electric field is confined. It is very useful to express the accelerating voltage in the form

$$V_{\mathrm{acc}} = V_0 T \tag{13}$$

where V_0 is the axial RF voltage or the voltage gain that a particle passing through a constant dc field equal to the field in the accelerating gap would experience, and T is the so-called transit-time factor.

The *transit time factor* is the ratio of the acceleration voltage to the (non-physical) voltage a particle with infinite velocity would see. It is defined as follows:

$$T = \frac{|V_{acc}|}{\left|\int E_z dz\right|} = \frac{\left|\int E_z e^{j\frac{\omega}{\beta c}z} dz\right|}{\left|\int E_z dz\right|} . \tag{14}$$

The transit time factor describes the reduction in the energy gain caused by the sinusoidal time variation of the field in the accelerating gap. Its value ranges between 0 and 1.

The *energy gain* of an arbitrary particle with charge q travelling through the accelerating gap is then given by

$$\Delta W = qV_0 T \cos\varphi . \tag{15}$$

Equation (15) is sometimes called Panofsky equation.

3.2 Acceleration efficiency figures of merit

Some of the power injected into the structure is lost or dissipated because of the electrical resistance in the cavity walls. The so-called *power losses*, or P_{loss}, are proportional to the *stored energy W*. In steady state, the total stored energy is

$$W = \iiint\limits_{V\mathrm{cavity}} \left(\frac{\varepsilon}{2}\left|\vec{E}\right|^2 + \frac{\mu}{2}\left|\vec{H}\right|^2 \right) \mathrm{d}V . \tag{16}$$

The energy in the cavity is stored in the electric and magnetic field. Since E and H are 90° out of phase, the stored energy continuously swaps from electric energy to magnetic energy. The (imaginary part of the) Poynting vector describes this energy flux.

The ohmic losses can be evaluated from the surface resistance R_{s} and the current density J_{s} as

$$\frac{dP}{dA} = \frac{1}{2} R_{\mathrm{s}} \left|J_{\mathrm{s}}\right|^2 . \tag{17}$$

The *surface resistance* is inversely proportional to the conductivity of the material σ and to the so-called skin depth δ, which represents the depth after which the field is attenuated by a factor 1/e:

$$R_{\mathrm{s}} = \frac{1}{\sigma\delta} . \tag{18}$$

The value of the skin depth depends on the conductor material properties and scales as the square root of the inverse of the frequency:

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} . \tag{19}$$

The cavity *quality factor Q* is defined as the ratio

$$Q = \frac{\omega_0 W}{P_{\mathrm{loss}}} = \frac{\omega_0}{\Delta\omega} . \tag{20}$$

It describes the number of cycles needed to fill up or empty a structure at the resonant frequency of ω_0. The ratio P_{loss} / W can also be identified as the full width half maximum of the resonance peak $\Delta\omega$. High-Q structures show, therefore, narrow band resonances and are thus typically very sensitive to frequency drifts due to temperature changes. It is important to notice that Q is proportional to the stored energy.

The relation between gap voltage and power is characterized by the so-called *shunt impedance*:

$$R = \frac{V_0^2}{P_{\text{loss}}} . \tag{21}$$

By defining the average axial electric field over the cavity length L as $E_0 = V_0 / L$, it is possible to introduce the *effective shunt impedance per unit length*:

$$ZT^2 = \frac{RT^2}{L} = \frac{E_0^2 T^2}{P_{\text{loss}} / L} . \tag{22}$$

Physically, it measures how well the RF power is concentrated in the useful region for acceleration. It is very important to underline that ZT^2 is *independent of the field level and cavity length*; it depends only on the cavity mode (frequency) and geometry (shape).

Another important quantity used to describe the structure efficiency is the ratio R/Q:

$$\frac{R}{Q} = \frac{\dfrac{V_0^2}{P_{\text{loss}}}}{\dfrac{\omega_0 W}{P_{\text{loss}}}} = \frac{V_0^2}{\omega_0 W} . \tag{23}$$

This quantity represents the proportionality constant between the square of the acceleration voltage and the stored energy. It is independent of cavity losses (it only depends on the geometry).

3.3 Field limiting quantities

High-gradient operation of linear accelerating structures is limited by a series of phenomena related to electrical discharges which are typically referred to as vacuum arcs or breakdowns. Although no theory or simulation method can predict breakdown performance of accelerating structures or high-power RF components during their technical design, a certain amount of experimental studies of the phenomena have been carried on in the past decades, mostly related to the development of high-gradient linear collider projects. From this experience, a series of scaling laws for high-gradient limiting quantities in terms of accelerating gradient and RF frequency has been discovered.

A quantity introduced in the late 1950s, but still used as a guideline for the design of structures at frequencies lower than 1 GHz, is the so-called Kilpatrik field limit. The Kilpatrik's criterion [2] for the determination of a threshold surface electric field defining the border between 'no vacuum sparking' and 'possible vacuum sparking' is described by

$$f = 1.64 \cdot E_{\text{K}}^2 \cdot \mathrm{e}^{-\frac{8.5}{E_{\text{K}}}} , \tag{24}$$

where f [MHz] is the RF frequency and E_{K} [MV/m] is the Kilpatrick limit for the surface electric field.

The value of the surface electric field on the structure wall is proportional to the average gradient E_0 in the structure and one can define the peak to average field ratio $E_{\text{S,max}}/E_0$ of a cavity as the ratio between the maximum surface electric field $E_{\text{S,max}}$ and the average axial electric field E_0. The typical value for electron linacs is 2, while for proton and hadron linacs, depending on the design of the cell, this ratio can take values up to 4–5.

For a long time, the surface electric field has been considered to be the only limiting quantity for high-gradient operations. More recently, experimental evidence supports the idea that a combination of electric and magnetic fields at the surface correlates well with the measured breakdown probability [3] and that the stress induced by the surface field on the crystalline structure of the conductive material could be used to explain the breakdown mechanisms [4].

3.4 Accelerating cavities optimization

When designing the RF structures of a linac, it is important to keep in mind the final goals and objectives, and the constraints.

For example, for the case of a linac for proton therapy, important design goals are the actual energy gain and the final dose rate. Such goals should be achieved by respecting some constraints. For example, the size of the machine should fit in an assigned space, or the power consumption should not exceed a certain amount. Other limitations can come from technical and very often economic considerations, such as, for example, a limited number of RF power sources to reduce the total cost, a minimum repetition rate (to obtain the desired dose rate) and a maximum repetition rate (due to thermal considerations and to limitation in the RF power systems), beam dynamics considerations of drift lengths between structures and space for focusing elements along the linac, etc. During the design optimization, the accelerator physicist, taking into account the goals and constraints, can propose design solutions that maximize some critical parameters of the machine.

An example is given by the optimization of the effective shunt impedance per unit length defined in (22). In fact, combining Eqs. (22) and (15), one can write the following relation:

$$\Delta W \propto \sqrt{ZT^2 \cdot P \cdot L}\ . \tag{25}$$

Equation (25) reveals that to achieve a certain energy gain one can reduce the power consumption of the linac P and/or the length of the machine L by increasing the effective shunt impedance. So, the maximization of the value of ZT^2 during the design phase results in a reduction of the cost of the machine. In order to increase the ZT^2, one can change the structure geometry.

For example, let us consider the inner shape of an accelerating structure with nose cones like the one sketched in Fig. 6. By changing the radius of the bore hole (*Rb*), one affects the concentration of the field on axis and therefore, the effective shunt impedance. Generally speaking, by reducing the bore hole aperture, one can increase the effective shunt impedance of the structure. This action, on the other hand, would result in the reduction of beam transmission due to the smaller aperture and therefore, would be in conflict with the constraints given by beam dynamics considerations.

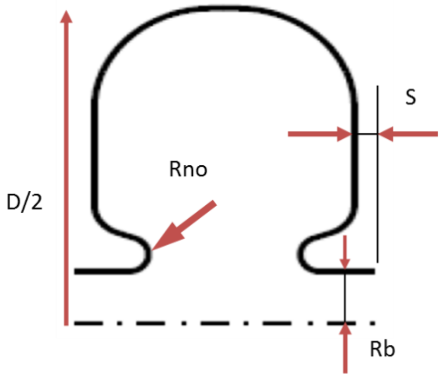


Fig. 6: Sketch of cell profile with parameters that can be used for the tuning and optimization

The results obtained during the optimization of the design for a structure resonating at 3 GHz, and designed for $\beta = 0.4$, are reported in Fig. 7. A reduction of 1 mm in the bore radius results in an increase of 10% of ZT^2, but the value of the bore radius should not go below a certain limit dictated by beam dynamics considerations. The same is true for a change of the septum thickness S. For a given cell length, by reducing the thickness of the walls between two cells, more space is left to the electric field and therefore, the ZT^2 increases. On the other hand, walls too thin would result in mechanical instabilities and problems in the cooling of the cells.

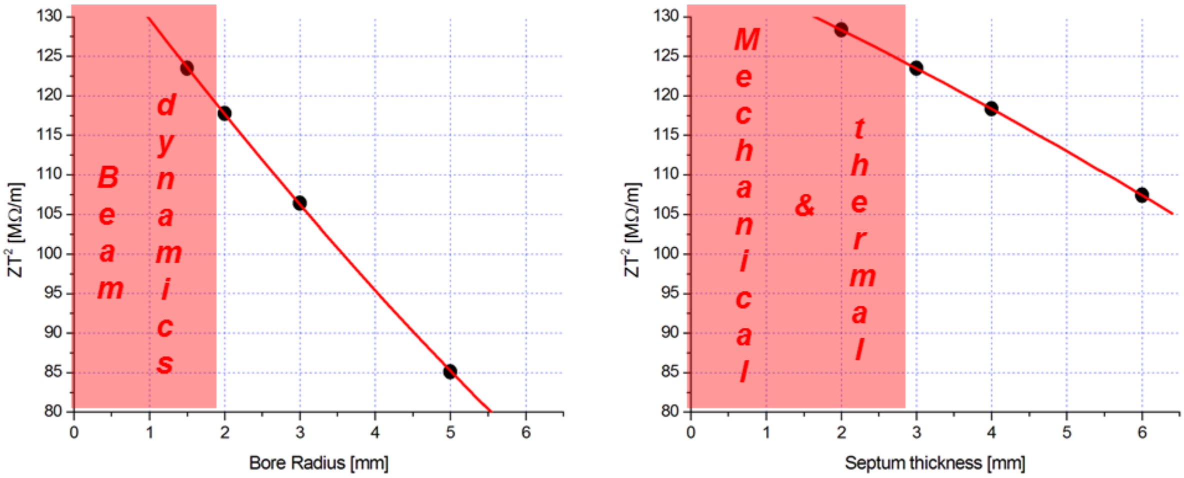


Fig. 7: Example of influence of the geometrical parameters on the effective shunt impedance

4 Example of structures

A large variety of structures are used in accelerator facilities worldwide. Typically, the structure design is adapted to the properties of the beam that is accelerated. A big difference exists between electron and proton or hadron linacs. In fact, due to the mass of 0.511 MeV/c^2, electron beams can be considered ultra-relativistic ($\beta \approx 1$)—already at a kinetic energy of a few MeV. On the contrary, linacs are used for protons and hadrons up to relatively low energies of a few hundred MeV/u, corresponding to $\beta \approx 0$–0.9. Since the accelerating efficiency is strongly dependent on the type of structure used and on the beam velocity, for proton and hadron linacs several structures are used in sequence to adapt to the increasing particle velocity, while electron linacs are typically made of the same type of structures.

Figure 8 shows the comparison of the values of effective shunt impedance calculated for eight different designs obtained in a joint study funded by EU called HIPPI (High Intensity Pulsed Power Injectors) [5]. The structures considered belong to two frequency ranges: 324–352 MHz, and its second harmonic 648–704 MHz. Similar results scaled towards larger shunt impedance are obtained with higher frequencies.

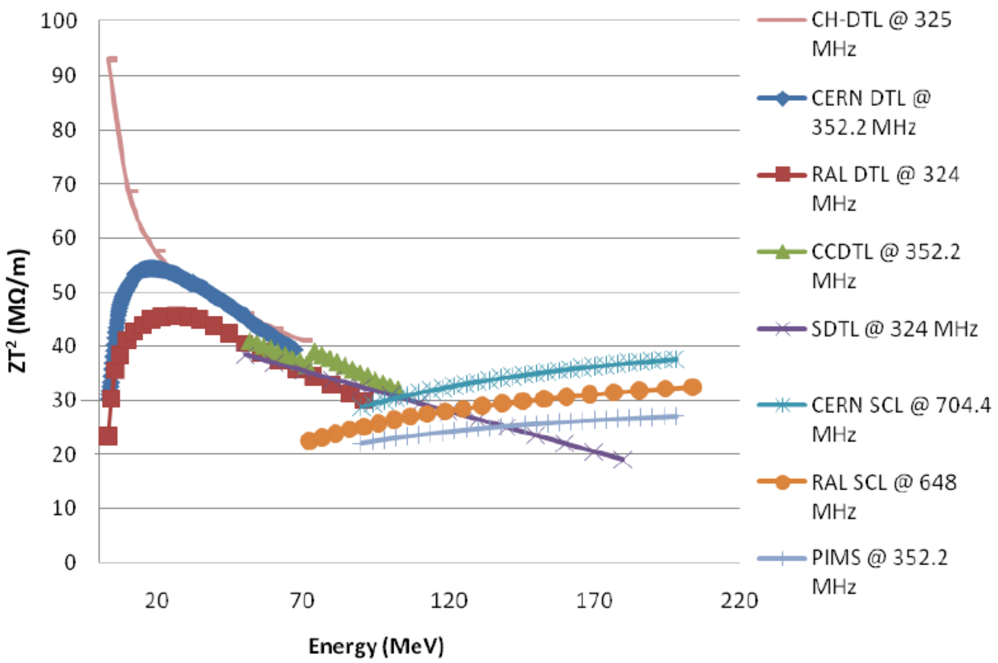


Fig. 8: Comparison of effective shunt impedance per unit length, for different types of structure, as a function of beam energy (taken from [5]).

Examples of structures using the TE mode are the Radio Frequency Quadrupole (RFQ) and the Interdigital and Crossbar H-mode (IH and CH) structures. Other structures using the TM mode are the Drift Tube Linac (DTL), the Cell Coupled Drift Tube Linac (CCDTL), the Cell Coupled Linac (CCL), and the elliptical cavities. In this section, several types of structures adapted to the acceleration of protons at different velocities are shown.

The curves of Fig. 8 present a strong dependence on beam energy. In the low-energy region, TE-mode-like structures are more efficient (CH structure), but above 20 MeV, their efficiency is comparable with that of TM-mode structures operating in 0-mode. Above 100 MeV, π-mode structures become more suitable.

4.1 TE-mode-like structures

The TE-mode-like structures (also called H-mode structures) are used at extremely low values of β. In this regime, focusing and bunching effects are more important than actual acceleration (see also the lecture on ‘Beam dynamics’ in these proceedings). The use of TE modes seems to be in contrast with what was discussed in Section 1, that only longitudinal electric field components are useful for acceleration. Indeed, pure TE modes have $E_z = 0$. In order to use them for acceleration, the field is forced to the longitudinal plane either by a longitudinal modulation (in the case of the RFQ) or by adding stems and drift tubes (in the IH or CH structures). Traditionally, H-mode structures have also been called Wideröe-linac structures.

4.1.1 The Radio Frequency Quadrupole

The RFQ concept was first proposed by Kapchinskiy and Tepliakov in 1969 [6] and then further developed both in the Soviet Union and in the USA. In this type of structure, a quadrupolar electric field pattern is obtained by the presence of four poles which concentrate the electric field lines as shown in Fig. 9.

There are two types of RFQ construction geometries: the four vanes and the four rods. The typical frequencies used go from 10 to 400 MHz. Machining tolerances and related frequency errors limit the scaling up in frequency of such accelerators. The RFQ is placed just after the ion source with three main functions:

1 bunch the beam adiabatically, so as to prepare it for the next stage of acceleration with minimum losses;

2 focus the beam transversally by means of the electric quadrupolar field (extremely important at low energy when space charge forces are stronger);

3 accelerate the beam up to the few MeV required for injection into the next acceleration stage (typically a DTL).

Typical transmissions obtained with the RFQ are of 90–95%, but the real estate gradient is very low (typically no more than 1–2 MeV/m).

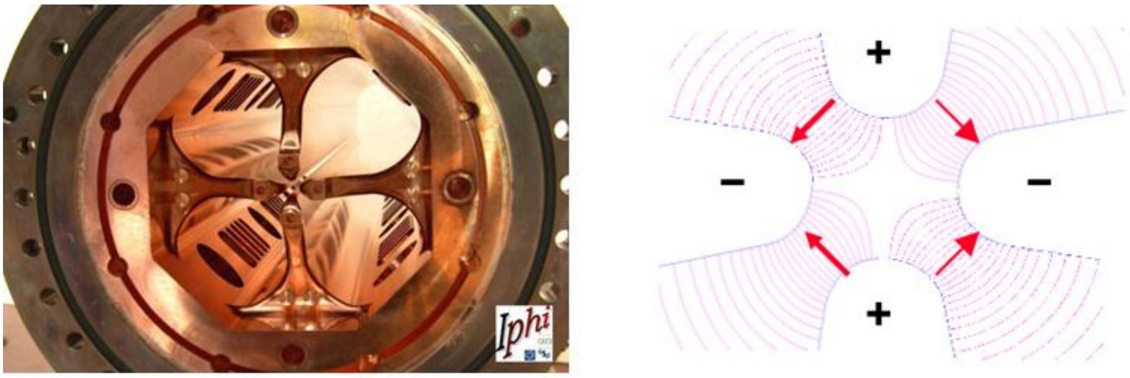


Fig. 9: An example of four vanes RFQ (left) with detail of the electric field lines close to the tips (right). The quadrupole field profile produces a focusing effect.

A novel design has been proposed at CERN for a high-frequency RFQ for medical application [7]. The RFQ is designed for an RF frequency of 750 MHz (almost double the maximum frequency used up to now), which makes it very compact compared to other similar machines (only 2 m for 5 MeV energy). It will be used as injector for proton-therapy linacs, but also its use as a booster for radio-isotope production is foreseen. A prototype is now being built at CERN which will be tested at the beginning of 2016.

4.1.2 IH and CH structures

After the first few MeV of acceleration, Interdigit and Crossbar H-mode structures (IH and CH) are more efficient. From the RF point of view, they are using, respectively, the TE_{110} and TE_{210} mode. However, the addition of the stems and the drift tubes forces the electric field (which in pure TE mode does not have a longitudinal component) along the beam axis, as is visible from the sketch in Fig. 10.

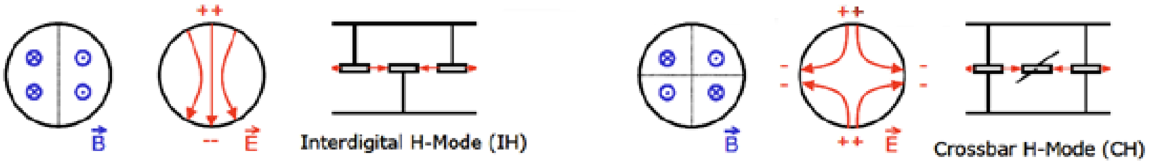


Fig. 10: Schematic representation of the mode excited in IH (left) and CH (right) structures (courtesy of F. Gerigk-CERN).

The H-mode structures have a good effective shunt impedance per unit length ZT^2 for β in the range 0.02–0.08 and typically work at low frequencies (about 200 MHz). They are particularly well suited for low-intensity beams and they are used, for instance, in linac injectors for synchrotrons used in hadron therapy. For example, Fig. 11(left) shows the inner geometry of the IH structure used at CNAO (Pavia-Italy). The stems holding the drift tubes are placed alternately on two sides of the structure. The place reserved for Permanent Magnetic Quadrupole (PMQ) triplets is also visible. This machine is part of the injector chain and accelerates the particle from 400 keV/u to 7 MeV/u. In Fig. 11(right), the design of a CH structure, working at 3 GHz, to be used as proton booster from 15 to 66 MeV is shown. This structure, named CLUSTER (Coupled-cavity Linac USing Transverse Electric Radial field) [8], has a higher shunt impedance than other structures in this energy range.

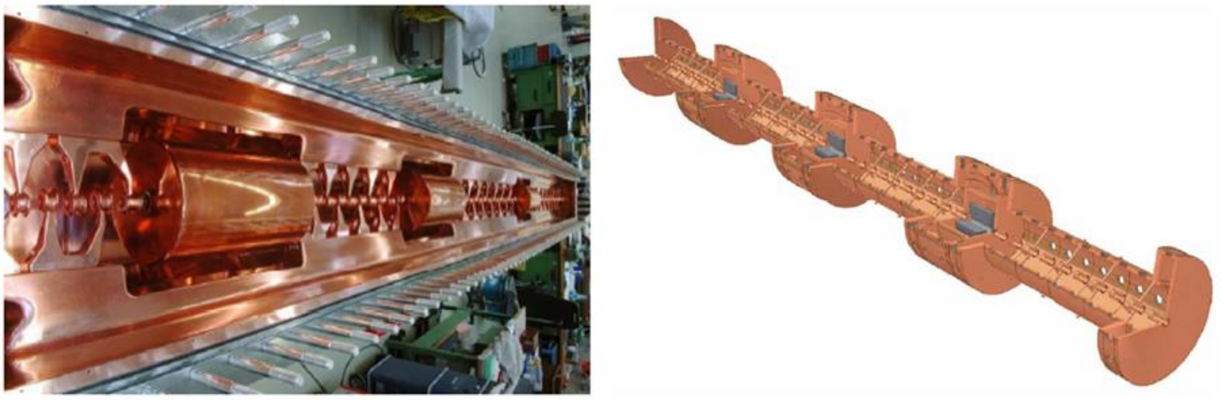


Fig. 11: Examples of IH (left) and CH (right) structures

In H-mode structures, the small drift tubes do not allow the insertion of quadrupoles. The focusing is provided by PMQ placed between tanks. A special beam dynamics approach is typically used (see also the lecture on ‘Beam dynamics in linacs’ in these proceedings).

4.2 TM-mode-like structures

The TM-mode structures, sometimes referred to as E-mode structures, are mostly used at intermediate β values ranging from a few MeV up to several hundred MeV.

4.2.1 The Drift Tube Linac

The DTL, also known as the Alvarez linac, is one of the simplest and oldest linac structures in use for the acceleration of protons and hadrons. It is a SW linac structure, used typically for β in the range 0.1–0.4 and with RF frequencies of 20–400 MHz.

The name comes from the presence of drift tubes, held by stems (as shown in Fig. 12). In a certain sense, the inner geometry is similar to an IH structure, but the DTL is working in the TM_{010} mode. The drift tubes are used to shield the particles from the RF field when this is in the decelerating phase. The length of the drift is such that when the particles are in the gaps between the drift tubes, they always see an accelerating field. It can be useful to consider the DTL as a multi-cell structure. The coupling between consecutive cells is at maximum, due to the fact that there are no walls with slots, so that the separation between cells is fully open. The 0 mode allows a long enough cell ($l = \beta\lambda$) to house focusing quadrupoles inside the drift tubes.

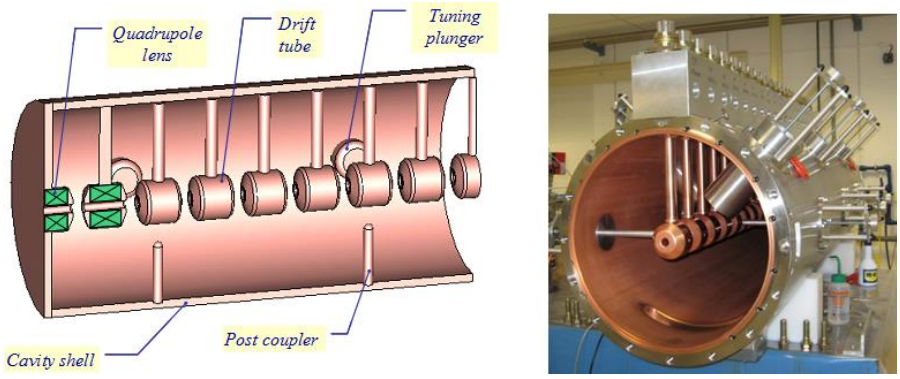


Fig. 12: Examples of the DTL module for LINAC4 at CERN (courtesy of M. Vretenar – CERN)

4.2.2 The Cell Coupled Linac

A CCL structure is made of a linear array of resonant cavities coupled together into a multi-cell structure.

In a normal CCL structure, the maximum frequency stability is obtained with the π/2 mode. This resonant mode is the least sensitive to frequency errors because it shows the largest neighbour mode separation. The synchronicity condition for acceleration is fulfilled for cavity length $l = \beta\lambda/4$. In this configuration (like the one shown in Fig. 13), each cavity has the same length, but only half of the cavities can be used for acceleration. The other half consist of unexcited cells. Consequently, the shunt impedance is lower than for a π-mode structure.

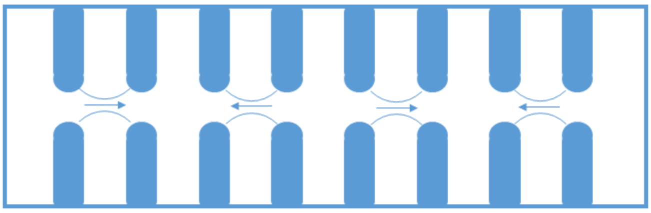


Fig. 13: Sketch of a CCL π/2-mode periodic structure

In order to keep the advantages of the π/2 mode in terms of frequency stability, and to achieve high shunt impedance, one can create a bi-periodic array of oscillators. For instance, one can reduce the space occupied axially by the unexcited cells—also called Coupling Cells (CCs)—and optimize the excited cells—or accelerating cells (ACs)—for higher shunt impedance. This is shown in Fig. 14 (left). This solution is typically called an *on-axis-coupled structure*.

A second type of geometry is the *side-coupled structure*, where the CCs are moved off-axis leaving all the axial space available for the ACs (Fig. 14 (right)). In this case the length of each AC is given by $l = \beta\lambda/2$.

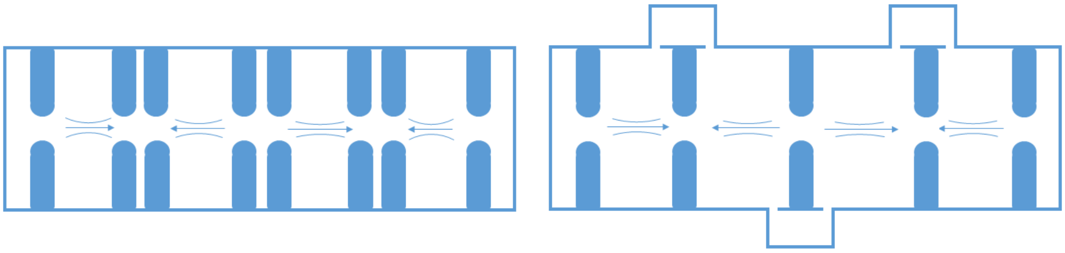


Fig. 14: Sketch of biperiodic CCL structures: on-axis-coupled structure (left) and side-coupled structure (right). The second case maximizes the space used for acceleration.

There are several types of CCL structures that are used to accelerate protons with β between 0.4 and 0.9. Depending on the application each of them has advantages and disadvantages. An example of a 3 GHz Side Coupled Linac (SCL) structure designed to accelerate protons from 62 to 74 MeV is shown in Fig. 15. This module, called LIBO, was designed, constructed, and tested by TERA Foundation, in collaboration with CERN and INFN, and was the first 3 GHz linac to be used for the acceleration of a proton beam.

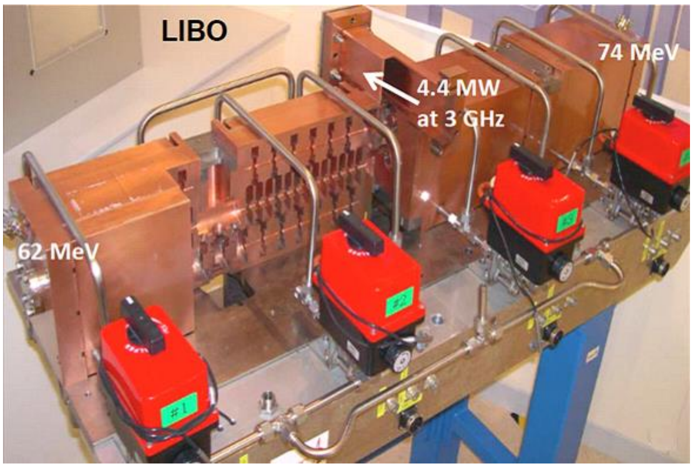
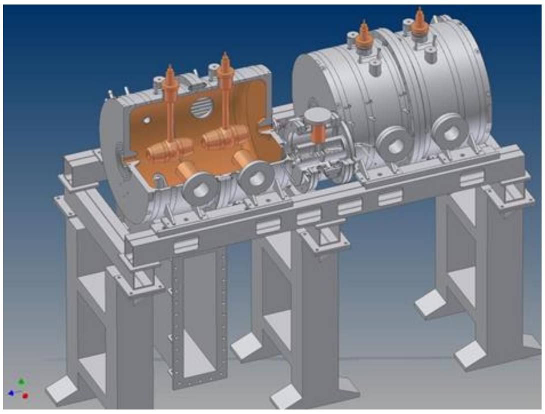


Fig. 15: The LIBO-module prototype is an example of 3 GHz SCL for proton therapy applications

4.2.3 The Cell Coupled Drift Tube Linac

A particular case of use of the CCL concept is given by the CCDTL. This is a kind of hybrid between a 0-mode DTL structure and a $\pi/2$-mode structure. In this case, short DTL tanks operating in 0 mode are coupled together via coupling cells, like in a CCL structure. The advantage of this structure, which is convenient for energies between 20 MeV and 100 MeV, is the longitudinal field stability given by the $\pi/2$ mode and the possibility of placing the focusing elements between the DTL tanks and not inside the drift tubes that, at low energies, are too short.

Two examples of CCDTL are shown in Fig. 16. Both are Side Coupled Drift Tube Linacs (SCDTL) due to the fact that the coupling cells are moved off-axis (on the side). On the left, a 352 MHz SCDTL module for the LINAC4 project at CERN, with two DTL tanks and one side coupling cell. On the right, the first module of a 3 GHz SCDTL designed by ENEA (Frascati-Italy) for a proton-therapy linac project.



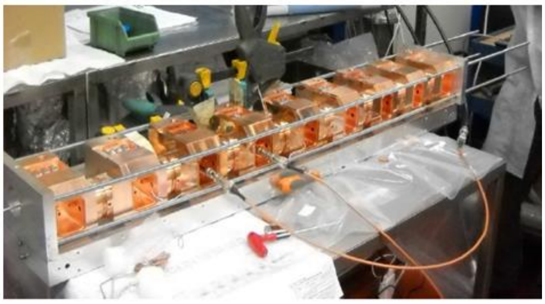


Fig. 16: Two examples of CCDTL structures: on the left is one of the 352 MHz SCDTL modules for LINAC4 (courtesy of M. Vretenar – CERN), and on the right is a 3 GHz SCDTL designed by ENEA for a proton therapy linac (courtesy of L. Picardi – ENEA).

4.3 Summary of linac structures

The following table (Table 1) summarizes the properties of the different linac structures discussed in this section for proton-therapy applications. This is given only as a general reference.

Table 1: Summary table of linac structures for protons

Structure type	Family	Mode	Typical energy [MeV]
Radio Frequency Quadrupole	TE	TE_{21}+vanes	0–2
Interdigit H-mode (Wideroe)	TE	TE_{110}	0.4–10
Crossbar H-mode (Wideroe)	TE	TE_{210}	0.4–10
Drift Tube Linac (Alvarez)	TM	0 mode	4–80
Cell Coupled Linac	TM	$\pi/2$ mode	80–400
Cell Coupled Drift Tube Linac	TM	Hybrid	20–100

5 RF linacs for medical applications

More than 15,000 electron linacs are used daily for radiotherapy treatment. They represent about 50% of all the accelerators with energies greater than 1 MeV and their number is continuously growing [9]. Electron linacs for radiotherapy can be considered industrial products.

A linac radiotherapy unit includes not only the accelerator, but is composed of the RF power source, the supporting structure that usually can rotate around the patient (gantry), the couch, the alignment system, and the control system, typically integrated with the treatment planning system. For what constitutes the accelerator part, electron tubes can be either SW or TW. For example, some manufacturers propose SCL type structures and others propose disc loaded waveguides. As RF sources, magnetrons or klystrons in the 5 MW range are used. The typical frequency used is 2.856 GHz in the USA and 2.998 GHz in Europe.

Recently, a CERN spin-off company called A.D.A.M. (Application of Detector and Accelerators to Medicine) has started the design and production of a high-frequency (3 GHz) linac for proton therapy, with the idea to commercialize it. As is shown in Fig. 17, a modular approach has been chosen, with a sequence of three types of linac structures discussed in the previous section: an RFQ, followed by an SCDTL, and an SCL.

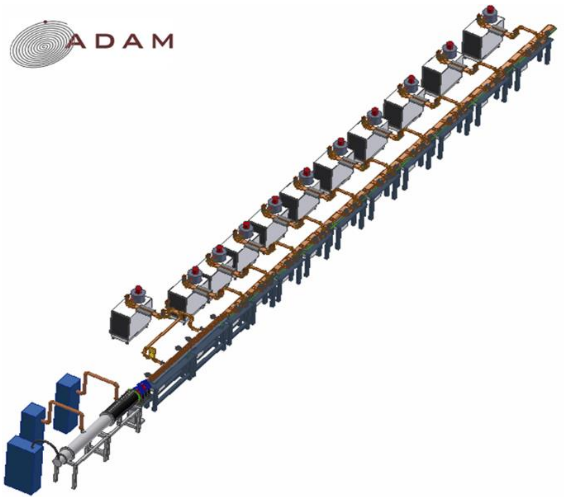


Fig. 17: Layout of a high-frequency linac for proton therapy, proposed by the CERN spin-off company A.D.A.M. (courtesy of A.D.A.M.).

Acknowledgements

Most of the material for this lecture comes from previous lectures of M. Vretenar, E. Jensen and F. Gerigk. My thanks to all of them.

References

[1] J. Gao, *Nucl. Instrum. Methods Phys. Res.* A **311**(3) (1992) 437. http://dx.doi.org/10.1016/0168-9002(92)90638-K

[2] W.D. Kilpatrick, *Rev. Sci. Instrum.* **28**(10) (1957) 824. http://dx.doi.org/10.1063/1.1715731

[3] A. Grudiev, S. Calatroni and W. Wuensch, *Phys. Rev. ST Accel. Beams* **12** (2009) 102001. http://dx.doi.org/10.1103/PhysRevSTAB.12.102001

[4] F. Djurabekova and K. Nordlund, *Phys. Rev. ST Accel. Beams* **15** (2012) 7 071002. http://dx.doi.org/10.1103/PhysRevSTAB.15.071002

[5] C. Plostinar (Ed.), Comparative Assessment of HIPPI Normal Conducting Structures, CARE-Report- 2008-071-HIPPI (2008).

[6] I. M. Kapchinskiy and V. A. Teplyakov, The linear accelerator with spatially uniform strong focusing, Preprint ITEP **673** (1969). http://www.slac.stanford.edu/pubs/slactrans/trans01/slac-trans-0099.pdf, last accessed 25 January 2016.

[7] M. Vretenar *et al.*, A compact high-frequency RFQ for medical applications, Proc. LINAC14 (2014) 935.

[8] U. Amaldi *et al.*, *Nucl. Instrum. Methods Phys. Res.* A **579**(3) (2007) 924. http://dx.doi.org/10.1016/j.nima.2007.05.208

[9] W. Maciszewski and W. Scharf, *Phys. Med.* **20** (2004) 137. http://dx.doi.org/10.1142/9789812702708_0060

Bibliography

P. Lapostolle, Proton linear accelerators: a theoretical and historical introduction, LA-11601-MS (1989), http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/20/072/20072193.pdf, last accessed 10 October 2015.

P. Lapostolle and M. Weiss, Formulae and procedures useful for the design of linear accelerators, CERN-PS-2000-001 (2000), https://cds.cern.ch/record/428133/files/ps-2000-001.pdf, last accessed 10 October 2015.T.P.

S. J. Orfanidis, *Electromagnetic Waves and Antennas* (2014), http://www.ece.rutgers.edu/~orfanidi/ewa/, last accessed 10 October 2015.

D.M. Pozar, *Microwave Engineering*, 4th ed. (Wiley, 2012).

T.P. Wangler, *RF Linear Accelerators*, 2nd ed. (Wiley-VCH, Weinheim, 2008). http://dx.doi.org/10.1002/9783527623426

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Beam Dynamics and Layout

A.M. Lombardi
CERN, Geneva, Switzerland

Abstract
In this paper, we give some guidelines for the design of linear accelerators, with special emphasis on their use in a hadron therapy facility. We concentrate on two accelerator layouts, based on linacs. The conventional one based on a linac injecting into a synchrotron and a all-linac solution based on high gradient high frequency RF cavities.

Keywords
Linac; medical; beam dynamics layout; performance.

1 Introduction

Accelerators for use in medical applications should produce a proton beam with an energy of about 250 MeV, a carbon-ion beam with an energy of about 450 MeV/u, or both, with an average current of around 30 nA and a peak microbunch current of 500 μA. Ideally, a medical accelerator should be compact, cheap, reliable, easy to operate, and modular.

There are two established designs of medical accelerators that need linacs, one based on a linac injecting into a synchrotron and another (all-linac solution) based on linear accelerators. Their respective merits and drawbacks are not for discussion here, the main difference being the possibility of changing the beam energy from pulse to pulse at a very rapid rate if the all-linac solution is chosen, thus considerably reducing the treatment time. In this paper we will discuss the choices for the layout of both options, which are sketched in Figs. 1 and 2.

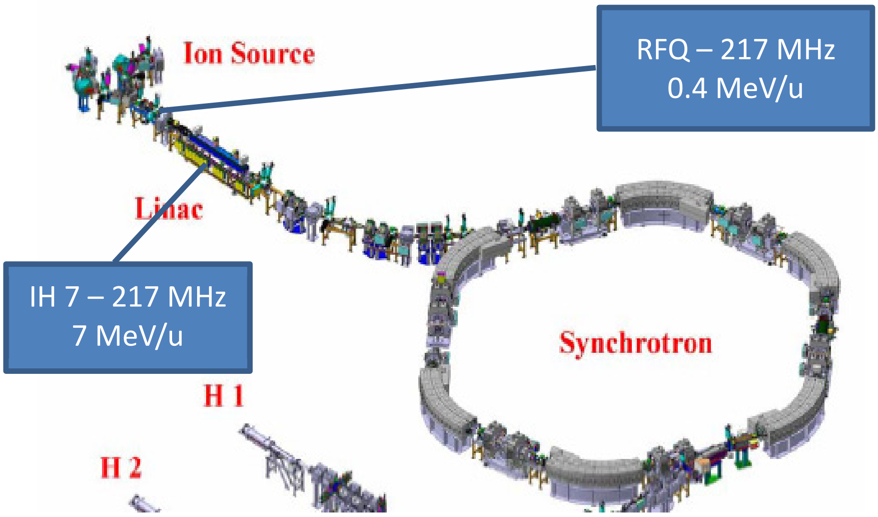


Fig. 1: Sketch of a medical facility based on a linac (5 m long) injecting into a synchrotron (10 m radius): typically the linac is composed of a Radio Frequency Quadrupole (RFQ) and a Interdigital H (IH) structure operating at a frequency of around 200 MHz with injection into the synchrotron at an energy of 7 MeV/u. This scheme can be used with protons and carbon ions, coming from two different sources funnelled into the RFQ at the low-energy end.

https://doi.org/10.23730/CYRSP-2017-001.109

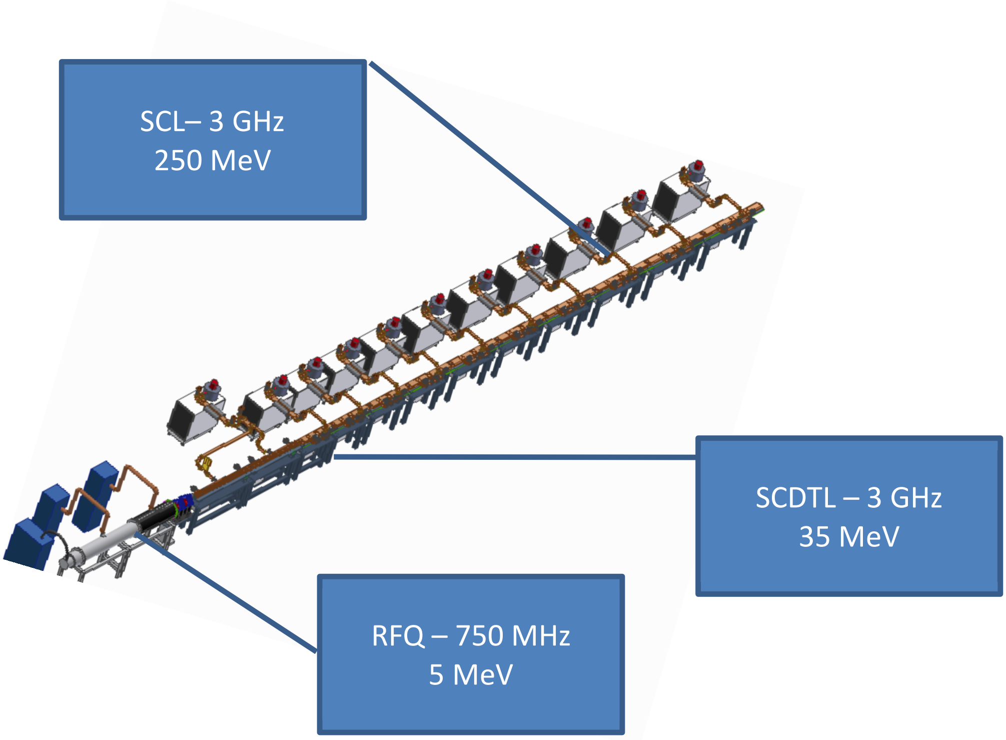


Fig. 2: Sketch of a medical facility based on high-frequency linacs, about 30 m in length, composed of a 750 MHz RFQ to bring a proton beam to 5 MeV and injecting into a Side-Coupled Drift Tube Linac (SCDTL) at 3 GHz. The final energy is reached with a standard Side-Coupled Linac (SCL) at 3 GHz. Such a facility can provide extremely fast (200 Hz) pulse-to-pulse energy and current variability for a proton beam.

2 Fundamentals of linear accelerators for medical applications

In this paper, we discuss guidelines for the design of a linear accelerator for medical applications, bearing in mind the two schemes shown above. For simplicity, and justified by the fact that the current in medical accelerators is rather modest, we treat the longitudinal and transverse planes separately.

2.1 Longitudinal plane: bunching and acceleration

A proton or carbon beam generated by a particle source is continuous on the scale of the radio frequency (RF) used in a linac. As it is not possible to transfer energy to a continuous beam by means of an RF field, it is necessary to prepare a beam from the source for RF acceleration. The section that achieves this is called a pre-injector. It is generally composed of an RF quadrupole, which also has the function of accelerating the beam to the input energy of the injector and of shaping the longitudinal emittance to match it to the acceptance of the injector. This manipulation of the beam to prepare for RF acceleration should be done whilst controlling losses and minimizing emittance growth. The pre-injector typically increases the energy of the beam to a few MeV over a few metres and is not very efficient.

The main function of the pre-injector is to bunch the beam on the scale of the wavelength. This operation is done by generating a velocity spread in the beam by passing it through an RF cavity and then letting the beam distribute itself around the particle with the average velocity. We can distinguish two extreme types of bunching: discrete and adiabatic.

Discrete bunching is shown in the sketch in Fig. 3, where the longitudinal phase space (phase–energy space) is shown at five different steps, starting from the top left. A beam from an ion source is continuous (1); this beam passes through an RF cavity, with the result that some particles are accelerated and some are decelerated (2). Notice that the average energy is not changed. After passing through the

RF cavity, the beam has a velocity distribution (2), which induces a change in the relative positions of the particles (3, 4). At the moment 5 the slowest, the average, and the fastest particle are at the same physical location: they are grouped around the average particle, as can be appreciated from the phase histogram of part 5 of the figure, shown in part 6.

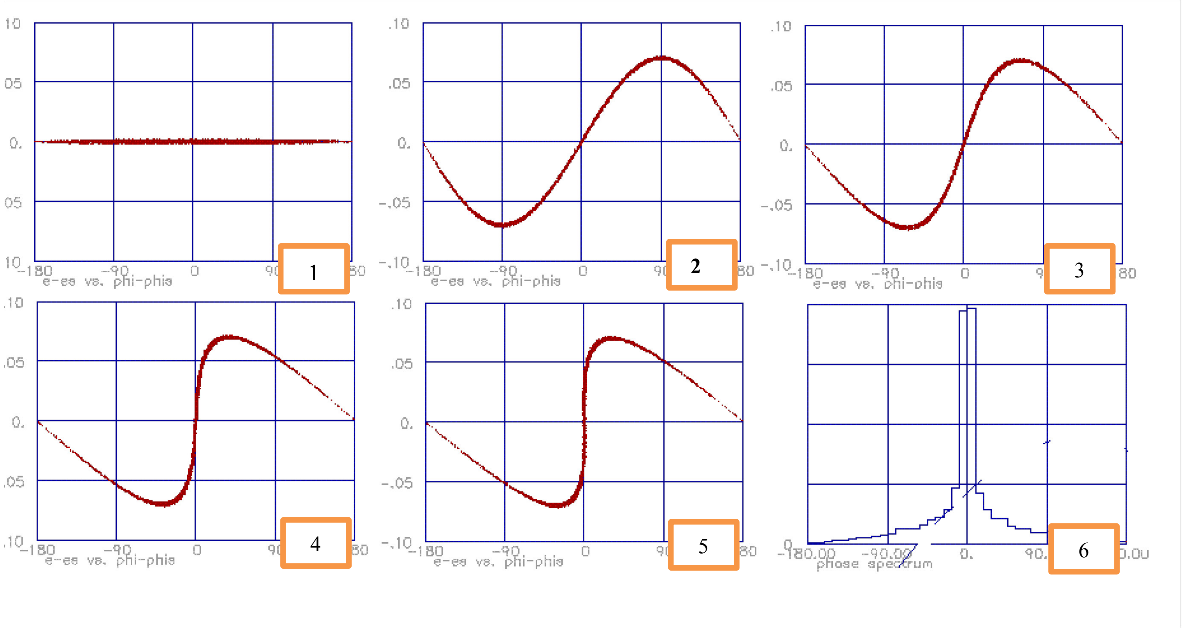


Fig. 3: Sketch of a discrete bunching process

The efficiency of discrete bunching is about 50%; namely, about 50% of the initial beam can be accelerated further by a chain of RF cavities. With a particularly well-designed system which includes higher harmonics of the frequency, an efficiency of 60–70% can be achieved. Nevertheless, a substantial fraction of the beam is lost. To improve the efficiency a different approach to bunching, called adiabatic bunching, can be used. Adiabatic bunching entails continuous bunching at a very low voltage, which gives the beam time to wrap around the synchronous phase. The concept is to generate a velocity spread continuously with a small longitudinal field and perform the bunching over several oscillation in the phase space (up to 100!). This allows better capture around the stable phase, achieving up to 95% capture. Adiabatic bunching is performed in the first few sections of an RFQ by slowly increasing the depth of the modulation along the structure, thus making it possible to bunch the beam smoothly and prepare it for acceleration. The beam needs to be kept bunched during the whole acceleration phase, i.e., it is necessary to provide a longitudinal restoring force. This is done by accurately choosing the phase of acceleration, as shown in Fig. 4: particles with less energy than the average should see a slightly higher accelerating field than the particle with the average energy, and the opposite should be true for particles with more energy than the average. This concept is known as the principle of phase stability.

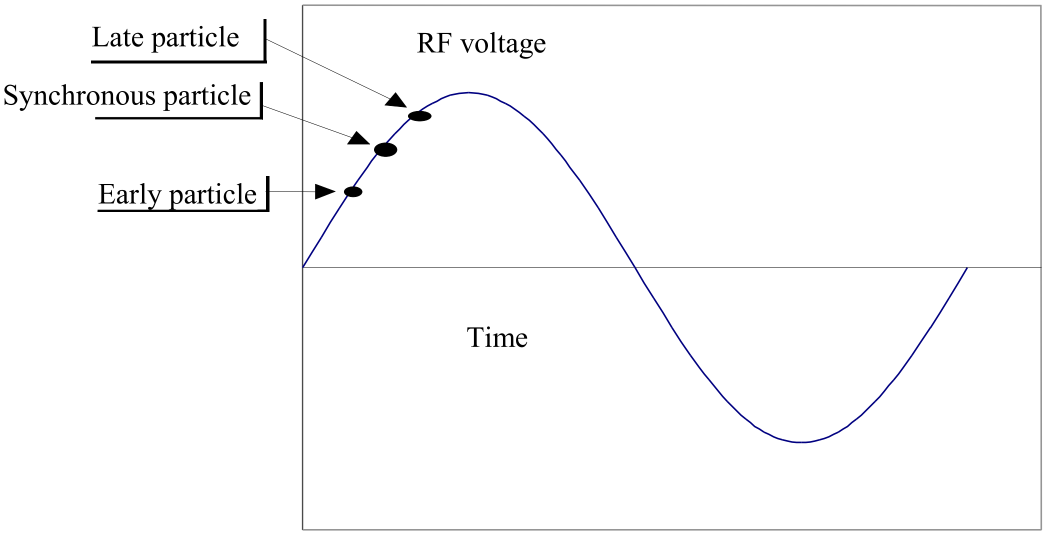


Fig. 4: Sketch of the principle of phase stability

Once the beam has been bunched and is ready to be accelerated, the average velocity of the beam will change as it traverses RF cavities. In the design phase, the layout is calculated around the average particle of the beam, often called the synchronous particle. The synchronous particle is the particle (possibly fictitious) used to calculate and determine the phase along the accelerator. It is the particle whose velocity is used to determine the synchronicity with the electric field. The idea is to design for the synchronous particle and provide longitudinal focusing so that other particles will perform small oscillations around it and remain bound to the centre of the bunch.

The length of each accelerating element determines the time at which the synchronous particle enters or exits an RF cavity, so for a given cavity length there is an optimum velocity (or beta = velocity/speed of light) such that a particle travelling at that velocity passes through the cavity in half an RF period. The difference in time of arrival between the synchronous particle and a particle travelling at a speed corresponding to the geometrical beta (the velocity of a particle which would traverse the cavity in half the RF period) determines the phase difference between two adjacent cavities. We can therefore adjust the phase between two adjacent RF cavities by changing the length of one of the cavities. In a synchronous structure, the geometrical beta is always equal to the synchronous-particle beta and each cell is different. A synchronous structure provides the best possible longitudinal beam dynamics and allows full control of the longitudinal phase space, but it implies that each cavity is different. For medical applications, it is possible to lift this constraint after the beam has become energetic enough (beta = 20%), to allow some standardization of the cavity length. To simplify construction and contain costs, cavities are not individually tailored to the evolution of the beam velocity but are constructed in blocks of identical cavities (tanks). Several tanks are fed by the same RF source. This simplification implies a ‘phase slippage’, i.e., motion of the centre of the beam around a stable phase. The phase slippage is proportional to the number of cavities in a tank, and it must be carefully controlled for successful acceleration.

Let us now turn to acceleration. We shall describe the motion of a particle in the longitudinal phase space and establish a relation between the energy and phase of the particle during acceleration.

If we write the energy gain of the synchronous particle as

$$\Delta W_s \quad = qE_0LT\cos(\varphi_s), \tag{1}$$

where the symbols are defined in the lecture “Overview of linacs” in these proceedings, then the energy gain of a particle with phase φ is

$$\Delta W = qE_0 LT\cos(\varphi), \tag{2}$$

and assuming a small phase difference Δφ = φ − φs, we can write the following equations for the energy and phase:

$$\frac{\mathrm{d}}{\mathrm{d}s}\Delta W = qE_0T\cdot[\cos(\varphi_\mathrm{s}+\Delta\phi)-\cos\varphi_\mathrm{s}], \tag{3}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\Delta\varphi = \omega\left(\frac{\mathrm{d}t}{\mathrm{d}s}-\frac{\mathrm{d}t_\mathrm{s}}{\mathrm{d}s}\right) = \frac{\omega}{c}\left(\frac{1}{\beta}-\frac{1}{\beta_\mathrm{s}}\right) \cong -\frac{\omega}{\beta_\mathrm{s}c}\frac{\Delta\beta}{\beta_\mathrm{s}} = -\frac{\omega}{mc^3\beta_\mathrm{s}^3\gamma_\mathrm{s}^3}\Delta W\ . \tag{4}$$

The expressions above are equations for the canonically conjugate variables phase and energy, with Hamiltonian (total energy of oscillation)

$$\frac{\omega}{mc^3\beta_\mathrm{s}^3\gamma_\mathrm{s}^3}\left\{\frac{\omega}{2mc^3\beta_\mathrm{s}^3\gamma_\mathrm{s}^3}(\Delta W)^2 + qE_0T\left[\sin(\varphi_\mathrm{s}+\Delta\varphi)-\Delta\varphi\cos\varphi_\mathrm{s}-\sin\varphi_\mathrm{s}\right]\right\} = H\ . \tag{5}$$

For each *H*, we have different trajectories in the longitudinal phase space. The equation of the separatrix (the line that separates stable from unstable motion) is

$$\frac{\omega}{2mc^3\beta_\mathrm{s}^3\gamma_\mathrm{s}^3}(\Delta W)^2 + qE_0T\left[\sin(\varphi_\mathrm{s}+\Delta\varphi)+\sin\varphi_\mathrm{s}-(2\varphi_\mathrm{s}+\Delta\varphi)\cos\varphi_\mathrm{s}\right] = 0\ , \tag{6}$$

from which we can deduce that the maximum energy excursion of a particle moving along the separatrix is

$$\Delta\hat{W}_\mathrm{max} = \pm 2\left[\frac{qmc^3\beta_\mathrm{s}^3\gamma_\mathrm{s}^3E_0T(\varphi_\mathrm{s}\cos\varphi_\mathrm{s}-\sin\varphi_\mathrm{s})}{\omega}\right]^{\frac{1}{2}}. \tag{7}$$

This is a very important and useful expression that gives the energy acceptance of an accelerator depending on the field level and the accelerating phase. The longitudinal acceptance of an accelerator has a characteristic shape, similar to a golf club. Particles falling into this area in phase space can be successfully accelerated. A practical example is shown in Fig. 5.

When we accelerate on the rising part of the positive RF wave, we have a longitudinal force which keeps the beam bunched. The force (of harmonic-oscillator type) is characterized by the longitudinal phase advance, expressed as

$$k_{0l}^2 = \frac{2\pi qE_0T\sin(-\varphi_\mathrm{s})}{mc^2\beta_\mathrm{s}^3\gamma^3\lambda}\left[\frac{1}{m^2}\right], \tag{8}$$

with the equation for the phase resulting in

$$\frac{\mathrm{d}^2\Delta\varphi}{\mathrm{d}s^2} + k_{0l}^2\left(\Delta\varphi - \frac{\Delta\varphi^2}{2\tan(-\varphi_\mathrm{s})}\right) = 0\ . \tag{9}$$

An exception is the KONUS beam dynamics used in the IH structure [1], where the particle beam is purposely accelerated outside the area of stability, i.e., outside the separatrix. This dynamics is very convenient for accelerating efficiently, but it requires periodic rebunching sections.

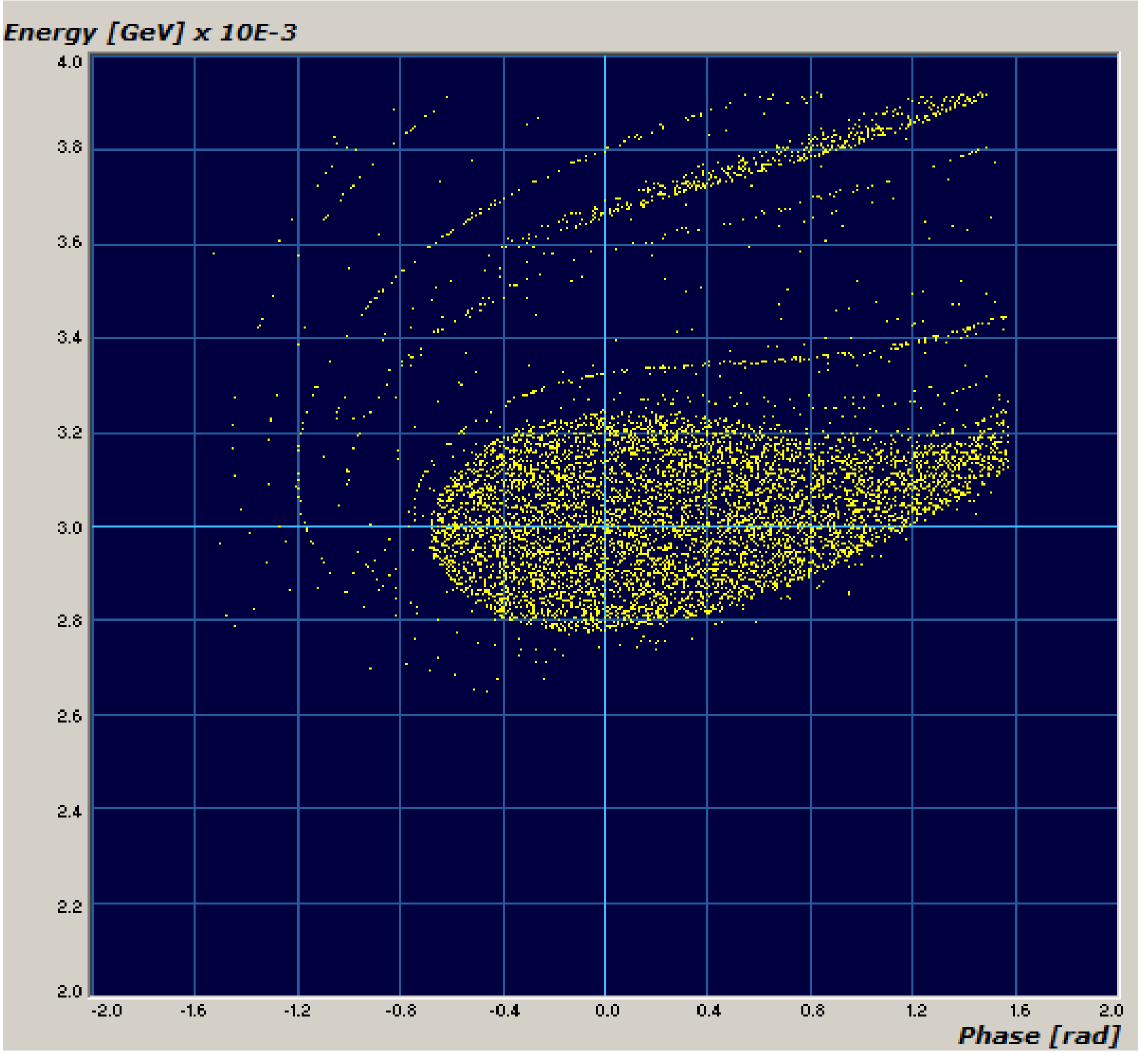


Fig. 5: Longitudinal acceptance of the CERN LINAC4 DTL (352 MHz, 3–50 MeV)

2.2 Transverse plane: focusing

All along the accelerator we need to provide a transverse force, to keep the beam confined transversely. Ideally we should apply a force towards the beam axis proportional to the distance from the axis, a linear force which would keep the beam confined without degrading the emittance. In the case of medical accelerators, the microbunch current is very low and space charge effects are generally negligible, thus simplifying the layout of the optics and allowing weak focusing. The only choice that can be made with respect to the focusing strength depends on the bore radius of the cavities (stronger focusing, smaller radius) and on the accelerating phase. As we will see in the following, at low energy there is substantial coupling between the transverse and longitudinal planes caused by the transverse defocusing RF effects, which is implied by the choice of a stable restoring force in the longitudinal plane to keep the beam bunched.

There are two types of focusing force: electric focusing and magnetic focusing. Electric focusing is independent of the particle velocity and is therefore generally used at the low-energy end. We have talked extensively about the electric focusing force in the lecture “Overview of linacs” in these proceedings.

A magnetic quadrupole (electromagnetic or composed of a special arrangement of permanent magnet material) provides a field B that can be expressed as

$$\begin{cases} B_x = G \cdot y \\ B_y = G \cdot x \end{cases}, \tag{10}$$

where G is the quadrupole gradient, i.e., the field on the pole tip divided by the quadrupole bore radius. A charged particle (with charge q) travelling at a velocity v through this field experiences a force proportional to its distance (x, y) from the axis, which can be written as

$$\begin{cases} F_x = -q \cdot v \cdot G \cdot x \\ F_y = q \cdot v \cdot G \cdot y \end{cases}. \tag{11}$$

The force in the expression above is focusing in the x plane and defocusing in the y plane for a positive gradient G.

In order to keep the beam confined along the accelerator axis, it is therefore necessary to interlace a series of quadrupoles of alternate polarity. This arrangement of quadrupoles is called a FODO channel. The beam dynamics in a FODO channel is shown in Fig. 6. The beam envelope in one plane is shown at characteristic locations along the channel, and the corresponding orientation of the beam emittance is shown below it. It should be appreciated that after one full focusing period (positions 1 and 7) the beam phase space is identical to what it was before. Such a channel can be extended throughout the whole accelerator, taking into account the fact that its length needs to be progressively adapted to the changing velocity of the beam.

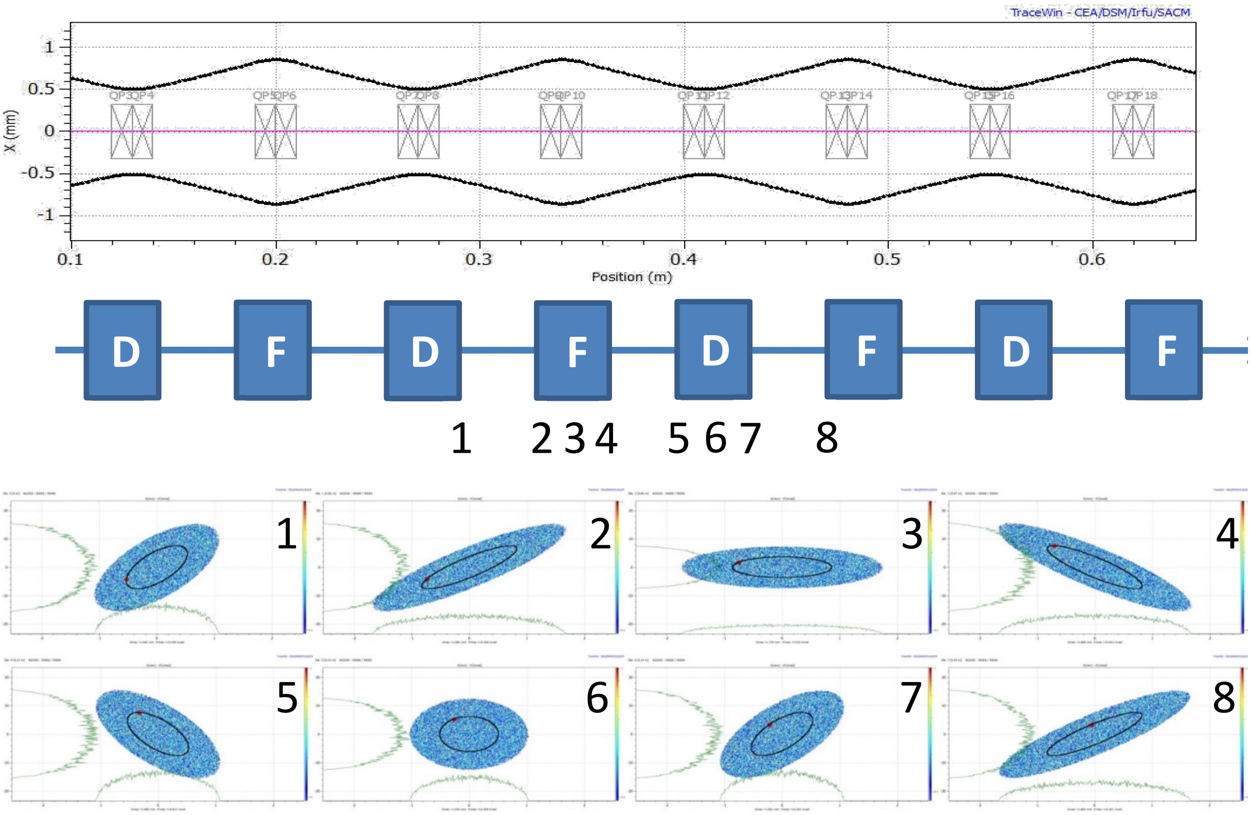


Fig. 6: Beam envelope (top) and beam phase space (bottom) in a FODO channel

We can write the solution of the equation of motion in a periodic channel as

$$X(s) = \sqrt{\varepsilon \beta(s)} \cos(\sigma_{0t}(s)), \tag{12}$$

where ε is the beam emittance, β is a periodic function with the periodicity of the focusing period, and σ, the transverse phase advance, is a measure of the strength of the focusing channel.

NB: in Eqs. (12) and (14), β is *not* the relativistic β.

In medical linacs, the overall force balance in the transverse plane is given by the static quadrupole focusing and the RF defocusing and can be expressed via the transverse phase advance at zero current,

$$\sigma_{0t} = \sqrt{\frac{\theta_0^4}{8\pi^2} + \Delta_{\mathrm{rf}}}\ . \tag{13}$$

The first term on the right-hand side depends on the strength of the quadrupoles and the particle velocity according to

$$\theta_0^2 = \frac{qG\lambda^2 N^2 \beta\chi}{m_0 c\gamma}, \tag{14}$$

where G is the magnetic quadrupole gradient (in units of T/m), N is the number of magnets in a period, and χ is as follows:

for + − ($N = 2$):

$$\chi = \frac{4}{\pi}\sin\left(\frac{\pi}{2}\Gamma\right); \tag{15}$$

for + + − − ($N = 4$):

$$\chi = \frac{8}{\sqrt{2}\pi}\sin\left(\frac{\pi}{4}\Gamma\right), \tag{16}$$

where Γ is the quadrupole filling factor (the quadrupole length relative to the period length).

The expression above allows one to determine the quadrupole gradient necessary to limit the beam size to a given value for a given quadrupole configuration and a given beam energy.

The RF defocusing, represented by the term Δ_{rf}, comes from the varying electric field in the RF cavities and is a consequence of the choice of the accelerating phase according to the principle of phase stability. If we write the Maxwell equation

$$\nabla \cdot E = 0 \tag{17}$$

as

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0\,, \tag{18}$$

we can see that a longitudinal restoring force

$$\frac{\partial E_z}{\partial z} > 0 \tag{19}$$

implies a transverse defocusing

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} < 0 \tag{20}$$

with an intensity equal to half of the longitudinal phase advance:

$$\Delta_{\mathrm{rf}} = \frac{1}{2}\sigma_{0l}^2 = \frac{1}{2}\beta^2\lambda^2 k_{0l}^2 = \frac{\pi q \lambda N^2 E_0 T \sin\phi_{\mathrm{s}}}{m_0 c^2 \beta\gamma^3} . \tag{21}$$

The RF defocusing is a phase-dependent defocusing term which is more important at the lowest energies. It is often the parameter that determines the focusing layout at energies up to 5 MeV and frequencies of the order of 200 MHz.

3 Putting it all together: a rough guide

In this final section we attempt to give some guidelines for the design of a linear accelerator for medical applications. Let us assume we have available a collection of RF cavities to increase the beam energy, such as an RFQ, an IH structure, a drift tube linac, a side-coupled linac, and some hybrid structure. Assume we have available solenoids, electromagnetic quadrupoles, and permanent magnet quadrupoles to keep the beam volume confined. Assume also that we have a green-field site on which to design an accelerator. The accelerator designer has to make some (difficult) starting choices which determine the layout of the accelerator. These choices are not generally straightforward, as each choice has different implications and there is not generally only one “right choice”. In the following we list a series of questions that are intended to trigger discussions and thoughts towards making an informed choice.

3.1 First basic questions and choices

Which frequency? The first and foremost important choice is the operating frequency. There are some standardized frequencies for an accelerator for which power sources are available. For an accelerator that is to be built, it is advised to choose a frequency for which a power source exists. The range of available frequencies is still rather large and the choice of the frequency has strong implications for the transverse size of the accelerator, the transverse acceptance, and the maximum accelerating field and duty cycle. The higher the frequency, the more compact the accelerator, but at the same time the smaller the acceptance. It can also be envisaged, for an all-linac solution, that one may have a frequency jump during the acceleration, thus optimizing the frequency for the energy range that is to be used.

At what energy does one make the transition between a TE structure (RFQ) to a TM structure (e.g., DTL)? This point should be evaluated when the design is already somewhat advanced, ideally after two structures overlapping in energy have been designed and several combinations of transition energies have been tried. In general, transitions are points of weakness in accelerators, especially at the low-energy end, and should be designed carefully (with enough variable elements to accommodate errors). A transition below the threshold for activation of neutron production in copper (about 3–5 MeV) is a good choice.

As we have seen, medical accelerators should be compact, so ideally we would like to chain as many RF cavities as possible together and reduce to a minimum the number of focusing elements. But *what is a sensible minimum number of focusing elements?* We should look at the beam size in a FODO channel: the maximum beam size increases as the length of the FODO period increases. So the smaller the number of quadrupoles, the larger the beam size and the larger the bore radius that we need in the RF cavities. A larger bore radius implies more power for the same accelerating gradient, so the answer to this question is a balancing exercise between real estate gradient, the bore aperture (and hence RF efficiency), and the transverse acceptance (source performance).

As we have seen, it is convenient to standardize the structures and have RF cavities that are all of the same length. It is economical to have as many RF gaps per cavity as possible, but *what is the maximum number of gaps that we can put together?* In this case also, we are faced with a balancing exercise between the acceptable phase slippage and RF power optimization and distribution, with implications for the longitudinal acceptance, i.e., the quality of the pre-injector.

What transverse acceptance should the accelerator provide? This choice is dictated mainly by cost. A large acceptance makes the operation of the accelerator easier and implies less stringent alignment and machining tolerances. On the other hand, a large acceptance means a larger bore radius and therefore less efficient acceleration.

Variable or fixed focusing system? The choice is between a fixed focusing system (using permanent magnet quadrupoles) and a more costly variable system. In general, for an all-linac solution, a fixed focusing system should be chosen because it is economical, it simplifies operation, and reduces the size of the accelerator. A variable focusing system is strictly needed only in transitions between structures and if different ion species need to be accommodated in the same accelerator.

Reference

[1] R. Tiede, U. Ratzinger, H. Podlech, C. Zhang and G. Clemente, Konus beam dynamics designs using h-mode cavities, Proc. Hadron Beam 2008, Nashville, TN, USA.

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Accelerators for Medical Applications—Radio Frequency Powering

E. Montesinos
CERN, Geneva, Switzerland

Abstract
This paper reviews the main types of radio-frequency powering systems which may be used for medical applications. It gives the essentials on vacuum tubes, including tetrodes, klystrons, and inductive output tubes, and the essentials on transistors. The basics of combining systems, splitting systems, and transmission lines are discussed. The paper concludes with a case study specific to medical applications, including overall efficiency and cost analysis regarding the various available technologies.

Keywords
Vacuum tube; tetrode; klystron; IOT; transistor; transmission lines.

1 Introduction

Cost is a very important factor for all projects, and this is particularly true for medical applications. Figure 1 summarizes, at a glance, the ratios that are currently applied when talking about radio frequency (RF)-power systems.

W	→	kW	→	MW
€	→	k€	→	M€

Fig. 1: Relationship between Watts and Euros in RF-power systems

We will describe, in this document, the technologies used to build RF-power amplifiers, and the high technicity involved explains such high numbers.

2 RF-power basics

When we talk about an RF-power system, one should understand the system amplifying a small RF signal from the generator, in the order of mW, up to the W, kW, or MW level at the Device Under Test input, that could be an accelerating cavity, or any other RF load, as described in Fig. 2.

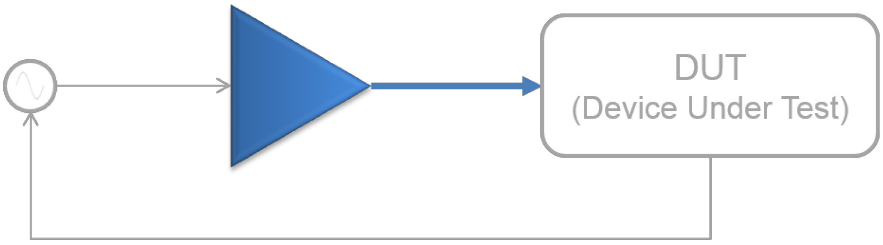


Fig. 2: A very simple representation of an RF-power system. It includes the RF-power amplifiers, the transmission lines and the Fundamental Power Coupler.

Some specific parameters characterize RF-power systems, such as wavelength, frequency, and Decibel (dB). We will describe these basic concepts in the following paragraphs.

https://doi.org/10.23730/CYRSP-2017-001.119

2.1 Wavelength and frequency

The wavelength (Fig. 3) and the frequency are linked with the following formulas:

$$\lambda = \frac{c}{\mathrm{f}\sqrt{\varepsilon}} \tag{1}$$

$$f = \frac{c}{\lambda\sqrt{\varepsilon}} \tag{2}$$

where

λ = wavelength in meters (m),

c = velocity of light (m/s)—(~300 000 000 m/s),

f = frequency in hertz (Hz), and

ε = dielectric constant of the propagation medium.

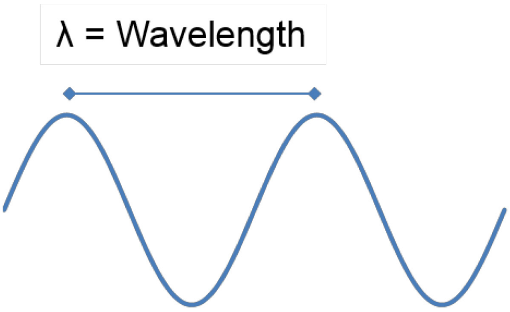


Fig. 3: The wavelength is the distance between two points reaching exactly the same potential of a sine wave oscillating at a given frequency.

One can see that the medium is important as the wavelength will be directly proportional to its propagation dielectric constant. In air or vacuum, ε is equal to 1, if the medium is PTFE, as in our coaxial cable from the antenna to the television, ε is around 2.2. The wavelength will then be shortened by $\sqrt{2.2}$. One will have to keep this in mind when designing an RF-power system.

2.2 Electromagnetic spectrum and radiofrequency spectrum

The RF spectrum is a fraction of the electromagnetic (EM) spectrum. It has been defined as starting at 30 kHz and ending at 300 GHz. Respectively, the wavelength is of 10 km reducing to 1 mm. Figure 4 shows the entire EM spectrum with some examples of current known applications, and Fig. 5 shows the RF spectrum with the uses in our current lives.

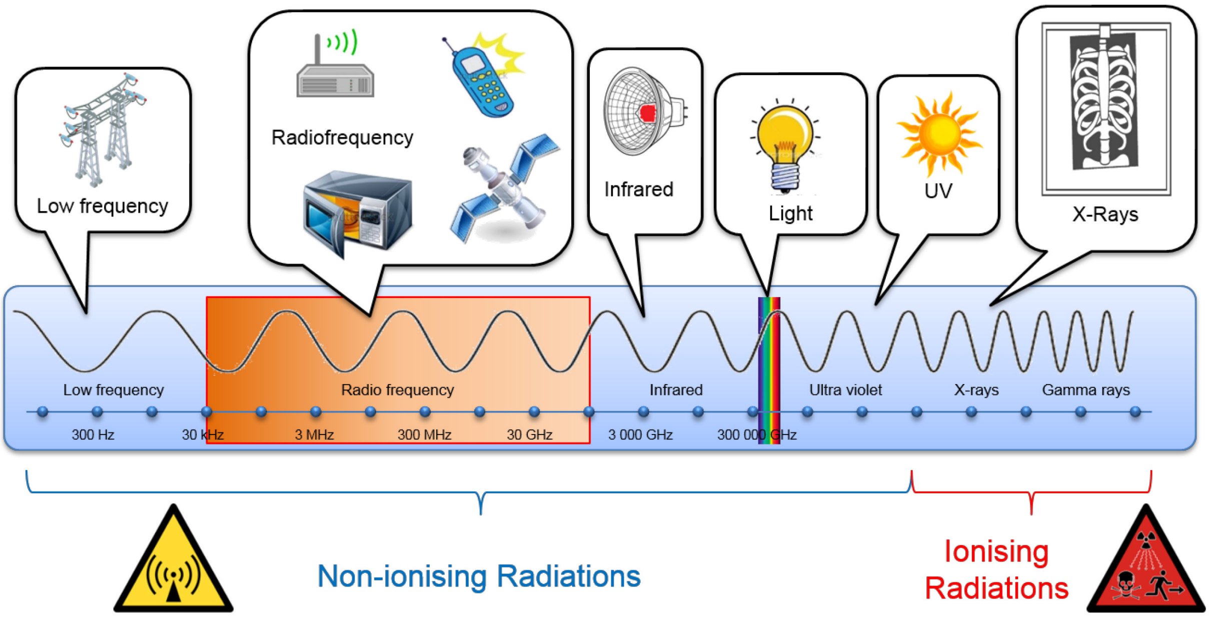


Fig. 4: The EM spectrum with respect to frequency range

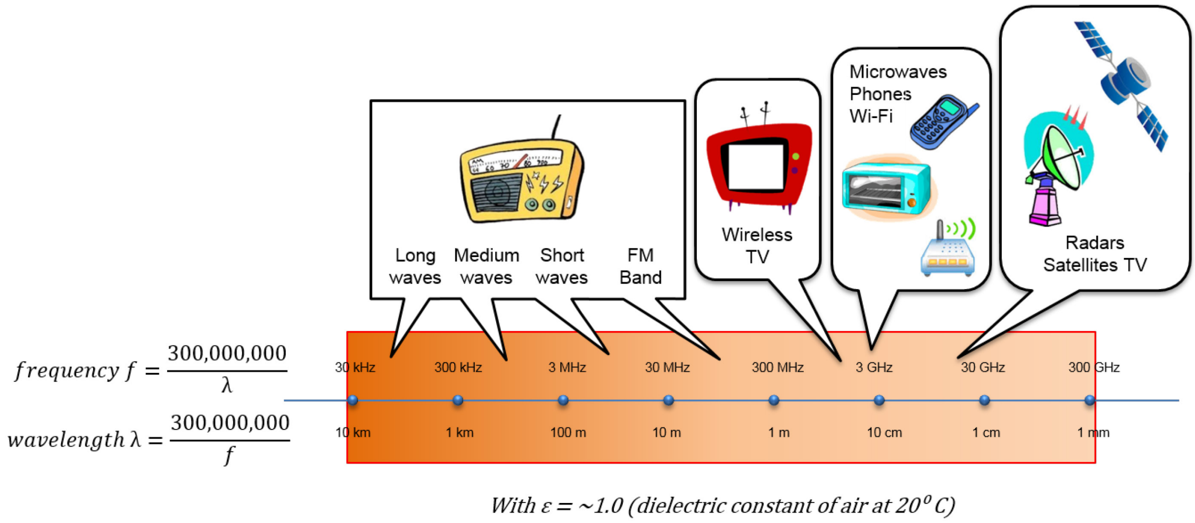


Fig. 5: The RF spectrum with relationship between frequency and wavelength

2.3 Decibel

A useful unit used with RF-power equipment is the Decibel. This unit allows us to sum the gain or subtract the attenuation instead of multiplying the gain and dividing the attenuation. It makes calculations of power systems easier. The dBm (3) is to quantify absolute values, and the dB (4) to quantify ratio. The definitions of the dB, commonly used in RF power, are:

$$\mathrm{dBm} = 10\ \log_{10}\left(\mathrm{P}_{\mathrm{mW}}\right), \tag{3}$$

$$\mathrm{dB} = 10\ \log_{10}\left(P_1 / P_2\right). \tag{4}$$

Several other definitions of the dB exist:

$$\mathrm{dB} = 20\log_{10}\left(V_1 / V_2\right), \tag{5}$$

$$\mathrm{dB}V = 20\ \log_{10}\left(V_{Vrms}\right), \tag{6}$$

$$\mathrm{dB}\mu V = 20\ \log_{10}\left(V_{\mu V_{\mathrm{rms}}}\right), \tag{7}$$

$$\mathrm{dB}c = 10\ \log_{10}\left(\frac{P_{\mathrm{carrier}}}{P_{\mathrm{signal}}}\right). \tag{8}$$

Some absolute values are useful to memorize (Fig. 6). One can switch from dBm to watt applying the following formula (9):

$$x_{dBm} = 10\ log_{10}\left(\mathrm{P}_{\mathrm{mW}}\right) \leftrightarrow P_{\mathrm{mW}} = 10\left(x_{dBm}/10\right). \tag{9}$$

0 dBm	=	1 mW
30 dBm	=	1 W
60 dBm	=	1 kW
90 dBm	=	1 MW

Fig. 6: Some known dBm to watt values

One can also switch from dB to ratio in power (Fig. 7) following formula (10):

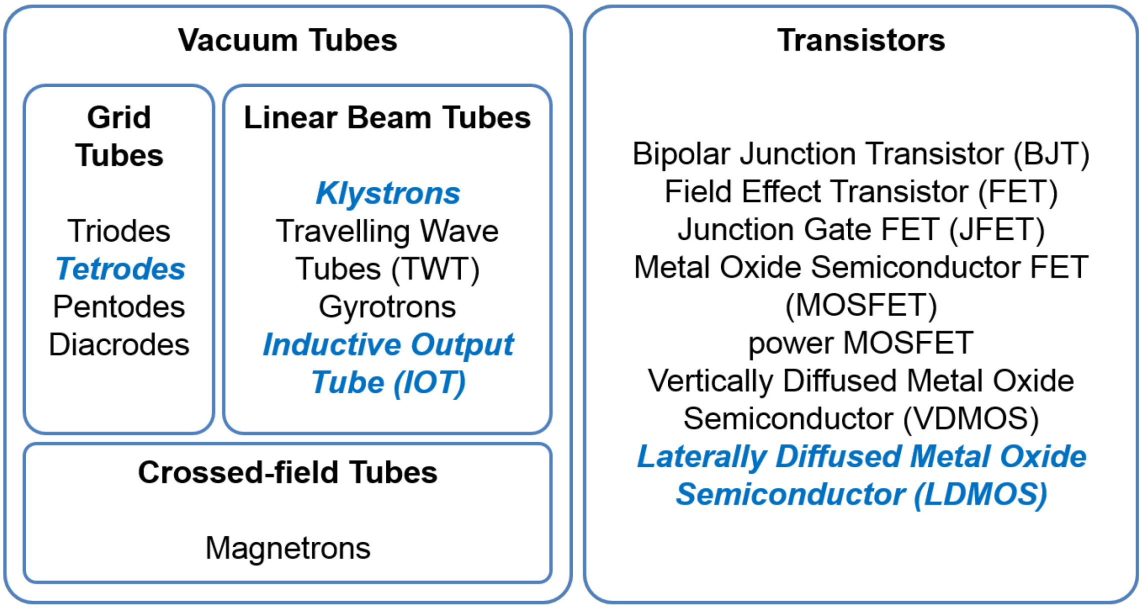
$$x_{dB} = 10\ log_{10}\left(P/\mathrm{P}_{\mathrm{ref}}\right) \leftrightarrow P/P_{\mathrm{ref}} = 10\left(x_{dB}/10\right). \tag{10}$$

x (dB)	**P/P$_{ref}$**	
+ 0.1	1.023	+ 2.5%
+ 0.5	1.122	+ 12%
+ 1	1.259	+ 25%
+ 3	1.995	2
- 0.1	0.977	- 2.5%
- 0.5	0.891	- 11%
- 1	0.794	- 20%
- 3	0.501	0.5

Fig. 7: Some known ratio values

3 RF-power amplifiers

RF-power amplifiers can be sorted into two main families: the vacuum tubes and the transistors. The vacuum tubes consist of three main families: the grid tubes, the linear beam tubes, and the crossed-field tubes. The transistor amplifiers are also named solid state amplifiers (SSA) or solid state power amplifiers (SSPA). In this document, we will look in further detail at the tetrodes, the klystrons, the inductive output tube (IOT), and the laterally diffuse metal oxide semiconductor (LDMOS). Figure 8 shows the different families.


Vacuum Tubes

Grid Tubes

Triodes
Tetrodes
Pentodes
Diacrodes

Linear Beam Tubes

Klystrons
Travelling Wave Tubes (TWT)
Gyrotrons
Inductive Output Tube (IOT)

Crossed-field Tubes

Magnetrons

Transistors

Bipolar Junction Transistor (BJT)
Field Effect Transistor (FET)
Junction Gate FET (JFET)
Metal Oxide Semiconductor FET (MOSFET)
power MOSFET
Vertically Diffused Metal Oxide Semiconductor (VDMOS)
Laterally Diffused Metal Oxide Semiconductor (LDMOS)

Fig. 8: Main list, non-exhaustive, of RF-power-amplifier families

3.1 Grid tubes

The grid tube story started more than a century ago in 1904 with the very first diode [1]. Hereunder, the list of the main milestones of the grid tube story. It is very interesting to note that most of the discoveries were made within the first quarter of the last century, even though, almost a century later in 1994, thanks to new fabrication methods, new tubes are still being developed.

1904 Diode, John Ambrose Fleming (Fig. 9) [1];

1906 Audion (first triode), Lee de Forest [2];

1912 Triode as amplifier, Fritz Lowenstein [3];

1913 Triode ‘higher vacuum’, Harold Arnold [4];

1915 first transcontinental telephone line, Bell [5];

1916 Tetrode, Walter Schottky [6];

1926 Pentode, Bernardus Tellegen [7];

1994 Diacrode, Thales Electron Devices [8].

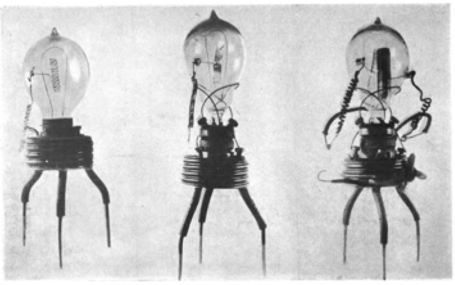


Fig. 9: Very first diode invented by John Ambrose Fleming in 1904

Vacuum tube history started with the diode [1]. Looking at Fig. 10, we can identify the heater and the cathode. In this illustration, we are looking at a heated cathode circuit. There is a separate heater and a cathode. There also exists a circuit with a direct-heated cathode. In that case, the cathode also includes the heater. The cathode system is a complex system composed of coated metal, doped with carbides, borides, and other specific components developed by the tube suppliers to ensure good electron emission. Once the cathode is heated, a thermionic emission starts and an electron cloud is generated around the cathode. If we now apply a high voltage on the anode side, these electrons will fly from the cathode to the anode. If we reverse the anode voltage to a negative value, the electrons will remain at the cathode level. We have here a diode.

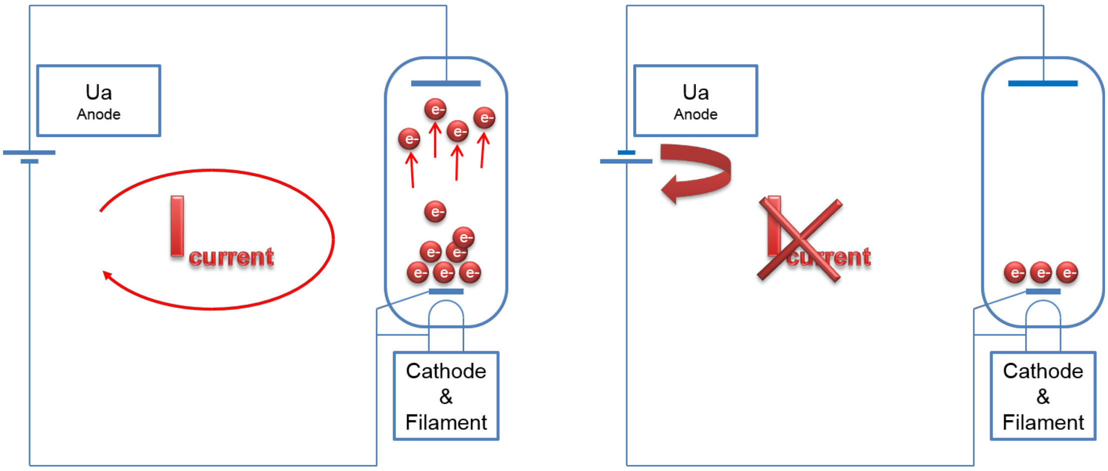


Fig. 10: On the left, with a positive voltage on the anode, electrons fly from the grid to the anode. On the right, with a negative voltage on the anode, electrons remain at the grid. This is the basic principle of the very first diode invented by John Ambrose Fleming in 1904 [1].

A few years later, in 1906, Lee de Forest [2] added a control grid in between the cathode and the anode, as showed in Fig. 11. By modulating the voltage applied to the grid, we proportionally modulate the anode current. This is the trans-conductance effect: voltage modulation at the grid is transformed into current modulation at the anode. Indeed, when the grid voltage is less negative than the cathode voltage, the electrons fly to the anode, and when the grid voltage is more negative than the cathode voltage, the electrons remain in the cathode. Unfortunately, there are some limitations in this system. The parasitic capacitor between the grid and the anode gives the system a tendency to oscillate.

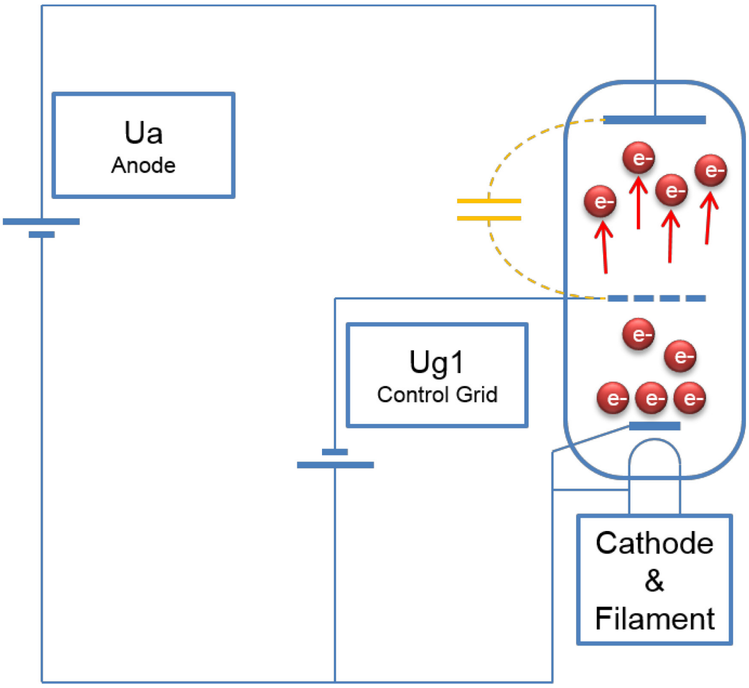


Fig. 11: A control grid is inserted in-between the cathode and the anode in order to modulate the electron flux. This is the basic principle of the triode invented by Lee de Forest in 1906 [2].

In order to suppress this tendency to oscillate, a second grid, the screen grid, has been added in between the control grid and the anode. With its positive voltage, lower than the anode, it provides two main advantages. It allows the parasitic capacitor between the control grid and the anode to be decoupled. It also provides better attraction of the electron, as it is close to the control grid and the cathode. This provides a better gain compared to the triode. Unfortunately, this also generates some additional limitations. As sketched in Fig. 12, some of the electrons are accelerated too much and once they hit the anode, they generate secondary electrons which fly from the anode to the control grid. To prevent this effect, tube manufacturers have developed a special treatment of the anode to reduce this secondary emission. Figures 13 and 14 show the CERN SPS based on RS2004 tetrode amplifiers.

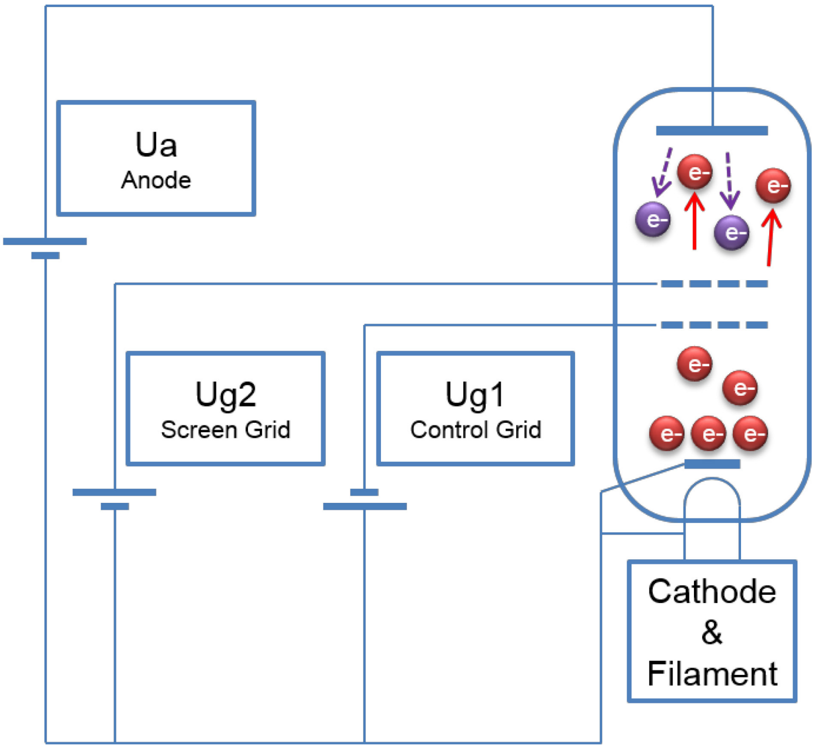


Fig. 12: A second grid, the screen grid, is inserted in between the control grid and the anode. This is the basic principle of the tetrode invented by Walter Schottky in 1916 [6].

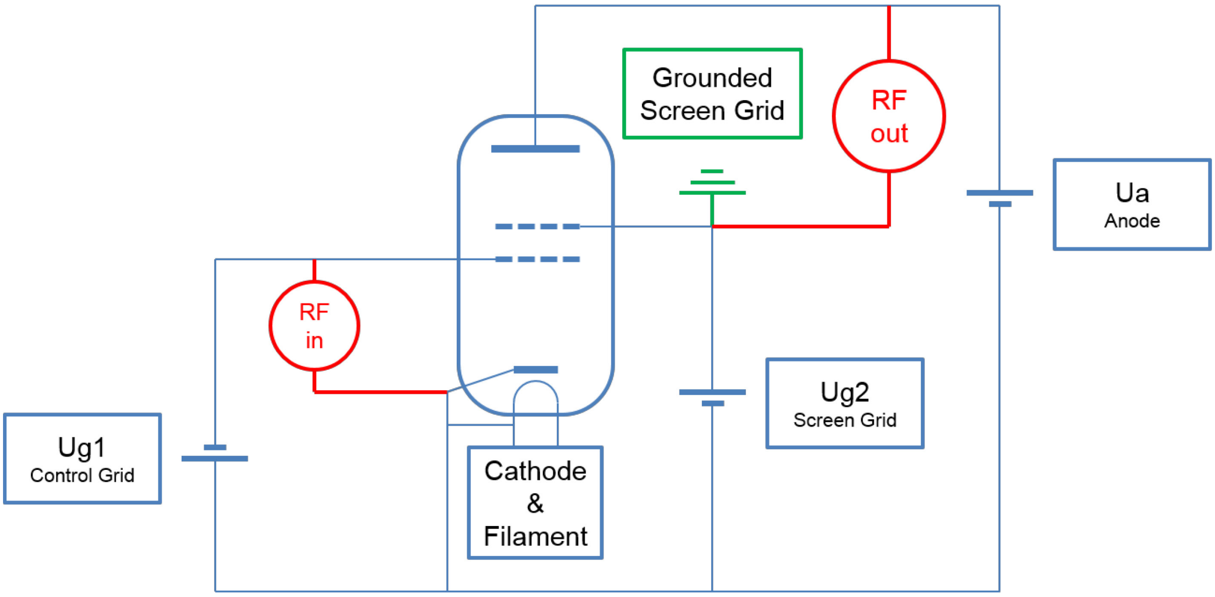


Fig. 13: CERN SPS, RS 2004 Tetrode (very) simplified bloc diagram

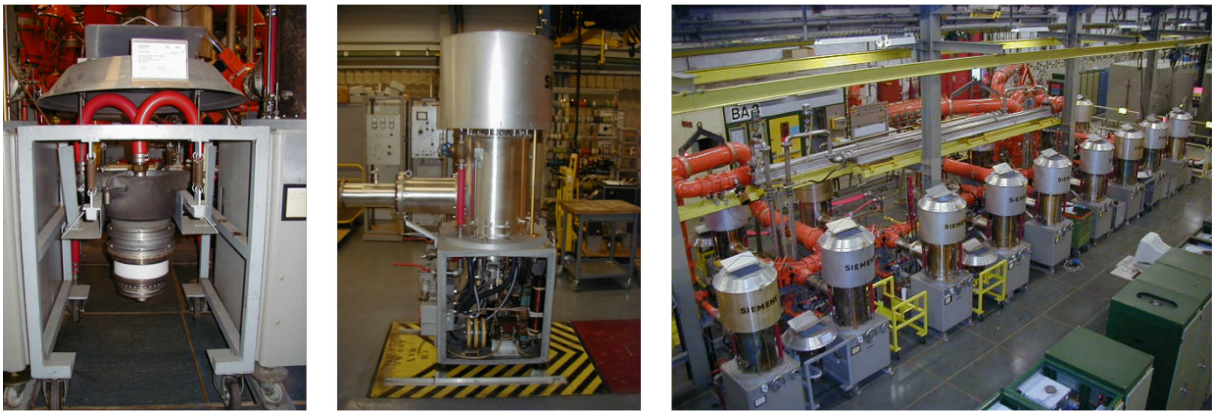


Fig. 14: CERN SPS, RS 2004 Tetrode, on the left a trolley (single amplifier), in the centre a transmitter (combination of four amplifiers) and on the right two transmitters (combination of eight amplifiers) delivering 2 x 1 MW @ 200 MHz, into operation since 1976.

An additional grid can be inserted. We then have a pentode. However, the construction complexity of such a tube limited its usage to lower-power systems.

More recently, technical fabrication improvements have been made allowing Thales to construct a Diacrode© [8]. This tube is equivalent to a double-ended tetrode, allowing even more power with a single tube.

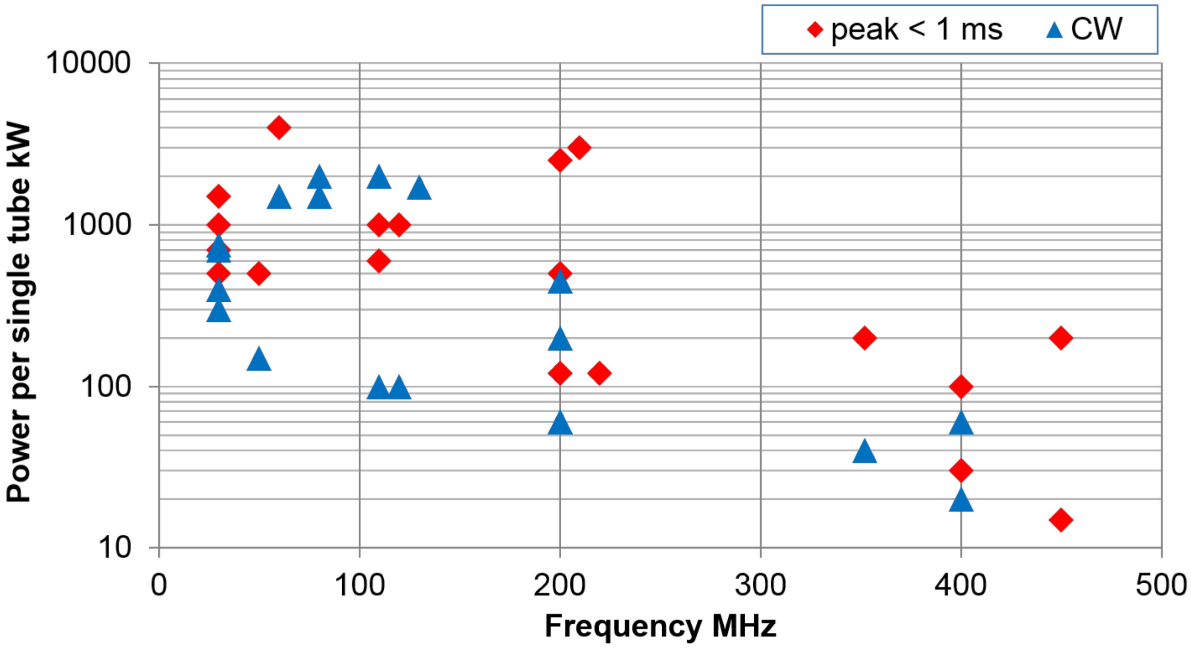


Fig. 15: Tetrodes and Diacrodes available from industry

Figure 15 summarizes all tubes currently available from worldwide suppliers. We note that the maximum power is over 1 MW at low frequency, power decreases with frequency, and the frequency range is from a few MHz to 400 MHz.

3.2 Linear beam tubes

The linear beam tube story started later, in 1937, with the very first klystron. Hereunder, the list of the main milestones of the linear beam tube story. It is very interesting to note that most of the discoveries have been made within a decade, from 1937 to 1948. Later, as for the grid tubes, thanks to the new fabrication methods, new tubes have been and are still developed:

1937	Klystron, Russell and Sigurd Varian [9];
1938	IOT, Andrew V. Haeff [10];
1939	Reflex klystron, Robert Sutton;
1940	Few commercial IOT;
1941	Magnetron, Randall and Boot [11];
1945	Helix travelling wave tube (TWT), Kompfner [12];
1948	Multi MW klystron;
1959	Gyrotron, Twiss and Schneider;
1963	Multi beam klystron, Zusmanovsky and Korolyov [13];
1980	High efficiency IOT.

3.2.1 Klystron

The klystron is built around a different concept than the grid tube. It uses the velocity modulation of an electron beam to generate high-power RF. The principle, here, is to convert the kinetic energy of the electrons into RF power. Looking at Fig. 16, we can identify the electron gun. It is composed of a thermionic cathode and an anode. We then have a drift space and, at the end, a collector. When applying all the voltages, an electron beam is generated and electrons fly with a constant speed from the gun to the collector through the drift space.

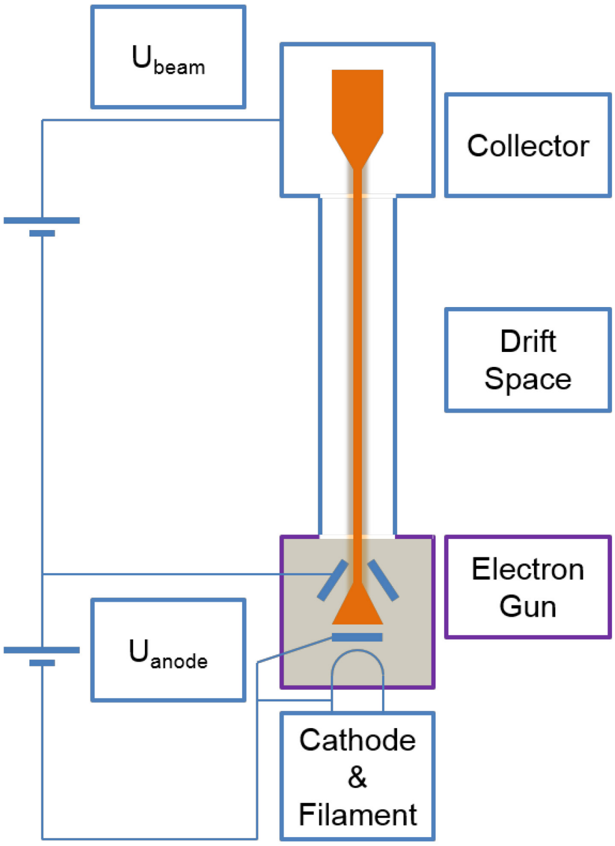


Fig. 16: The DC sketch of a klystron

In order to convert this constant electron flux into an RF-power generator, we add cavity resonators. The RF input cavity is the Buncher and the RF output cavity is the Catcher (Fig. 17).

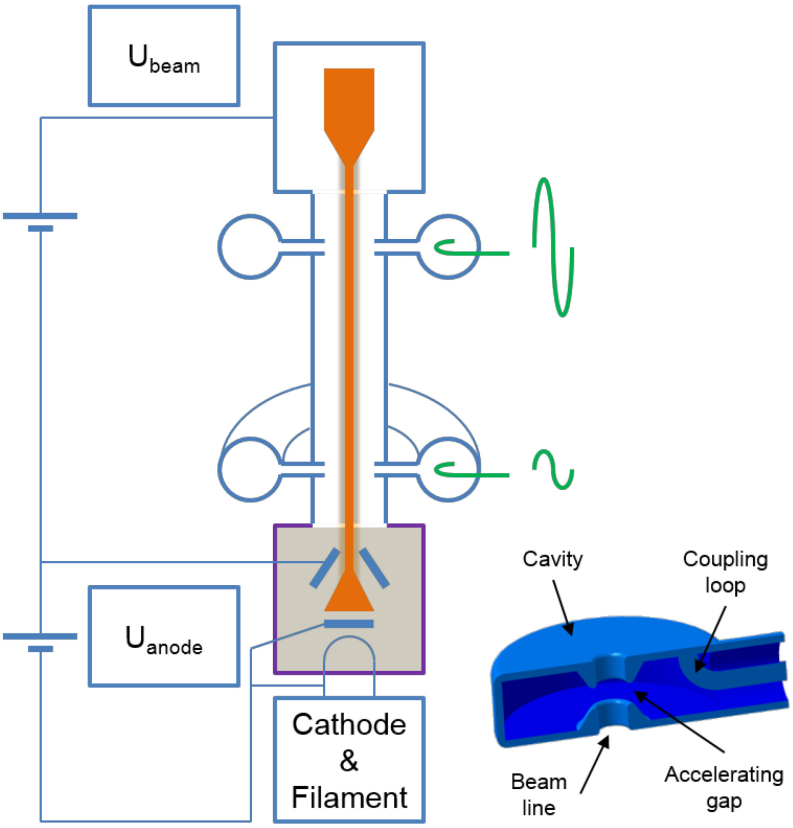


Fig. 17: The RF sketch of a klystron. The input cavity is the Buncher. The output cavity is the Catcher

The principle is the following. By applying RF on the Buncher, we modulate the speed of the electrons. Some electrons are accelerated, some are neutral, and some are decelerated. Figure 18 illustrates how we obtain the bunching of the electrons.

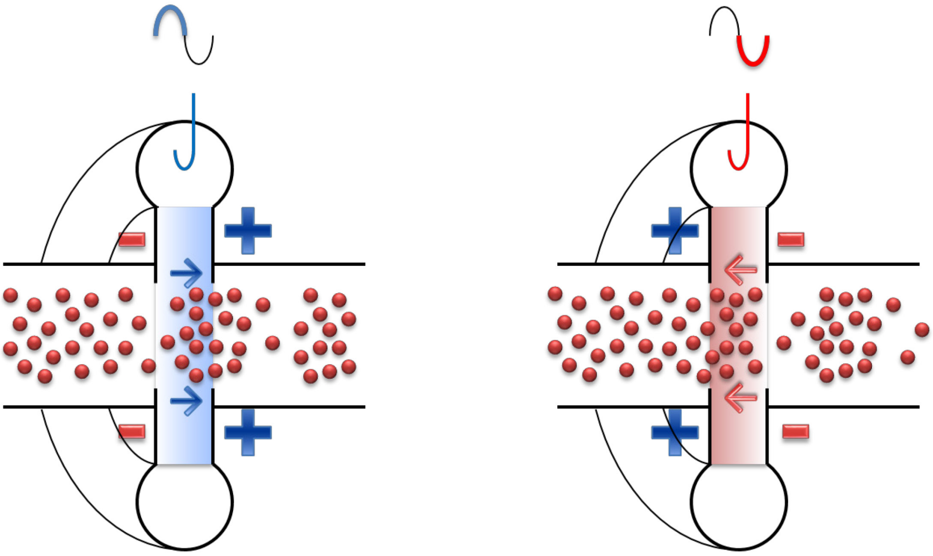


Fig. 18: Bunching of the electrons. On the left, when the voltage seen by the electrons at the Buncher cavity gap is positive, electrons are accelerated. On the right, when the voltage seen by the electrons at the Buncher cavity gap is negative, electrons are decelerated.

At the end of the electrons' journey, the Catcher cavity resonates at the same frequency as the input cavity. It is designed to be at the exact position with the maximum number of electrons. The kinetic energy of all these electrons is then converted into voltage and extracted from the output cavity. We then have an RF-power amplifier as shown in Fig. 19. Figure 20 shows the drift space distance between the Buncher cavity and Catcher cavity and how they must be spaced in order to maximize the efficiency of the tube.

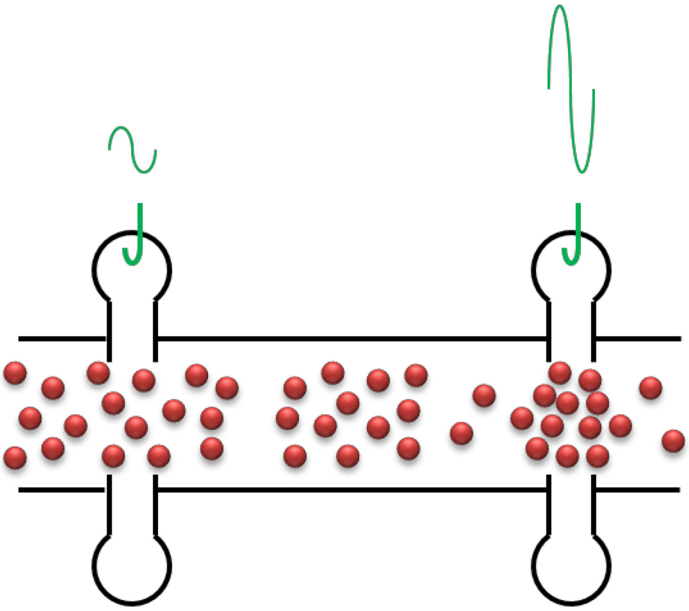


Fig. 19: Buncher and Catcher cavities. A constant electron flux before the Buncher cavity is transformed into bunched electrons at the Catcher cavity. The kinetic energy of these bunched electrons is converted into voltage and extracted from the Catcher cavity.

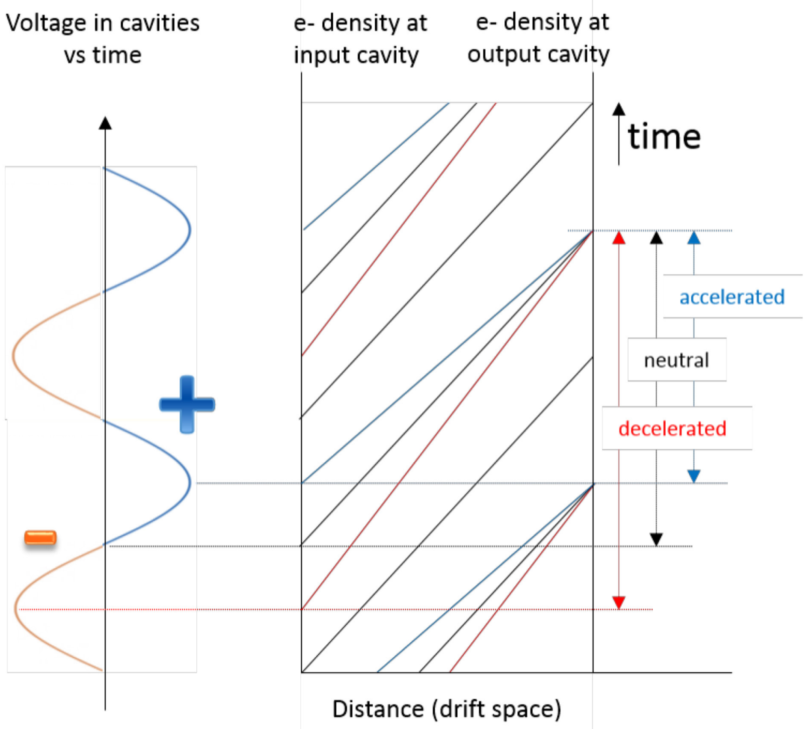

Voltage in cavities vs time
e- density at input cavity
e- density at output cavity
time
accelerated
neutral
decelerated
Distance (drift space)

Fig. 20: Bunching of electron beam in a klystron. Distance of the drift space allows for maximum electron density at the Catcher cavity plan.

In order to increase the gain of this two-cavity klystron, additional cavities, resonating with the pre-bunched electron beam, are added. These additional cavities generate additional accelerating and decelerating fields. They provide a better bunching, and it is commonly acknowledged that they provide around 10 dB gain per additional cavity.

In order to keep the beam correctly focused in the drift space, focusing magnets are mandatory. They ensure that the electron beam is maintained, as expected and where expected. Figures 21 and 22 show the CERN LHC based on TH2167 klystron amplifiers.

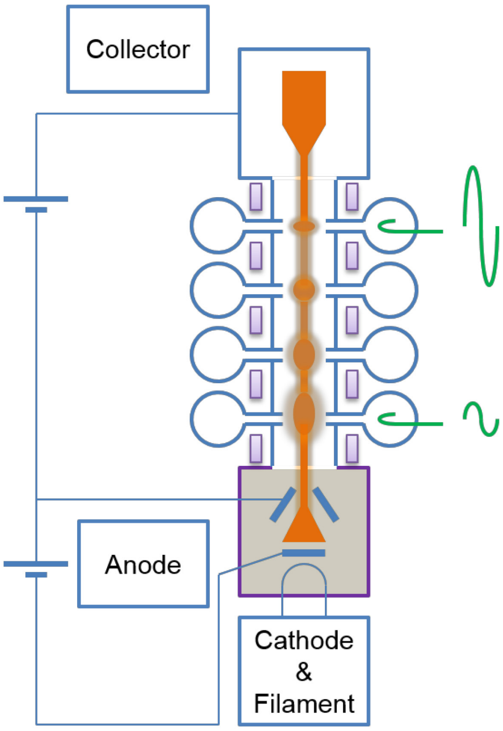


Fig. 21: Sketch of a klystron with four cavities and its focusing magnets. The gain of such a device would be around 40 dB.

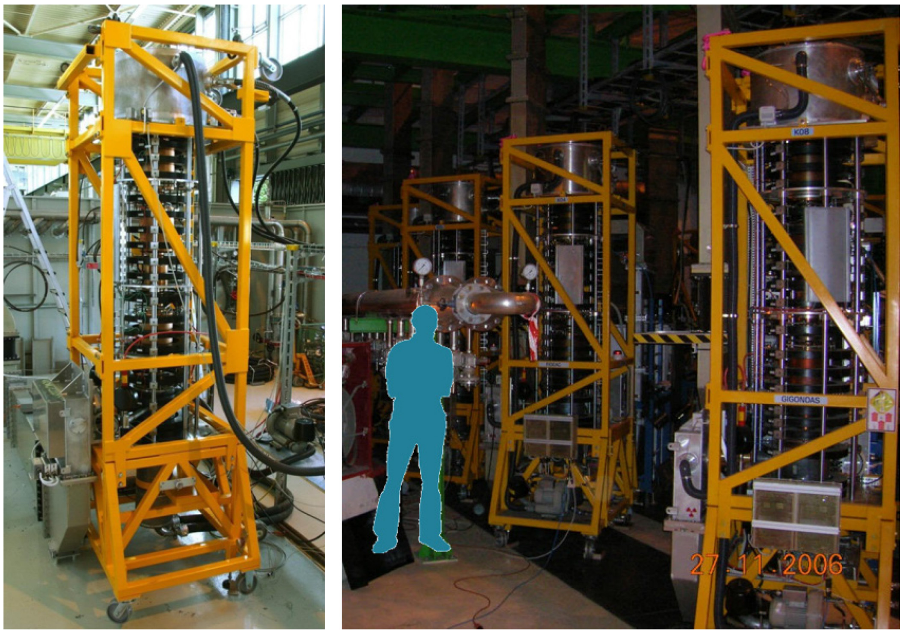


Fig. 22: CERN LHC, TH 2167 klystron. On the left, in lab, and on the right in UX45 LHC cavern,16 klystrons delivering 330 kW @ 400 MHz, in operation since 2008.

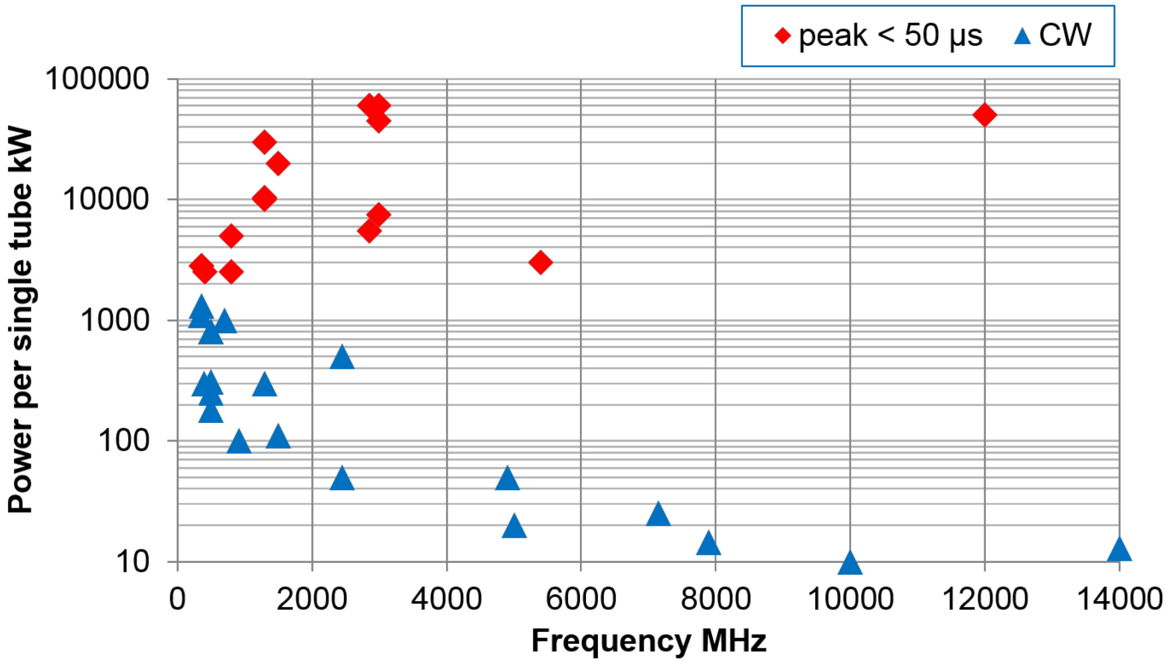


Fig. 23: Klystrons available from industry

Figure 23 summarizes all klystrons currently available from worldwide suppliers. We note that the maximum peak power is over 10 MW at low frequency. Continuous wave (CW) power decreases with frequency, and the frequency range is from a few MHz to several GHz.

3.2.2 Inductive output tube

The IOT is a mix between a triode and a klystron. Here, the principle is to modulate the density of an electron beam with a triode input. We recognize the thermionic cathode and the control grid that modulates the electron emission. On the output side, we have a simplified klystron circuit. We recognize the anode that accelerates the electron beam. Then, we have a short drift space, the Catcher cavity, and the collector. We also have a magnet to keep the beam as expected. Even though the IOT was invented at approximately the same time as the klystron, they were not used before the 1990s. Indeed, IOT gain is lower than the klystron, being approximately in the order of 23 dB, much lower than a five-cavity klystron that will have approximately 50 dB gain. In addition, it also requires a high-voltage power supply, as for the klystron, of around 30 kV to 50 kV, much higher than the 10 kV to 15 kV needed with a tetrode. For all these reasons, it was considered for a long time that IOTs were the sum of all the disadvantages of the tetrodes and of the klystrons. However, with the recent improvement in solid state amplifiers allowing a higher-power driver without tubes, they recently turned out to be an elegant solution where a single RF medium power source is needed. Since the beginning of the new century, they have been implemented in several laboratories around the world. Figure 25 shows the CERN SPS based on TH795 IOT amplifiers.

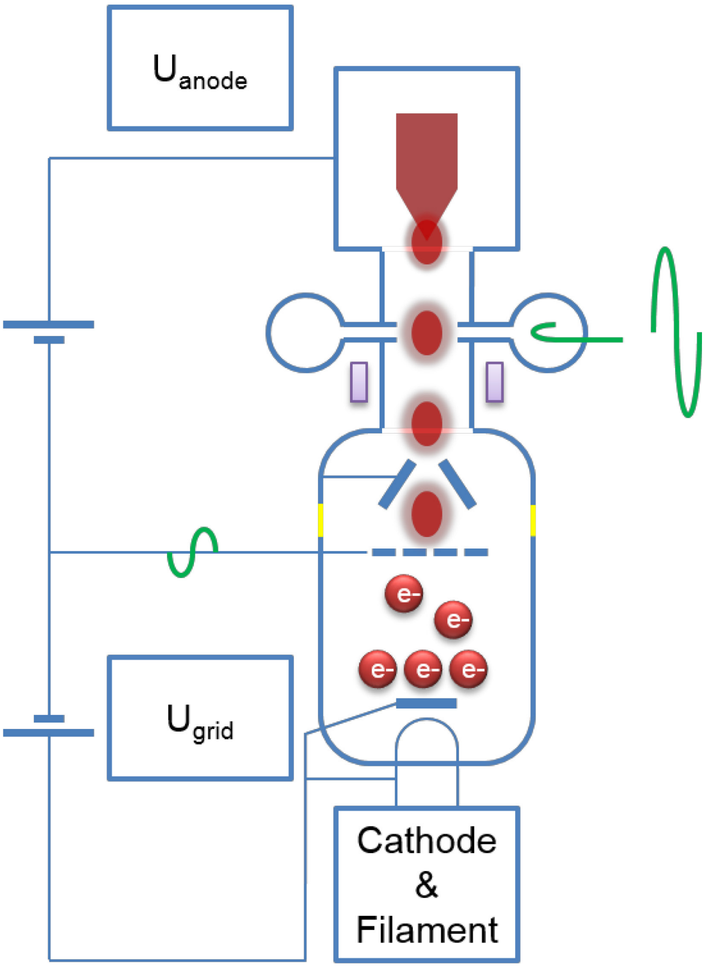
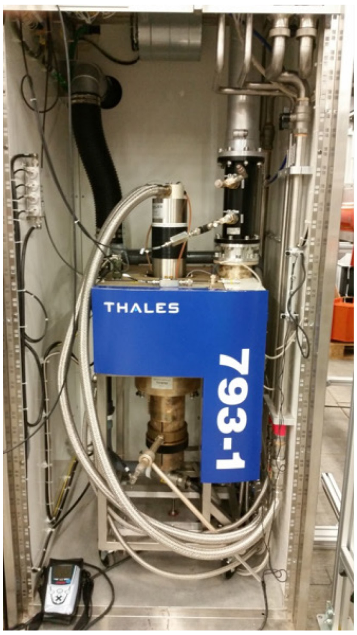


Fig. 24: Sketch of an IOT



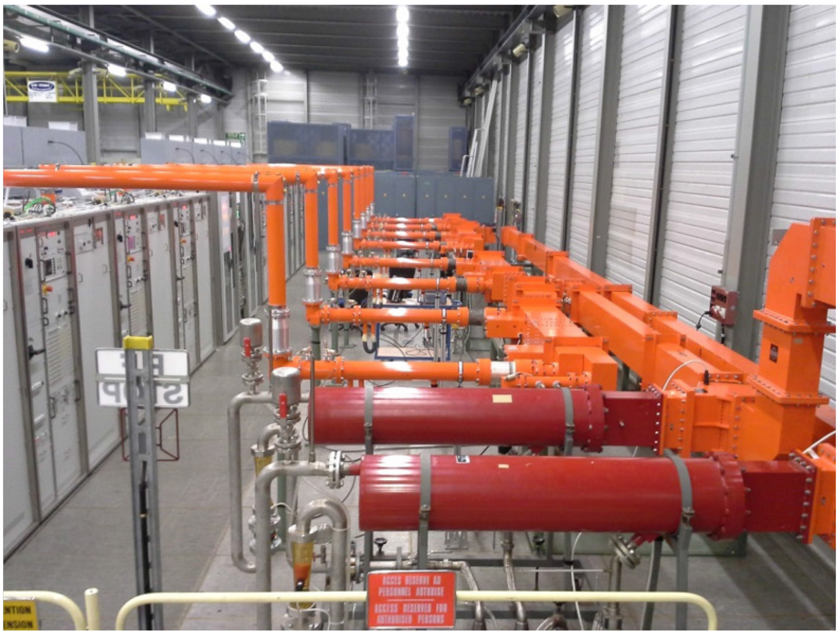


Fig. 25: CERN SPS, TH 795 IOT, trolley (single amplifier), and transmitter (combination of amplifiers). Two transmitters of four tubes delivering 2 x 240 kW @ 801 MHz, in operation since 2014.

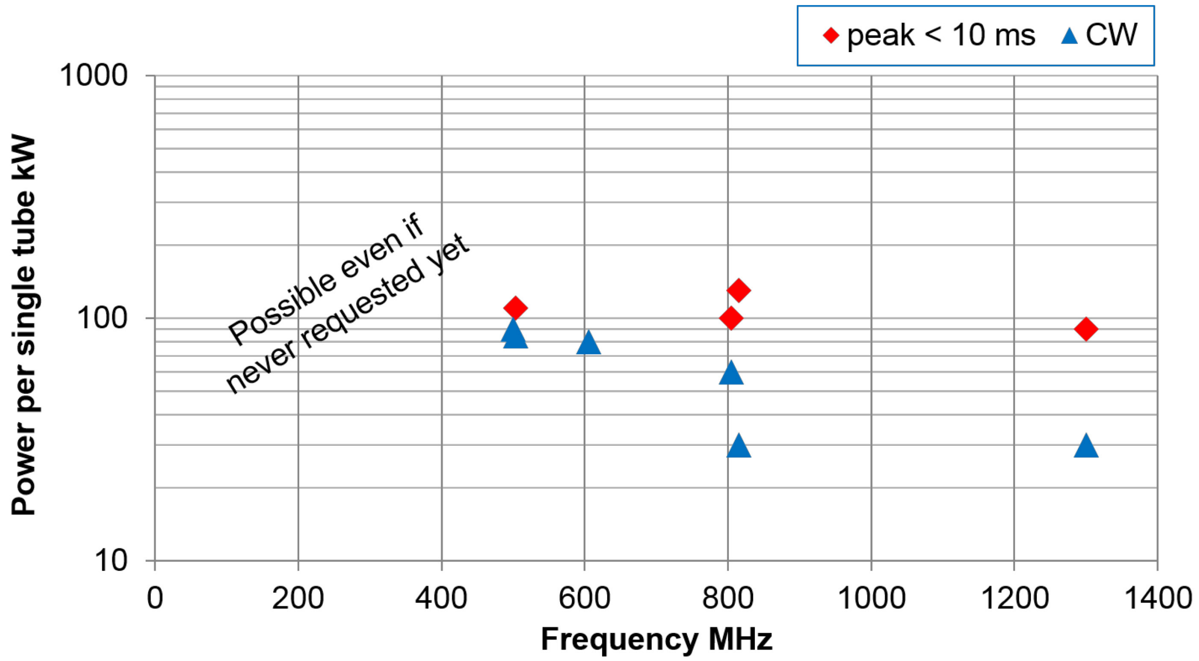


Fig. 26: IOTs available from industry

As we can see in Fig. 26, there are not a lot of IOTs available on the market. However, they offer a lot of possibilities in the range of few a MHz to 1.5 GHz with a power level of 20 kW to 100 kW.

3.3 Transistors

The last technology discussed in this paper is the SSA (Fig. 27). Although the theory was developed at almost the same time as the tubes, the construction capability came much later. Since the middle of the last century, transistors have never stopped improving. With the arrival of the mobile telephone and the digital TV broadcast, transistors have been developed in huge quantities and with increased power capabilities. They are still improving considerably, and new materials are very promising, allowing incredibly high power levels per single unit to be reached. The following list contains the main developments:

1925	Theory, Julius Edgar Lilienfeld [14];
1947	Germanium US first transistor, John Bardeen, Walter Brattain [15];
1948	Germanium European first transistor, Herbert Mataré and Heinrich Welker [16];
1953	first high-frequency transistor, Robert Wallace [17];
1954	Silicon transistor [18];
1960	Metal Oxide Semiconductor (MOS) [19];
1966	Gallium arsenide (GaAs) [20];
1980	Vertical Diffused MOS (VDMOS) [21];
1989	Silicon-Germanium (SiGe);
1997	Silicon carbide (SiC) [22];
2004	Carbon graphene [23].

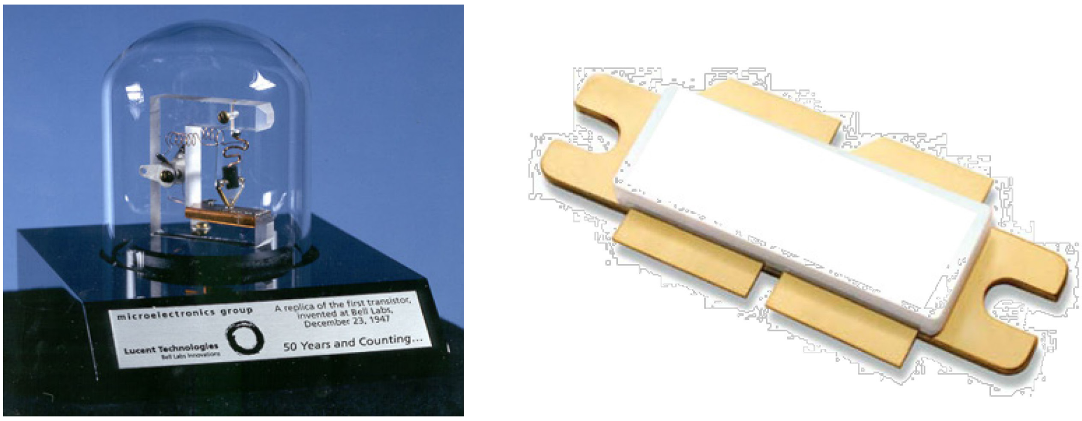


Fig. 27: From the first Germanium transistor in 1947 to recent LDMOS transistors in the 2000s

The conventional amplifier circuitry with transistors is the push–pull amplifier. In a push–pull circuit, the RF signal is applied to two devices. One of the devices is active on the positive voltage swing and off during the negative voltage swing. The other device works in the opposite manner so that the two devices conduct half the time. The full RF signal is then amplified. One of the main difficulties in making an RF amplifier, with this circuitry, is the use of two different types of device, one negative-positive-negative (NPN) transistor and one positive-negative-positive (PNP) transistor. Intrinsic differences in the devices will naturally introduce disturbances. Figure 28 describes such a circuit.

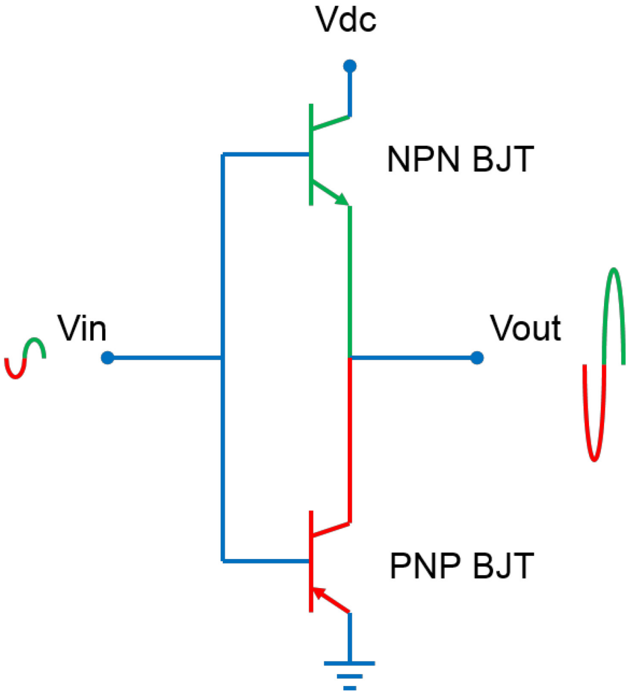


Fig. 28: Bipolar junction transistor (BJT) push–pull circuit

Another push–pull configuration is the balun (balanced–unbalanced) circuit (Fig. 29). Such a circuit acts as a power splitter, equally dividing the input power between the two transistors. The balun keeps one port in phase and inverts the second port in phase. As the signals are out of phase, only one device is on at a time. This configuration is easier to manufacture since only one type of device is required, and so, if the balun is correctly calculated, no disturbances are generated. This is the way most of the RF SSA are designed.

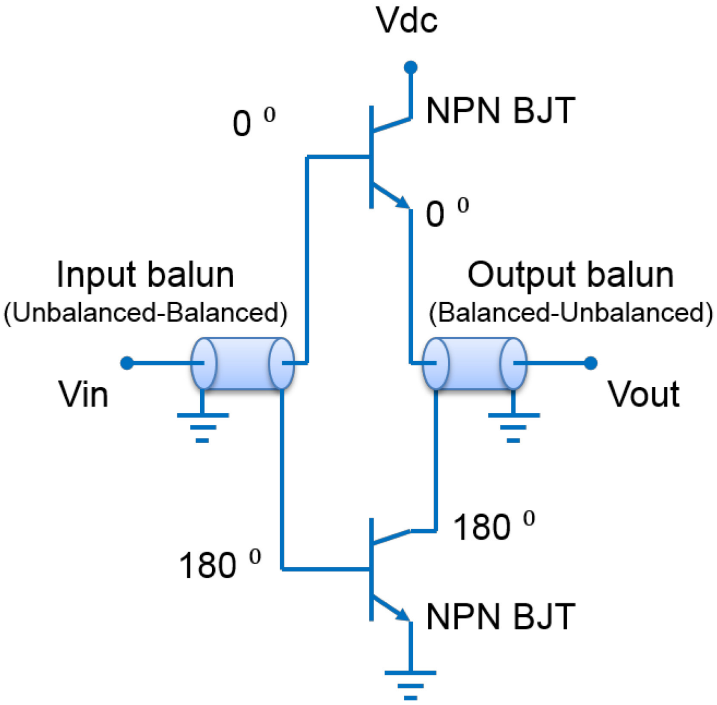


Fig. 29: Transistor balun circuit

The power level per unit is quite small compared to vacuum tubes. Figure 30 shows the transistors available on the market, and we can see that the frequency range starts as with the tubes and extends further compared to the tetrode, and even compared to the IOT.

If we want to build an RF high-power amplifier, we will have to combine transistors together (see Section 4). A comparison of 100 transistors with grid tubes and IOTs shows that the power level per unit is within the same power-level range. Figure 31 shows the achievable power levels from 100 combined transistors.

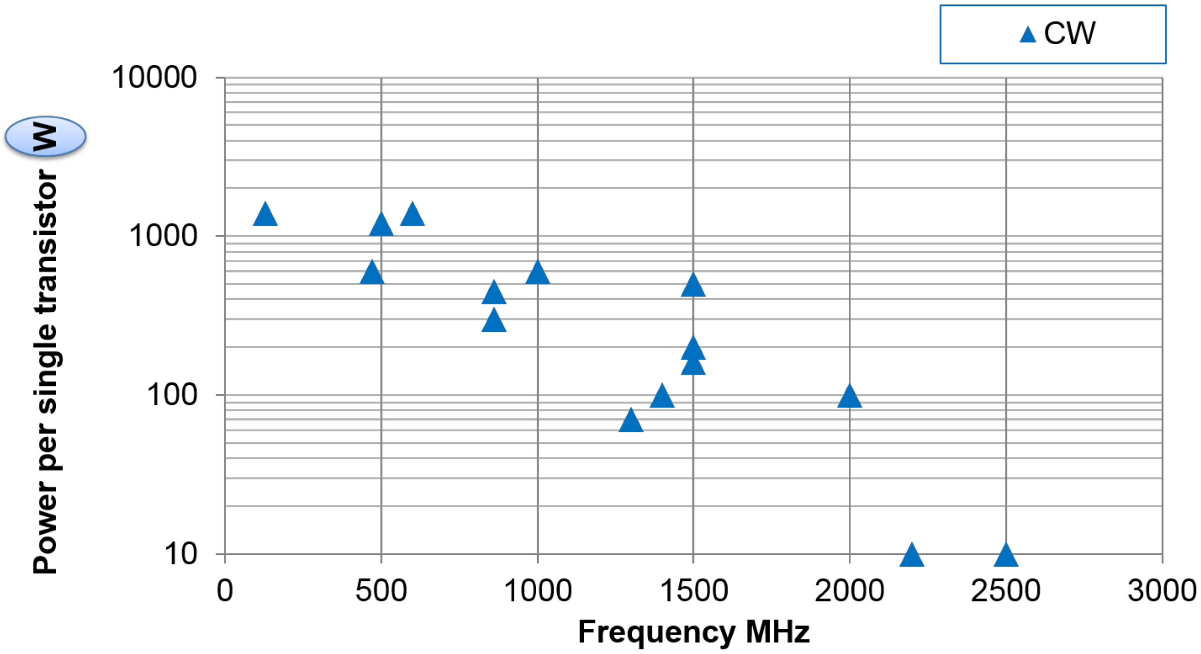


Fig. 30: Transistors available from industry

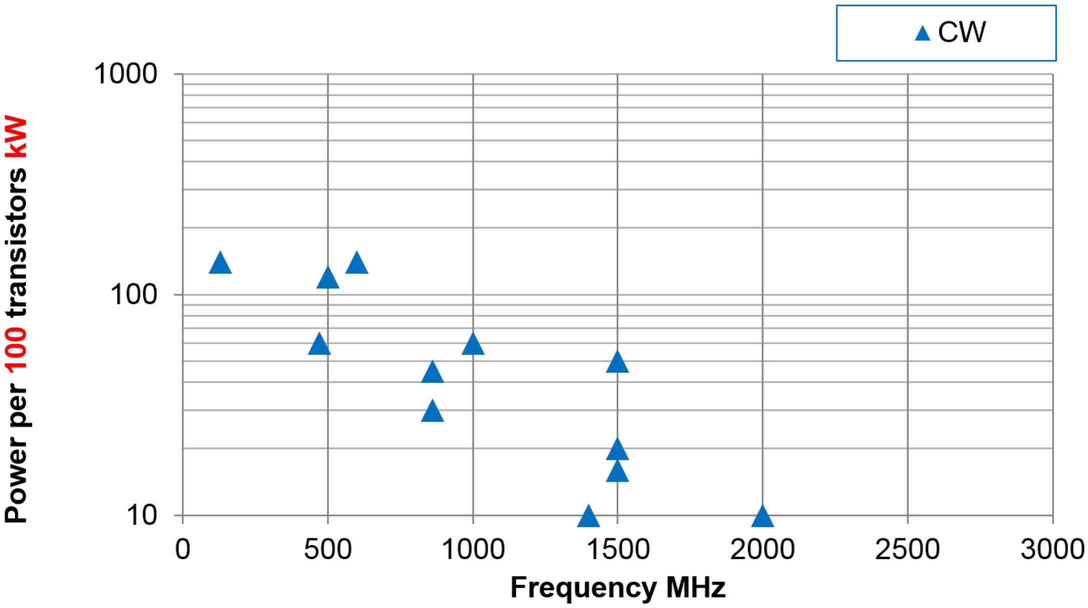


Fig. 31: Power level available from combining 100 transistors

3.4 Power overhead

Once we have defined a power system, overheads will have to be taken into consideration. Indeed, losses induced by the transmission lines, discrepancies between single units, the need for Low-Level RF (LLRF) regulations, and the fluctuation of electrical mains must be anticipated. It is commonly acknowledged that klystrons are operated at 30% (approximately –1 dB to –1.5 dB) below their maximum ratings in order to avoid running in saturation mode. Because SSPA are operated at only 10% below their maximum ratings, thanks to the granularity they offer, a fault does not affect their overall parameters so much. Tetrodes and IOTs are operated at 20% below their maximum ratings, keeping in mind that these grid tubes can be operated well over the nominal characteristic levels in pulsed mode. Figure 32 illustrates these limitations.

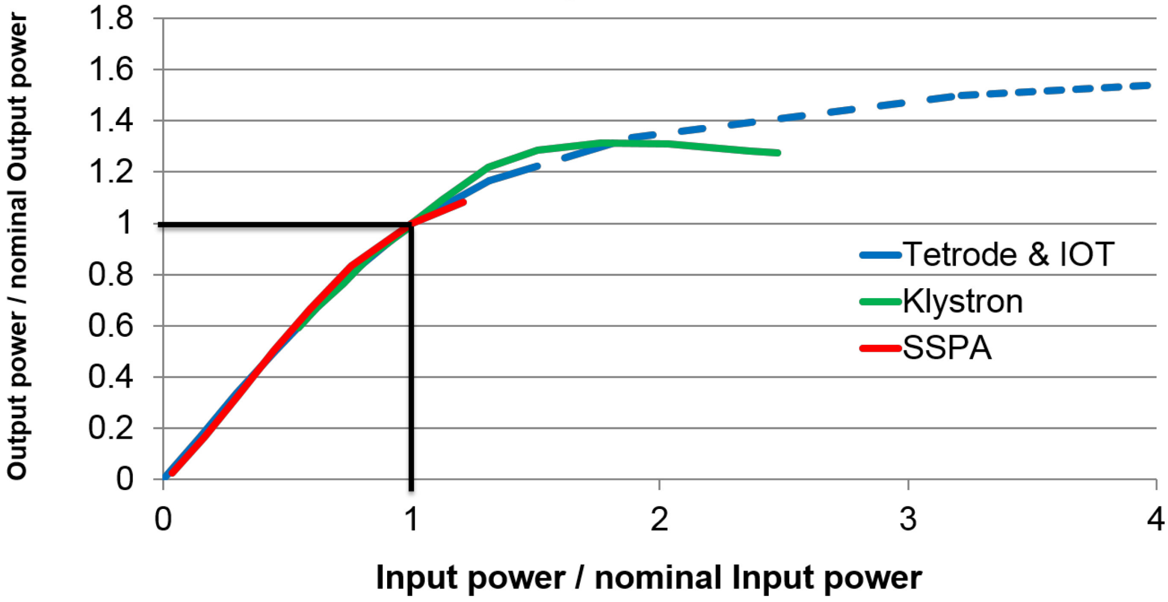


Fig. 32: Normalized power capability over nominal power level. SSPA cannot be operated much above their nominal ratings without being damaged. Klystrons saturate above 30 % above their nominal ratings. Grid tubes, including IOTs, allow operation above nominal ratings, even much higher in short-pulse mode.

3.5 Combiners and splitters

Once the RF-power amplifier source has been selected, it could be necessary to sum, or to divide, the output power of the device (Fig. 33). In RF, most of the power-combiner and power-splitter devices are reversible.

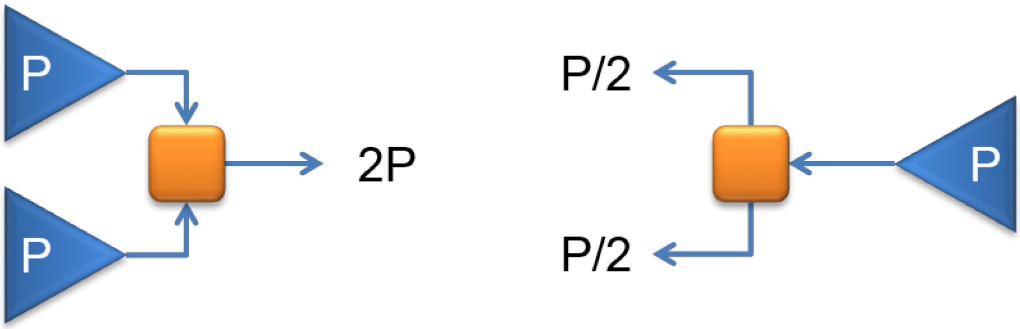


Fig. 33: The same device can be used as a power combiner or as a power splitter

The cheapest and easiest combiners to build are the resistive power splitters and combiners. In order to keep the correct impedance seen by all ports, they are built from resistors. Unfortunately, they are not really suited for high-power applications due to the power limitation of the resistor and to losses induced by the resistors.

The commonly used devices for power applications are the hybrid combiners. They are built from RF transmission lines and provide low losses. The power limitation of such combiners and splitters is mainly governed by the size of the lines themselves. A perfect 3 dB phase combiner (Figs. 34 and 35), with correct input phases, will allow the same power applied on each input port to be summed:

$$\Sigma = \frac{P_1 + P_2}{2} + \sqrt{P_1 P_2} \, , \tag{11}$$

$$\Delta = \frac{P_1 + P_2}{2} - \sqrt{P_1 P_2} \, . \tag{12}$$

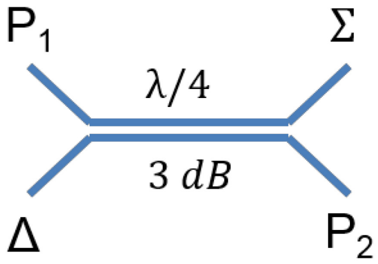


Fig. 34: Configuration of the 3 dB phase combiner

Correctly adjusting the phase and the gain, $P_1 = P_2 = P$:

$$\Sigma = \frac{P + P}{2} + \sqrt{PP} = 2\,P \, , \tag{13}$$

$$\Delta = \frac{P+P}{2} - \sqrt{PP} = 0\,. \tag{14}$$

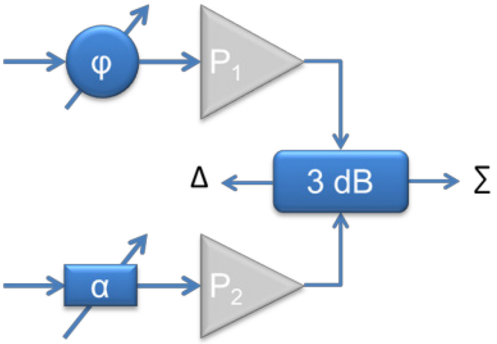


Fig. 35: With a phase shifter on one input line and an attenuator on the second input line, phase and gain can be adjusted to obtain a perfect 3 dB combiner.

Figure 36 shows one of the CERN SPS combiner operating at 200 MHz.

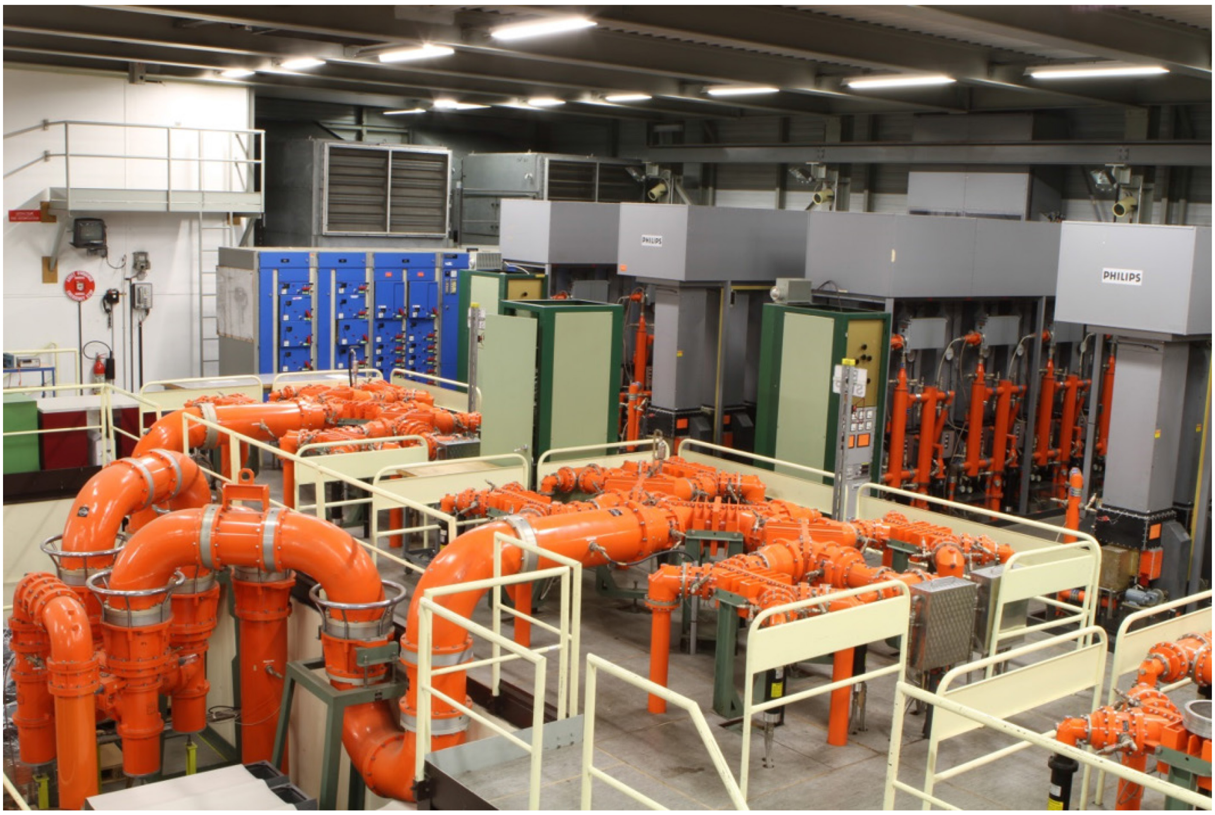


Fig. 36: CERN SPS 64 to 1 combiner @ 200 MHz

Another way to make a combiner (or a splitter) is the low-loss T-Junction as shown in Fig. 37. With $Z_{\lambda/4} = Z_c\sqrt{N}$, we obtain an N-ways splitter.

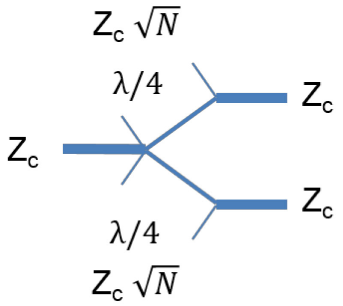


Fig. 37: The T-junction configuration

4 RF power lines

Once we have the required RF output power, it has to be transported from the RF-amplifier output to the load. Several transmission lines exist. The main lines that are used in high-power RF are the rectangular waveguides (Fig. 38) and the coaxial lines.

4.1 Rectangular waveguides

The main advantage of the waveguides is that waveguides provide propagation with low loss.

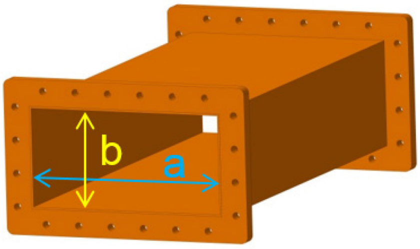


Fig. 38: Example of a rectangular waveguide. The size ***a*** is the width, and the size ***b*** is the height

The main parameters of a rectangular waveguide are given by the following formulas:

waveguide wavelength
$$\lambda_{\mathrm{g}} = \frac{\lambda}{\sqrt{1-\left(\frac{\lambda}{2a}\right)^2}}, \tag{15}$$

cut-off frequency dominant mode
$$f_{\mathrm{c}} = \frac{c}{2a}, \tag{16}$$

cut-off frequency next higher mode
$$f_{c2} = \frac{c}{4a}, \tag{17}$$

usable frequency range
$$1.3\ f_{\mathrm{c}} \text{ to } 0.9\ f_{c2}. \tag{18}$$

From the transmission line, Eqs. (15)–(18), we can see that the waveguides are usable only over certain frequency ranges. For very low frequencies, the waveguide dimensions become impractically large. For very high frequencies, the waveguide dimensions become impractically small and the manufacturing tolerance becomes a significant portion of the waveguide size. Figure 39 lists some of the currently used waveguide sizes. The EIA standard names the waveguides with respect to their width, so a WR2300 waveguide has a width of 23.00 inches. This is very convenient to quickly identify the size and the reference of the waveguides. It is also common to have a half-height waveguide, when the RF power is not too high. Indeed, as it can been seen within Eqs. (15)–(17), the height is not providing any limitation.

Waveguide name			Recommended frequency band of operation (GHz)	Cutoff frequency of lowest order mode (GHz)	Cutoff frequency of next mode (GHz)	Inner dimensions of waveguide opening (inch)
EIA	RCSC	IEC				
WR2300	WG0.0	R3	0.32 — 0.45	0.257	0.513	23.000 × 11.500
WR1150	WG3	R8	0.63 — 0.97	0.513	1.026	11.500 × 5.750
WR340	WG9A	R26	2.20 — 3.30	1.736	3.471	3.400 × 1.700
WR75	WG17	R120	10.00 — 15.00	7.869	15.737	0.750 × 0.375
WR10	WG27	R900	75.00 — 110.00	59.015	118.03	0.100 × 0.050
WR3	WG32	R2600	220.00 — 330.00	173.571	347.143	0.0340 × 0.0170

Fig. 39: Some of the commonly used standard waveguides

The peak-power limitation for such waveguides is given by the following formula:

$$P = 6.63\times10^{-4}\ E_{\max}{}^{2}\sqrt{b^{2}\left(a^{2}-\frac{\lambda^{2}}{4}\right)}, \tag{19}$$

with

P = Power in watts,

a = width of waveguide in cm,

b = height of waveguide in cm,

λ = free space wavelength in cm, and

$E_{\max}$ = breakdown voltage gradient of the dielectric filling the waveguide in V/cm (for dry air 30 kV/cm, for ambient air 10 kV/cm).

Looking at this formula, we can see that each waveguide size will have its own characteristics. Figure 40 illustrates the peak power limitation per waveguide size and Fig. 41 illustrates the attenuation per waveguide size.

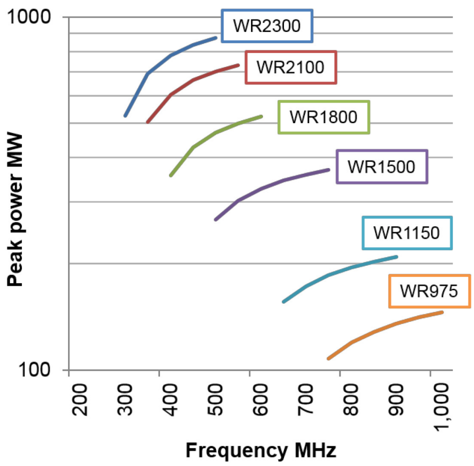


Fig. 40: RF frequency range and peak power per waveguide size

The attenuation of the line is dependent on the geometry, including the height, of the material:

$$\text{attenuation} = \frac{4a_0}{a}\frac{\sqrt{c/\lambda}}{\sqrt{1-(\lambda/2a)^2}}\left(\frac{a}{2b}+\frac{\lambda^2}{4a^2}\right), \tag{20}$$

with

$a_0 = 3\times10^{-7}$ dB/m, for copper,

a = width of waveguide in meters,

b = height of waveguide in meters, and

λ = free space wavelength in meters.

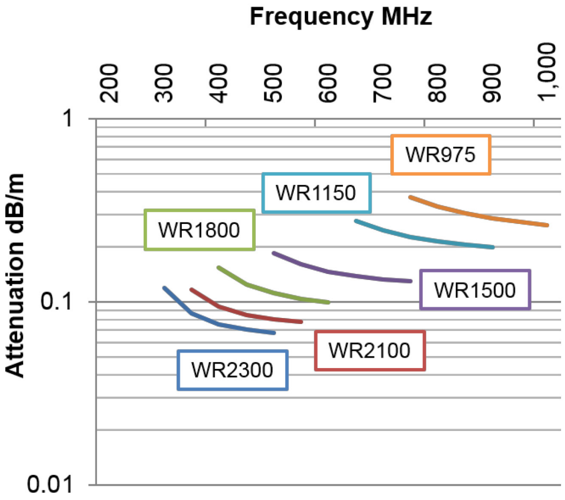


Fig. 41: Attenuation of a copper waveguide full height size

4.2 Coaxial lines

As we have seen, for lower frequencies, the waveguide dimensions become impractically large. The coaxial lines are then one of the most commonly used solutions. The characteristic impedance of a coaxial line is given by the following formula:

$$Z_\mathrm{c} = \frac{60}{\sqrt{\varepsilon_\mathrm{r}}}\ln\left(\frac{D}{d}\right), \tag{21}$$

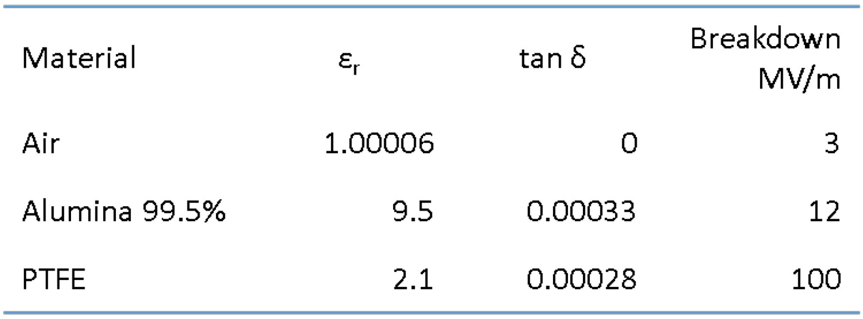
with

D = inner dimension of the outer conductor,

d = outer dimension of the inner conductor, and

ε_r = dielectric characteristic of the medium.

The dielectric characteristic of the medium plays a very important role in a coaxial line and it is not to be neglected. Indeed, coaxial cables often have PTFE foam to keep concentricity, flexible line spacers helicoidally placed all along the line, and rigid lines made of two rigid tubes with supports to maintain concentricity. Regarding the spacers used (Fig. 42), the size of the inner and outer diameters will have to be compensated.



Material	ε_r	tan δ	Breakdown MV/m
Air	1.00006	0	3
Alumina 99.5%	9.5	0.00033	12
PTFE	2.1	0.00028	100

Fig. 42: Some of the commonly used standard coaxial lines

Power handling of a coaxial line is related to the medium (Fig. 43) and to the breakdown field *E*. The peak-power limitation for such coaxial lines is given by the following formulas:

$$V_{\text{peakmax}} = E\frac{d}{2}\ln\left(\frac{D}{d}\right), \tag{22}$$

$$P_{\text{peakmax}} = \frac{V_{\text{peakmax}}{}^{2}}{2Z_{\text{c}}}, \tag{23}$$

$$P_{\text{peakmax}} = \frac{E^2\ d^2\sqrt{\varepsilon_{\text{r}}}}{480}\ln\left(\frac{D}{d}\right), \tag{24}$$

with

E = breakdown strength of air ('dry air' *E* = 3 kV/mm, commonly used value is *E* = 1 kV/mm for ambient air),

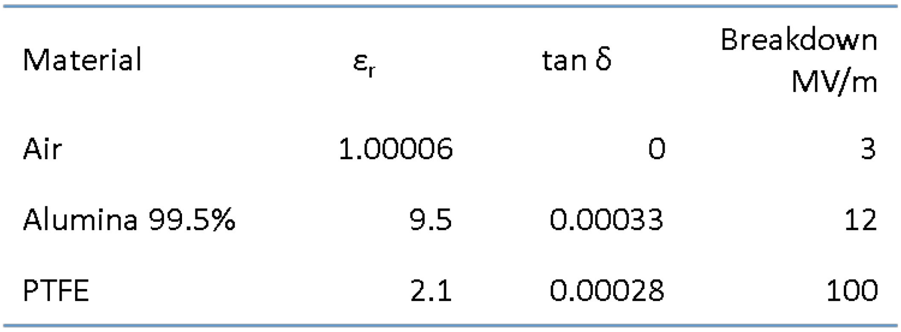
D = inside electrical diameter of outer conductor in mm,

d = outside electrical diameter of inner conductor in mm,

Z_{c}= characteristic impedance in Ω,

ε_{r} = relative permittivity of dielectric, and

f = frequency in MHz.



Material	ε_r	tan δ	Breakdown MV/m
Air	1.00006	0	3
Alumina 99.5%	9.5	0.00033	12
PTFE	2.1	0.00028	100

Fig. 43: Peak power capability of coaxial lines strongly depends on the medium

The attenuation of a coaxial line can be approximated with the following expression:

$$\alpha = \left(\frac{36.1}{Z_{\text{c}}}\right)\left(\frac{1}{D}+\frac{1}{d}\right)\sqrt{f} + 9.1\ \sqrt{\varepsilon_{\text{r}}}\ \tan\delta\, f\ , \tag{25}$$

where

α = attenuation constant, dB/m,

Z_c= characteristic impedance in Ω,

f = frequency in MHz,

D = inside electrical diameter of outer conductor in mm,

d = outside electrical diameter of inner conductor in mm,

ε_r = relative permittivity of dielectric, and

tan δ = loss factor of dielectric.

When selecting the coaxial line, the best compromise has to be made with respect to the needs of the project between size, peak power capability, and attenuation. It is always very important to take all the parameters into consideration and to remember that the power limitations given by the suppliers must be carefully, and strictly, followed. Damage to a coaxial line, by mechanical deformation or by overheating, will change its impedance characteristic, and this will considerably reduce its power-handling capability.

4.3 Reflection from load and circulator

A major phenomenon that has to be considered when selecting the correct transmission line is the matching of the impedance. The standing wave ratio (SWR) is a measure of impedance matching of the Device under test (DUT). A wave is partly reflected when a transmission line is terminated with anything other than a pure resistance equal to its characteristic impedance (Figs. 44 and 45).

The reflection coefficient is defined by

$$\Gamma = \frac{V_r}{V_f} \quad . \tag{26}$$

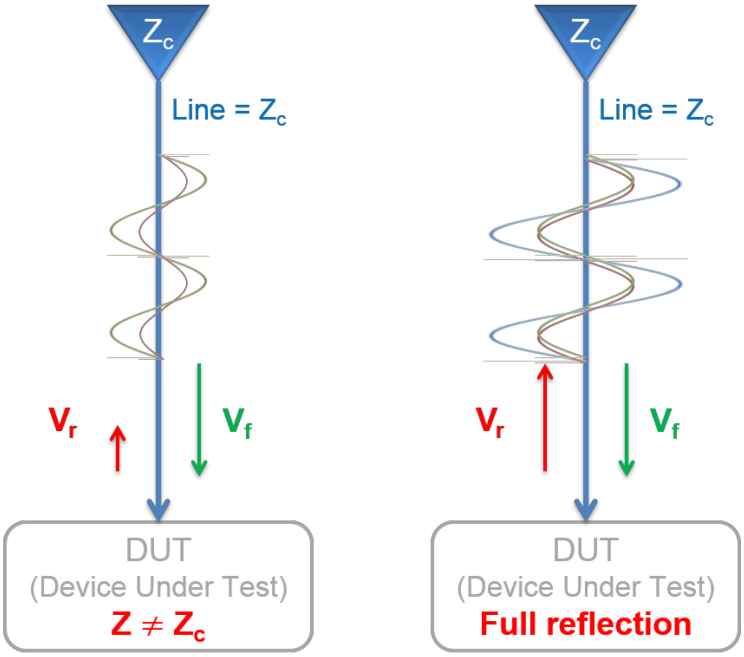


Fig. 44: Forward and reflected waves regarding the impedance of the DUT

Γ = -1	when the line is short-circuited complete negative reflection
Γ = 0	when the line is perfectly matched, no reflection
Γ = 1	when the line is open-circuited complete positive reflection

Fig. 45: SWR regarding the impedance of the DUT

At some points along the line the forward and reflected waves are exactly in phase, and then

$$|V_{\max}| = |V_{\mathrm{f}}| + |V_{\mathrm{r}}| = |V_{\mathrm{f}}| + |\Gamma V_{\mathrm{f}}| = \left(1+|\Gamma|\right)|V_{\mathrm{f}}|. \tag{27}$$

In the case of full reflection,

$$|V_{\max}| = 2\,|V_{\mathrm{f}}|. \tag{28}$$

At other points the forward and reflected waves are 180° out of phase, and then

$$|V_{\min}| = |V_{\mathrm{f}}| - |V_{\mathrm{r}}| = |V_{\mathrm{f}}| - |\Gamma V_{\mathrm{f}}| = \left(1-|\Gamma|\right)|V_{\mathrm{f}}|. \tag{29}$$

In the case of full reflection,

$$|V_{\min}| = 0. \tag{30}$$

The voltage standing wave ratio (VSWR) is defined by

$$\mathrm{VSWR} = \frac{|V_{\max}|}{|V_{\min}|} = \frac{1+|\Gamma|}{1-|\Gamma|}. \tag{31}$$

So, we have seen that in case of full reflection (28), $V_{\max} = 2\ V_{\mathrm{f}}$. This means that $P_{\max}$ is equivalent to 4 P_{f}. Here, we have to be careful, as RF-power people are often stating that the maximum power 'is' four times greater in the transmission line, in the case of full reflection. This is an abuse of language, as it is only the voltage that is doubled, and the current that is doubled, but both are not in phase under these full reflection conditions. We must remain very careful and clearly state that maximum power is equivalent to four times the forward power. An explanation is that all the datasheets from the suppliers are given in power, not in voltage. So, if we want to operate a system that has to sustain full reflection, we have to select the correct line from the supplier taking into account four times the forward power.

In any case, RF-power amplifiers will not like this reflected wave. Klystron output cavities are disturbed and grid tube, IOT, and transistor voltage capabilities have to be fully respected, so as not to damage the devices. A swift protection, if $P_{\mathrm{r}} > P_{\mathrm{rmax}}$, has to be implemented. Unfortunately, this solution, as illustrated in Fig. 46, makes the system non-operational, which is not always acceptable.

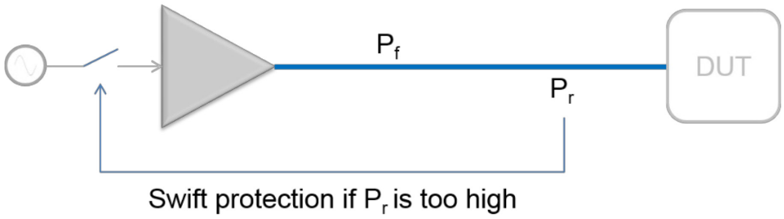


Fig. 46: Swift protection to protect the DUT

In order to protect our lines and our amplifiers from this reflected power, a specific component, the circulator, has been developed. This is a passive, non-reciprocal, three-port device. As shown in Fig. 47, the signal entering any port is transmitted only to the next port in rotation. The best place to insert it is close to the reflection source. The lines between the circulator and the DUT shall sustain four times P_f in case of full reflection. A load of P_f is needed on port 3 to absorb P_r.

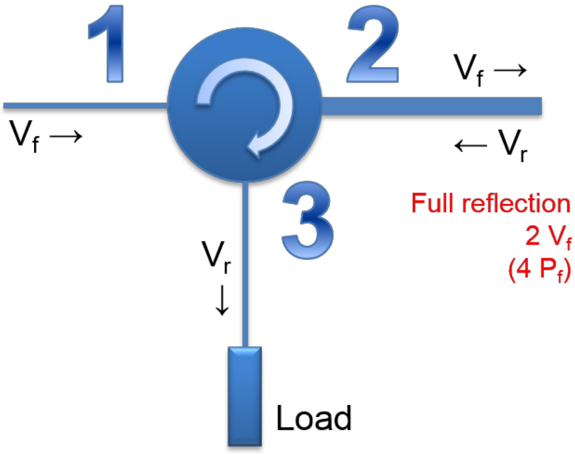


Fig. 47: Circulator basic principle. A device to correctly protect the DUT

Even in the case of full reflection, V_{max} = 2 V_f, and so P_{max} is equivalent to 4 P_f, the RF-power amplifiers will not see the reflected power and will not be affected. The lines between circulator and DUT must, at least, be designed for 4 P_f and the load must be designed for P_f. The main advantage of such a configuration, shown in Fig. 48, is that the system always remains operational.

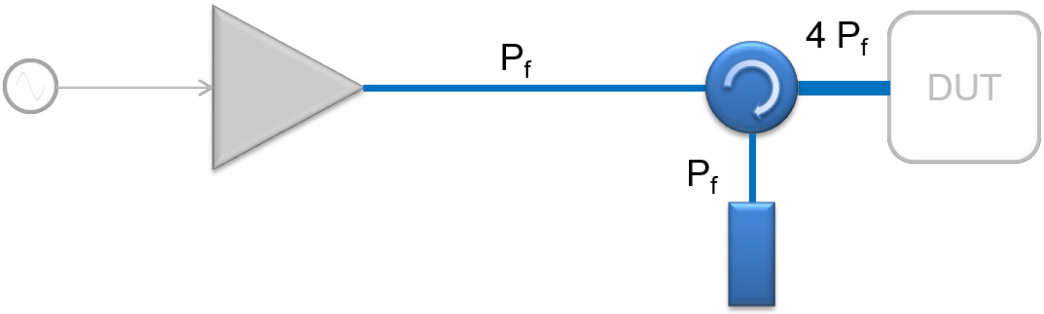


Fig. 48: A system protected with a circulator remains operational at any time

5 Fundamental Power Coupler

To complete the description of the transmission power chain, there is a final device that ensures the transfer of power from the line to the DUT: the Fundamental Power Coupler (FPC). The FPC is the connecting part between the RF transmission line and the RF cavity. It is a specific piece of transmission line that also has to provide the vacuum barrier for the beam vacuum. FPCs are one of the most critical parts of the RF-cavity system in an accelerator. A good RF design, a good mechanical design, and a high-quality fabrication are essential for efficient and reliable operation. Illustrated in Fig. 49 are some of the FPCs recently designed at CERN.

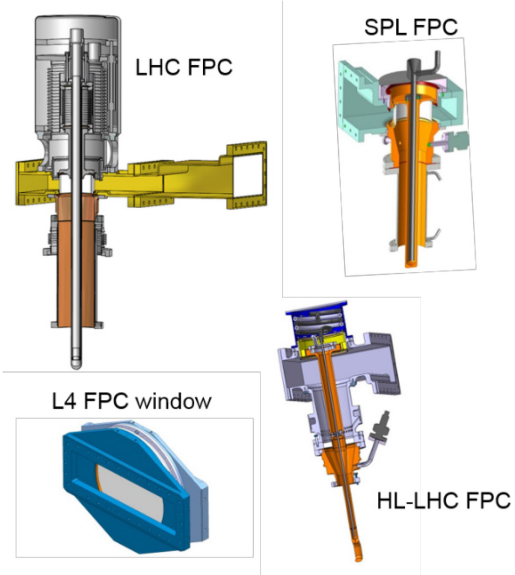


Fig. 49: Various CERN FPCs developed in the past few years

It is a large topic that deserves an entire document to address it. More details can be found in previous CAS listed in the References.

6 Case study for medical application

In order to help students to define what would be the best RF-power system, here are few illustrations to summarize the main parameters that are:

- frequency;
- overhead, peak, and average power;
- efficiency;
- rough cost estimate.

One of the first questions to address, that will define the choice of the technology, is what will be the operating frequency, or frequencies, of your machine. Figure 50 sums all the technologies available within a single plot.

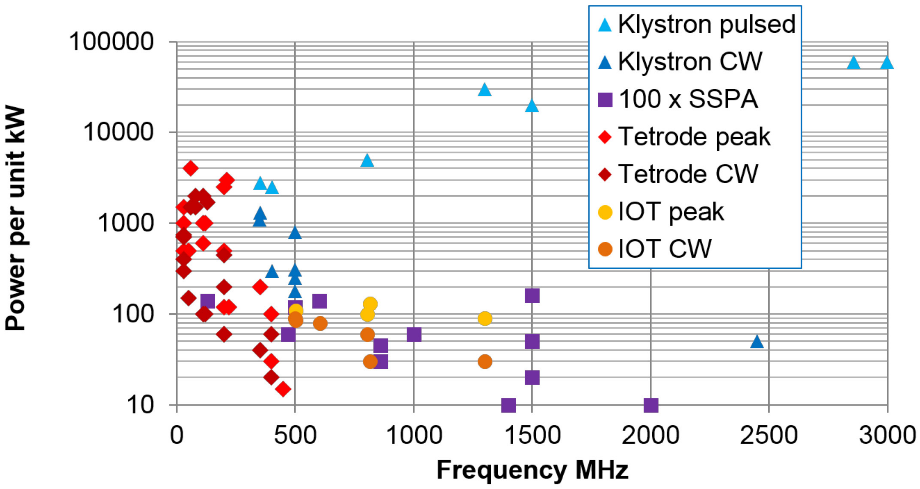


Fig. 50: All available RF-power sources

The second question to be answered is the overhead needed for your project, keeping in mind that one additional dB will directly add an additional 25% to the cost of an RF-power station. Peak or CW operation will also drive the selection of the RF-power source. Grid tubes, including IOTs, offer the capability of a pulsed system at much lower overhead cost, especially in short-pulse-mode operation. Figure 32 shows the overhead ratio commonly applied in the RF-power station definition for reliable operation.

Another very important issue is the overall efficiency. Usually the following numbers are applied.

- $P_{\mathrm{RF_{in}}} \simeq$ 1 to 5% $P_{\mathrm{RF_{out}}}$ as the gain of the last amplifier stage is usually high compared to its driver stage.
- $\eta_{\mathrm{RF/DC}} \simeq$ 65% including overhead.
- $\eta_{PAC/PDC} \simeq$ 95% to98 % as power converters are of very high-quality nowadays. If the system is a pulsed system, with klystrons, this parameter has to be corrected to a lower value. Indeed, high voltage will have to be modulated and an additional fraction of the powering will be lost as the voltages must be established before the RF is applied. Figure 52 describes the phenomenon.
- Amplifier cooler $\simeq$ 15% $P_{\mathrm{RF_{out}}}$. This reduces considerably the overall efficiency, and it shall not be neglected.
- Building cooler $\simeq$ 30% $P_{\mathrm{RF_{out}}}$. This is one of the most costly parameters.

Finally the overall efficiency, also described in Fig. 51, is given by the following formula:

$$\text{overall efficiency} = \frac{P_{\mathrm{RF_{out}}}}{P_{\mathrm{RF_{in}}} + P_{\mathrm{AC_{in}}} + P_{\mathrm{coolers}}} \simeq \frac{P_{\mathrm{RF_{out}}}}{P_{\mathrm{RF_{out}}}\left(0.05+1.62+0.45\right)} \simeq 45\% \qquad (32)$$

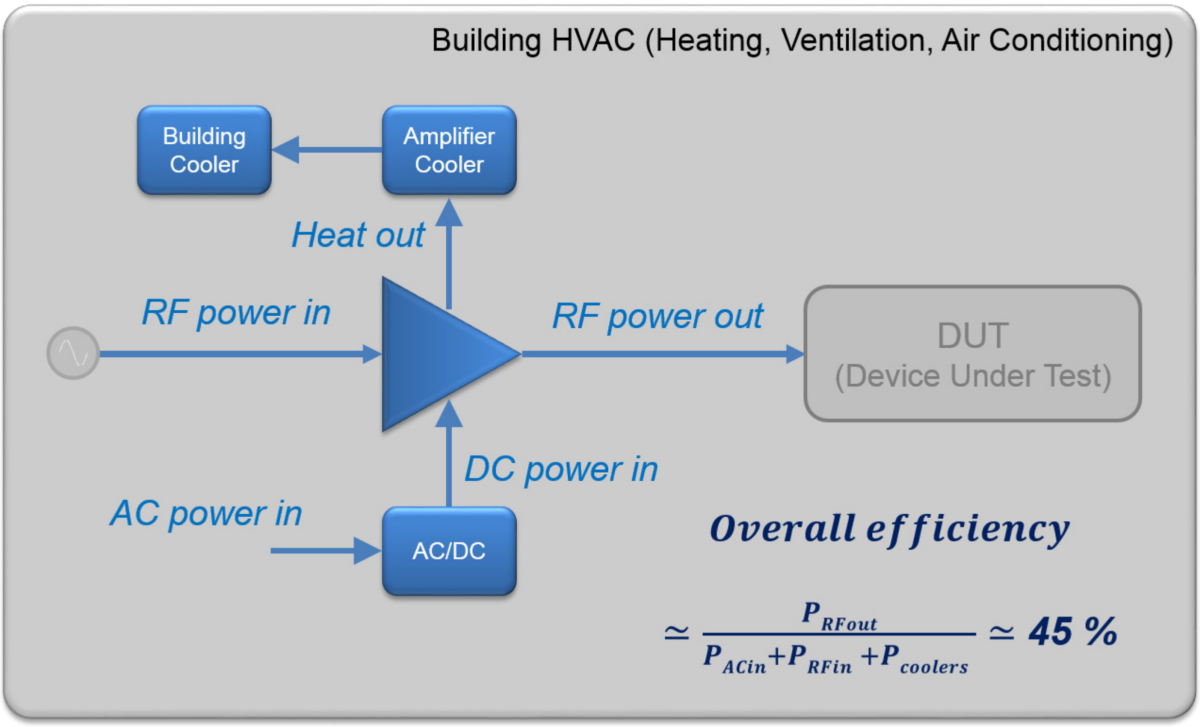


Fig. 51: Overall efficiency, RF power amplifier is only a fraction of the parameters to be considered

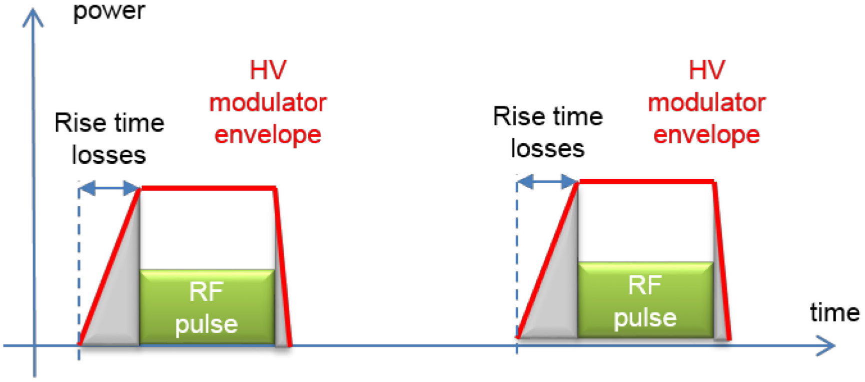
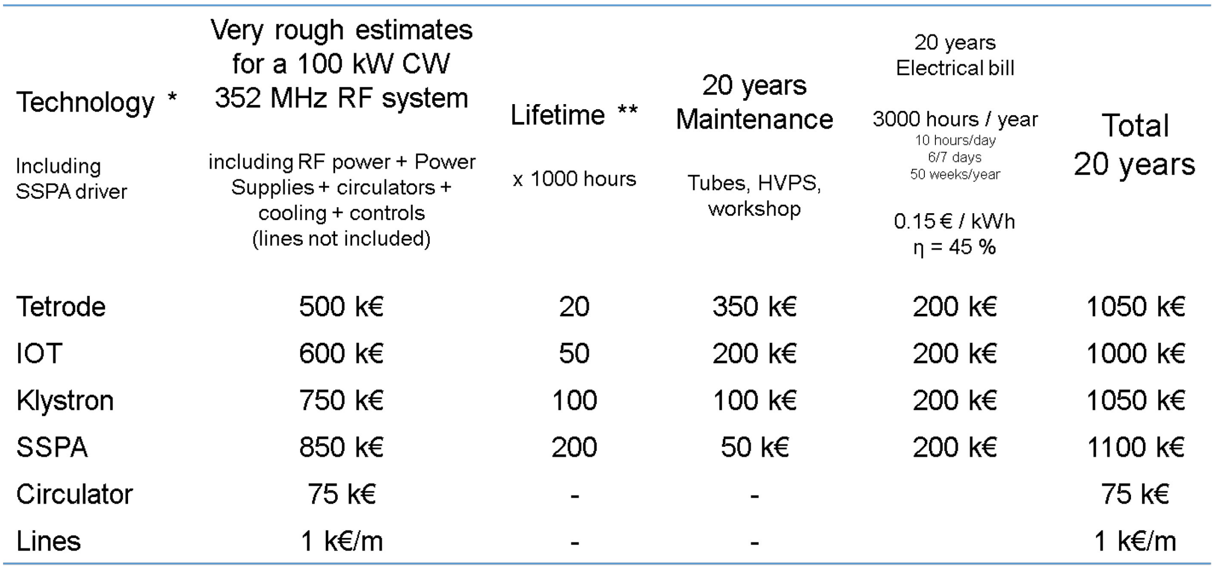


Fig. 52: In case of pulsed system, $\eta_{P_{\mathrm{AC}}/P_{\mathrm{DC}}}$ has to be corrected depending the accepted rise time and consequent loss of efficiency.

Finally, operational costs, maintenance costs, and electrical consumption costs will have to be studied in detail in order to validate the proposed RF-power system. Figure 53 provides some numbers for a given frequency and a given power level that allows all technologies to be compared.



Technology * Including SSPA driver	Very rough estimates for a 100 kW CW 352 MHz RF system including RF power + Power Supplies + circulators + cooling + controls (lines not included)	Lifetime ** x 1000 hours	20 years Maintenance Tubes, HVPS, workshop	20 years Electrical bill 3000 hours / year 10 hours/day 6/7 days 50 weeks/year 0.15 € / kWh η = 45 %	Total 20 years
Tetrode	500 k€	20	350 k€	200 k€	1050 k€
IOT	600 k€	50	200 k€	200 k€	1000 k€
Klystron	750 k€	100	100 k€	200 k€	1050 k€
SSPA	850 k€	200	50 k€	200 k€	1100 k€
Circulator	75 k€	-	-		75 k€
Lines	1 k€/m	-	-		1 k€/m

Fig. 53: Comparison of various technologies costs with a given frequency and power level. Notice that the circulator and lines are at very low cost compare to the overall costs. They must be carefully sized.

In conclusion, to design an RF-power system, you will have to carefully consider:

- your infrastructure which leads into additional overall costs;
- what power specialists are available—this will also drive your technology choice;
- the size of the transmission lines—this point is, unfortunately, often neglected, leading to deep difficulties for the whole project because of a ‘simple’ sub-system;
- the need, or not, for a circulator;
- your HVAC system, as this will dominate your wall–plug efficiency ratio.

Acknowledgements

The author would like to thank Erk Jensen, CERN RF group leader, for his constant support and advice; Hans-Peter Kindermann, his CERN RF team former supervisor for his teaching; and all members of his current team for the daily achievements.

References

[1] J.A. Fleming, Improvements in instruments for detecting and measuring alternating electric currents, GB patent 24850, 1904.

[2] L. de Forest, Device for amplifying feeble currents, US patent 841387, 1907.

[3] F. Lowenstein, Telephone Relay, US patent 1231764, 1912.

[4] H. Arnold, Vacuum tube device, US patent 1354939, 1918.

[5] Phone to Pacific from the Atlantic, The New York Times, 1915. http://www.nytimes.com/learning/general/onthisday/big/0125.html

[6] W. Schottky, Grid-amplifier valve based upon a space-charge and screen grid principle, German patent 300617, 1916.

[7] G. Holst and B.D.H. Tellegen, Means for amplifying electrical oscillations, US Patent 1945040, 1934.

[8] C. Robert, Diacrode TH628, Vacuum Electronics Conference IVEC'07 (IEEE International, 2007), p. 1. http://dx/doi.org/10.1109/IVELEC.2007.4283298

[9] R. Varian, Electrical translating system and method, US patent 2242275, 1937.

[10] A.V. Haeff, Electron discharge device, US patent 2239421, 1940.

[11] J. Randall and Henry Boot, Magnetron, US patent 2648028, 1947.

[12] R. Kompfner, Traveling wave tube amplifier, US patent 2804511, 1953.

[13] Korolyov, Multiple-beam klystron amplifiers performance parameters and development trends, IEEE Transactions on Plasma Science **32 (3)** (2004) 1109. http://dx.doi.org/10.1109/TPS.2004.828807

[14] J.E. Lilienfeld, Method and apparatus for controlling electric currents, US patent 1745175 A, 1926.

[15] J. Bardeen and W. Brattain, Three-electrode circuit element utilizing semiconductive materials, US patent 2524035 A, 1948.

[16] H. Mataré and H. Welker, Semiconductor for control purposes, US patent 2683840 A, 1948.

[17] R. Wallace, High frequency transistor circuit, US patent 2695930 A, 1952.

[18] M. Tanenbaum, R. Logan and A. Peters, Fabrication of silicon devices, US patent 2879190 A, 1957.

[19] M. Attala, Semiconductor devices having dielectric coatings, US patent 3206670 A, 1960.

[20] L. De Vaux, Method for making a gallium arsenide transistor, US patent 3245848 A, 1963.

[21] L. Goodman and K. Smith, Vertical MOSFET with reduced turn-on resistance, US patent 4366495 A, 1979.

[22] H. Cooke, Microwave semiconductor device, US patent 3753056 A, 1971.

[23] J.-S. Moon, K. Gaskill and P.Campbell, Graphene Transistors and RF Applications, InTech Europe, Croatia, 2011.

Bibliography

Proceedings of the CAS-CERN Accelerator School: RF for Accelerators, Ebeltoft, Denmark, 8-7 June 2010, edited by R. Bailey, CERN-2011-007 (CERN, Geneva, 2011), http://dx.doi.org/10.5170/CERN-2011-007.

Proceedings of the CAS-CERN Accelerator School: Radio Frequency Engineering, Seeheim, Germany, 8-16 May 2000, edited by J. Miles, CERN-2005-003 (CERN, Geneva, 2005), http://dx.doi.org/10.5170/CERN-205-003.

Proceedings of the CAS-CERN Accelerator School: RF Engineering for Particle Accelerators, Oxford, United Kingdom, 3-10 April 1992, edited by T. Stuart, CERN-1992-003 (CERN, Geneva, 1992), http://dx.doi.org/10.5170/CERN-1992-003.

M. Valkenburg, *Reference Data for Radio Engineers*, 8th ed. (Focal Press, 1993).

HÜTTE des ingenieurs taschenbuch, (Wilhelm Ernst & Sohn, Berlin, 1955).

H. Meinke and F. Gundlach, *Taschenbuch der Hochfrequenz-technik*, Meinke (Springer, Berlin, 1968).

Web-sites

Thales Group: https://www.thalesgroup.com/en/worldwide/security/rf-sources-medical-accelerators

e2v : http://www.e2v.com/products/rf-power/

CPI : http://www.cpii.com/division.cfm/1

L-3 communications : http://www2.l-3com.com/edd/

Toshiba: http://www.toshiba-tetd.co.jp/eng/tech/index.htm

NXP: http://www.nxp.com/products/bipolar_transistors/

Freescale: http://www.freescale.com/

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Future Trends in Linacs

A. Degiovanni
CERN, Geneva, Switzerland

Abstract
High-frequency hadron-therapy linacs have been studied for the last 20 years and are now being built for dedicated proton-therapy centres. The main reason for using high-frequency linacs, in spite of the small apertures and low-duty cycle, is the fact that, for such applications, beam currents of the order of a few nA and energies of about 200 MeV are sufficient. One of the main advantages of linacs, pulsing at 200–400 Hz, is that the output energy can be continuously varied, pulse-by-pulse, and a moving tumour target can be covered about ten times in 2–3 minutes by deposing the dose in many thousands of 'spots'. Starting from the first proposal and the on-going projects related to linacs for medical applications, a discussion of the trend of this field is presented focussing, in particular, on the main challenges for the future, such as the reduction of the footprint of compact 'single-room' proton machines and the power efficiency of dual proton and carbon-ion 'multi-room' facilities.

Keywords
Linacs; spot scanning; single-room facility; multi-painting; hadron therapy.

1 Introduction

During this lecture, a short overview of linac technology for medical applications is given, with an emphasis on existing and future projects. In particular, the case of linear accelerators for hadron therapy is discussed.

The use of protons and hadrons for the therapy of deep-seated tumours was first proposed in 1946 by Wilson in a famous paper [1]. Since then, hadron therapy has developed in the last 70 years as an advanced technique in radiation therapy, allowing non-invasive and precise irradiation of solid tumours, with the advantage of sparing the surrounding healthy tissues. This is due to the presence of the Bragg peak in the depth-dose profile of charged hadrons, compared to the exponential decay of the dose with depth typical of X-rays (Fig. 1). The overlap of many Bragg peaks, obtained by adjusting the proton beam energy, allows the production of a flat dose distribution over the depth of the tumours, the so-called Spread-Out Bragg Peak (SOBP) shown in Fig. 1. The integral dose delivered to the surrounding healthy tissues with protons and charged hadrons is 3–4 times smaller than with X-rays.

In the past decades, more than 100,000 patients have been treated with proton beams, and more than 10,000 with carbon ions. Other species, such as helium, oxygen or neon, have been used in a reduced number of cases. The typical energies used for the treatment of tumours seated at a maximum depth of 27 cm are 200 MeV for protons and 400 MeV/u for fully stripped carbon ions. In terms of beam currents, typical values are 1 nA for protons and 0.1–0.2 nA for carbon ions, corresponding to a typical dose rate of 2 Gy/l in one minute.

X-rays have two main problems in the treatment of deep seated tumours: (i) they deposit unwanted dose in the critical organs close to the target volume (see Fig. 1) and (ii) they cannot cure the so-called 'radio-resistant' tumours (about 5% of the total) that are less sensitive to radiation than the surrounding normal tissues.

https://doi.org/10.23730/CYRSP-2017-001.151

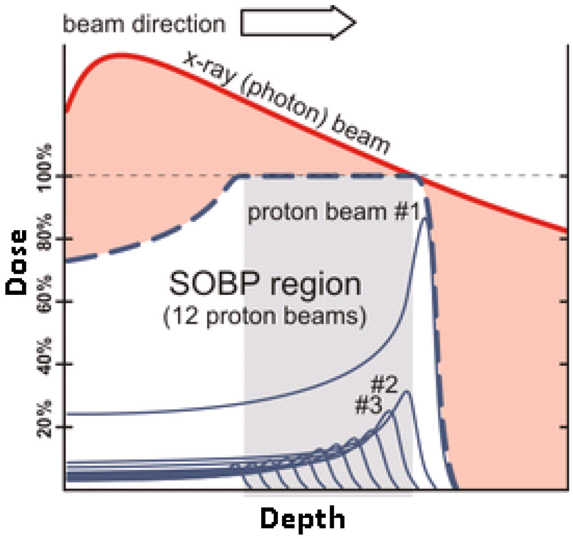


Fig. 1: Depth-dose profiles for X-rays and proton beams

In this respect, protons and charged hadrons, in general, can improve the treatment outcome: first of all, because they spare normal tissues, due to the lower dose delivered in the entrance channel and the practically zero dose delivered in the exit channel (a very small amount of dose is still present due to nuclear fragmentation effects of the treatment beam in the body of the patient); and secondly, because carbon ions have a higher radio-biological effectiveness (RBE) than protons, allowing better control of radio-resistant tumour cells. Indeed, it has been observed that a beam of carbon ions produces, along its track, a great number of clustered, unrepairable 'double strand breaks' on the DNA of the traversed cells. This is related to the six-times-larger charge and the associated 25-times-larger ionization density compared to protons of the same range.

2 The first proposals

Linac technology is widely used for radiotherapy. In the world, more than 15,000 electron linacs are used by radiation oncologists, representing about 50% of existing accelerators with energies larger than 1 MeV. The typical energy of the electron beam used for the production of X-rays is in the range 6–20 MeV. Most of the electron-radiotherapy linacs are normal conducting structures working at a frequency of 3 GHz. They are powered by a single power source (magnetron or klystron) and they are typically short and light enough to be mounted on a rotating support, allowing treatment from different angles.

The use of linacs for hadron therapy was proposed for the first time in the late 1980s. Of course, the energy needed for protons is much larger, so proton-therapy linacs cannot be as compact as radiotherapy ones. On the other hand, the main advantage of using a linac for hadron therapy is the possibility of energy modulation, obtained by switching off some units.

2.1 The all-linac proposal

In 1991, the first proposal of using a linac for proton therapy was published [2]. The proposed design consisted of a sequence of linear accelerator structures able to accelerate protons up to 250 MeV. A radio frequency quadrupole (RFQ) followed by a drift tube linac (DTL) operating at 499.5 MHz were used as injectors to reach an energy of 70 MeV. Then, a 3 GHz Cell-Coupled Linac (CCL) was added to accelerate the beam from 70 to 250 MeV, as shown in Fig. 2. This type of solution based on an RFQ, a DTL, and a CCL has been named the 'all-linac' approach.

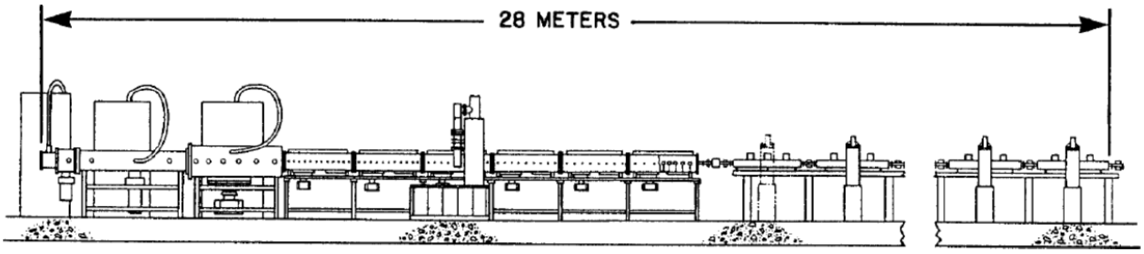


Fig. 2: Layout of the PL-250 proton-therapy linac designed in 1991 (taken from [2])

2.2 The cyc-linac solution

The so-called 'cyc-linac' solution, first proposed in 1993 [3], is based on the combination of a high-intensity low-energy cyclotron and a high-frequency linac. In this approach, a cyclotron pre-accelerates the particles up to a typical energy of 24–30 MeV. The following high-frequency CCL is used to boost the energy up to 200–230 MeV. In this case, all the accelerating units have the same CCL structure (instead of the three types of linacs needed for the all-linac approach). Furthermore, the cyclotron can be used at night and during the weekends for the production of the radio-isotopes needed for imaging and even for therapy, making the centre a multi-disciplinary environment for physicists, radio-chemists and medical doctors. The first proposed cyc-linac facility is shown in Fig. 3.

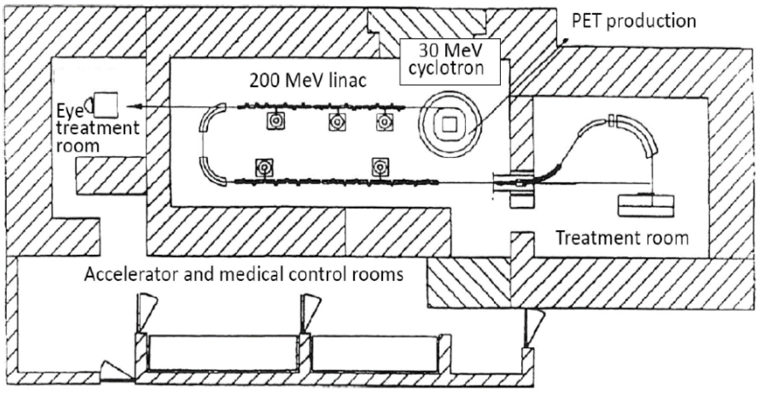


Fig. 3: Layout of the first cyc-linac proposal. This type of facility can host both proton-therapy treatment and radio-pharmacy production.

2.3 The rationale for proton and carbon-ion tumour therapy with linacs

Up to now, only cyclotrons or synchrotrons have been used for hadron therapy. Most of the proton centres in operation are equipped with cyclotrons while, for carbon therapy, only synchrotrons are used.

Generally, cyclotrons used in proton therapy can reach an energy of 230–250 MeV, are quite compact (with a typical diameter of about 2.5 m), and are continuous wave (CW) machines, so that the beam—bunched by the radio-frequency (RF) cavities at 50–100 MHz—is always present during a patient irradiation. On the other hand, the beam energy can only be varied by passive means, i.e. with motors that introduce absorbers of various thicknesses into the beam. With this method, the time needed for an energy change is of the order of 50–100 ms.

Synchrotrons are more complex and larger than cyclotrons (with a typical diameter of 7–8 m for proton machines and 20–25 m for carbon-ion machines). The beam is extracted in spills of a few seconds with a time separation of 1–2 s. The energy can be varied actively, without needing movable absorbers, by adjusting the number of turns in the machine, but each energy change needs to wait for a new spill. Only very recently, a novel fast extraction technique has been implemented at HIMAC, allowing for multi-energy extraction within one spill [4].

In a linac running at 200 Hz and composed of a large number of accelerating units (typically 10–12) singly powered by independently controlled klystrons, the final beam energy can be varied continuously from pulse to pulse, i.e. every 5 ms, by adjusting the amplitude and/or the phase of the klystron drive signals [5]. The intensity of the beam can also be adjusted on a pulse-to-pulse basis by acting on the source. The particular time structure of a linac pulsed beam is shown in Fig. 4. The number of particles per pulse N can be adjusted from pulse to pulse between the maximum N_m and $N_m/10$, while the energy E can be changed between the minimum energy E_{min} (typically 70 MeV for protons) and E_{max} (230 MeV for protons).

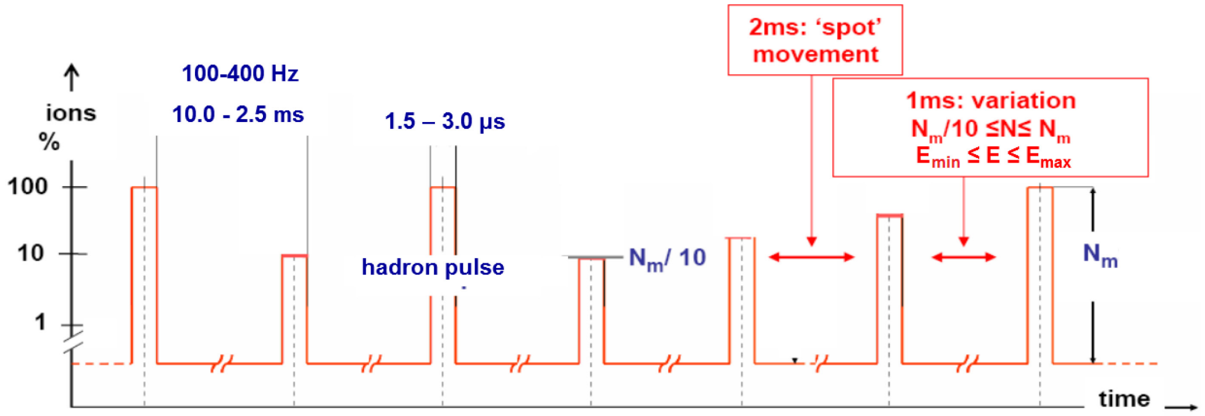


Fig. 4: Typical time structure of a high-frequency pulsed linac for hadron therapy. Both energy and intensity of the beam can be varied actively in a few milliseconds.

The 'active' (i.e. obtained by electronic means without needing absorbers) energy and intensity modulation is a unique feature of the linacs discussed in this lecture and makes possible the implementation of the active spot scanning technique with tumour multi-painting, which is considered the best possible way for treating moving organs [6].

2.4 Another possible application: linacs for proton imaging

The concept of a linac booster can also be applied to a different application than therapy. Indeed, a proton beam of 350–400 MeV could be used for imaging purposes in so-called proton radiography. In order to fully exploit the benefits of proton therapy, precise imaging of the tumour is needed. In fact, one of the biggest issues is the precise outline of the target volume and the surrounding healthy tissues. Furthermore, the treatment plan for proton beams is typically calculated on images taken with X-rays, and one of the critical issues is the exact conversion from CT (Computed Tomography) numbers to proton-stopping power. At present, such problems are mitigated by the addition of margins to the treatment target volume.

The typical energy of a proton-therapy machine is 230–250 MeV, but in order to have protons traversing the body of the patient (like X-rays used for radiography and CT) a minimum energy of 350–380 MeV is needed. A high-frequency linac coupled to the typical cyclotron used in commercial proton-therapy centres could provide a valuable option to boost the energy of the beam by about 100 MeV, reaching energies of interest for proton imaging. Such types of linac booster were studied by TERA Foundation in collaboration with PSI [7]. A 5 m long linac could be installed along the transfer line going to the last treatment room of cyclotron-based centres, and could be used as a booster when needed or by-passed during normal treatment operation.

3 The present

Typical frequencies used for proton linacs in high-energy physics laboratories are 350 or 700 MHz. This is mostly due to the fact that large currents are typically needed. In proton therapy, the requirement for beam current is much less demanding and smaller bore apertures can be used. For this reason, higher frequencies can be used. Furthermore, the availability of S-band power sources already developed by industry for radiotherapy machines has driven the choice of the frequency of the first proton-therapy linac designs. The proof of principle of the use of a 3 GHz linac for acceleration of protons was achieved in 2002 with the LInac BOoster (LIBO) prototype. Since then, several studies and projects with the aim of developing and realizing high-frequency linacs for proton therapy have started.

3.1 The LIBO prototype

A 3 GHz Side-Coupled Linac (SCL) unit called LIBO has been designed and built under the leadership of Mario Weiss (CERN) by TERA Foundation in collaboration with CERN and the INFN sections of Milan and Naples. The accelerating unit, shown in Fig. 5, was made of four tanks coupled together by three 'bridge couplers'. In each tank, 13 accelerating cells allow an accelerating gradient of 16 MV/m to be achieved. In the space between each tank, small Permanent Magnet Quadrupoles (PMQs) are installed to create a FODO lattice that focuses the narrow beam through the accelerating structure.

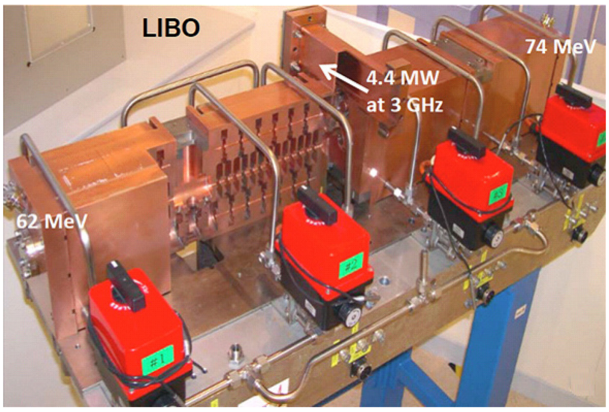


Fig. 5: The LIBO-module prototype was the first 3 GHz linac to accelerate protons [8]

The first unit of LIBO has accelerated protons from 62 to 74 MeV at the same 3 GHz frequency of electron linacs. The LIBO project gave birth to other high-frequency linac designs for proton therapy.

3.2 The project IMPLART by ENEA

The ENEA group, led by Luigi Picardi and Concetta Ronsivalle, has proposed and built a Side-Coupled Drift Tube Linac (SCDTL) which is better suited than a SCL (because of the larger shunt impedance as discussed in the lecture on 'Accelerating Structures' in these proceedings) to accelerate protons from a few MeV to 40–70 MeV. The SCDTL is made of DTL cells coupled together by coupling cells placed off axis (side-coupled cells) and is designed to work at 3 GHz. The SCDTL structure installed in the ENEA laboratories in Frascati (Rome, Italy) is shown in Fig. 6.

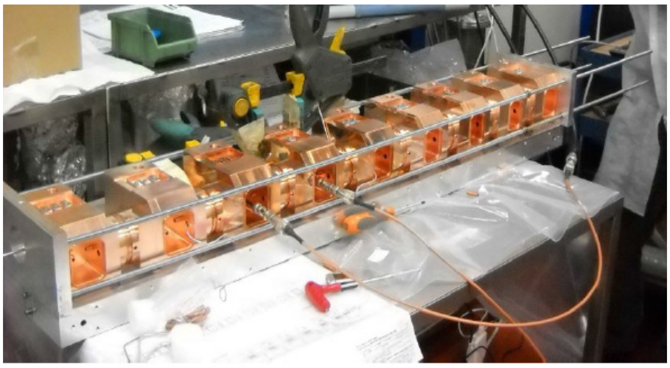


Fig. 6: The first SCDTL module tested at CECOM (Guidonia, Italy) before the installation in the ENEA laboratories in Frascati.

This is part of the Intensity Modulated Proton Linear Accelerator for Radio Therapy (IMPLART) which is expected to accelerate protons up to 150 MeV, and after its technical validation at ENEA will be transferred to IFO hospitals in Rome. The all-linac design envisages a commercial RFQ–DTL system working at 425 MHz used as the injector, followed by four SCDTL modules bringing the proton energy up to 35 MeV, and completed by four CCL units to boost the energy up to 150 MeV. In the Frascati ENEA laboratories, the first SCDTL module has been successfully tested up to 11.7 MeV with a proton beam. This is the first time that an SCDTL at such high frequency has been used for proton acceleration [9].

3.3 LIGHT by A.D.A.M.

The CERN spin-off company A.D.A.M. (Application of Detector and Accelerators to Medicine), founded in 2007 with the aim of industrializing novel detectors and high-frequency linacs for medical applications, has developed in the past few years a linac for proton therapy, based on the TERA design of LIBO, called LIGHT (Linac for Image Guided Hadron Therapy). The LIGHT design (Fig. 7) is composed of three linear accelerating sections: an RFQ up to 5 MeV, an SCDTL up to 37.5 MeV and a CCL up to 230 MeV.

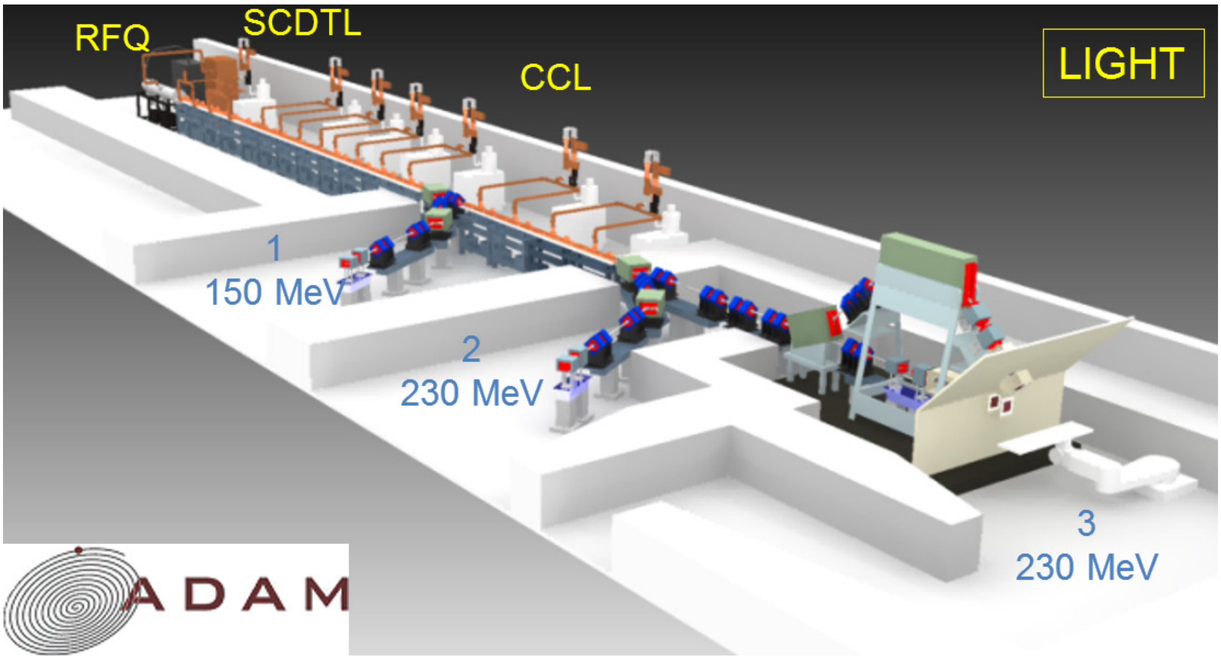


Fig. 7: The all-linac design, proposed by A.D.A.M., for LIGHT

The 5 MeV RFQ is based on a novel design at 750 MHz, made by CERN [10]. This is almost double the maximum frequency typically used for this type of machine. A very compact modular solution of only 2 m has been found. A prototype is now being built and is expected to be tested with a beam in the spring of 2016. The second section is based on the 3 GHz SCDTL design by ENEA. The use of the fourth sub-harmonic of 3 GHz for the RFQ makes the matching between the first two sections easier and allows for a compact design. Finally, a 3 GHz CCL section is added to reach 230 MeV. The design of the accelerating units of LIGHT (Fig. 8) has been improved compared to the LIBO prototype, in particular, by reducing the size of the bridge couplers and allowing for an open space between each tank, where the PMQs used to focus the beam can easily be placed.

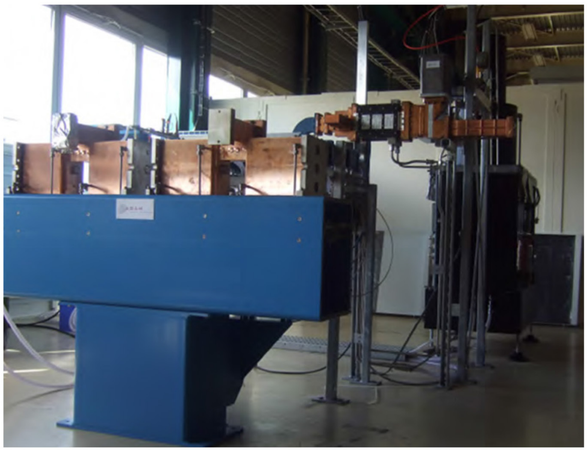


Fig. 8: Prototype of the LIGHT first unit built by A.D.A.M.

The first prototype of LIGHT, up to 90 MeV, will be assembled by A.D.A.M. on CERN premises in 2016 and, after an acceleration test, will be mounted in a hospital, followed by the last units needed to accelerate the protons from 90 to 230 MeV.

3.4 The novel cyc-linac TERA design

Based on the experience developed after the design and test of LIBO, TERA Foundation has improved the cyc-linac scheme and proposed a novel design for PERLA (Protontherapy and Exotic Nuclei from Linked Accelerators), shown in Fig. 9.

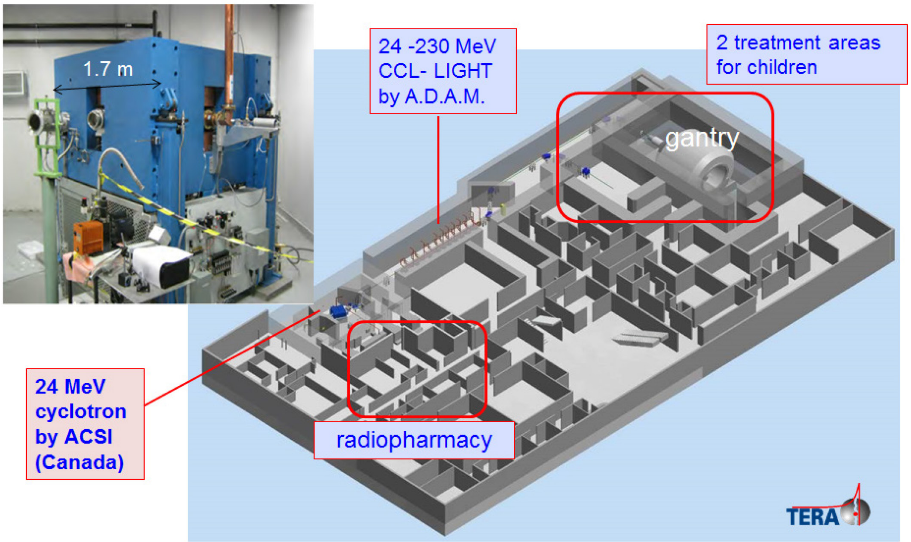


Fig. 9: Design of the PERLA accelerator. In this complex, radio-pharmacy production and therapy are combined in the same centre.

The design, based on the cyc-linac concept, combines a 24 MeV cyclotron with a 3 GHz linac, allowing both radio-pharmacy production and proton therapy in the same complex. A preliminary characterization of the beam properties of the TR24 cyclotron produced by ACSI (Canada) has been performed to study the matching line between the two machines [11].

3.5 ACLIP design by INFN

After the test of LIBO, the INFN sections of Naples, Milan and Bari worked on a new design to improve the efficiency of SCLs in the energy range from 30 to 62 MeV. A prototype of the first module of ACLIP was built in 2007 and, at the end of 2008, underwent high-power tests in the UK on the premises of the e2V company (Fig. 10). A 4 MW magnetron was used to power the module that was conditioned up to 5.4 μs pulses at 120 Hz [12]. In 2009, a test with a beam was also performed at LNS in Catania, confirming the possibility of using 3 GHz SCL structures for protons, even at 30 MeV.

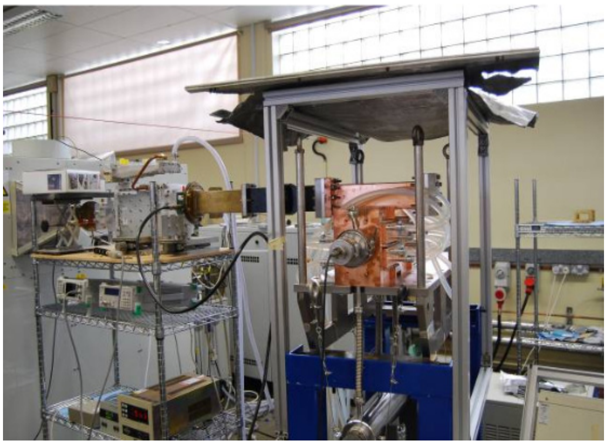


Fig. 10: The ACLIP module high-power RF test set

4 The future

The main issues that make hadron-therapy linacs more difficult to build and more expensive than typical electron-radiotherapy linacs are the fact that: (i) we want to accelerate hadrons (much heavier than electrons) and (ii) these need to be to an energy of at least 200–230 MeV for protons or 400–430 MeV/u for carbon ions (compared to the maximum of 20 MeV for electrons).

The approximately 45 proton-therapy centres running in the world are 'multi-room' facilities, in the sense that one accelerator typically feeds three treatment rooms. This approach makes good use of the accelerator, but requires long displacements of many patients because the facility serves more than 5 million people. Many experts are convinced that 'single-room' facilities linked to a big hospital and serving 1.5 million people will have a place in the future development of hadron therapy [13].

The radio-biological properties of carbon ions make them suitable for the treatment of radio-resistant tumours and extremely encouraging results have been obtained with more than 10,000 patients treated. Despite this fact, at present there are only eight carbon-ion centres treating patients in the world, and most of them are in Japan [14]. An accelerator for carbon-ion therapy is larger and more expensive than for proton therapy, due to the higher energies needed (400 MeV/u compared to 200 MeV) and the larger beam rigidity (almost a factor of three larger). The use of high-efficiency linacs, with a reduced power consumption and a high-duty cycle, can be envisaged for future dual proton and carbon-ion machines.

4.1 Studies for the future: high-gradient hadron structures

The accelerating gradient used in electron linacs is of the order of 20 MeV/ 1 m = 20 MV/m. With the same gradient inside the accelerating structures, the effective accelerating gradient in a structure for low-energy protons—taking into account the transit-time factor (roughly 0.85) and the synchronous phase of the RF field (typically –15°)—would be 16 MV/m. The active length of a 230 MeV proton linac would then be 230 MeV / 16 MV/m = 14.5 m. On top of this, it is important to recall that hadron linacs need focusing elements, whose length, as a rule of thumb, is about 40% of the active length. The actual distance needed for a 230 MeV proton linac—not considering extra space for beam matching or the fact that in the initial acceleration stage, with an RFQ, the gradient is much lower—would then be: 14.5 m × 1.4 = 20 m. This rough estimate gives an idea of the difference in size of proton linacs compared to radiotherapy electron linacs. Of course, for other types of ions, the length becomes a factor of two larger, due to the typical charge to mass ratio Z/A=1/2 for the accelerated particles.

In order to make linacs more compact, a larger accelerating gradient is needed. However, an increase in gradient can be limited by several considerations.

First of all, since hadrons are much heavier than electrons, at the energies of interest for therapy, the speed of the particles are typically in the range 0.1 to 0.6 c. To increase the acceleration efficiency for such low β values, 'nose cones' are typically added along the axis. This results in a better focalization of the field in the accelerating gaps and thus an increase of the transit-time factor. On the other hand, the addition of nose cones enhances the value of the surface electric field E_s (Fig. 11). This is why, in hadron linac structures, the ratio of surface to average field E_s/E_0 typically takes values of 4–5 (while for electron linacs E_s/E_0 is around 2). In other words, for the same accelerating gradient, the electric surface field in hadron-linac structures is two-times larger. But high surface fields are related to increased probability of breakdowns. Secondly, the power dissipated in the structures scales with the square of the accelerating gradient. A higher gradient means more RF power from the power sources and more heat load that needs to be cooled in order to keep constant the resonant frequency of the structures, and therefore, increased costs.

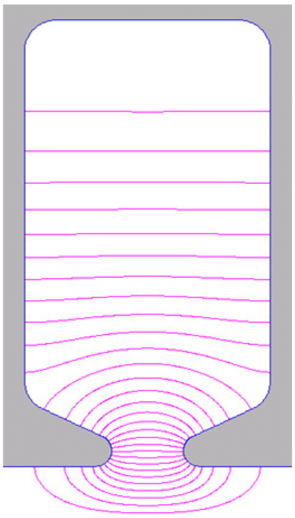


Fig. 11: Typical inner shape and electric field lines profile of an accelerating cavity in a CCL hadron linac. The maximum surface electric field is close to the point of the 'nose'.

The CLIC study at CERN has developed a novel accelerating scheme for an electron–positron collider. In the past years, an extensive test program demonstrated the feasibility of 100 MV/m acceleration in normal conducting 12 GHz RF structures [15]. TERA Foundation, in collaboration with the CLIC RF structure development group led by Walter Wuensch, has performed extensive studies on high-gradient accelerating structures for applications in hadron therapy [16]. Even though the aim and the geometry of the TERA and CLIC structures are different, both share the same operational limits in terms of maximum surface electric field E_{max} (~200 MV/m) and of maximum break-down rate (BDR) expressed in breakdown per pulse (bpp) over a certain length (~10^{-6} bpp/m). The first limit is strictly related to the cell geometry, which defines the ratio between peak surface electric field and average accelerating gradient E_{max} / E_0. The second limit, concerning linac boosters, corresponds to one breakdown every two treatment fractions (of about 2–3 minutes) at a maximum repetition rate of 200 Hz for a 20 m long linac, which is considered to be acceptable for medical applications.

Two single-cell standing-wave accelerating structures, one at 3.0 GHz and one at 5.7 GHz, have been designed, built and high-power tested [17]. Measurements of the performance in terms of BDR were conducted and compared with the results of single-cell standing-wave X-band accelerating cavities tested at SLAC (Fig. 12).

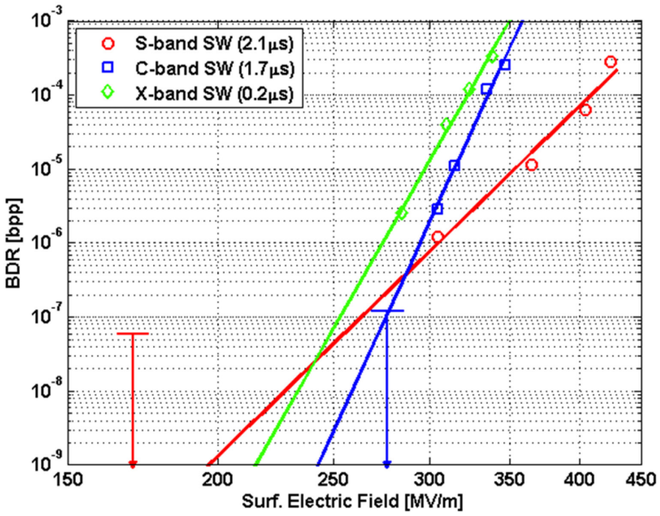


Fig. 12: Comparison of results of BDR measurements performed at different frequencies [18]

From these results, it has been concluded that the surface fields in normal conducting structures at 3.0 and 5.7 GHz can be pushed to more than 200 MV/m with a probability of breakdown per cell of less than 10^{-8}. But even more importantly, the experimental results show that the maximum modified Poynting vector S_c—introduced in [19]—describes the BDR measurements in the 3–12 GHz frequency range better than the maximum surface electric field E_s.

4.2 High-gradient linac for single-room facilities

The project TULIP (Turning LInac for Protontherapy), patented by TERA, envisages a linac, mounted on a rotating gantry, used as a booster for protons previously accelerated by a cyclotron [20]. In the tanks, the maximum average gradient is E_0 = 30 MV/m. As is shown in Fig. 13, the RF power transmission is made possible by high-power rotating joints developed in collaboration with the CLIC group.

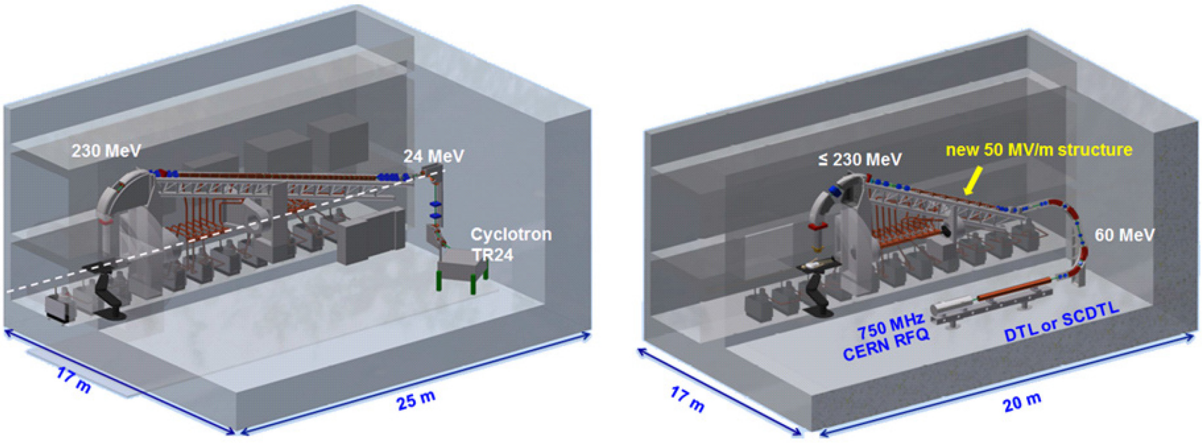


Fig. 13: Two possible linac-based solutions for the proton single-room facility TULIP

After several studies, a novel design—based on the optimization of the modified Poynting vector, and using the high-frequency RFQ designed by CERN, followed by an SCDTL as an injector—has been proposed. In this design, average gradients E_0 of 50 MV/m in the structures have been considered in order to obtain a more compact solution (as is shown in Fig. 13 (right)).

4.3 High-efficiency design for dual proton and carbon-ion facilities

When considering the design of a linac for carbon-ion therapy, the power efficiency of the whole system becomes a challenge. The intrinsic properties of carbon ions are such that:

- the 'spot' size is transversely two-times smaller than the proton spot size (5 mm instead of 10 mm, if the pencil beam has a FWHM= 4 mm) because of the sharper lateral fall-off. In order to keep the treatment time short enough (a few minutes), a higher repetition rate of 300–500 Hz rather than 100–200 Hz is needed;
- concerning acceleration, each proton carries a 'useless' neutron (Z/A = 1/2), which means that for 100 MeV/u acceleration, a voltage of 200 MV is needed.

As a consequence, much more average and peak power from the RF sources is required. This is crucial for a carbon-ion facility based on a linac, such as the CABOTO (CArbon BOoster for Therapy in Oncology) design proposed and patented by TERA Foundation [21]. In Figs. 14 and 15, two schematic layouts of CABOTO are presented. The scheme of Fig. 14 is based on a cyc-linac solution. Several designs have been studied for the injection in a linac [22]. As an example, the cyclotron can be similar to the one built by VECC in Kolkata [23], which accelerates light Z/A = 1/2 ions up to 70 MeV/u. The following linac—33 m long and pulsed at 300–400 Hz—can boost fully stripped carbon ions from 70 MeV/u to 400 MeV/u. With these figures, the voltage gain of the particles is very large: (400–70) x 2 = 660 MV. The use of very high gradients to reduce the total length (as the ones foreseen for TULIP) would bring the overall power consumption to unacceptable levels (more than 2 MW). This is why the linacs shown in Fig. 14 are designed for a gradient in the structures of 30 MV/m.

Recent developments have enabled significant improvement of the klystron efficiency. Multi-Beam Klystrons (MBKs) at 3 GHz, producing 7 MW RF peak power in pulses of 5 μs at 500 Hz with an efficiency of better than 60% are now available. Similar klystrons, which are the object of a CERN tender launched by the CLIC group, need only 60 kV so that the modulators are oil-free and small. This item will soon be tested at CERN.

With this new type of RF source, the 32 CCL units of CABOTO running at a duty cycle of 1–1.5 10^{-3} (300–400 Hz with 3.5 μs RF pulses) could consume 700 kW. In a full linac configuration, as shown in Fig. 15, with an RFQ, SCDTL and CLL, the maximum average power consumption would be 800 kW. Considering extra ancillaries and power supplies for beam transfer lines, the objective is to run the accelerator complex with a plug-power not larger than 1.2 MW. As a final remark, it has to be underlined that this is the highest power consumption related to the accelerators. While RFQ and SCDTL represent a fixed power consumption, the CCL contribution (which is more than 85%) strongly depends on the depth of the treated tumours.

Apart from the smaller power consumption compared to synchrotrons (1.2 MW at maximum compared to 2.5–3 MW for synchrotrons), the footprint of a dual proton–carbon-ion linac complex can also be smaller than a synchrotron one. In the layouts of Fig. 14, the three treatment rooms and the distribution of the close-by rooms are copied from CNAO in Pavia. The footprints of the accelerators and the power supplies of the two versions of CABOTO are 750 m^2 and 1000 m^2, to be compared with the 1600 m^2 of CNAO.

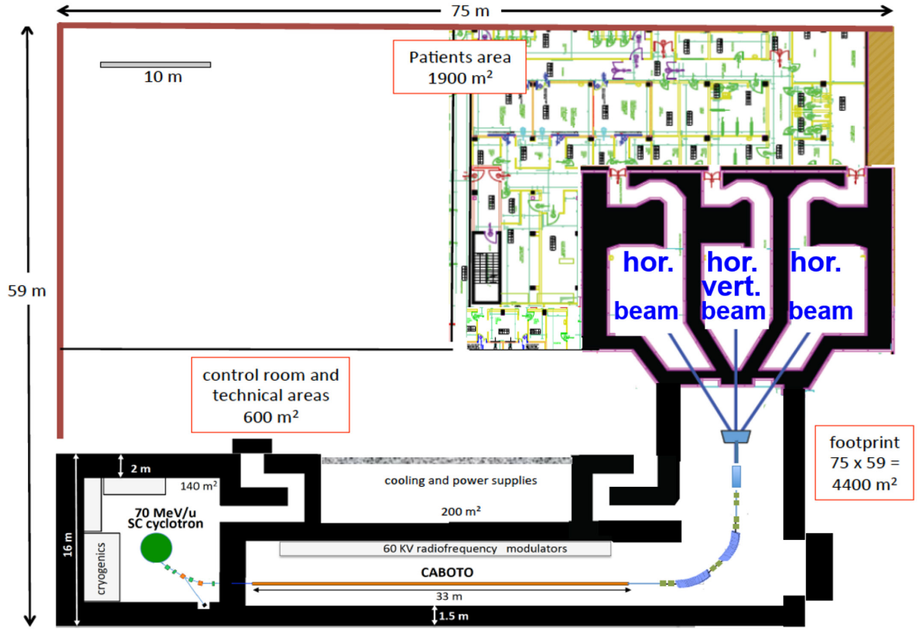


Fig. 14: Preliminary layout of a cyc-linac carbon-ion facility based on high-efficiency design

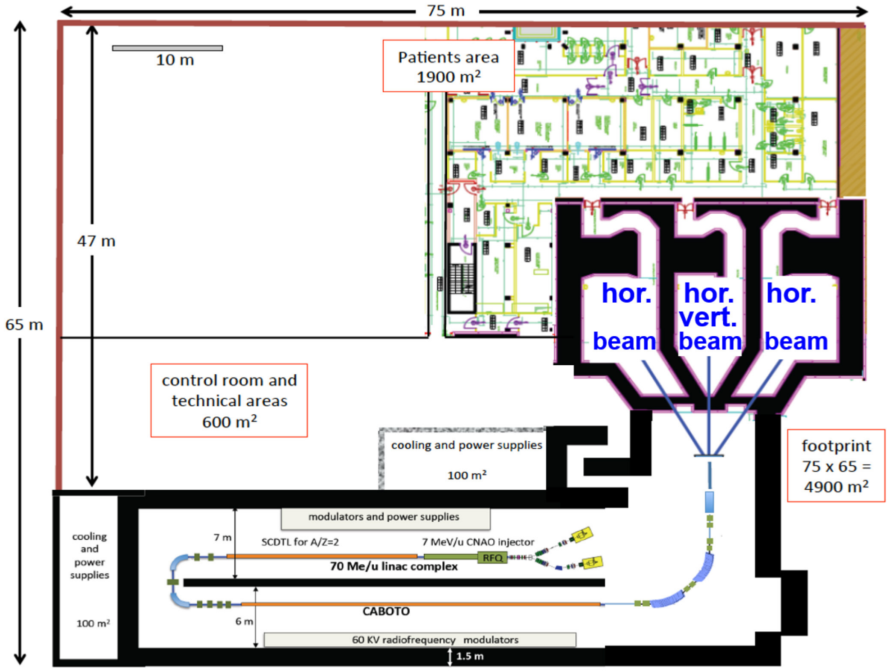


Fig. 15: Preliminary layout of an all linac carbon-ion facility based on high-efficiency design

5 Summary

High-frequency RF linacs can produce hadron beams that are well suited to treating moving organs with the multi-painting spot scanning technique. The field is now moving from the first prototypes and RF tests to full-scale commercially built linacs.

At present, several groups are working on construction projects. Low-velocity SCDTL and high-velocity CCL accelerating structures have been built and tested by ENEA, TERA, and INFN. At CERN, a new 750 MHz RFQ is being built with the support of the CERN medical application office. The CERN spin-off company A.D.A.M. is building, at CERN, an all-linac facility that will be the first commercial item of this type and will be transferred to a hospital to treat patients.

Future challenges include high-gradient and high-efficiency structures. TERA and the CERN CLIC group have been collaborating for several years on this subject, with the support of the Knowledge Transfer group of CERN. In future, this will lead to TULIP, a compact proton linac rotating around the patient, and to CABOTO, a high-efficiency linac for the therapy of deep-seated radio-resistant tumours with carbon ions.

Acknowledgement

I would like to thank Prof. Ugo Amaldi for being the motor of many of the developments described in this lecture and for his contagious passion for physics and health. His comments and the discussions about linacs for medical applications have always been stimulating and enlightening.

References

[1] R.R. Wilson, *Radiology* **47**(5) (1946) 487. http://dx.doi.org/10.1148/47.5.487

[2] R.W. Hamm, K.R. Crandall and J.M. Potter, Preliminary design of a dedicated proton therapy linac, Proc. PAC'90, Vol. 4 (1991) 2583. http://dx.doi.org/10.1109/pac.1991.165037

[3] U. Amaldi and B. Larsson (Eds.), *Hadron Therapy in Oncology*, (Elsevier, Amsterdam, 1994), p. 45.

[4] Y. Iwata *et al.*, *Nucl. Instrum. Methods Phys. Res.* A **624**(1) (2010) 33.
http://dx.doi.org/10.1016/j.nima.2010.09.016

[5] U. Amaldi, S. Braccini and P. Puggioni, *Rev. of Acc. Sc. and Tech. RAST* **02**(01) (2009) 111. http://dx.doi.org/10.1142/S179362680900020X

[6] C. Bert *et al.*, *Med. Phys.* **34**(12) (2007) 4768. http://dx.doi.org/10.1118/1.2815934

[7] M. Schippers *et al.*, A next step in proton therapy: boosting to 350 MeV for therapy and radiography applications, Proc. PTCOG 51 Conf., Seoul, 2012.
http://www.ptcog.ch/index.php/past-programs-and-talks

[8] U. Amaldi *et al.*, *Nucl. Instrum. Methods Phys. Res.* A **521**(2–3) (2004) 512.
http://dx.doi.org/10.1016/j.nima.2003.07.062

[9] L. Picardi *et al.*, Experimental results on SCDTL structures for protons, Proc. IPAC'14, Dresden, 2014, p. 3247.
http://accelconf.web.cern.ch/AccelConf/IPAC2014/html/author.htm

[10] M. Vretenar *et al.*, A compact high-frequency RFQ for medical applications, Proc. LINAC'14, 2014, p. 935.
http://accelconf.web.cern.ch/AccelConf/LINAC2014/html/author.htm

[11] A. Degiovanni *et al.*, Emittance measurements at the Strasbourg TR24 cyclotron for the addition of a 65Mev linac booster, Proc. CYC'13, 2013, p. 329.
http://accelconf.web.cern.ch/AccelConf/CYCLOTRONS2013/html/author.htm

[12] V.G. Vaccaro *et al.*, RF high power tests on the first module of the ACLIP linac, Proc. PAC'09, 2009, p. 4950.
http://accelconf.web.cern.ch/AccelConf/PAC2009/html/author.htm

[13] U. Amaldi *et al.*, *Nucl. Instrum. Methods Phys. Res.* A **620**(2–3) (2010) 563.
http://dx.doi.org/10.1016/j.nima.2010.03.130

[14] http://www.ptcog.ch/index.php/facilities-in-operation, last accessed 14 October 2015.

[15] W. Wuensch *et al.*, High-gradient test results from a CLIC prototype accelerating structure: TD26CC, Proc. IPAC'14, 2014, p. 2285.
http://accelconf.web.cern.ch/AccelConf/IPAC2014/html/author.htm

[16] A. Degiovanni *et al.*, *Nucl. Instrum. Methods Phys. Res.* A **657**(1) (2010) 55.
http://dx.doi.org/10.1016/j.nima.2011.05.014

[17] S. Verdú Andrés, Ph.D. thesis, Univeristat de València, 2013.

[18] A. Degiovanni, Ph.D. thesis, EPFL, 2014.

[19] A. Grudiev *et al.*, *Phys. Rev. ST Accel. Beams* **12** (2009) 102001:1.
http://dx.doi.org/10.1103/PhysRevSTAB.12.102001

[20] A. Degiovanni *et al.*, Design of a fast-cycling high-gradient rotating linac for protontherapy, Proc. IPAC'13, 2013, p. 3642.
http://accelconf.web.cern.ch/AccelConf/IPAC2013/html/author.htm

[21] U. Amaldi *et al.*, Ion acceleration system for hadrontherapy, Patent US 7423278.

[22] A. Garonna *et al.*, Comparison of superconducting 230 MeV/u synchro- and isochronous cyclotron designs for therapy with cyclinacs, Proc. CYC'13, 2013, p. 108.
http://accelconf.web.cern.ch/AccelConf/CYCLOTRONS2013/html/author.htm

[23] C. Mallik and R. K. Bhandari, Commissioning status of Kolkata superconducting cyclotron, Proc. CYC'10, 2010, p. 2664.
http://accelconf.web.cern.ch/AccelConf/Cyclotrons2010/html/author.htm

Bibliography

A Multi-TeV linear collider based on CLIC technology, CLIC Conceptual Design Report, edited by M. Aicheler, P. Burrows, M. Draper, T. Garvey, P. Lebrun, K. Peach, N. Phinney, H. Schmickler, D. Schulte and N. Toge, CERN-2012-007 (CERN, Geneva, 2012). http://dx.doi.org/10.5170/CERN-2012-007.

U. Amaldi, *Particle Accelerators: From Big Bang Physics to Hadron Therapy* (Springer International Publishing, Switzerland, 2015).

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Cyclotrons for Particle Therapy

J.M. Schippers
Paul Scherrer Institut, Villigen, Switzerland

Abstract
In particle therapy with protons a cyclotron is one of the most used particle accelerators. Here it will be explained how a cyclotron works, some beam dynamics aspects, its major subsystems, as well as the advantages and disadvantages of a cyclotron for this application are discussed. The difference between the standard isochronous cyclotron and the synchrocyclotron is explained. New developments are presented and especially those which aim to reduce the size of the accelerator.

Keywords
Cyclotron, RF system; ion source; extraction system; superconducting coils; synchrocyclotron; isochronous cyclotron.

1 Introduction

In particle therapy, cyclotrons and synchrotrons are the accelerators currently used. The choice depends on treatment method, price, and local conditions, such as available expertise and available space. Synchrotrons are discussed in another chapter in these proceedings. Here, cyclotrons for particle therapy are described, but is important to note that, with both machines, good clinical results have been obtained. The major differences are the footprint of the accelerator, the need for a degrader to set the beam energy of cyclotron beams, and the continuous beam from cyclotrons versus the spill structure of synchrotrons.

Certain characteristics and parameters of the accelerator will depend on the method of the dose-delivery process at the patient. There are two major techniques for applying the dose to the patient. After aiming the beam from the desired direction by setting the gantry (this is a beam-transport system mounted on a rotating mechanical structure), the beam, which has a typical diameter of 1 cm, must be spread in the lateral direction to match the cross section of the tumour as seen from the incoming beam direction. This is done either by the so-called ‘scattering technique’ or by the so-called ‘scanning technique’. In the scattering method, the beam cross section is increased by sending the beam through a scattering system, consisting of one or more foils of material with a high atomic number Z, by which the beam diameter is increased to match to the maximum lateral tumour cross section (a ‘passive’ technique). In the scanning method, a pencil beam of less than 1 cm in cross section is ‘actively’ scanned in the transverse plane over the tumour cross section. This motion is done in steps and the applied ‘spot’ dose is varied per step (‘spot scanning’), or scanning is performed by continuously shifting the beam along lines in the tumour, during which the beam intensity is varied to deliver the correct dose along the line (‘continuous scanning’). Until now, the scattering technique has been used most commonly. However, for several years, the scanning technique has been regarded as the optimal technique (i.e. the technique that applies the best possible dose distribution) currently feasible in practice, and almost all new facilities are designed to employ this technique. Therefore, in this chapter the focus will be on the application of cyclotrons for scanning techniques.

2 Basic concept of the cyclotron

Acceleration of charged particles is achieved by means of an electric field. Since electric fields are limited in strength, use is made of the repetitive crossing of electric-field regions in gaps between

https://doi.org/10.23730/CYRSP-2017-001.165

electrodes. The particles are accelerated when crossing this gap and continue their path through the electrode. Within the electrode, there is no electric field, so the sign of its voltage can be changed without any effect on the particles within. This voltage change must be performed in phase with the particle position; at a phase at which the particle again experiences an accelerating field when leaving the electrode and crosses the gap to the next one. So, only if this voltage change is performed ‘in phase’, another acceleration step is made. A row of such electrodes is the basic idea of the linear accelerator [1]. Ernest Lawrence investigated the use of a homogeneous magnetic field to bend the particles along a circular trajectory, so that the same gap is crossed repeatedly. The motivation was to reduce the dimensions needed to reach a high energy in a linear accelerator. The centripetal force needed to produce a circular orbit, with radius r, of a particle with mass m, speed v, and charge q, is then made by the Lorentz force obtained from the magnetic field B:

$$\frac{mv^2}{r} = Bqv\,. \tag{1}$$

Using this relation, Lawrence realized that the time T a particle needs to make a full circle is

$$T = \frac{2\pi \mathrm{r}}{v} = \frac{2\pi m}{Bq}\,, \tag{2}$$

from which it follows that the revolution frequency $1/T$ does not depend on radius of the particle’s circular orbit and, thus, not on the particle energy. All particles have to cross the gap at the same phase, so that they are all, approximately, at the same azimuth (=angular position) in the cyclotron. The frequency at which the electric field in the gap between the electrodes is varying must be in phase with the particle revolution time. In 1929, Ernest Lawrence realized that, since the frequency of the RF signal must match the particle’s revolution frequency, it is also independent of the radius and energy of the particle orbit. From Eq. (2), it follows that, for a given particle mass and charge, the RF only depends on the magnetic field. This is the basic principle of the cyclotron operation [2].

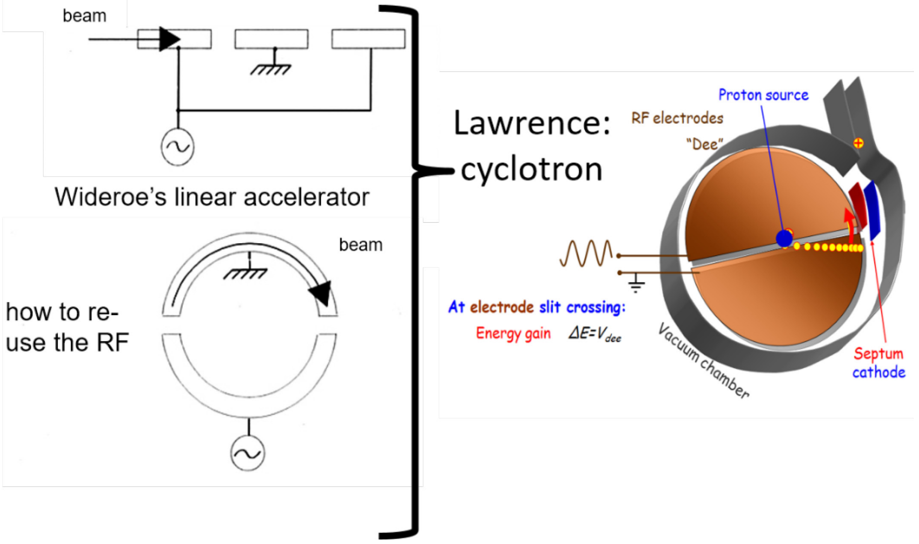


Fig. 1: The cyclotron is a combination of the idea of a linear accelerator and a circular version of it, to use the RF repetitively.

As shown in Fig. 1, the major components of a typical cyclotron for therapy are:

- an RF system providing a strong electric field accelerating the protons between electrode plates;
- a strong magnet that confines the particle trajectories into a spiral-shaped orbit, so that they can be accelerated repeatedly by the RF voltage between the electrodes;

– a proton source in the centre of the cyclotron, in which hydrogen gas is ionized and from which the protons are extracted;

– an extraction system that guides the particles that have reached their maximum energy out of the cyclotron into a beam-transport system.

The cyclotron has become a common work horse as an accelerator used in nuclear-physics laboratories and isotope production. The pioneering work of Ernest, and his brother and medical doctor John Lawrence, created interest for the use of energetic beams of heavy charged particles (protons and ions such as He, Ne and C) for therapeutic applications. The 1946 publication of Robert Wilson, showing the advantageous properties of the dose distribution of protons in tissue [3], was a breakthrough in this field. In 1950, John Lawrence was the first to treat patients with cancer by means of a beam of energetic ions [4] from the 60 in. cyclotron in Berkeley (CA, USA), which started its operation in 1939. In 1957, these treatments were successfully duplicated at the cyclotron in Uppsala (Sweden) [5] and, in 1961, at the Harvard cyclotron in Boston (MA, USA) [6]. From then on, the worldwide number of facilities has increased slowly, mostly at facilities located at physics and accelerator laboratories that applied the therapy program in parallel to the physics research. Many cyclotrons, giving proton energies of 60–100 MeV, have been used for treatments of melanoma in the eye. The Harvard Cyclotron was converted to proton-therapy operation in 1949 and has been used solely for proton therapy since 1961. Since then, the group working with this cyclotron has played a major role in the development of proton-therapy techniques. Apart from cyclotrons, synchrotrons also started to play their role in particle therapy. In 1991, the first hospital-based proton-therapy facility came into operation in Loma Linda (CA, USA) [7]. Here, a specially developed synchrotron is used as proton accelerator. The first dedicated, commercially provided, proton-therapy centre using a cyclotron as an accelerator, came into operation at NCC Kashiwa (Japan) in 1998. In 2001, the first patient was treated at the second dedicated, commercially achieved cyclotron, which is based at the proton-therapy facility at the Mass. General Hospital in Boston (MA, USA). This facility was acquired in the first official commercial tender (1992) for a proton-therapy centre. Since then, several commercial companies have been developing cyclotrons for this application. The designs have been based on a simplification of existing cyclotrons by fixing many operational parameters, since therapy requires much less variation of parameter values than applications in physics research. Therefore, only a limited amount of time is needed for tuning a clinically used cyclotron. Furthermore, this simplification increases the reliability, which, together with price, easy operation, and short services, are essential requirements for operation in a clinical environment. In 2007, the first cyclotron using superconducting coils came into operation for a proton-therapy facility at PSI in Switzerland [8]. This Varian machine was the next step in the continuously ongoing process to reduce the size of the accelerator and, thus, the related price of the facility.

Modern cyclotrons, dedicated to proton therapy, accelerate protons to a fixed energy of 230 or 250 MeV. Compared to the classical cyclotrons in accelerator laboratories, the new cyclotrons are rather compact with a magnet height of approximately 1.5 m and a typical diameter between 5 m (200 tons) and 3.5 m (100 tons), when equipped with room temperature coils or with superconducting coils, respectively. Usually, some extra space is needed above and/or below the cyclotron for the support devices of the ion source, RF coupling to the Dees (see Section 3), and equipment to open the machine.

Currently, all operating cyclotrons for particle therapy are accelerating protons. Developments of cyclotrons for acceleration of heavier particles, such as helium or carbon ions, are in progress.

3 The RF system

The RF system usually consists of two or four electrodes (often called '*Dee*' due to their shape in the first cyclotrons built) which are connected to an RF generator, driving an oscillating voltage between 30 and 100 kV with a fixed frequency. This RF equals the orbital frequency $1/T$ multiplied by an integer ('harmonic number' h, with e.g. h=2) and is somewhere in the range of 50–100 MHz. Each Dee consists of a pair of copper plates on top of each other, with a few centimetres in between. The Dees are mounted between the magnet poles, which are at ground potential, as is the whole magnet. When a proton crosses the gap between a Dee and the neighbouring pole, which is at ground potential, it experiences acceleration towards the grounded region when the Dee voltage is positive. When it approaches the Dee at the negative voltage phase, the proton is accelerated into the gap between the two plates. During its trajectory within the electrode or in the ground potential, the electrodes change sign. The magnetic field forces the particle trajectory along a circular orbit, so that it crosses the acceleration gap, the Dees, and the ground, four or eight times during one turn, in the case of an RF system with two Dees or four Dees, respectively.

If the Dee voltage is, for example, 50 kV at the moment of gap crossing in a four-Dee RF system, the proton gains $\Delta E = 0.40$ MeV per turn. Due to the energy gain, the radius of the proton orbit increases, so that it spirals outward. The maximum energy, $E_{\max}$, (for proton-therapy cyclotrons, typically, 230 or 250 MeV) is reached at the outer radius of the cyclotron's magnetic field, after approximately $E_{\max}/\Delta E = 625$ turns. To limit the number of turns and, thus, the risk of beam losses, it is an advantage to have two, three, or four Dees, so that a high energy gain per turn is achieved at not too high a Dee voltage. Then, the Dees are mounted in the so-called valleys between the pole hills, so that the gap between the poles can be minimized. This pole shape will be explained later.

As mentioned earlier, it is extremely important that the RF is in phase with the azimuthal (angular) position of the particles in the cyclotron. Therefore, such a cyclotron is called an isochronous cyclotron. In a so-called synchrocyclotron, this is achieved by adapting the frequency of the RF signal to a decreasing particle orbit frequency. This will be discussed later.

4 Central region of the cyclotron

The protons start their acceleration process in the centre of the cyclotron. In most cyclotrons, an ion source is located here. External sources are of interest, especially when other particles also need to be accelerated. The internal sources operate by exploiting the Penning effect [9, 10]: the ionization of gas by the energetic electrons created in an electrical discharge.

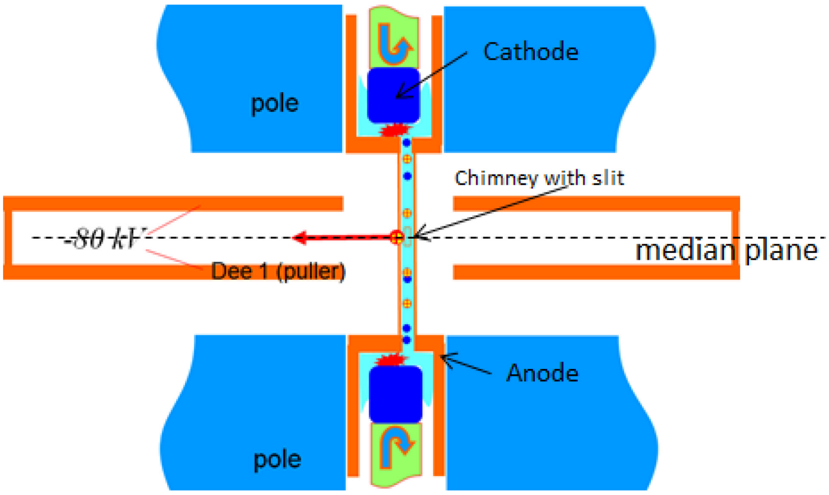


Fig. 2: The internal ion source between the magnet poles in the center of the cyclotron

Different configurations of the Penning source are used in cyclotrons, but, in principle, they all operate by the same principles. In the cyclotron at PSI, the ion source consists of two cathodes at a negative voltage of the order of 1 kV, located above and below the median plane of the cyclotron, at each end of a vertical hollow cylinder at ground potential (chimney), see Fig. 2. Hydrogen gas is flushed in between the cathodes and their opposing grounded anodes. Free electrons are created by spontaneous electron emission from the cathodes in the strong electric field between cathode and ground (‘cold cathode source’). The electron emission can be stimulated by heating the cathode or by a filament. In the electric field, the electrons are accelerated towards the anodes and they ionize the gas. The ionized gas atoms bombard the cathodes, so that they emit even more electrons. The electrons and ions are confined in gyroscopic orbits along the vertical magnetic-field lines and bounce up and down between the cathodes, thus further ionizing the gas.

The ions (H^+, H_2^+, H^-, etc.) and electrons form a plasma that fills the volume in the chimney between the cathodes. Protons and other ions that have diffused to a little hole in the chimney wall experience the electric field from the nearest Dee edge (the puller). When this Dee is at negative potential, protons that escape from the plasma accelerate towards the Dee. If they arrive at the right phase, they will be accelerated further. Due to the narrow acceptance windows, in time (i.e. RF phase) and in the further acceleration path, only a fraction of the protons leaving the source are actually accelerated.

Important parameters describing the ion source are: the total proton current extracted from the source within not too large an emittance, the stability of the intensity, and the time between services, since these go on cost of operational time of the cyclotron. Modest operational conditions and a careful material choice (heat properties, electron emission, sputtering resistance) are of importance to obtain an acceptably long time interval between source services. Currently, typical time intervals between services are 1–3 weeks.

The maximum beam intensity extracted from the cyclotron (i.e. dose rate at the patient) is determined by the ion-source output. Variation of the source output is done by modifying the gas flow, the cathode voltage, the heating power, or a combination of these. This is a relatively slow process, however, which takes, of the order of, milliseconds. In several scanning processes, there is a need to vary the intensity more rapidly. In addition to adjustment of the source intensity, in several cyclotrons the beam-intensity regulation is performed much faster by intercepting a controlled part of the beam from the source. This is done by a system mounted in the central region of the cyclotron. It consists of an electrostatic deflection plate and intercepting collimators. The source is then operated at a constant intensity, determining the upper limit of the beam intensity extracted from the cyclotron. By deflecting the beam over the slit-shaped collimator aperture, the intensity of the beam passing the aperture can be controlled with high accuracy and within several tens of microseconds. The beam passing the slit will be accelerated. At the slit, the beam energy is still low enough to prevent high power losses and activation.

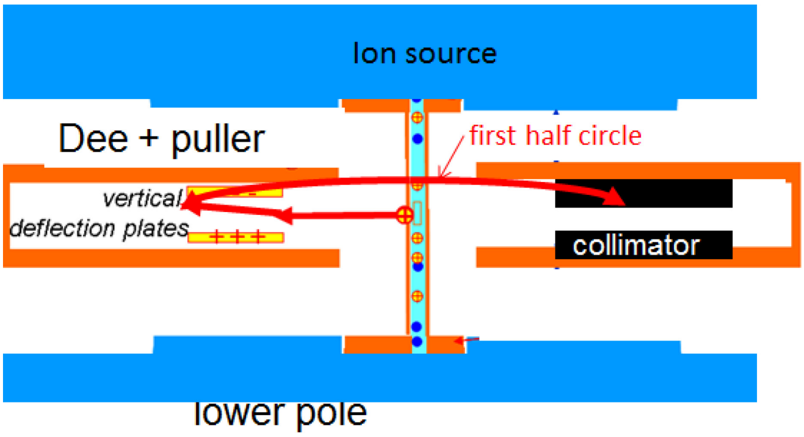


Fig. 3: With an adjustable vertical field between vertical deflection plates, the proton beam can be deflected upwards and will be stopped at a collimator jaw. The intensity of the beam, behind the slit between the jaws, is a function of the voltage difference between the vertical deflector plates.

The spiral orbit must be well centred in the magnetic field. A combination of this off-centring and horizontal (radial) betatron oscillations will lead to overlapping turns and to an increase of the horizontal betatron-oscillation amplitude. Due to coupling resonances between horizontal and vertical (axial) betatron oscillations, this may also induce large vertical-oscillation amplitudes. Due to the limited space between the magnet poles and the Dee plates, such large amplitudes could easily cause severe beam losses. Apart from a lower extracted beam intensity, this will also yield neutron production and activation of the cyclotron. This makes the service more complicated.

Centring of the beam is performed by small shifts of the orbit positions, due to a small controlled local variation of the magnetic field in the central region of the cyclotron. This can be done by means of correction coils mounted at the pole surface, or by means of iron pieces that can be shifted, to adjust the pole gap and thus the magnetic field, locally. If one measures the beam intensity as a function of radius in the cyclotron (using a radial probe), an optimal centring is reached by adjusting the magnetic field corrections such that the intensity fluctuations as a function of radius are minimized, as illustrated in Fig. 4.

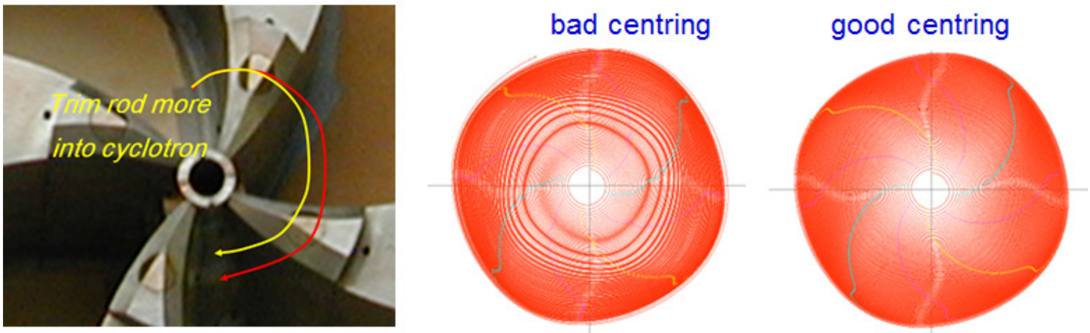


Fig. 4: Left: the beam is centred in the first few turns in the cyclotron by adjusting some iron pieces to make some local field fluctuations. A well-centred beam will have a very smoothly varying intensity profile in the radial direction (right). A badly centred beam will show a very irregular radial intensity pattern (middle), since turns with different centres overlap each other at certain radii.

5 Vertical focusing during in acceleration in a cyclotron

Vertical focusing is essential during the relatively long distance that that particles have to cover during their acceleration in the cyclotron. Above energies of a few MeV, the focusing due to the electric fields in the acceleration gaps can be neglected. So, focusing is mostly done by the magnetic field. Two vertical focusing processes can be distinguished: *weak and strong focusing*. Weak focusing is used in cyclotrons in which the field decreases as a function of radius. As will be discussed later, often this is the case in cyclotrons that use a very strong magnetic field, necessary to obtain a small machine radius. As shown in Fig. 5, the outward curvature of the field lines causes a component of the Lorentz force in the vertical direction, towards the median plane. It acts on particles that are either above or below the median plane. In most cyclotrons, however, the field is increasing with radius in the cyclotron to compensate the relativistic mass increase of the protons at high energies. This will cause vertical defocusing.

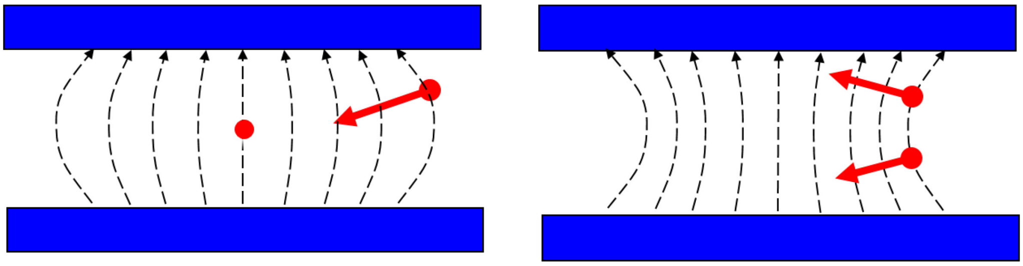


Fig. 5: The field lines of the magnetic field and their effect on the vertical focusing. Left: a field which decreases in strength with radius in the cyclotron. Right: a field that increases in strength with radius.

To compensate for vertical defocusing, a stronger vertical focusing is added by an azimuthal variation in the magnet poles by means of ‘hills’ and ‘valleys’. Then, the particles experience a varying field: the field is stronger between the hills and weaker between the valleys. Due to this varying field, the orbits are no longer following an exact circular shape. The orbit has different radii of curvature when crossing a hill or valley, and the boundaries between hills and valleys are crossed non-perpendicular. A non-perpendicular crossing of a change in magnetic field always results in a vertical focusing or defocusing. However, a repeated focusing and defocusing of equal strengths will yield an effective focusing, which is called *strong focusing*. To compensate for the increasing energy with radius and the increasing defocusing by the main field, the angle by which the hill–valley boundary is crossed is increased by making the hill–valley structure spirally shaped, as shown in Fig. 6.

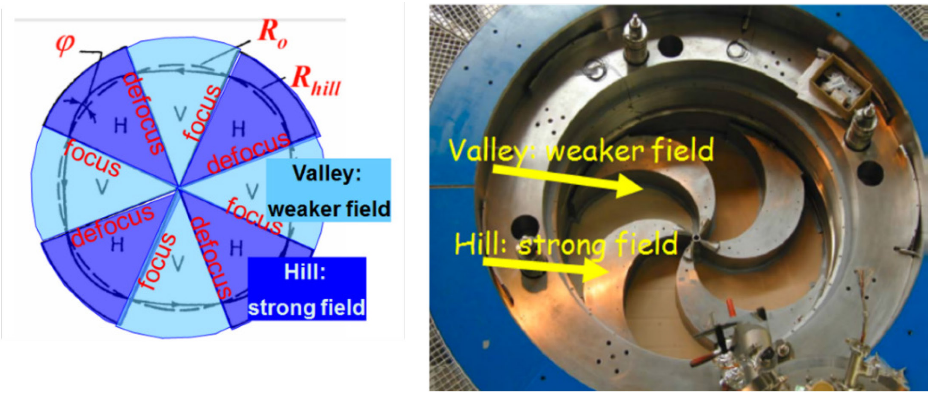


Fig. 6: A structure of hills and valleys on the surface of the poles creates an azimuthally fluctuating magnetic-field strength along the almost circular orbit of the particles. This provides the vertical focusing.

6 Synchrocyclotron

In order to reduce the diameter of a cyclotron, one must use a stronger magnetic field. This can be achieved with superconducting coils in the magnet. But, as has been mentioned before, in cyclotrons with a very strong magnetic field (4–10 T), the field will decrease as a function of the radius within the cyclotron, due to saturation of the iron at such strong fields and the coil geometry. However, according to Eq. (2), the time to make one turn will increase at lower field strength. Then, the particles will cross the acceleration gaps at too late a phase and will obtain a lower and lower energy increase per turn. At a certain radius, they will be lost. In order to deal with this, the frequency of the RF signal is decreased as a function of time. First, the frequency is matched to the revolution time in the central orbits. Then, the frequency is decreased synchronous to the increasing revolution time at the larger radii. In this way, a group of particles is accelerated from source to extraction, but, in the meantime, particles at other radii cannot be accelerated, since their revolution time does not match with the frequency of the RF signal .This shifting of the frequency of the RF signal is repeated at, approximately, a few hundred Hertz (maximum 1 kHz). As shown in Fig. 7, the beam intensity from such a so-called synchrocyclotron is thus pulsed with this frequency.

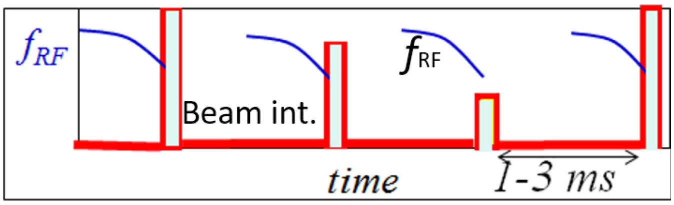


Fig. 7: The RF and the beam intensity as a function of time, as extracted from a synchrocyclotron

Currently, several synchrocyclotrons are in development to reduce the size (and price) of a cyclotron for therapy. Since 2010, a synchrocyclotron with a mass of approximately 20 tons, produced by Mevion, has been in operation. It has been mounted on a gantry rotating around the patient [11]. This system is applying the scatter technique, however. In a system recently developed by the company IBA, a small superconducting synchrocyclotron [12] is directly coupled to a spot-scanning gantry in a system as compact as possible. Spot scanning can be performed by applying spots at the frequency at which the frequency of the RF signal is varied. To apply the correct dose, the pulse intensity from the ion source must be set accurately at this frequency (for an ion source this is quite a fast reaction). To obtain sufficient accuracy of the total dose per spot, multiple irradiations of each spot are expected to be necessary.

7 Extracting the beam from a cyclotron

When the particles have reached the outer radius of the cyclotron, they have to be extracted from the strong magnetic field. This is achieved by means of a septum: a foil between the last turn in the cyclotron and one of the particles that have to be extracted. At just a few-millimetre-larger radius in the cyclotron, an electrode bar is mounted, parallel to the foil. A strong electric field between the bar at negative potential and the grounded foil pulls the orbit that is to be extracted to a larger radius, where the magnetic field is lower. Using focusing elements, the extracted beam is then guided out of the cyclotron through a channel in between the coils of the cyclotron magnet to the entrance of the beam line.

As can be derived from Eq. (1), the increase per turn of the orbit radius will be smaller at higher energies. Therefore, the orbit separation is rather small at extraction radius and orbits may even overlap. This would cause beam losses at the septum, which should be prevented because of undesired high power loss, activation, and neutron production due to scattered particles. Several methods are used to increase the orbit separation at the extraction septum. The most straightforward methods are to make the cyclotron not too small and to use a high Dee voltage (therefore, a high energy gain per turn). Further, one could exploit extra orbit shifting by a local decrease of the magnetic field, for example, by means of a groove in the pole hill at extraction radius. Another method to increase the radial separation is the so-called resonant extraction, in which a radial betatron resonance is excited. In Fig. 8, it can be seen that the radial spacing between the turns is approximately 1 mm at extraction radius (in this cyclotron at ~0.8 m). But, due to the resonance excited by a small field bump, an orbit has been shifted so that at a certain azimuth, the orbit of next higher energy will be located at a smaller radius. Here, a large turn separation of almost 5 mm occurs at a radius of 81.5 cm, which is of course where the septum should be placed. Using such methods, it is possible to obtain an extraction efficiency (=extracted beam intensity divided by beam intensity in cyclotron centre) of approximately 80%.

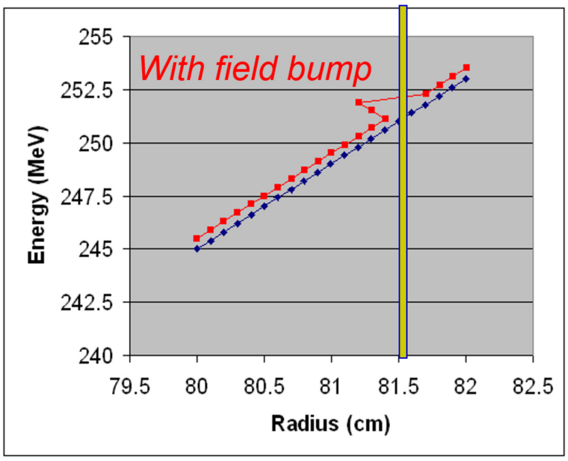


Fig. 8: The beam energy as a function of radius. Each point represents a turn. Due to a field bump, a betatron resonance is excited and, at the azimuth where this data is taken, several turns are located at a smaller radial position, due to the orbit shift associated with the excited resonance. Therefore, a large turn separation occurs between the shifted and non-shifted orbits. The location of the septum is indicated.

The energy of the beam at an extraction radius that can be reached in a cyclotron is determined by its magnetic field, its radial profile, and the extraction radius. Although some groups are working on this, currently these parameters are not easy to change sufficiently quickly. Therefore, cyclotrons in therapy are fixed-energy machines. The energy of the beam transferred to the patient is decreased from the energy at extraction to the desired value by means of an adjustable degrader in the beam line, just after extraction. The emittance of the beam extracted from the cyclotron is of the order of a few π mm mrad—usually the vertical emittance is slightly larger than the horizontal emittance. But, since the degrader system increases the emittance so much, the initial emittance is insignificant. Therefore, the asymmetry and exact beam shape of the extracted emittance only play a minor role. Since the emittance of the degraded beam increases with decreasing energy, the transmission of the beam through the beam-transport system to the patient is strongly energy dependent. Therefore, an energy change in the degrader could be coupled to an intensity change of the extracted beam, to get a more or less energy-independent beam intensity at the patient. However, since an error in this method could unexpectedly yield far too high a beam intensity, developments are still in progress in several groups to do this with the ion source and/or vertical deflector plate in the cyclotron.

8 Relevant cyclotron characteristics for therapy

When considering a cyclotron for particle therapy, several key parameters need to be considered. A list of these parameters and the issues on which they have an effect, is presented below (Table 1). Here, no numbers are given, since these are extremely dependent on conditions such as how the dose is applied to the patient (i.e. using scattering technique, spot-scan technique, or continuous scanning) and on price.

Table 1. Cyclotron parameters that are relevant for therapy applications

Parameter	Most effect on
Energy and its stability	Range, sharpness of dose fall off
Beam size (emittance)	Transmission to patient, pencil-beam size
Beam intensity: structure, stability (kHz), adjustability (range, speed)	Irradiation dose rate, acceptable instantaneous intensity for dosimetry equipment, average intensity—determines the time a treatment takes and affects how spot scanning is done, (im)possibility of continuous scanning
Extraction efficiency	Lifetime of components, activation of cyclotron, time needed for service, personal dose
Modular control systems + comprehensive user interface	Safety, easy control, fast error tracing, help in decisions in case of 'no beam'
Reliability	Number of treatments per year, patient waiting time
Activation level (person dose per year)	Personnel safety, time to access machine
Maintenance interval, maintenance time, maintenance effort	Number of patients, types of treatments
Needed start up time after 'beam off' and after 'cyclotron open' for service	Number of patients, patient waiting time
When using ions: time to switch ion species	Number of patients, patient waiting time
Choice between synchrocyclotron or isochronous cyclotron	Pulses (<1 kHz) versus continuous beam, scanning possibilities, time a treatment takes

Both dose application techniques (scattering and scanning) can be applied by means of a cyclotron as the accelerator. However, to improve efficiency of the delivery process when using scanning, it makes sense to provide the cyclotron with several features that can control the beam intensity quickly, accurately, and reliably. For application of the scattering technique, the beam intensity of the cyclotron should not be too low.

9 Summary

Currently cyclotrons are used in the majority of proton-therapy facilities. Advantages of cyclotrons are:

- continuous beam (however, pulsed when using a synchrocyclotron);
- ‘any’ (low or high) beam intensity;
- very quickly and accurately adjustable intensity;
- great reliability (few components);
- relatively small footprint.

But, of course, there are also some disadvantages:

- due to various beam losses the cyclotron gets radioactive, especially when it has low extraction efficiency;
- one and fixed energy, so one needs a degrader in the beam line to set the desired energy;
- activation of components near degrader;
- no carbon ions (yet).

Of course, the accelerator choice strongly depends on many things, such as the type of dose application, available space, local accelerator experience, and, last but not least, the price. It is very interesting to follow the different innovative programs, both in synchrotrons and in cyclotrons.

Acknowledgement

I would like to thank the numerous colleagues at other particle-therapy institutes and at several companies for the countless discussions at conferences and workshops and the detailed information given at site visits.

References

[1] R. Widerøe, *AEU-Arch. Elektron. Ub.* **21**(4) (1928) 387. http://dx.doi.org/10.1007/BF01656341

[2] E.O. Lawrence and N.E. Edlefsen, *Science* **72** (1930) 376.

[3] R. Wilson, *Radiology* **47** (1946) 487. http://dx.doi.org/10.1148/47.5.487

[4] J.H. Lawrence, *Cancer* **10** (1957) 795.
http://dx.doi.org/10.1002/1097-0142(195707/08)10:4<795::AID-CNCR2820100426>3.0.CO;2-B

[5] B. Larsson, *Br. J. Radiol.* **34** (1961) 143. http://dx.doi.org/10.1259/0007-1285-34-399-143

[6] R. Wilson, *A Brief History of the Harvard University Cyclotron* (Harvard University Press, Harvard, 2004).

[7] J.M. Slater et al., *Int. J. Radiat. Oncol. Biol. Phys.* **22** (1992) 383. http://dx.doi.org/10.1016/0360-3016(92)90058-P

[8] M. Schillo et al., AIP Conf. Proc. **600** (2001) 37. http://dx.doi.org/10.1063/1.1435191

[9] B. Wolf (ed.), *Handbook of Ion Sources* (CRC Press, 1995), ISBN 0-8493-2502-1

[10] I.G. Brown (ed.), *The Physics and Technology of Ion Sources* (Wiley-VHC, Freiburg, 2004), ISBN 3-527-40410-4. http://dx.doi.org/10.1002/3527603956

[11] Website Mevion, http://www.mevion.com

[12] W. Kleeven et al., The IBA superconducting synchrocyclotron project S2C2, Proc. Cyclotrons, 2013, MO4PB02, 119.

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015,
edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Cyclotrons: Magnetic Design and Beam Dynamics

W. Kleeven and S. Zaremba
Ion Beam Applications, Louvain-La-Neuve, Belgium

Abstract

Classical, isochronous, and synchro-cyclotrons are introduced. Transverse and longitudinal beam dynamics in these accelerators are covered. The problem of vertical focusing and iscochronism in compact isochronous cyclotrons is treated in some detail. Different methods for isochronization of the cyclotron magnetic field are discussed. The limits of the classical cyclotron are explained. Typical features of the synchro-cyclotron, such as the beam capture problem, stable phase motion, and the extraction problem are discussed. The main design goals for beam injection are explained and special problems related to a central region with an internal ion source are considered. The principle of a Penning ion gauge source is addressed. The issue of vertical focusing in the cyclotron centre is briefly discussed. Several examples of numerical simulations are given. Different methods of (axial) injection are briefly outlined. Different solutions for beam extraction are described. These include the internal target, extraction by stripping, resonant extraction using a deflector, regenerative extraction, and self-extraction. Different methods of creating a turn separation are explained. Different types of extraction device, such as harmonic coils, deflectors, and gradient corrector channels, are outlined. Some general considerations for cyclotron magnetic design are given and the use of modern magnetic modelling tools is discussed, with a few illustrative examples. An approach is chosen where the accent is less on completeness and rigorousness (because this has already been done) and more on explaining and illustrating the main principles that are used in medical cyclotrons. Sometimes a more industrial viewpoint is taken. The use of complicated formulae is limited.

Keywords
Cyclotron; extraction; injection; medical applications; magnetic design; synchro-cyclotron.

1 Different types of cyclotron

1.1 The basic equation of the cyclotron—the classical cyclotron

Consider a particle with charge q and mass m that moves with constant velocity v in a uniform magnetic field B. Such a particle moves in a circle with radius r; the centripetal force is provided by the Lorentz force acting on the particle:

$$\frac{mv^2}{r} = qvB \,. \tag{1}$$

The angular velocity is given by

$$\omega = \frac{v}{r} = \frac{qB}{m} \,. \tag{2}$$

https://doi.org/10.23730/CYRSP-2017-001.177

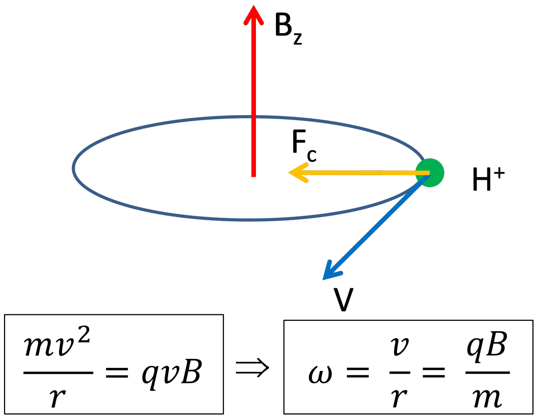

B_z
F_c
H^+
V
$\boxed{\frac{mv^2}{r} = qvB} \Rightarrow \boxed{\omega = \frac{v}{r} = \frac{qB}{m}}$

Fig. 1: At first order, a particle in a cyclotron rotates at constant frequency, independent of radius or energy

This is illustrated in Fig. 1. Thus, the angular velocity is constant: it is independent of radius, velocity, energy (in the non-relativistic limit), or time.

There are a few very important consequences of this feature:

1. particles can be accelerated with an RF system that operates at constant frequency;
2. the orbits start their path in the centre (injection) and spiral outward to the pole radius (extraction);
3. the magnetic field is constant in time;
4. the RF structure and the magnetic structure are completely integrated: the same RF structure will accelerate the beam many times (allowing for a compact, cost-effective accelerator);
5. the operation of the accelerator and thus the beam is a fully continuous wave.

The cyclotron was invented in 1932 by Lawrence and Livingston [1]. This type (quasi-uniform magnetic field) is called the classical cyclotron. The principle of the cyclotron is illustrated in Fig. 2. The frequency of the RF-structure and the magnetic field are related as

$$f_{\mathrm{RF}} \approx 15.2\, h\, \frac{Z}{A}\, B\, . \tag{3}$$

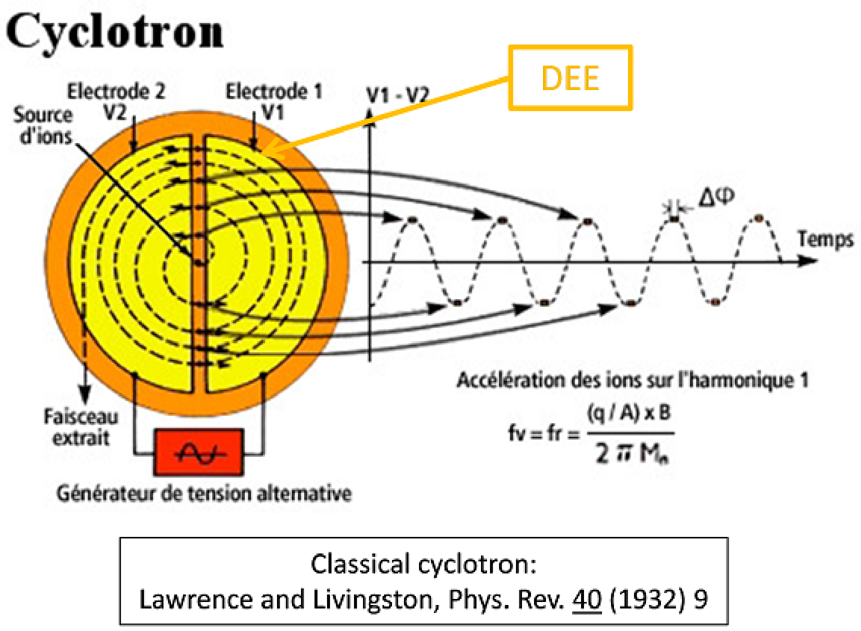
Here Z and A are the charge number and mass number of the particle, h is the harmonic mode of the acceleration $h = f_{\mathrm{RF}} \,/\, f_{\mathrm{ion}}$; f_{RF} is expressed in megahertz and B in tesla.

There is a fundamental problem with the classical cyclotron, which can be seen as follows.

1. In a uniform magnetic field there is no vertical focusing (the motion is meta-stable).
2. During acceleration, the relativistic mass increases; therefore, the angular velocity is actually not constant but gradually decreases (see Eqs. (2) and (4)). A loss of synchronism occurs between the RF and the beam (loss of isochronism).
3. Simply increasing the magnetic field with radius is not possible, because the motion then becomes vertically unstable.

The particle angular velocity taking into account the relativistic mass increase is given by

$$\omega = \frac{qB}{m_0}\sqrt{1-\left(\frac{v}{c}\right)^2}\, . \tag{4}$$


Cyclotron
DEE
Electrode 2
V2
Electrode 1
V1
Source d'ions
V1 - V2
ΔΦ
Temps
Faisceau extrait
Générateur de tension alternative
Accélération des ions sur l'harmonique 1
fv = fr = (q / A) x B / 2 π Mn

Classical cyclotron:
Lawrence and Livingston, Phys. Rev. 40 (1932) 9

Fig. 2: Synchronism in a cyclotron between the particle rotation and the RF accelerating wave (courtesy of Frederic Chautard.)

Here m_0 is the particle rest-mass and c is the speed of light.

Let us make a small sidestep and see how much energy can be achieved with the classical cyclotron. Assume a magnetic field with a small negative gradient, such that some vertical focusing is provided. The magnetic field as a function of radius is given by

$$B(r) = B_0 \left(\frac{r}{r_0} \right)^{-n} . \tag{5}$$

Here, r_0 is some reference radius and B_0 is the field at that radius. The field index n is defined as

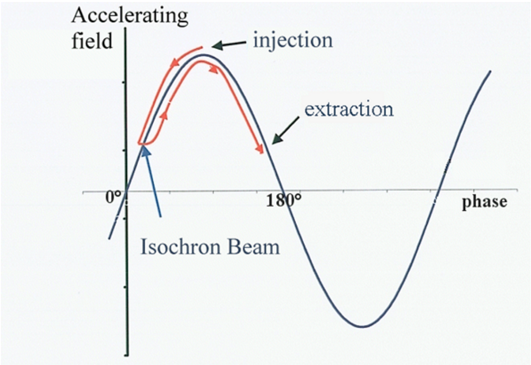
$$n = -\frac{\mathrm{d}B}{\mathrm{d}r} \cdot \frac{r}{B} .$$

Vertical tuning is related to the field index as $\nu_z = \sqrt{n}$. During the acceleration, the particles gradually run out of phase with respect to the RF. However, the RF frequency can be tuned such that, in the cyclotron centre, the magnetic field is too high. Here the particles are extracted from the ion source at an RF phase of approximately 90°, but then the phase will decrease because the RF frequency is too low. Since the magnetic field decreases with radius, there will be, after some number of turns, a moment where the revolution frequency and RF frequency are exactly the same. Beyond that point, the RF phase will start to increase, because the RF frequency is now too high. This RF phase motion is illustrated in Fig. 3.

The longitudinal motion can be studied using a simple Excel model. The energy and phase of the accelerated particle are found by integrating the following equations:

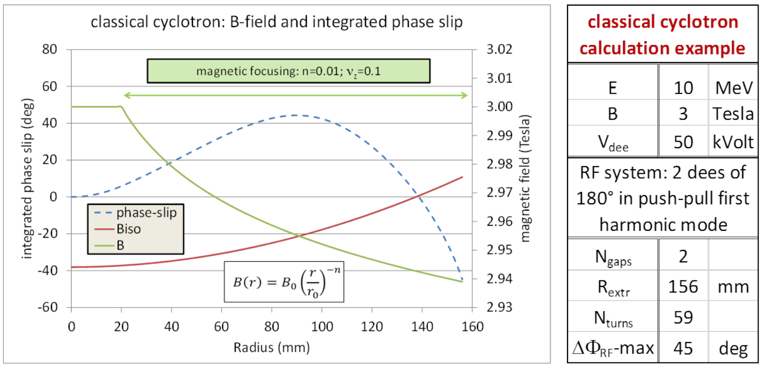
$$\Delta E_n = q N V_{\mathrm{dee}} \cos \Phi_n , \tag{6}$$

$$\Delta \Phi_n = 2\pi h \frac{B - B_{\mathrm{iso}}}{B_{\mathrm{iso}}} . \tag{7}$$



Courtesy Frédéric Chautard

Fig. 3: Owing to the relativistic mass increase and the small negative magnetic field gradient, the particle in a classical cyclotron runs steadily out of phase with respect to the RF. Nevertheless, acceleration can be obtained during a limited number of turns (courtesy of Frederic Chautard.)



classical cyclotron calculation example		
E	10	MeV
B	3	Tesla
V_{dee}	50	kVolt
RF system: 2 dees of 180° in push-pull first harmonic mode		
N_{gaps}	2	
R_{extr}	156	mm
N_{turns}	59	
$\Delta\Phi_{RF}$-max	45	deg

Fig. 4: A simple Excel model can show how much energy can be reached in a weak-focusing rotational symmetric cyclotron, depending on the dee voltage, the field index, and the magnetic field value.

Here, ΔE_n is the energy gain at turn n, V_{dee} is the maximum dee voltage, N is the number of accelerating gaps, Φ_n is the RF phase at turn n, $\Delta\Phi_n$ is the RF phase slip in the turn n, h is the harmonic mode, and B_{iso} is the isochronous magnetic field corresponding to the given RF frequency and taking into account the relativistic mass increase.

An example of such a calculation is shown in Fig. 4. This cyclotron is a candidate for a small superconducting machine for isotope production for positron emission tomography. The main parameters are shown in the same figure. It can be seen that quite a high dee voltage is needed to limit the number of turns and thus the RF phase slip. In this case, an RF system with two 180° dees in push–pull mode was assumed. In such a system, the two opposite dees oscillate 180° out of phase, such that the total maximum energy gain per turn is four times the dee voltage. It is seen that protons of 10 MeV can be obtained in a field of 3 T, at an extraction radius of 156 mm, with a dee voltage of 50 kV, and about 60 turns in the cyclotron. At such a low energy, it is possible to accelerate H^- without substantial losses by magnetic stripping. Such a cyclotron is currently under construction in the CIEMAT Institute in Madrid, Spain [2].

1.2 Another solution: the synchro-cyclotron

A solution for the energy and vertical focusing limitations of the classical cyclotron has been introduced independently by Veksler [3] and McMillan [4]. (Note that the synchrotron was also invented independently by Veksler and McMillan and is described in the same papers.) This solution, the synchro-cyclotron, differs in the following ways from the classical cyclotron:

1. the magnetic field gradually decreases with radius in order to obtain weak vertical focusing:

$$n = -\frac{r}{B}\frac{\mathrm{d}B}{\mathrm{d}r} \Rightarrow \nu_z = \sqrt{n}\,; \tag{8}$$

2. the RF frequency gradually decreases with time, to compensate for the decrease in magnetic field and the increase in particle mass (see Eq. (2)).

This type of cyclotron brings about several important consequences.

1. Much higher energies can be obtained, in the range 100 MeV to 1 GeV.
2. The RF is pulsed but the magnetic field is constant in time (which is not the case in a synchrotron).
3. The beam is no longer continuous wave but is modulated (pulsed) in time.
4. The average beam current is much lower than for a continuous-wave machine (OK for proton therapy).
5. There is a longitudinal beam dynamics similar to that of the synchrotron.
6. The beam can only be captured in the cyclotron centre during a short time-window.
7. The timing between the RF frequency, RF voltage, and ion source needs to be well defined and controlled.
8. A more complicated (but not necessarily more expensive) RF system is needed to obtain the required frequency modulation.
9. The RF frequency cannot be varied very quickly (rotating capacitor) and therefore the acceleration must be slow. This implies the following:
   (a) low energy gain per turn;
   (b) many turns up to extraction;
   (c) low RF voltage and low RF power needed.
10. There is only a very small turn separation at extraction. Therefore a special extraction method (called a regenerative extraction) is needed to get the beam out of the machine.

Recently, IBA has developed a 230 MeV superconducting synchro-cyclotron (S2C2) for proton therapy. The advantage of such a solution, as compared with compact superconducting isochronous cyclotrons, is that the average magnetic field can be increased to substantially higher values, because there is no concern about lack of vertical focusing. Figure 5 shows some properties of this cyclotron. The graph on the right shows the average magnetic field and the vertical focusing frequency (the passive extraction system was not installed in this case). The magnetic field in the centre is about 5.8 T and the extraction radius is about 450 mm; the pole radius is 500 mm. Also shown is the vertical focusing frequency; in this weak-focusing machine, the vertical focusing is produced solely by the negative field gradient. The graph on the left illustrates the time structure of the RF. The pulse length is 1 ms and the corresponding pulse rate is 1 kHz. The RF frequency varies from about 88 MHz (when the beam is captured in the cyclotron centre) to about 63 MHz (when the beam begins to be extracted at $r = 450$ mm). The total acceleration time is about 600 µs, and the number of turns in this cyclotron is greater than 45 000.

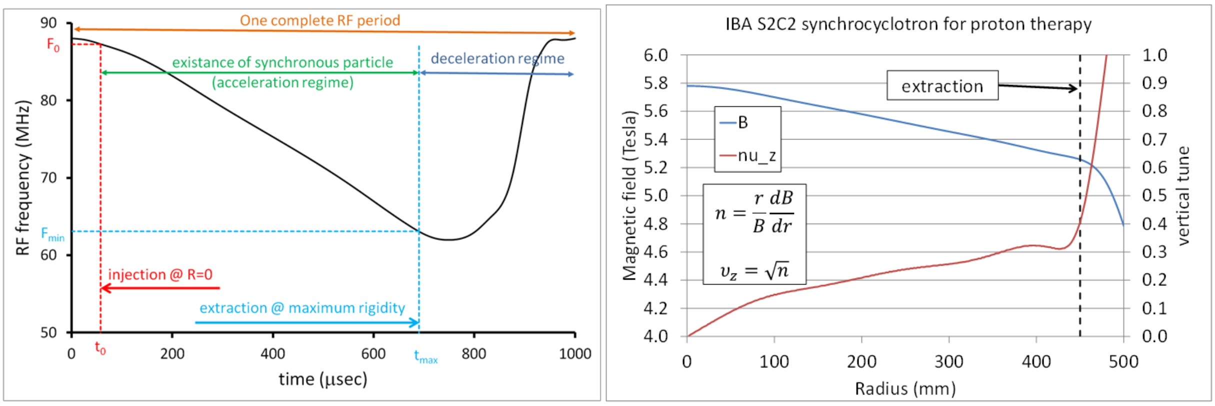
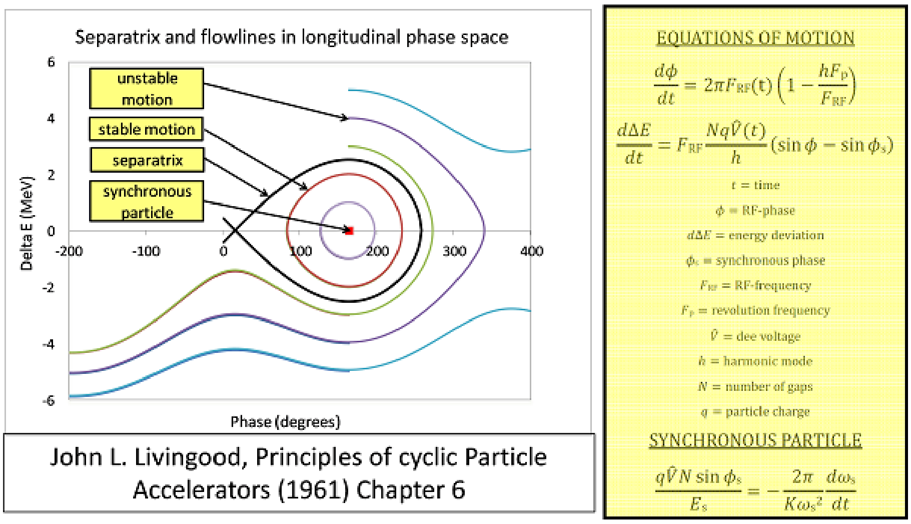


Fig. 5: Some properties of the IBA superconducting synchro-cyclotron (S2C2). Left: modulation of the RF frequency as a function of time within one pulse. Right: average magnetic field and vertical tuning as a function of radius.


Separatrix and flowlines in longitudinal phase space
unstable motion
stable motion
separatrix
synchronous particle
6
4
2
0
-2
-4
-6
Delta E (MeV)
-200
-100
0
100
200
300
400
Phase (degrees)
John L. Livingood, Principles of cyclic Particle Accelerators (1961) Chapter 6

EQUATIONS OF MOTION

$$\frac{d\phi}{dt} = 2\pi F_{\mathrm{RF}}(\mathrm{t})\left(1 - \frac{hF_{\mathrm{p}}}{F_{\mathrm{RF}}}\right)$$

$$\frac{d\Delta E}{dt} = F_{\mathrm{RF}}\frac{Nq\hat{V}(t)}{h}(\sin\phi - \sin\phi_{\mathrm{s}})$$

t = time

ϕ = RF-phase

$d\Delta E$ = energy deviation

ϕ_{s} = synchronous phase

F_{RF} = RF-frequency

F_{p} = revolution frequency

$\hat{V}$ = dee voltage

h = harmonic mode

N = number of gaps

q = particle charge

SYNCHRONOUS PARTICLE

$$\frac{q\hat{V}N\sin\phi_{\mathrm{s}}}{E_{\mathrm{s}}} = -\frac{2\pi}{K{\omega_{\mathrm{s}}}^2}\frac{d\omega_{\mathrm{s}}}{dt}$$

Fig. 6: Longitudinal beam dynamics in a synchro-cyclotron. Left: flow lines and separatrix in longitudinal phase space. Right: equations of motion that govern this phase space.

Figure 6 illustrates the longitudinal beam dynamics in a synchro-cyclotron such as the S2C2. The graph on the left shows the longitudinal phase space. For the synchronous particle, the angular velocity is (by definition) always the same as the RF frequency at all radii in the machine. A non-synchronous particle executes an oscillation around this synchronous particle. The horizontal axis is the RF phase and the vertical axis is the energy difference of the particle with respect to the synchronous particle. For small excursions, particles execute elliptical (symmetric) oscillations around the synchronous point. For larger excursions, owing to the non-linear character of the dynamics, the flow lines start to deform. The separatrix separates the stable zone from the unstable zone. Inside the separatrix, there remains, on average, a resonance between the RF frequency and the particle revolution frequency, and the particle will be accelerated. Outside the separatrix, there is no longer a resonant acceleration of the particle and it will stay close to a fixed radius in the cyclotron. The right panel of Fig. 6 shows the equations of motion that govern the longitudinal phase space. More explanations of this can be found elsewhere, e.g., in the textbook by Livingood [5].

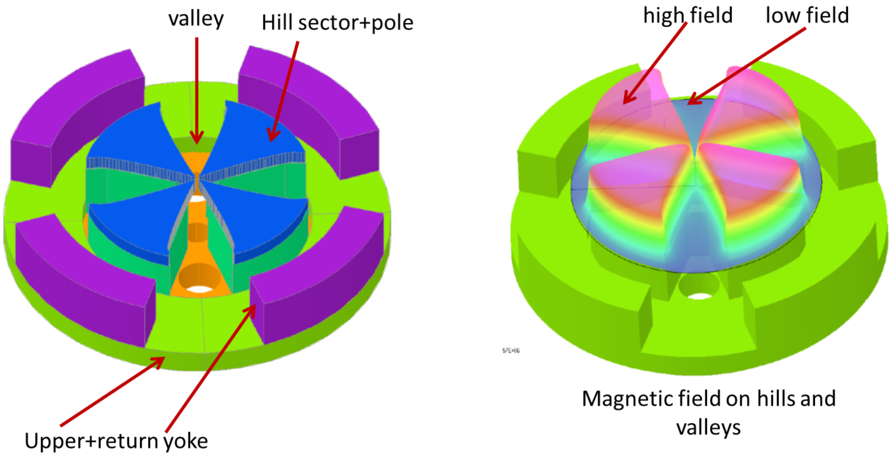


Fig. 7: Magnet of a compact azimuthally varying field cyclotron. Left: the hill sectors and poles, the valleys, and the different parts of the yoke. Right: histogram of the magnetic field.

1.3 The isochronous cyclotron

In the isochronous cyclotron, an additional resource of vertical focusing is introduced by allowing the magnetic field to vary with azimuth along a circle. This additional focusing is so strong that it dominates the vertical defocusing arising from a radially increasing field. The radial increase can be made such that the revolution frequency of the particles remains constant in the machine, even for relativistic energies (for which the mass increase is significant). This new resource of vertical focusing was invented by Thomas [6]. In the next section, this type of cyclotron is discussed in more detail.

2 More about compact azimuthally varying field cyclotrons

2.1 Vertical focusing in cyclotrons

To better understand the vertical focusing in a cyclotron, consider the vertical component of the Lorentz force, F_z:

$$F_z = q(\vec{v} \times \vec{B})_z = -q\left(v_\theta B_r - v_r B_\theta\right) \ . \tag{9}$$

The first term, $v_\theta B_r$, is obtained in a radially decreasing, rotationally symmetric magnetic field, such as for the classical cyclotron or the synchro-cyclotron. If only this term is present, this would correspond to the case of weak focusing. The second term, $v_r B_\theta$, requires an azimuthal modulation of the magnetic field. If such a modulation exists, it will, by itself, also generate a radial component of the velocity.

The azimuthal field modulation can be produced by introducing high-field sectors (hills), separated by low-field regions (valleys). This is illustrated in Fig. 7, which shows the magnet of a compact four-fold symmetrical azimuthally varying field (AVF) cyclotron with four hills and four valleys. The hill sectors are mounted on upper and lower plates of the yoke and surrounded by a return yoke placed in between the upper and lower plates. The plates contain circular holes in the valleys, which are used for vacuum pumping or installation of RF cavities. The right panel shows a histogram of the magnetic field in the median plane superimposed on the geometry. This field map was computed using the 3D finite-element software package Opera-3d from Vector Fields Cobham Technical Services [7].

Figure 8 illustrates the vertical focusing in such an AVF cyclotron. The drawing on the left shows how a computed closed orbit oscillates around a reference circle to produce a scalloped orbit. The upper

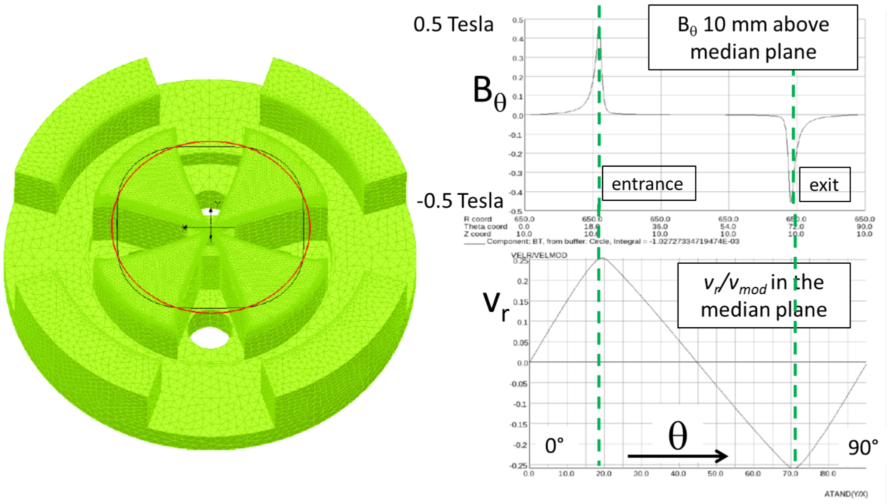


Fig. 8: Vertical focusing in an azimuthally varying field cyclotron. Left: scalloping of the orbit with respect to the geometrical circle. Right: the radial velocity component and the azimuthal field component 10 mm above the median plane. The product of these gives the stabilizing vertical Lorentz forces directed towards the median plane.

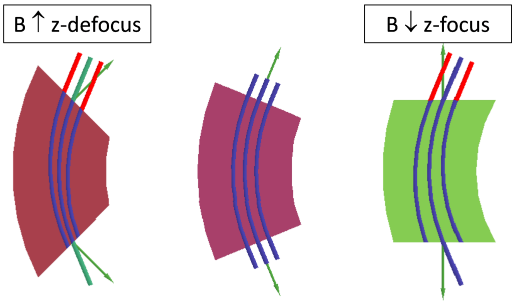


Fig. 9: Vertical focusing in a cyclotron relates closely to vertical focusing at the entrance and exit of a dipole magnet.

graph on the right shows the azimuthal component of the field in a circle 10 mm from the median plane. It can be seen that B_θ is strongly peaked at the entrance and exit of the sector. The lower graph on the right shows the (normalized) radial component of the velocity v_r. The maximum of this component is also at the entrance and exit of the sector. The product of both terms is positive at both the sector entrance and exit, indicating that the vertical focusing is concentrated at these azimuthal locations and is always positive (not alternating).

The vertical focusing in a cyclotron with straight sectors is of the same nature as the edge focusing that occurs at the entrance and exit of the dipole bending magnets. This is shown in Fig. 9. To find the sign of the vertical focusing at an edge, one should draw the normal vector on the orbit, pointing away from the orbit centre. If the magnetic field along this direction is decreasing, then the edge will be vertically focusing. Otherwise, it will be vertically defocusing (see, for example, the TRANSPORT manual [8]).

It is interesting to note that Thomas [6] invented sector focusing (Thomas focusing) in 1938, several years before the invention of the synchro-cyclotron and the synchrotron. However, his solution could not be applied immediately, owing to the increased complexity of the magnetic structure. This is why synchro-cyclotrons have been used at the birth of proton therapy.

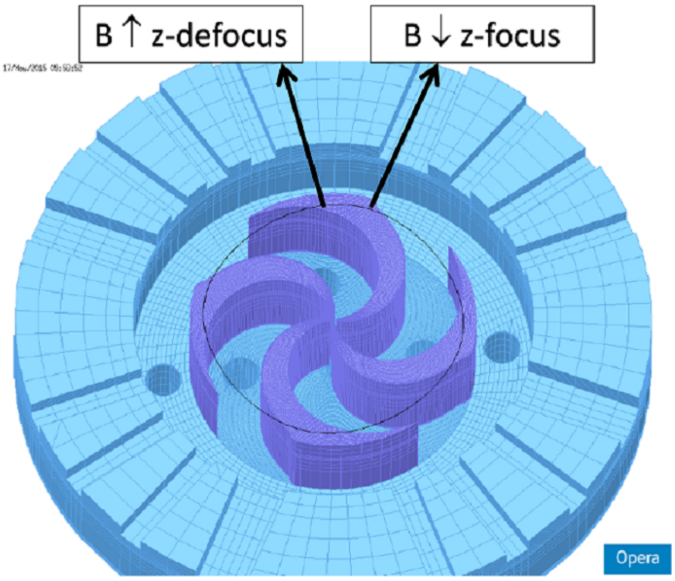


Fig. 10: Vertical focusing in an azimuthally varying field cyclotron can be strongly increased by spiralling the hill sectors.

The vertical focusing created in an AVF cyclotron can be strongly increased if the shape of the sectors is changed from straight to spiral. Figure 10 shows an Opera-3d preprocessor model of a compact cyclotron with spiralled sectors. By drawing the normal vector on the closed orbit at the sector edges, it can be seen that the angle between the orbit and the edge can be made rather large (choosing a large spiral); thus, generating a strong vertical (de-)focusing effect. However, it can also be seen that the direction of the vertical force changes sign between entrance and exit of the sector. Thus, the spiralling of the sectors creates a sequence of alternating focusing, which can become relatively strong. This strong (alternating) focusing was invented by Christofilos [9] and Courant *et al.* [10].

We note that in many compact cyclotrons, the vertical focusing is not only concentrated at the sector edges, but can be more distributed along the closed orbit:

1. edge focusing occurs at the entrance and exit of the hill sectors;
2. for spiral sectors, this focusing starts to alternate and can be made stronger;
3. in the middle of a hill sector, there can be a positive field gradient (e.g., by application of an elliptical pole gap), creating vertical defocusing;
4. in the middle of the valley, there is often a negative field gradient, creating vertical focusing.

The strength of the azimuthal field variation in a cyclotron is expressed in the flutter function F. This function is defined as

$$F(r) = \frac{\overline{B^2} - (\overline{B}\,)^2}{(\overline{B}\,)^2} \,. \tag{10}$$

Here, $\overline{B}$ is the average of the median magnetic field over the azimuthal range from 0° to 360° and $\overline{B^2}$ is the average over the square of this field. The median plane magnetic field can be represented in a Fourier series as

$$B(r,\theta) = \overline{B}(r)\left[1 + \sum_{n=1}^{\infty} A_n(r)\cos n\theta + B_n(r)\sin n\theta\right] \,, \tag{11}$$

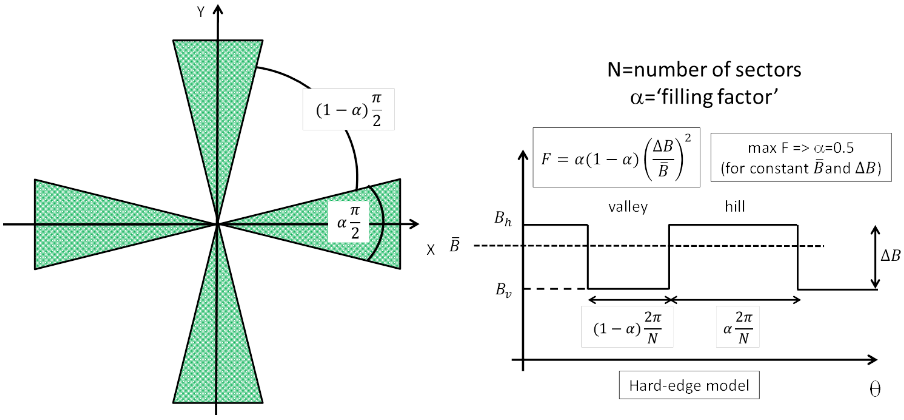

Y
X
$(1-\alpha)\frac{\pi}{2}$
$\alpha\frac{\pi}{2}$
N=number of sectors
α='filling factor'
$F = \alpha(1-\alpha)\left(\frac{\Delta B}{\bar{B}}\right)^2$
max F => α=0.5
(for constant $\bar{B}$ and ΔB)
valley
hill
B_h
$\bar{B}$
B_v
ΔB
$(1-\alpha)\frac{2\pi}{N}$
$\alpha\frac{2\pi}{N}$
Hard-edge model
θ

Fig. 11: Simple model to estimate flutter in a hard-edge approximation

where A_n and B_n are the normalized Fourier harmonics of the field. With this representation of the field, the flutter can be written as

$$F = \sum_{n=1}^{\infty} \frac{A_n^2 + B_n^2}{2} \; . \tag{12}$$

Often, in a compact cyclotron, a hard-edge model of the magnetic field can be used to estimate the flutter. This is illustrated in Fig. 11 for a cyclotron with four-fold symmetry. The drawing on the left defines the hill angle $\alpha\pi/2$ and the valley angle $(1-\alpha)\pi/2$. The parameter α is a kind of filling factor. The drawing on the right shows the hard-edge field approximation, with B_v the field in the valley and B_h the field in the hill. The parameter N is the number of the symmetry periods in the magnet. It is easily seen that for such a model, the flutter takes the form

$$F = \alpha(1-\alpha)\left(\frac{\Delta B}{B}\right)^2 \; , \tag{13}$$

where $\Delta B = B_h - B_v$. Thus, the maximum flutter is obtained for $\alpha = 0.5$, where the hills and the valleys have the same width.

The flutter is a useful quantity, because the betatron oscillation frequencies can be expressed quite precisely in terms of F. The expressions for the vertical (ν_z) and the radial (ν_r) tunings are given by:

$$\nu_z^2 = k \quad + \quad \frac{N^2}{N^2-1} F(1 + 2\tan^2\xi) \; , \tag{14}$$

$$\nu_r^2 = (1-k) \quad + \quad \frac{3N^2}{(N^2-1)(N^2-4)} F(1 + \tan^2\xi) \; . \tag{15}$$

Here, k is the field index and ξ is the spiral angle of the pole. This angle is defined in Fig. 12

We note that Eqs. (14) and (15) are approximations. A better approximation has been obtained by Hagedoorn and Verster [11]; they express the tuning functions in terms of the Fourier components A_n and B_n of the magnetic field.

2.2 Major milestones in cyclotron development

We have seen the main differences between the three types of cyclotron that have been invented, starting in the 1920s. Now, we can make a brief overview of the major milestones that have been achieved in the

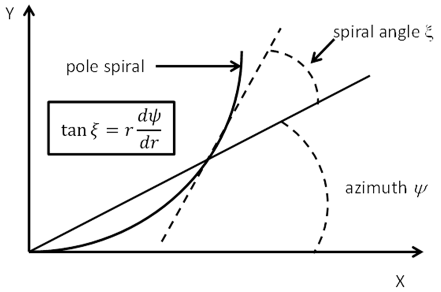


Fig. 12: Spiral angle ξ of the pole

development of the cyclotron. Note that some of the features in this list will be discussed later on in this course.

1. Classical cyclotron (Lawrence and Livingston [1]):
   (a) uniform magnetic field ⇒ loss of isochronism due to relativistic mass increase ⇒ limited energy;
   (b) continuous wave but weak focusing ⇒ low currents.
2. Synchro-cyclotron (McMillan–Veksler [3,4]):
   (a) decreasing $B(r)$ but time-varying RF frequency ⇒ high energies achievable;
   (b) pulsed operation and weak focusing ⇒ very low currents.
3. The isochronous AVF cyclotron (Thomas focusing):
   (a) azimuthally varying magnetic fields with hills and valleys;
   (b) allows both isochronism and vertical stability;
   (c) continuous-wave operation, high energies, and high currents;
   (d) radial sectors ⇒ edge focusing [6];
   (e) spiral sectors ⇒ alternating focusing [9,10].
4. The separate sector cyclotron (Willax [12]):
   (a) no more valleys ⇒ hills constructed from separate dipole magnets;
   (b) more space for accelerating cavities and injection elements;
   (c) example: PSI cyclotron at Villigen, Switzerland;
   (d) very high energy (590 MeV) and very high current (2.5 mA) ⇒ 1.5 MW beam power.
5. H^- cyclotron (TRIUMF, Richardson [13]):
   (a) easy extraction by H^- stripping;
   (b) low magnetic field (centre 3 kG) because of electromagnetic stripping;
   (c) TRIUMF is the largest cyclotron in the world (17 m pole diameter).
6. Superconducting cyclotron: Fraser, Chalk River, Blosser, MSU [14,15]:
   (a) high magnetic field (up to 5 T) ⇒ high energies at compact design.
7. Superconducting synchro-cyclotrons (Wu–Blosser–Antaya [16]):
   (a) very high average magnetic fields (9 T (Mevion) and almost 6 T (IBA));
   (b) very compact ⇒ cost reduction ⇒ future proton therapy machines.

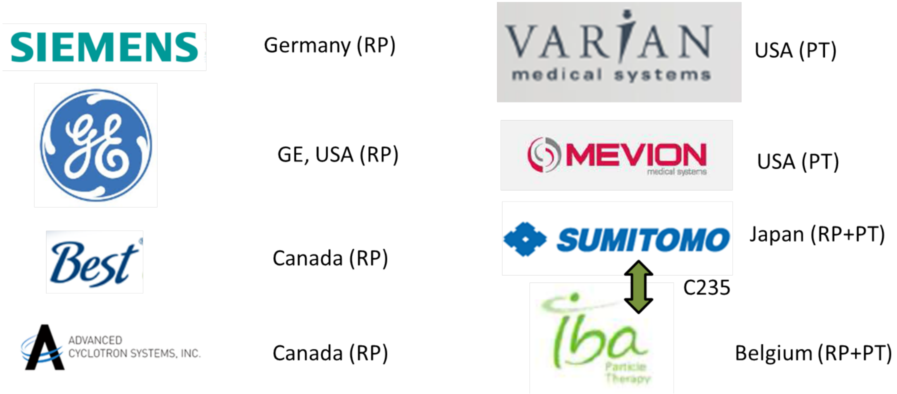


Fig. 13: The most important cyclotron vendors or manufacturers. RP, radiopharmaceuticals (mostly isotope production); PT, proton therapy.

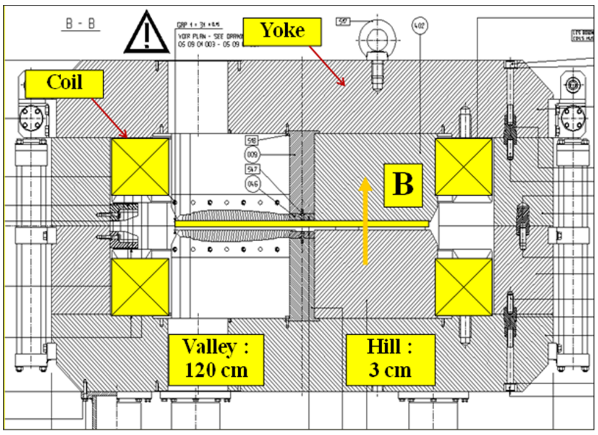
2.3 Commercial cyclotrons for medical applications

The company IBA was founded in 1986. Since then, more than 300 cyclotrons for isotope production have been sold by IBA, as well as about 40 cyclotrons for proton therapy for cancer treatment. There are a few reasons why cyclotrons are so successful in the medical market, for radiopharmaceutical applications such as isotope productions and also for proton therapy applications:

1. cyclotrons are very cost-effective machines for achieving:
   (a) the required energies (<100 MeV for radiopharmaceuticals and <250 MeV for proton therapy);
   (b) sufficient beam current (up to 1-2 mA for radiopharmaceuticals and <1 µA for proton therapy).
2. efficient use of RF power $\Rightarrow$ the same RF cavities are used many (N_{turns}) times;
3. very compact design:
   (a) the magnetic and RF-structures are fully integrated into system;
   (b) the system is single stage $\Rightarrow$ no injector accelerator is needed.
4. for radiopharmaceuticals, the required energies can be achieved with moderate magnetic fields (1–2 T), allowing for conventional magnet technology;
5. simple RF system:
   (a) constant RF frequency (10–100 MHz), allowing for continuous-wave operation;
   (b) moderate RF voltages (10–100 kV).
6. easy injection into the cyclotron (internal ion source or by axial injection);
7. for radiopharmaceuticals, simple extraction based on stripping of H^- ions.

Globally, there are not so many cyclotron vendors and manufacturers. The largest are listed in Fig. 13. Note that Sumitomo and IBA collaborated in the development of the C235 cyclotron.

A classical example of a compact industrial isochronous cyclotron for medical isotope production is the C30, developed by IBA in 1986. The magnet of this machine is shown in Fig. 14 and the cyclotron itself in Fig. 15. It is characterized by the very large ratio of valley pole gap to hill pole gap (so-called deep-valley design). This produces a large magnetic flutter and thus relatively strong vertical focusing. As a consequence, the vertical beam size remains small, so the pole gap in the hills can be small too (30 mm). In a non-saturated magnet, the magnetic field is inversely proportional to the hill gap; thus,



An industrial cyclotron

deep valley
⇓
large flutter
⇓
strong focusing
⇓
small beam size
⇓
small vertical gap
⇓
little resistive coil losses

Place RF cavities in the valleys
Acceleration of H^-

Fig. 14: Deep-valley cyclotron; the concept is used in many compact isotope production and proton therapy cyclotrons.

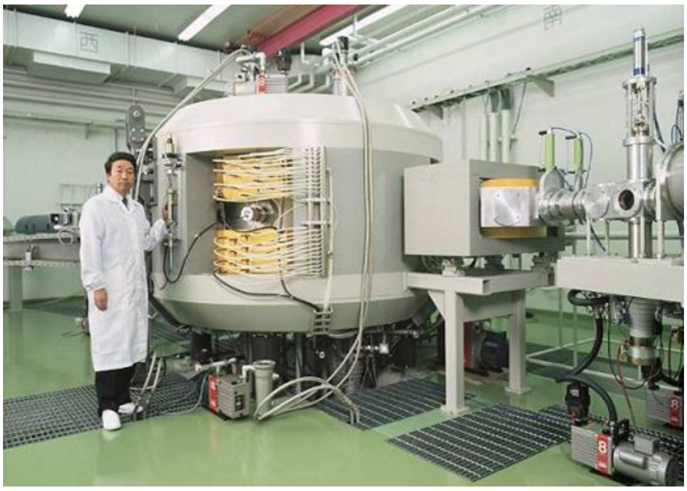


Fig. 15: IBA C30 cyclotron

with such a small hill gap, the ampere turns required in the main coil can be strongly reduced. The success of the different types of cyclotron for isotope production stems from the following features, which were integrated in the design:

1. the deep-valley design, allowing for low electric power dissipation in the coils;
2. the four-fold rotational symmetry:
   (a) allowing a compact design with two RF cavities placed in two opposite valleys;
   (b) two remaining valleys remaining for pumping and diagnostics, ion sources, etc.
3. acceleration of negative ions (H^- or D^-):
   (a) simple extraction by stripping, with almost 100% extraction efficiency.
4. simple injection by the use of an internal Penning ionization gauge (PIG) source or external injection with a spiral inflector.

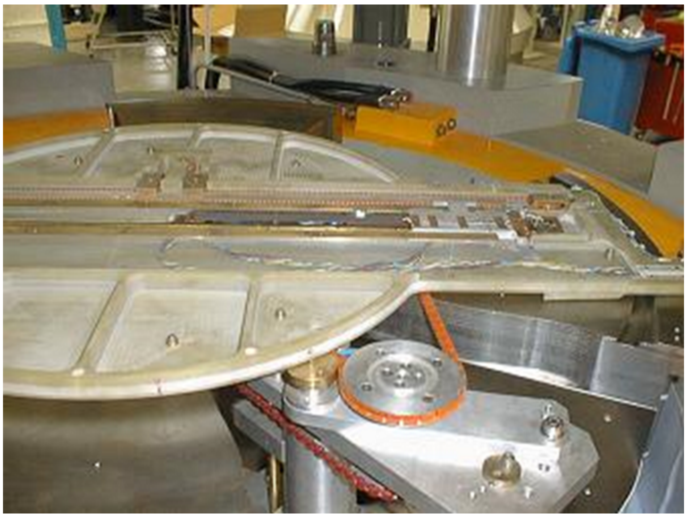


Fig. 16: Mapping system carrying a Hall probe that moves on a polar grid in the cyclotron median plane

2.4 Isochronization of the magnetic field

The cyclotron is perfectly isochronous if the particle angular velocity is constant everywhere in the cyclotron, independent of the energy of the particle. To achieve this, the average magnetic field needs to be correctly shaped as a function of radius. It is impossible to obtain perfect isochronism just from the design of the magnet. The required precision of the average magnetic field is of the order of 10^{-4} to 10^{-5}. To assess the field error, a precise mapping of the magnetic field in the median plane of the cyclotron is needed. This is done with an automated and computer controlled mapping system, such as shown in Fig. 16. The mapping system moves a Hall probe (or a search coil) on a 2D polar or Cartesian grid in order to obtain a full field map. The probe positioning can be pneumatic (compressed air) or motorized. The Hall probes or search coils need to be precisely calibrated against NMR, and possible temperature effects need to be compensated. The field map is analysed by computing equilibrium orbits and determining the revolution frequency as a function of particle energy. The iron of the hill sectors must be shimmed, to improve the isochronism.

Some essential information is obtained from a cyclotron field map.

1. The level of isochronism $\Rightarrow$ the RF phase slip (per turn and accumulated).
2. Information about transverse optical stability $\Rightarrow$ the tune functions of the betatron oscillations.
3. Possible crossing of dangerous resonances $\Rightarrow$ the tune operating diagram.
4. Magnetic field errors $\Rightarrow$ first- and second-harmonic errors may be drivers for resonances.
5. The optical functions (Twiss parameters) of a closed orbit can also be obtained. This may be useful when matching the extracted beam to a beam line or target.

We note that besides the harmonic errors, median plane errors may also exist in a cyclotron. Such errors can push the beam out of the median plane. These errors are very difficult to measure because a possible radial magnetic field component is always much smaller than the main vertical field component; thus, an almost perfect alignment of the Hall probe with respect to the median plane is needed.

An analysis of a magnetic field map can be done at different levels:

1. by Fourier analysis and inspection of the average magnetic field $\bar{B}(r)$ and harmonic field errors;
2. by a static orbit analysis $\Rightarrow$ acceleration is turned off and a series of closed orbits and their properties are determined at the relevant range of energies;

3. by computation of accelerated orbits as needed in specific cases, such as:
   (a) central region studies and design;
   (b) extraction studies;
   (c) studies of resonance crossing.

Closed orbits in a cyclotron are computed by solving the static (non-accelerated) motion [17, 18] of the particle. Two types of closed orbit exist.

1. Equilibrium orbits have the same N-fold symmetry as the cyclotron. They are obtained in the ideal magnetic field map where errors have been removed.
2. Periodic orbits have a periodicity of 2π and are obtained in a real (measured) field map with errors.

Different dedicated programs are available, such as CYCLOPS [18] and EOMSU. At IBA, we use a custom-written program. These programs solve the equations of motion and determine the proper initial conditions, such that the orbit closes in itself.

The closed orbit code computes the phase slip per turn on each orbit. However, the integrated (accumulated) phase slip will depend on the energy gain per turn. For a larger dee voltage, there will be less turns and thus less integrated phase slip. Conversely, the energy gain per turn depends on the RF phase slip that was already accumulated. To take this into account, a self-consistent formula is needed, as follows [18]:

$$\Phi(E) = \arcsin\left(\frac{2\pi h}{f_{\mathrm{RF}}}\int_0^E \frac{\Delta f(E')}{\Delta E_0(E')}\mathrm{d}E'\right) . \tag{16}$$

Here, Φ is the integrated phase slip, h is the harmonic mode, f_{RF} is the RF frequency, Δf is the closed orbit frequency error, and E_0 is the nominal energy gain per turn.

2.5 Different ways to isochronize a cyclotron

Often, cyclotrons for medical isotope production or proton therapy are fixed-field, single-particle machines. In such a case, the isochronization of the magnetic field can be achieved by shimming the iron of the pole sectors. This is illustrated in Fig. 17. Each pole contains an (easily) removable pole edge that can be shimmed. For a rough estimate of the shimming δ that is needed to compensate a certain field error ΔB, a hard-edge model can be used, as illustrated in the lower left panel of Fig. 17. This gives

$$\Delta\bar{B} = \frac{\delta}{2\pi}\Delta B , \tag{17}$$

where ΔB is the difference between the hill field and the valley field. Care must be taken that not too much iron is removed. Several iterations are often needed, with some safety margin applied each time. A hard-edge model is not so precise and does not take into account the effect that magnetic flux is not completely removed together with the iron that has been cut, but may redistribute to radii other than where the cut was made. This may particularly occur when the iron is saturated. A better estimate can be obtained by using a 3D finite-element code (such as Opera-3d) to calculate the effect of a multitude of individual small shims at gradually increasing radius on the pole edge. Then, for each pole cut, the modification of the average magnetic field as a function of radius is obtained. This is illustrated in the right panel of Fig. 17. From this, a shimming matrix is obtained, which relates the change of the average field $\Delta\bar{B}(r_1, r_2)$ at the radius r_2 due to a small cut at radius r_1. Such a matrix needs to be calculated once for a given (prototype) cyclotron and can then be used to speed up the isochronization of all following cyclotrons of the same type. Figure 18 shows, as an example, the IBA C235 cyclotron for proton therapy. There are three removable pole edges on each pole.

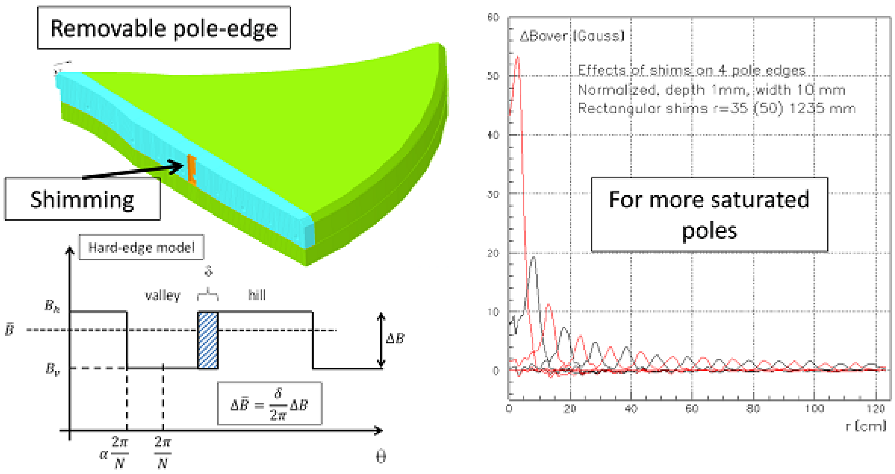


Fig. 17: Upper left: isochronization of a cyclotron magnetic field is achieved by machining the profile of a removable pole edge. Lower left: a rough estimate of the shimming effect (change of the average magnetic field) can be obtained with a hard-edge model. Right: a better prediction of the shimming can be obtained by calculating (with a 3D finite-element code) the change of the radial profile of the average magnetic field due to a well-defined pole cut and repeating this for several cuts at varying radii.

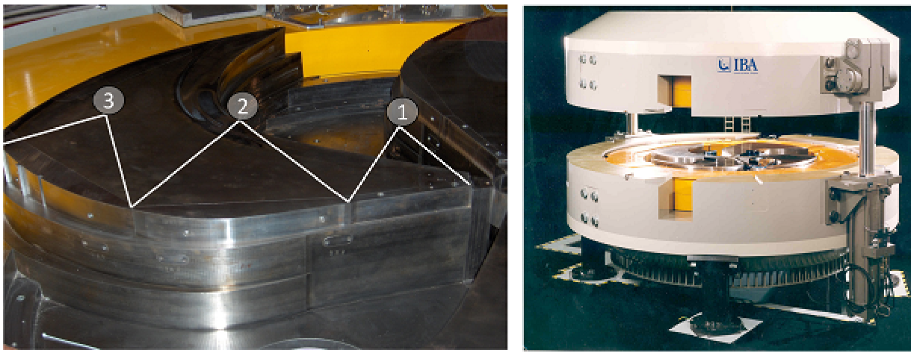


Fig. 18: Left: IBA C235 cyclotron for proton therapy. Right: in this cyclotron there are three removable pole edges for isochronization of the magnetic field.

Modern cyclotrons for the production of PET isotopes can often accelerate two types of particle, namely H^- and D^- ions. For D^- ions, about half the energy of H^- ions can be obtained. The extraction is achieved by stripping and the energy can be varied by moving the radial position of the stripping foil; the magnetic field remains fixed. However, the relativistic field correction needed for H^- ions is about four times as large as for D^- ions. Thus, two different isochronous field maps need to be made in the machine. In the IBA cyclotrons, this is done with the so-called ‘flaps’; these are movable iron wedges that are placed in two opposite valleys in the cyclotron. For the H^- ion field, the flaps are moved vertically to a position close to the median plane. In this configuration, the average magnetic field increases approximately 2% (for the IBA C18/9 cyclotron) in order to create the H^- isochronous field shape. For D^- ions, the flaps are moved farther away from the median plane, such that their contribution to the field is strongly reduced. In the cyclotron, there are still removable pole edges on the hills, which can be shimmed to create an isochronous field shape for the deuterons. The wedge shapes of the flaps are optimized, to create the isochronous field shape for the protons. This optimization needs to be done only once (for the prototype cyclotron). The geometry of the C18/9 cyclotron is shown in Fig. 19, together

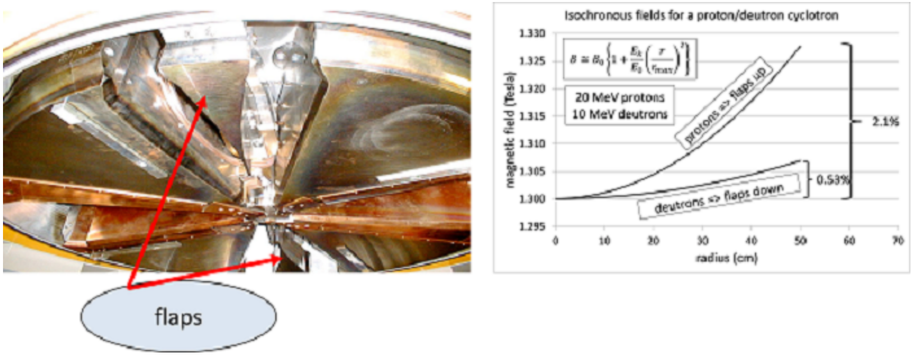


Fig. 19: Left: upper poles of the IBA C18/9 cyclotron. The movable iron wedges (flaps) are used to change the average magnetic field profile from protons (flaps close to the median plane) to deuterons (flaps farther away from the median plane). Right: relativistic correction of the magnetic field, as needed for protons (about 2.1%) and for deuterons (about 0.53%).

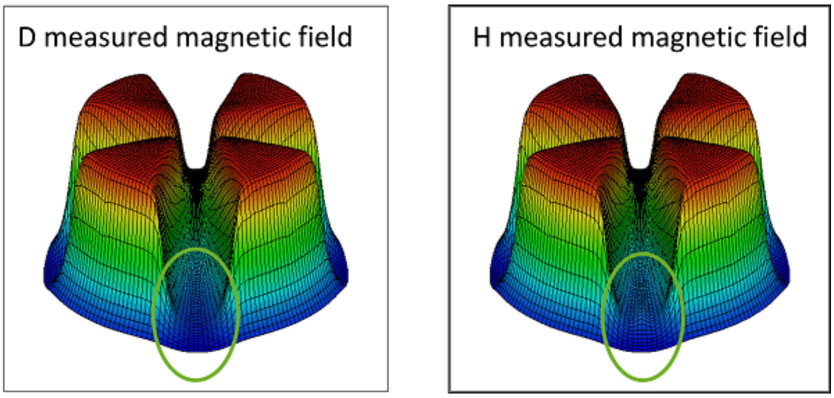


Fig. 20: Finite-element Opera-3d simulation of the effect of the flaps in the IBA C18/9 cyclotron. Left: the neutron field with the flaps farther away from the median plane. Right: the proton field with the flaps close to the median plane.

with an illustration of both the proton and deuteron isochronous field profiles. Figure 20 shows a finite-element Opera-3d simulation of the effect of the flaps in the IBA C18/9 cyclotron. The right panel shows the proton field, where the flaps are close to the median plane, and the left panel shows the deuteron field, where the flaps are farther away from the median plane.

The flaps, as discussed, cannot so easily be applied to higher-energy cyclotrons (e.g., 70 MeV p and 35 MeV d machine). For such a cyclotron, a larger relativistic correction is needed (about 7% for 70 MeV p), which cannot be produced by the 'floating flaps' geometry. A way to solve this is to connect the flaps with an iron cylinder to the base of the valley, as illustrated in Fig. 21. Here, much more magnetic flux is guided towards the median plane, owing to higher magnetization of the iron of the flaps. To produce both (p and d) field maps one could again move the flaps vertically, close to (or away from) the median plane, but in this case the cylinder, attached to the flaps, also has to move into a circular hole machined into the return yoke. A much simpler solution would be not to move the flaps at all but to place a solenoid coil around the cylinder; in this way, the flux guidance from the base of the valley towards the median plane can be set (and optimized) by the DC current in the solenoid (see Fig. 21).

Yet another method is applied in the IBA C70XP cyclotron [19]. This cyclotron accelerates four different particle types, namely H^- to 70 MeV and ions of D^-, H_2^+, and $^4He^{2+}$ to 35 MeV. There

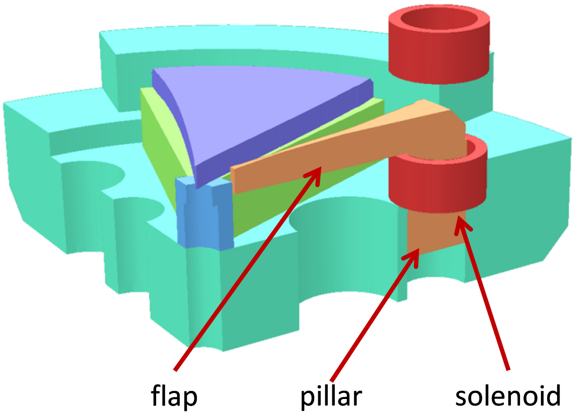


Fig. 21: Alternative method for isochronization of a dual-particle cyclotron. In this case, the flaps are in a fixed position and connected magnetically to the base of the return yoke by a solid iron cylinder. The magnetic flux into the flaps is controlled by the solenoidal coil around the cylinder. This method has not yet been realized but has been patented by IBA.

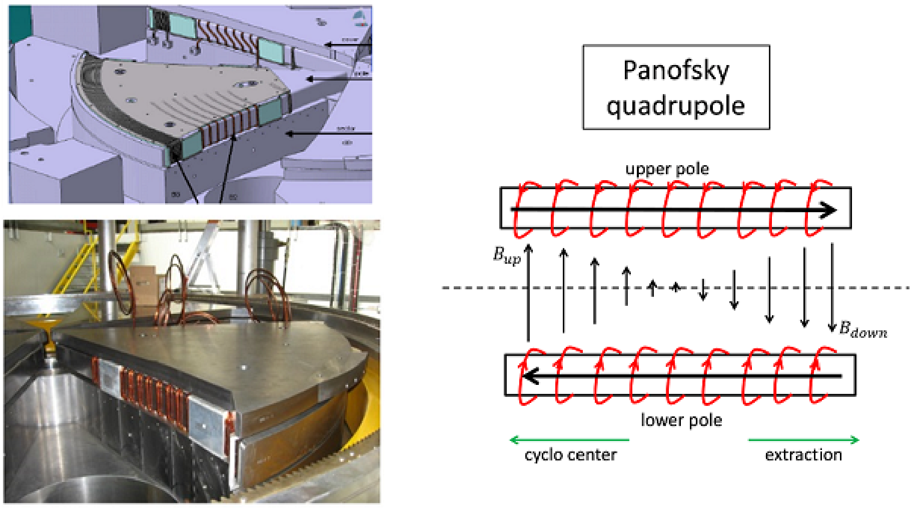


Fig. 22: Yet another method for isochronization of a dual-particle cyclotron is to split the pole into three layers and to wind coils around the middle layers. As explained in the right figure, these coils push magnetic field from the outer pole region towards the centre or vice versa, depending on the polarity of the coil current. Note that in the upper left figure, the cover has been removed.

are two different isochronous field shapes: the first is for the $q/m = 1/1$ particle (H^-) and the other for the $q/m = 1/2$ particles (D^-, H_2^+ and $^4He^{2+}$). The hill sectors are divided into three layers (lower = sector, middle = pole and upper = cover) with an air gap above and below the middle layer). This enables winding of a coil around this pole, as illustrated in Fig. 22. With this coil, magnetic flux can be pushed from the extraction region towards the centre or vice versa, thus modifying the profile of the average magnetic field. In the actual machine, there is not one coil but three independent coils wound around each pole, which are needed to shape the average field with sufficient precision.

A similar, more general, method is applied in the AGOR superconducting cyclotron [20, 21] (Fig. 23). This is a variable-energy, multiparticle, superconducting cyclotron that has been mainly used

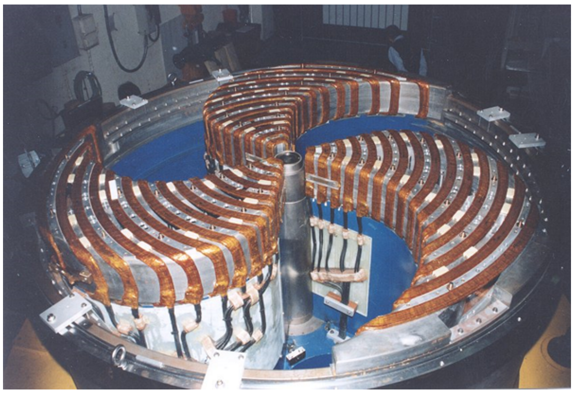


Fig. 23: A method similar to the one outlined in Fig. 22 is used in the AGOR superconducting cyclotron [20, 21]

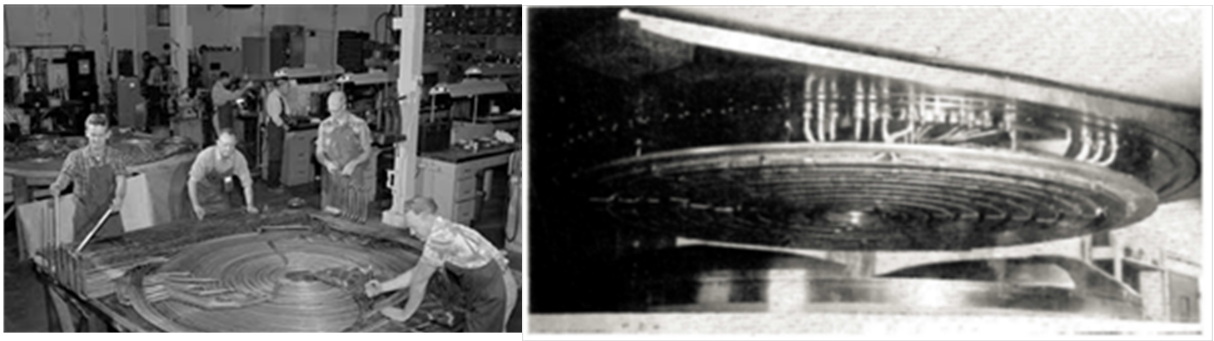


Fig. 24: A more general method is used to isochronize multiparticle variable-energy cyclotrons. Here, an array of concentric circular coils is placed on the pole and each coil uses its own independent power supply. In this way, there is a lot of flexibility in creating the required average magnetic field profile. Left: correction coils for the Berkeley 88-inch cyclotron [24]. Right: correction coils of the Philips 30 MeV azimuthally varying field cyclotron [23].

for nuclear physics research. It requires a much broader range of magnetic field maps and adjustments. To obtain this kind of flexibility, there are 15 independent Panofsky-type correction coils placed around each pole.

Another general method, used for variable-energy multiparticle cyclotrons is to place a set of independent circular coils on the pole of the cyclotron. This is, for example, applied in the Berkeley 88-inch cyclotron [22] and the Philips 30 MeV variable-energy cyclotron [23], as illustrated in Fig. 24.

2.6 Example: A 70 MeV industrial cyclotron for isotope production

Recently at IBA, a 70 MeV cyclotron for the production of medical radioisotopes has been designed and constructed. This cyclotron is under commissioning at the time of writing (end of 2015). The C70 is a high-intensity (750 μA), four-fold symmetrical H^- cyclotron with axial injection and dual beam extraction by stripping. The magnetic structure is given in Fig. 25. In this single-particle, fixed-field machine, isochronization is ensured by shimming the removable pole edges. It can be seen that a small amount of pole spiral has been applied to increase and fine tune the vertical betatron frequency ν_z. The right panel of Fig. 25 shows the shape of the average magnetic field and the flutter in this cyclotron.

The average magnetic field increases by approximately 1% per 10 MeV up to 7% at the maximum radius. It is seen that the flutter goes to zero in the centre of the cyclotron. To provide some additional

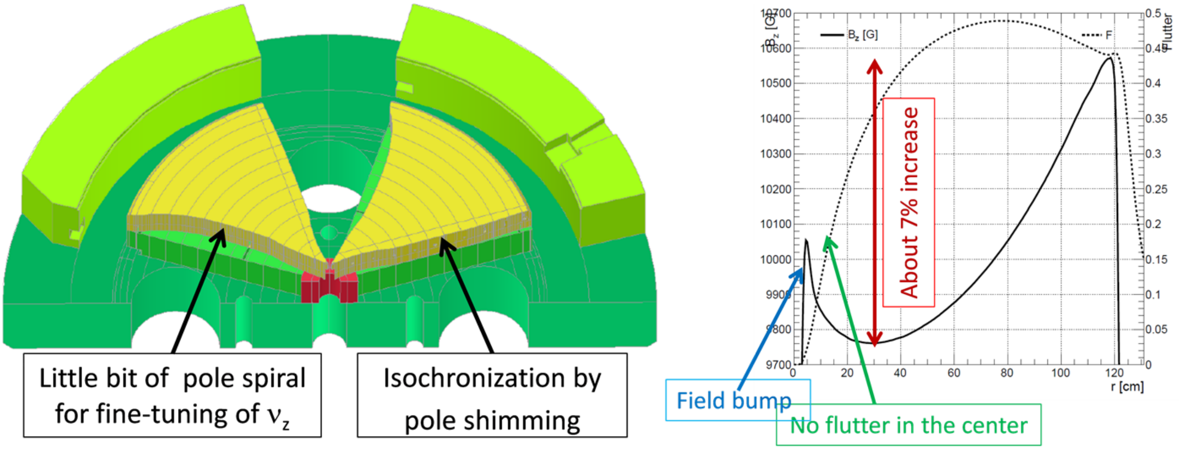


Fig. 25: Left: the IBA C70 as an example of an industrial cyclotron for medical isotope production, showing the small spiral in the pole to adjust vertical focusing and the removable pole edge for isochronization. Right: average magnetic field and flutter of this cyclotron.

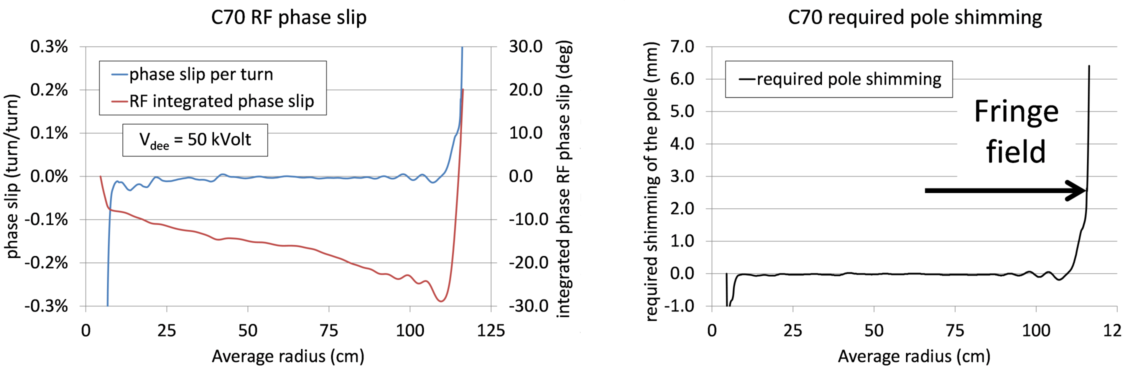


Fig. 26: Left: RF phase slip per turn and integrated RF phase slip in the IBA C70 cyclotron. Right: calculated phase slip per turn can be related to the required shimming of the removable pole edges.

vertical magnetic focusing in this region, a field bump is provided, which generates a negative field index. This zone is not isochronous and thus generates some RF phase slip, but not so much because it is crossed in a few turns. Some additional vertical electrical focusing is obtained at the first few accelerating gaps (as discussed in Section 3). The sharp drop of the magnetic field towards the centre of the cyclotron is due to the axial hole in the return yoke needed for the axial injection.

Figure 26 shows the phase slip per turn and the integrated RF phase slip in this cyclotron. The horizontal axis is the average radius of the closed orbits. These orbits are found up to an energy of 71.4 MeV. The highest energy orbits enter the radial fringe area of the poles where a large part of the accumulated phase slip is generated. The RF frequency is tuned such that the (negative) minimum and (positive) maximum of the integrated phase slip are equal. In this case, the maximum phase slip of 30° is considered acceptable; the actual shift will be smaller because particles are extracted at energies less than 70 MeV. The right panel of Fig. 26 shows the amount of shimming of the pole that would be needed to isochronize the field perfectly. The negative values correspond with cutting of the iron. The sharp rise at the highest radii is due to the fringe field of the magnet. This error is intrinsic to the magnet design and cannot be corrected.

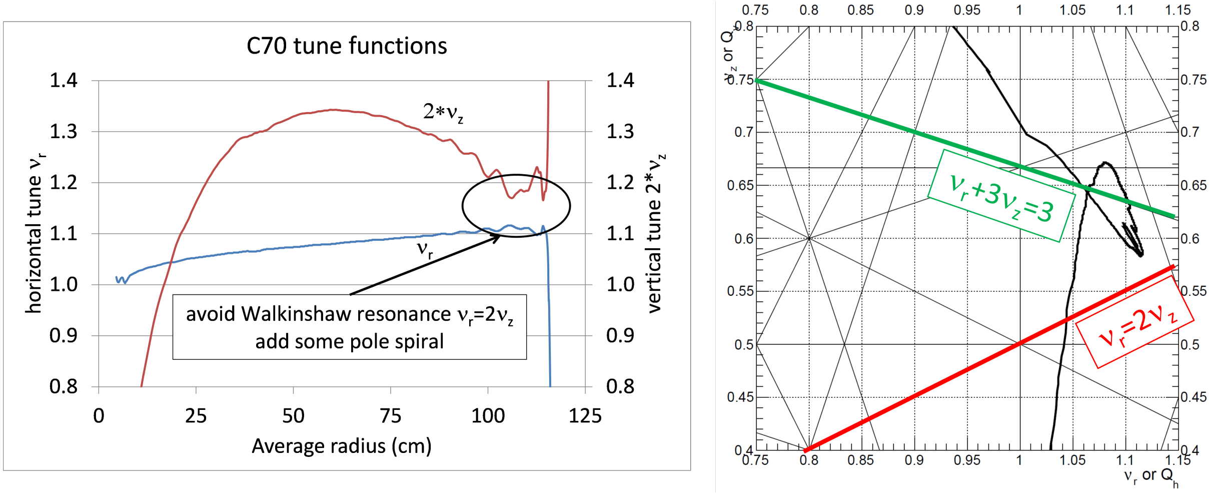


Fig. 27: Left: radial (ν_r) and vertical (ν_z) tune function of the IBA C70 cyclotron as a function of radius. Right: related tuning diagram, showing the working curve and some important resonance lines.

Figure 27 shows the tuning functions and tuning operating diagram of the C70. In the left panel, the vertical tuning, ν_z, has been multiplied by a factor of two, to clearly visualize the situation of the Walkinshaw resonance $\nu_r = 2\nu_z$. This is considered as a resonance that might be dangerous in the region of extraction, especially if the radial beam size is large. In the design of the magnet, the (slight) spiralling of the pole was introduced in an effort to avoid this resonance. This spiral increases the value of ν_z and lifts the red curve in Fig. 27 above the blue one. It is seen in the resonance diagram that besides the Walkinshaw, the resonance $\nu_r + 3\nu_z = 3$ is crossed a few times. However, this resonance is considered much less dangerous, because it is of higher order (four, as compared with three for the Walkinshaw) and also because it is not a structural resonance (it is driven by a field error of harmonic three and not by some intrinsic harmonic of the magnetic field).

2.7 The limit of a three-fold rotational symmetry cyclotron

In this section, we make a small sidestep and ask whether a cyclotron with three-fold symmetry can be used for proton therapy at about 230 MeV. The simple formula ($\nu_r \approx \gamma$) saying that the radial betatron tuning, ν_r, is about equal to the relativistic γ is not valid when the resonance

$$2\nu_r = 3 \tag{18}$$

is approached. This resonance is driven by quadrupole terms (radial field gradients) of symmetry 3. For the cyclotron considered, this is the basic symmetry of the magnetic field; therefore, this resonance is a structural resonance. For a proton therapy cyclotron of 230 MeV, we have $\gamma \approx 1.25$, and one could think that this is still far enough away from the resonance value $\nu_r = 1.5$. However, this is not the case. At IBA, we conducted a small study, to see whether variable-energy stripping extraction could be made in a compact H_2^+ cyclotron with symmetry $N = 3$. The left panel of Fig. 28 shows the Opera-3d magnetic model of this cyclotron. The flutter in the high-field (assumed superconducting) cyclotron is limited by the fact that it is produced only by the iron poles. The right panel of Fig. 28 shows the radial tune function as a function of the radius. The red curve corresponds to the simple approximation $\nu_r \approx \gamma$. The green curve is obtained from the analytic formula derived by Hagedoorn and Verster [11], which is closely related to Eq. (15) (but more precise than it). The black curve is obtained from the closed orbit calculations in the actual field map. Here, it is clearly seen that the theory is no longer valid when the resonance is approached. At an energy of 185 MeV, the beam optics enters the stop band of

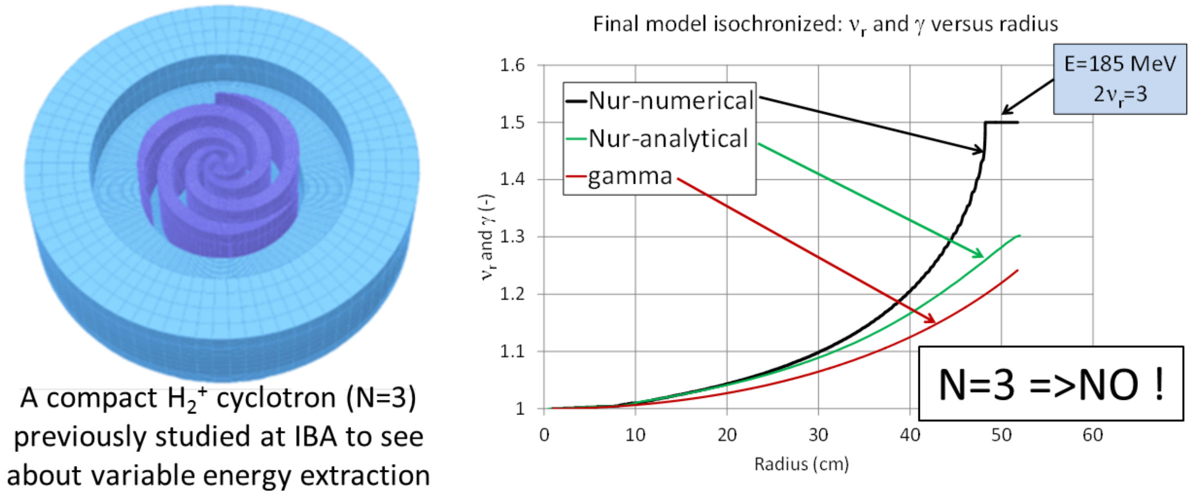


Fig. 28: Energy limitation of a cyclotron with three-fold rotational symmetry due to the $2\nu_r = 3$ resonance. The flutter is produced by iron poles and is therefore limited in a high-field cyclotron. Left: Opera-3d model of the compact H_2^+ cyclotron used in the simulations. Right: ν_r curve as a function of radius, showing the $2\nu_r = 3$ stop band.

the resonance. At this energy, the beam would no longer be stable and thus 185 MeV is the maximum energy that can be achieved.

To better understand this, we repeated the derivation of the Hamiltonian made by Hagedoorn and Verster [11], but instead of studying the radial motion in a coordinate frame rotating at the (dimensionless) frequency equal to 1 (which is a good approximation when $\nu_r \approx 1$), we studied the motion in a coordinate frame rotating at a frequency equal to 1.5 (which should give a better approximation when $\nu_r \approx 1.5$). We note that Hagedoorn's theory is very precise, as long as the flutter is not too large. The results of this analysis are summarized in Fig. 29. The horizontal axis gives the flutter of the magnetic field. The left vertical axis gives the value of the relativistic γ at the $2\nu_r = 3$ resonance. The right vertical axis gives the value of the vertical betatron frequency, ν_z, in the magnetic field. The magnetic field index is chosen such that the cyclotron is isochronous. Different lines in Fig. 29 correspond to different values of the pole spiral angle. The coloured solid lines must be read on the left axis and show at what value of the flutter and what value of the spiral angle, the $2\nu_r = 3$ resonance sets in. It can be seen that, for a given flutter, the maximum energy decreases with increasing spiral angle. Similarly, for a given spiral angle, the maximum energy decreases with increasing flutter. The dashed lines represent the vertical tuning ν_z, which must be read on the right axis. Here, it can be seen that ν_z increases with both increasing flutter and increasing spiral angle. For a stable cyclotron, it is necessary that one remains below the inset of the resonance and also that the vertical tuning is positive. This limit of stability, for both radial and vertical motion, is found by looking for the points on the solid coloured curves for which the vertical tune is exactly zero. Connecting these points gives the bold black line in Fig. 29. This line represents the stability limit of a compact three-fold symmetric AVF cyclotron (with small flutter). It can be seen that in all cases, the maximum relativistic γ is almost the same and corresponds to a maximum proton energy of about 185 MeV.

2.8 The notion of the orbit centre and the magnetic centre in a cyclotron

Betatron oscillations in a cyclotron can be represented by the usual amplitude and phase, but also by the coordinates of the orbit centre. The latter can be more convenient, because the orbit centre oscillates slowly ($\nu_r - 1$) compared with the betatron oscillation itself (ν_r). In the orbit centre representation,

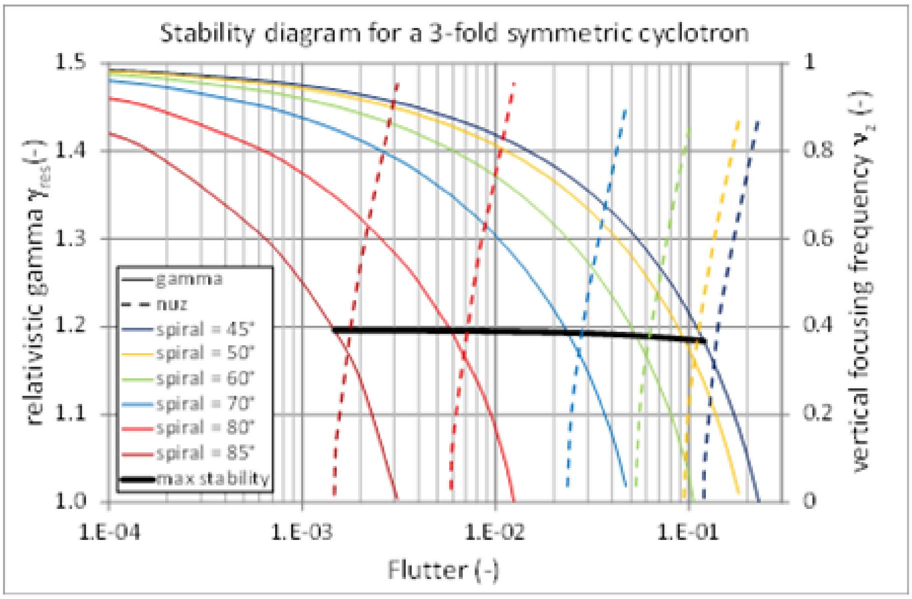


Fig. 29: Stability diagram of a cyclotron with three-fold rotational symmetry, assuming that the flutter is relatively small, as it is produced by the iron poles. The solid black line shows the maximum kinetic energy (relativistic γ) that can be obtained, taking into account the requirements of vertical focusing, as well as the $2\nu_r = 3$ stop band.

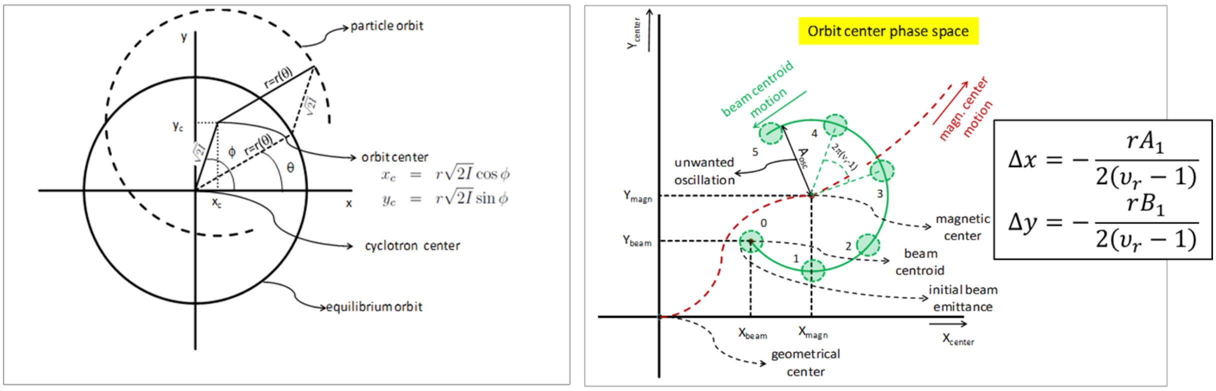


Fig. 30: Left: in a cyclotron, the radial betatron oscillation can be represented as a (slow) motion of the orbit centre. Right: the betatron oscillations take place with respect to the magnetic centre of the cyclotron. Owing to a first-harmonic field error, the magnetic centre shifts away from the geometric centre of the cyclotron.

the equations of motion can be simplified using approximations that make use of the slowly varying character of this motion and the integration can be done much faster. This may be especially useful in a synchro-cyclotron where the particle makes many turns (approximately 50 000 for the S2C2) and full orbit integration from the source to extraction is almost impossible. The left panel of Fig. 30 illustrates the radial betatron oscillation around the equilibrium orbit in terms of the orbit centre. The real orbit can be reconstructed from the orbit centre coordinates and the equilibrium orbit radius $r(\theta)$. In this illustration, a circular equilibrium orbit (synchro-cyclotron) is shown. For AVF cyclotrons, this will be a scalloped orbit.

Particles execute a betatron oscillation around the magnetic centre. A first-harmonic field error displaces the magnetic centre of the cyclotron relative to the geometrical centre. This displacement is given by:

$$\Delta x = -\frac{rA_1}{2(\nu_r - 1)} , \tag{19}$$

$$\Delta y = -\frac{rB_1}{2(\nu_r - 1)} . \tag{20}$$

When there is acceleration, the magnetic centre itself is also moving and the total motion is a superposition of the two separate motions. This is illustrated in the right panel of Fig. 30. The beam quality (emittance) degrades when the beam centroid is not following the magnetic centre. This may occur in two ways.

1. A beam centring error at injection.
2. Accelerating through a region where the gradient of the first harmonic is large. This is a non-adiabatic effect (which will not occur in a synchro-cyclotron where the acceleration is very slow).

The coherent amplitude $A_{\rm osc}$ of the betatron oscillation is a good measure of the harmful effect of the centring error. The numbers in Fig. 30 indicate subsequent turns. Hagedoorn and Verster [11] have derived a Hamiltonian description of the dynamics of the orbit centre. Their theory includes linear and non-linear motion (separatrix), as well as the influence of field errors.

2.9 Harmonic field errors in a map

As discussed in the previous section, a first-harmonic field error will de-centre the closed orbit. This effect becomes large when $\nu_r \simeq 1$. This happens in the cyclotron centre and in the radial fringe region of the pole. An off-centred orbit might become sensitive to other resonances. Excessively high harmonic errors must be avoided. However, a localized first-harmonic field bump may be used to create a coherent beam oscillation, enabling extraction of the beam from the cyclotron (precessional extraction).

The gradient of the first harmonic can drive the $2\nu_z = 1$ resonance. In the stop band of this resonance, the motion becomes vertically unstable. This may lead to amplitude growth and emittance growth. The stop band of this resonance is defined by [25]:

$$4\nu_z \left|\nu_z - \frac{1}{2}\right| < \sqrt{r\frac{\mathrm{d}C_1}{\mathrm{d}r}} , \tag{21}$$

where $C_n = \sqrt{A_n^2 + B_n^2}$.

The gradient of a second harmonic can drive the $2\nu_r = 2$ resonance. In the stop band of this resonance, the horizontal motion becomes unstable. This problem may occur, for example, when second-harmonic iron shims are used to isochronize a dual-particle (proton–deuteron) cyclotron. The stop band of this resonance is given by [25]

$$4|\nu_r - 1| < \left|2C_2 + r\frac{\mathrm{d}C_2}{\mathrm{d}r}\right| . \tag{22}$$

We note that the same resonance is used in the synchro-cyclotron for beam extraction (regenerative extraction). Figure 31 shows the first few harmonic errors, as measured in the IBA C70 cyclotron (discussed in Section 2.6). The large peaks observed in the cyclotron centre and near the pole radius are artefacts due to small to small radial alignment errors of the mapping system. Such errors produce a large effect in the regions where the radial gradients are high. Often, in general, harmonic errors smaller than 5–10 G are considered acceptable in such industrial H^- cyclotrons.

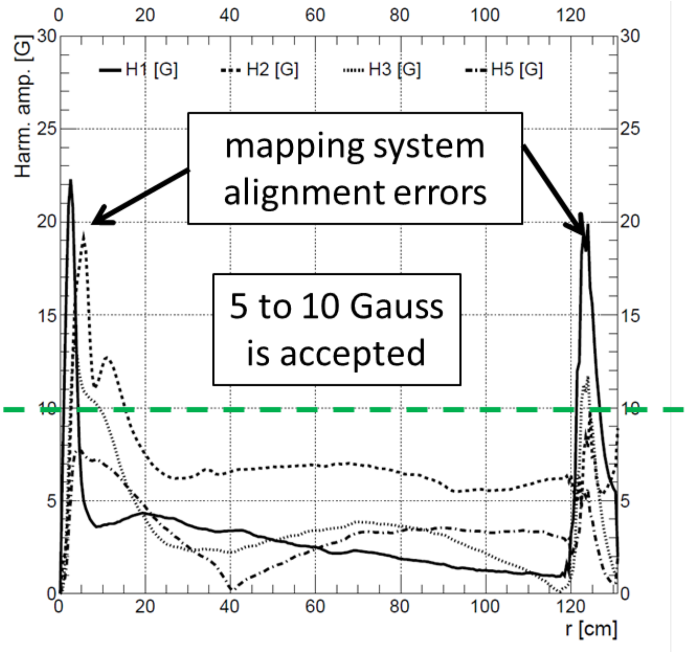


Fig. 31: Amplitudes of the (most important) harmonic field errors in the measured field map of the IBA C70 cyclotron.

2.10 The revival of the synchro-cyclotron

In an isochronous cyclotron, vertical focusing is generated by the flutter of the magnetic field. This flutter can be created by the iron of the magnet (not by the solenoidal coils). The maximum achievable field modulation with the iron is about 2 T. If the average magnetic field is pushed too far up (using superconducting coils), the flutter will steadily decrease and, at a certain point, insufficient vertical focusing will be achieved. In a synchro-cyclotron, this problem does not occur. Thus, in a synchro-cyclotron one can fully exploit the potential offered by superconductivity.

In 2007, the company Still River Systems (today, Mevion Medical Systems, Inc.) began manufacturing a superconducting synchro-cyclotron for proton therapy based on the patent of Dr T. Antaya from the Massachusetts Institute of Technology. This accelerator (left panel of Fig. 32) has a central magnetic field of 9 T; this high field is obtained with a $\mathrm{Nb_3Sn}$ superconducting coil, cooled by cryo-coolers. The unique feature of this extremely compact cyclotron is that it is mounted on a gantry rotating around the patient. The proton beam is extracted at a fixed energy of 250 MeV. As with other superconducting magnets, the large magnetic forces acting on the superconducting coil impose the presence of a special former around the coil. The total consumed power is about 120 kW.

In 2008, IBA began development of the compact superconducting synchro-cyclotron S2C2 [27] (see right panel of Fig. 32). With a superconducting NbTi coil, the magnetic field in the centre of the cyclotron is 5.7 T and the size of the cyclotron is reduced to a diameter of 2.5 m. The total weight of the S2C2 is about 45 tonnes and the beam energy is constant at 230 MeV. The beam from the S2C2 is pulsed at 1 kHz and pulses are about 10 ms long. This results from the synchro-cyclotron concept by which the RF frequency is reduced synchronously with the accelerated proton beam. The first S2C2 has been installed at the Centre de Protonthérapie Antoine Lacassagne in Nice, France and is commissioned at the time of writing (end of 2015).

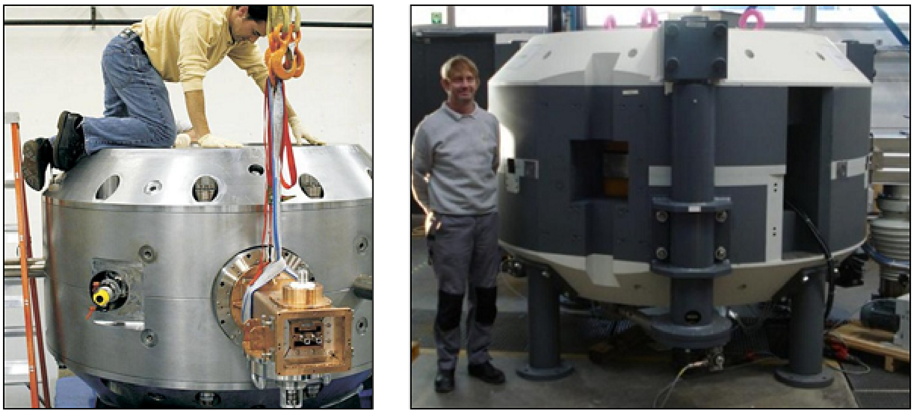


Fig. 32: Left: Monarch S250 synchro-cyclotron from Mevion (Couresy Mevion medical systems Ref. [26]). Right: superconducting synchro-cyclotron (S2C2) from IBA.

3 Injection into a cyclotron

In this section, the main design goals for beam injection are explained and special problems related to a central region with internal ion source are considered. The principle of a PIG source is addressed. The issue of vertical focusing in the cyclotron centre is briefly discussed. Several examples of numerical simulations are given. Axial injection is also briefly outlined.

The topic of cyclotron injection has already been covered in earlier CAS proceedings of the general accelerator physics course [28], as well as in CAS proceedings of specialized courses [29, 30]. An overview of issues related to beam transport from the ion source into the cyclotron central region has been given by Belmont at the 23rd ECPM [31]. Since then, not so many substantial changes have occurred in the field, especially if one only considers small cyclotrons that are used for applications. For this reason, an approach is chosen where the accent is less on completeness and rigorousness (because this has already been done) but more on explaining and illustrating the main principles that are used in medical cyclotrons. Sometimes a more industrial viewpoint is taken. The use of complicated formulae is avoided as much as possible.

Two fundamentally different injection approaches can be distinguished, depending on the position of the ion source. An internal ion source is placed in the centre of the cyclotron, where it constitutes an integrated part of the RF accelerating structure. This may be a trivial case, but it is the one that is most often implemented for compact industrial cyclotrons, as well as for proton therapy cyclotrons. The alternative is the use of an external ion source, where some kind of injection line with magnets for beam guiding and focusing is needed, together with some kind of inflector to kick the beam onto the equilibrium orbit. This method is used for higher-intensity isotope production cyclotrons and also in the proposed IBA C400 cyclotron for carbon therapy.

3.1 Design goals

Injection is the process of particle beam transfer from the ion source, where the particles are created, into the centre of the cyclotron, where the acceleration can start. When designing an injection system for a cyclotron, the following main design goals must be identified.

1. Horizontal centring of the beam with respect to the cyclotron centre. This is equivalent to placing the beam on the correct equilibrium orbit given by the injection energy.
2. Matching (if possible) of the beam phase space with respect to the cyclotron eigenellipse (acceptance).

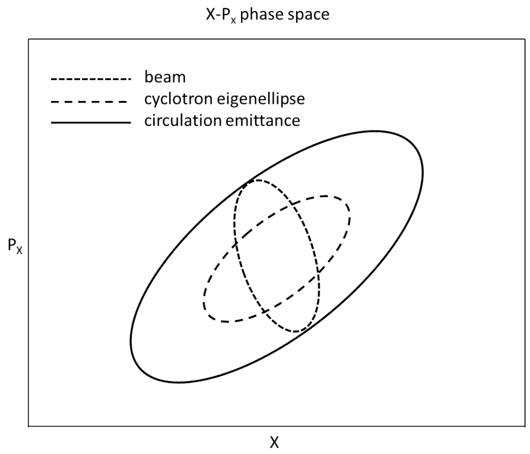


Fig. 33: Mismatch between the cyclotron eigenellipse and the beam ellipse injected into the machine leads, after many turns, to an increase of the circulating emittance.

3. Vertical centring of the beam with respect to the median plane.
4. Longitudinal matching (bunching), i.e., compressing the DC beam from the ion source into shorter packages at the frequency of the RF.
5. Minimization of beam losses and preservation (as much as possible) of the beam quality between the ion source and the cyclotron centre.

The requirement of centring of the beam with respect to the cyclotron centre is equivalent to requiring that the beam is well positioned on the equilibrium orbit corresponding with the energy of the injected particles. The underlying physical reasons for the first three requirements are the same. A beam that is not well centred or is badly matched will execute coherent oscillations during acceleration. In the case of off-centring, these are beam centre-of-mass oscillations. In the case of phase-space mismatch, these are beam envelope oscillations. After many turns , these coherent oscillations smear out and directly lead to an increase in the circulating beam emittances (see Fig. 33). Consequently, beam sizes will be larger, the beam is more sensitive to harmful resonance, the extraction will be more difficult, and the beam quality of the extracted beam will be lower.

The last two requirements directly relate to the efficiency of injection into the cyclotron. Longitudinal matching requires a buncher, which compresses the longitudinal DC beam coming from the ion source into RF buckets. A buncher usually contains an electrode or small cavity that oscillates at the same RF frequency as the cyclotron dees. It works like a longitudinal lens that introduces a velocity modulation of the beam. After a sufficient drift, this velocity modulation transfers into longitudinal density modulation. The goal is to obtain an RF bucket phase width smaller than the longitudinal cyclotron acceptance. For a compact cyclotron without flat-top dees, the longitudinal acceptance is usually around 10–15%. With a simple buncher, a gain of a factor of two to three can easily be obtained. However, at increasing beam intensities, the gain starts to drop, owing to longitudinal space charge forces that counteract the longitudinal density modulation. The issue of beam loss minimization also occurs, for example, in the design of the electrodes of an axial inflector. Here, it must be ensured that the beam centroid is well centred with respect to the electrodes. This is not a trivial task, owing to the complicated 3D orbit shape in an inflector. An iterative process of 3D electric field simulation and orbit tracking is required.

3.2 Constraints

It should be kept in mind that the design of the injection system is often constrained by requirements at a higher level of full cyclotron design. Such constraints can be determined, for example by:

1. the magnetic structure:
   - (a) magnetic field value and shape in the centre;
   - (b) space available for the central region, inflector, ion source, etc.
2. the accelerating structure:
   - (a) the number of accelerating dees;
   - (b) the dee voltage;
   - (c) the RF harmonic mode.
3. the injected particle:
   - (a) charge-to-mass ratio of the particle;
   - (b) number of internal ion sources to be placed (one or two);
   - (c) injection energy.

3.3 Cyclotrons with an internal ion source

The use of an internal ion source is the simplest and certainly also the least expensive solution for injection into a cyclotron. Internal ion sources are used in proton therapy cyclotrons as well as in isotope production cyclotrons. The internal ion source is also used in high-field (6–9 T) superconducting synchro-cyclotrons for proton therapy (see, for example, Ref. [27]). Besides the elimination of the injection line, a main advantage lies in the compactness of the design. This opens up the possibility of placing two ion sources in the machine simultaneously. In many small PET cyclotrons, an H^- and a D^- source are included, to be able to accelerate and extract both protons and deuterons. These two particles are sufficient to produce four well-known and frequently used PET isotopes ^{11}C, ^{13}N, ^{15}O, and ^{18}F. However, an internal ion source brings several limitations: (i) often only low to moderate beam intensities can be obtained; (ii) only simple ion species such as, for example, H^+, H^-, D^-, ^{3}He, or ^{4}He can be obtained; (iii) injected beam manipulation, such as matching or bunching is not possible; (iv) there is a direct gas leak into the cyclotron, which might be especially limiting for the acceleration of negative ions because of vacuum stripping; (v) high beam quality is more difficult to obtain; and (vi) source maintenance requires venting and opening the cyclotron.

3.3.1 The Penning ionization gauge ion source

A cold-cathode PIG [32] ion source is often used as an internal source. The PIG source contains two cathodes that are placed at each end of a cylindrical anode (Fig. 34). The cathodes are at negative potential relative to the anode (the chimney). They emit electrons that are needed to ionize the hydrogen gas and create the plasma. The cyclotron magnetic field must be along the axis of the anode. This field is essential for the functioning of the source, as it enhances confinement of the electrons in the plasma and therefore the level of ionization of the gas. The electrons oscillate up and down as they are reflected between the two cathodes and spiralize around the vertical magnetic field. To initiate the arc current, the cathode voltage must initially be raised to a few kilovolts. Once a plasma is formed, the cathodes are self-heated by ionic bombardment and the arc voltage will decrease with increasing arc current. Usually, an operating voltage of a few hundred volts is obtained. This is sufficient to ionize the gas atoms. The ions to be accelerated are extracted from the source via a small aperture called the slit. This extraction is obtained by the electric field that exists in between the chimney and the so-called puller. This puller is

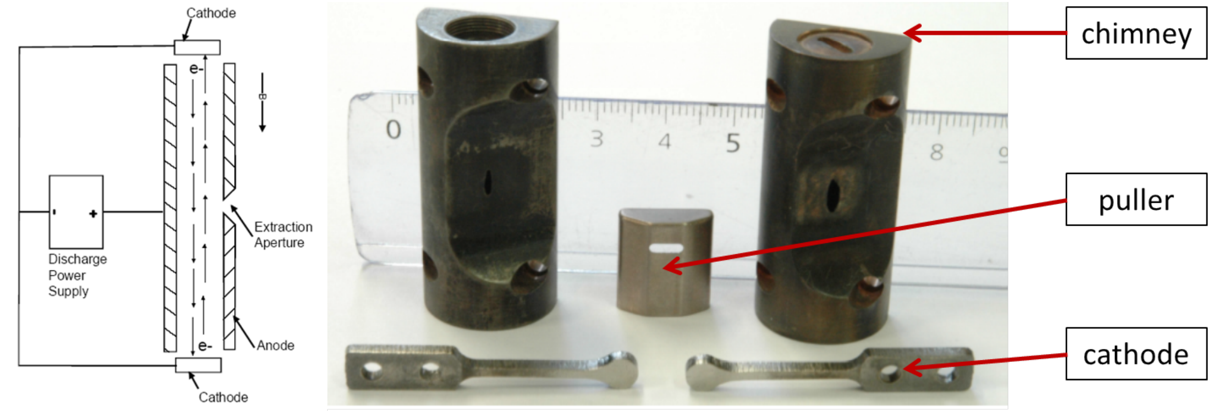


Fig. 34: Left: cold-cathode Penning ion gauge source. Right: two chimneys, two cathodes, and a puller. The chimney on the right shows an eroded slit.

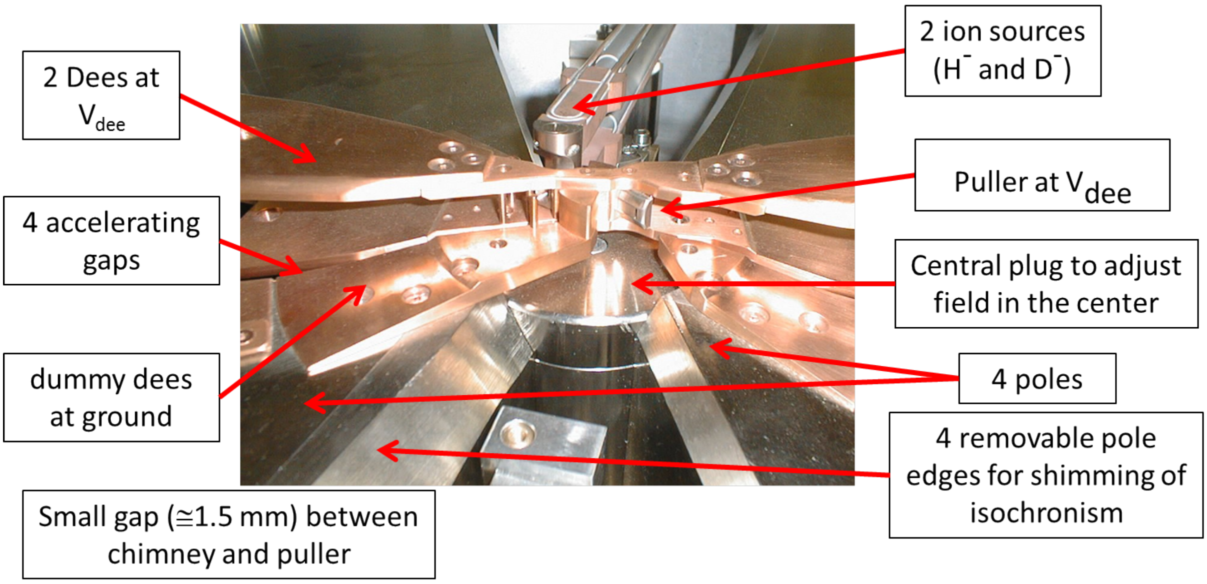

2 Dees at V_{dee}
4 accelerating gaps
dummy dees at ground
Small gap (≅1.5 mm) between chimney and puller
2 ion sources (H^- and D^-)
Puller at V_{dee}
Central plug to adjust field in the center
4 poles
4 removable pole edges for shimming of isochronism

Fig. 35: Central region of the IBA C18/9 cyclotron showing the dees and dummy dees. One of the two ion sources has been removed to show the puller. The figure also shows the four hill sectors. The removable circular disc underneath the central region is the central plug; it is used to fine tune the magnetic field bump in the centre.

at the same electric potential as the RF accelerating structure, as it is mechanically connected to the dees (see Fig. 35).

The right panel of Fig. 34 shows chimneys and cathodes used in compact IBA cyclotrons. The chimneys are made of a copper-tungsten alloy, which has good thermal properties and good machining properties. The cathodes are fabricated from tantalum, because of its good thermal properties and its low work function for electron emission.

3.3.2 Some guidelines for central region design

Figure 35 shows a typical design of a central region for a compact PET-cyclotron with two internal ion sources; one for H^- and one for D^-.

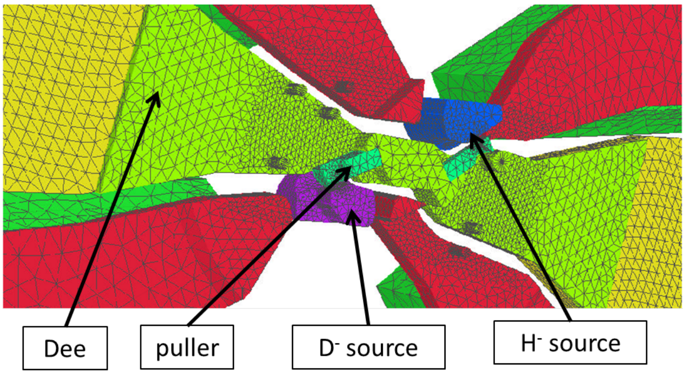


Fig. 36: Example of a 3D finite-element model of the IBA C18/9 central region. Fine meshing is used in regions with small geometrical details, such as the source–puller gap. The graded mesh size allows modelling of the full dee-structure. Complete parameterization of the model is used for fast modification and optimization.

The design of such a central region with an internal ion source is a tedious task that requires precise numerical calculation and often many iterations before good beam centring, vertical focusing, and longitudinal matching are obtained. Some general guidelines for such a design process can be given.

1. Start with a crude model and refine it step by step. Begin with a uniform magnetic field and assume a hard-edge uniform electric field in the gaps. Initially, only consider orbits in the median plane. Try to find the approximate position of the ion source and the accelerating gaps that will centre the beam with respect to the cyclotron centre and centre the beam RF phase with respect to the accelerating wave (longitudinal matching). This may be done by drawing circular orbit arcs by hand, using analytical formulae [29], or even using a pair of compasses (for drawing circular arcs between gaps) and a protractor (for estimating RF phase advance between gaps). Transit time effects should be taken into account, especially in the first gap [33]. The starting phase for particles leaving the source should be roughly between -40° and -10°, assuming that the dee voltage is described by a cosine function.
2. Once an initial gap layout has been found, the model should be further refined by using an orbit integration program. Here, the electric field map can still be generated artificially by assuming, for each gap, an electric field shape with a Gaussian profile that only depends on the coordinate that is normal to the gap and not on the coordinate that is parallel to the gap. Empirical relations may be used to find the width of the Gauss function in terms of the gap width and the vertical dee gap [34]. For the first gap between the source and the puller, a half-Gauss function should be used. The advantage of this intermediate step is that the layout can be easily generated and modified. At this stage, the vertical motion can be taken into account.
3. Create a full 3D model of the central region and solve the Laplace equation to calculate the 3D electric field distribution. An electrostatic map can be used, as long as the wavelength of the RF is much larger than the size of the central region. Several 3D codes exist, such as RELAX3D from TRIUMF [35], or the commercial code TOSCA, from Vector Fields. The latter is a finite-element code that allows modelling of very fine details as part of a larger geometry without the need of very fine meshing everywhere (see Fig. 36). This enables modelling of the full accelerating structure and orbit tracking from the source to the extraction. If possible, the model should be fully parametric, to allow for fast modifications and optimizations. With the availability of 3D computer codes, it is no longer necessary to measure electric field distribution as has been achieved in the past by electrolytic tank measurements [36, 37] or a magnetic analogue model [34].
4. Track orbits in the calculated electric field and in the realistic magnetic field (obtained from field mapping or from 3D calculations). Fine tune the geometry further for better centring, vertical

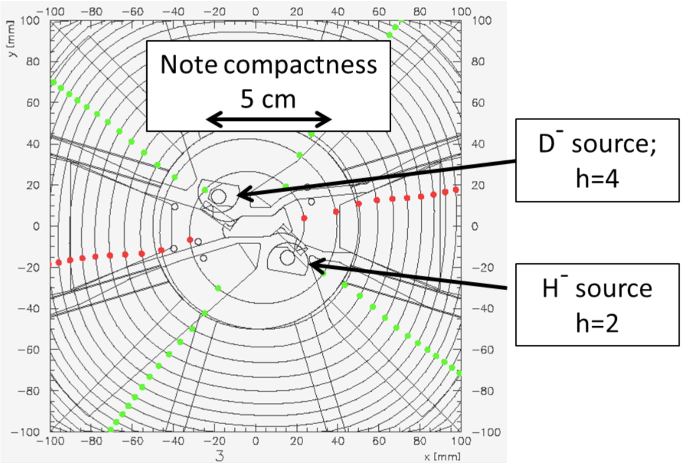


Fig. 37: Calculated orbits in the IBA C18/9 central region. The D^- source is shifted farther outward because its orbit is larger. Note that parts of the D^- chimney have been cut away, to give sufficient clearance for the H^- beam. Assessment of longitudinal matching is illustrated: the red dots and the green dots give the particle position when the dee voltage is zero and a maximum, respectively. Ideally, the red dots should be on the dee centre line. Electric fields are obtained from Opera. Magnetic fields from Opera or from a measured map.

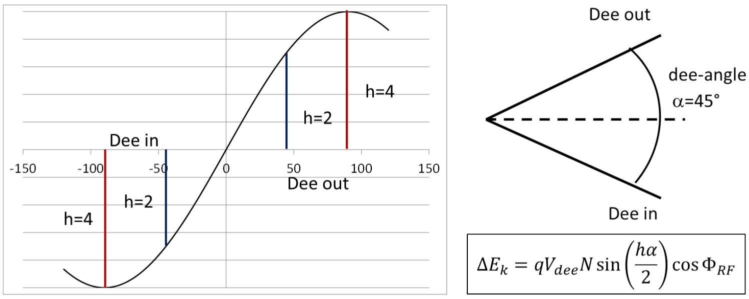


Fig. 38: Energy gain per turn depends on several parameters, such as the dee voltage V_{dee}, the number of dees, N, the harmonic mode, h, the dee angle, α, and the RF phase Φ_{RF}. For a dee angle of 45°, the harmonic mode $h = 4$ is most efficient.

focusing, and longitudinal matching (RF phase centring). An example of such a calculation is given in Fig. 37.

5. Track a full beam (many particles) to find beam losses and maximize the beam transmission in the central region (see, for example, Ref. [38]).

The energy gain per turn in the cyclotron is given by:

$$\Delta E_k = qV_{\mathrm{dee}}N\sin\left(\frac{h\alpha}{2}\right)\cos\Phi_{\mathrm{RF}} \ . \tag{23}$$

Here, q is the charge of the particle, V_{dee} is the dee voltage, N is the number of accelerating gaps, h is the harmonic mode, α is the dee opening angle and Φ_{RF} is the RF phase of the particle. Maximum acceleration is obtained if the RF phase advance between the dee entrance and the dee exit is just 180° ($h\alpha = \pi$). This is illustrated in Fig. 38. For example, if the dee angle is 45° and $h = 4$, the energy gain is 100%, but for $h = 2$, it is only 71%.

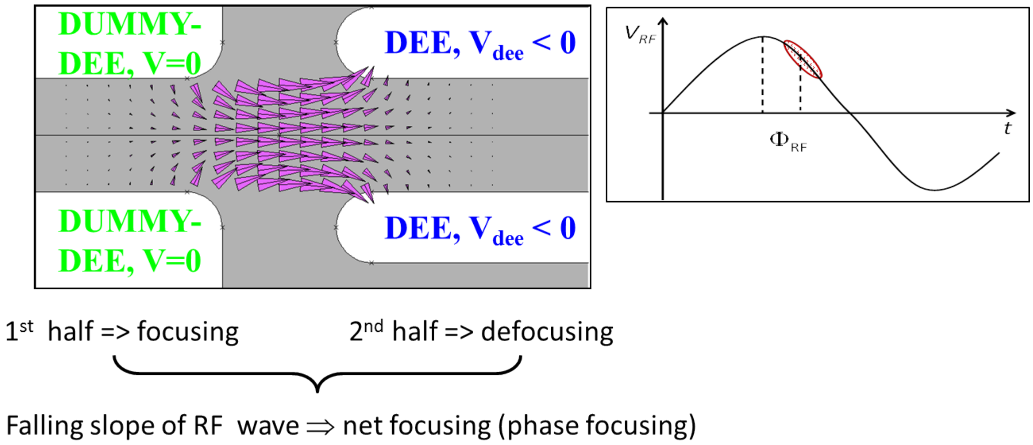


Fig. 39: Vertical electrical focusing in the cyclotron accelerating gap. This focusing is important only in the cyclotron centre.

3.3.3 Vertical focusing in the cyclotron centre

The azimuthal field variation goes to zero in the centre of the cyclotron; therefore, this resource for vertical focusing is lacking. There are two remedies to restore the vertical focusing.

1. Add a small magnetic field bump (a few hundred gauss) in the centre. The negative radial gradient of this bump provides some vertical focusing. The bump must not be too large, to avoid too large an RF phase slip. In small IBA PET cyclotrons, the bump is fine tuned by modifying the thickness of the central plug (see Fig. 35).
2. Fully exploit the electrical focusing provided by the first few accelerating gaps.

If an accelerating gap is well positioned with respect to the RF phase, it may provide some electrical focusing. Figure 39 illustrates the shape of the electric field lines in the accelerating gap between a dummy dee (at ground potential) and the dee. The particle is moving from left to right and is accelerated in the gap. In the first half of the gap, the vertical forces point towards the median plane and this part of the gap is vertically focusing. In the second half of the gap, the vertical forces have changed sign and this part of the gap is vertically defocusing. If the dee voltage were DC, there would already be a net focusing effect of the gap for two reasons: (a) a focusing and de-focusing lens, one behind the other, provide some net focusing in both planes and (b) the defocusing lens is weaker because the particle has a higher velocity in the second part of the gap. This is comparable to the focusing obtained in an Einzel lens. However, the dee voltage is not DC but varying in time and this may provide an additional focusing term that is more important than the previous two effects (phase focusing). This is obtained by letting the particle cross the gap at the moment that the dee voltage is falling (instead of accelerating at the top). In this case, the defocusing effect of the second gap half is additionally decreased. To achieve this, the first few accelerating gaps must be properly positioned azimuthally. This forms part of the central region design.

Vertical electrical focusing is also illustrated in Fig. 40. This figure shows the (normalized) vertical electrical force acting on a particle during the first five turns in the cyclotron (IBA C18/9). A minus sign corresponds with focusing (force directed towards the median plane). The dee crossings (two dees) as well as the gap crossings (four gaps) are indicated. It can be seen that each gap is focusing at the entrance and defocusing at the exit. It can also be seen from the amplitude of the force that the electrical focusing rapidly falls with increasing beam energy. However, after a few turns, the magnetic focusing becomes sufficient.

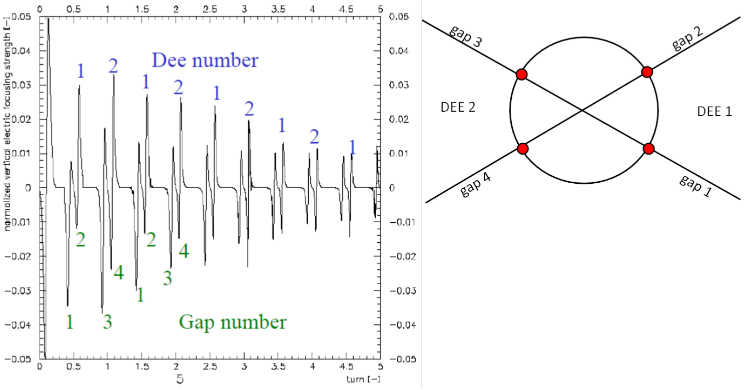


Fig. 40: Normalized vertical electrical forces obtained from orbit tracking of a particle during five turns in a central region with two dees (four accelerating gaps). Dee crossings are indicated at the top of the curve (in blue) and gap crossings (two per dee) at the bottom of the curve (in green). Each gap is focusing at the entrance and defocusing at the exit.

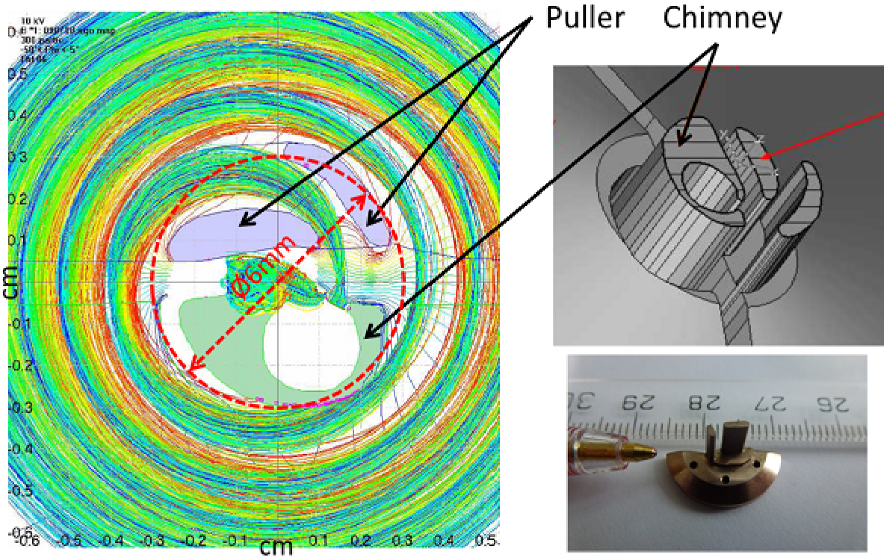


Fig. 41: Left: calculated orbits in the central region of the IBA S2C2. Right: 3D view of the Penning ion gauge source and puller used in this central region.

3.3.4 The central region of a superconducting synchro-cyclotron

The internal cold-cathode PIG ion source is also used in superconducting synchro-cyclotrons for proton therapy. In such a cyclotron, the magnetic field in the centre is very high (5–9 T) and the energy gain per turn is low, with a dee voltage of about 10 kV. In such a case, the central region needs to be very compact. This is illustrated in Fig. 41, which shows the central region of the IBA S2C2. The source diameter is <5 mm and the diameter of the first turn in the cyclotron is ≈6 mm. The vertical dee gap in the centre is only 6 mm. The first 100 turns are within a radius of only 30 mm.

The precise position of the ion source in a synchro-cyclotron is of the utmost importance, in order to obtain very good beam centring and the highest beam quality at the extraction. The central region and the ion source of the S2C2 can be removed as one subsystem for easy maintenance and precise alignment

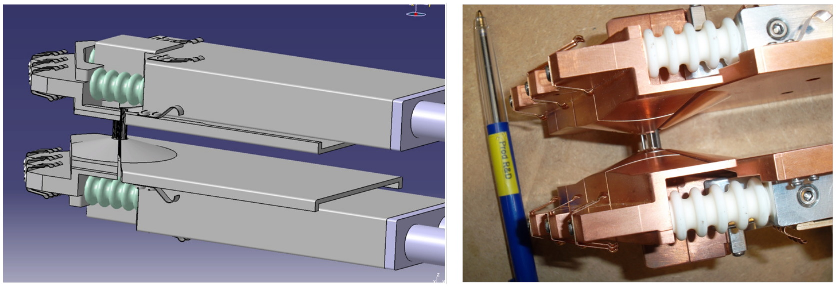


Fig. 42: The ion source and central region can be extracted from the cyclotron as one assembly, for easy maintenance and precise repositioning.

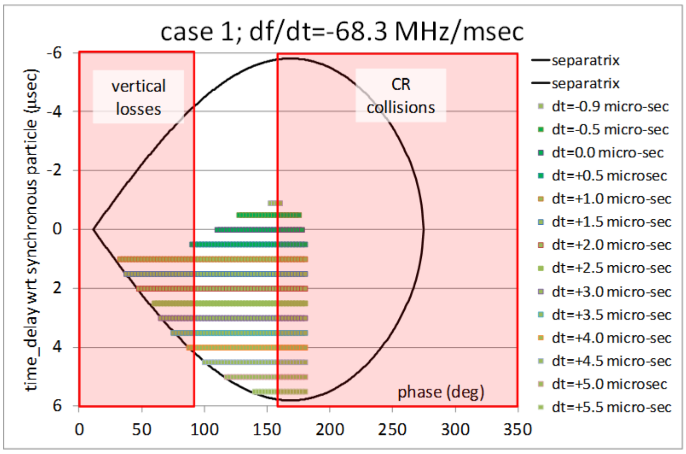


Fig. 43: Simulation of the beam capture process in the S2C2 central region (CR).

and realignment after reassembly (see Fig. 42). To suppress the multipactor, both the dee and the counter dee are biased at a DC voltage of 1 kV.

As mentioned in Section 1.2, in a synchro-cyclotron, there is only a short time in which the beam can be captured into phase-stable orbits [39]. Simulation of the beam capture requires a combined study of the orbit dynamics in the cyclotron central region and the subsequent acceleration. Here, particles are started at the ion source at different time points and at different RF phases. Only a subset of the started particles are captured. Particles outside of the acceptance window fall out of synchronism with respect to the RF and are decelerated back towards the ion source, where they are lost. Superimposed on that, there are the usual additional transverse (radial or vertical) losses due to the collisions with the geometry of the central region. This is illustrated in Fig. 43.

3.3.5 Burning paper

When a new central region has been designed and is being tested in the machine, it does not always immediately function correctly and it may happen that the beam is lost after a few turns. Owing to space limitations, it is not always possible to have a beam diagnostic probe that can reach the centre of the cyclotron and it may be difficult to find out why and where the beam is lost. In such cases, it may help to put small bits of thin paper in the median plane; they will change colour, owing to the interaction with the beam. This is illustrated in Fig. 44 , which shows the central region layout of the IBA self-extracting cyclotron. Seven bits of burned paper have been fixed on the central region design drawing. In this way, the position of each turn and the corresponding beam sizes are nicely indicated.

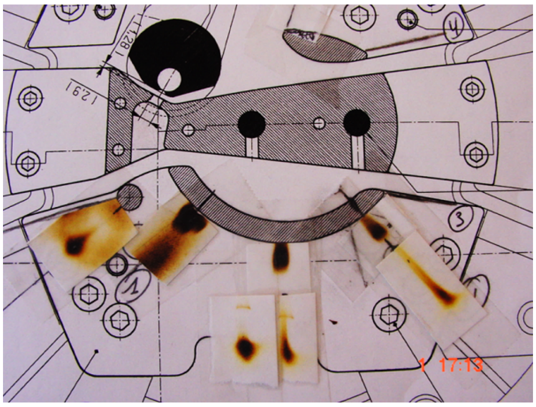


Fig. 44: Small bits of thin paper are placed in the pole gap and burned by the beam, to find the beam position and size during the first few turns.

3.4 Cyclotrons with an external ion source

In many cases, the ion source is placed outside the cyclotron. There may be different reasons for this choice: (i) higher beam intensities are needed, which can only be produced in a more complex and larger ion source than the simple PIG source, (ii) special ion species, such as heavy ions or highly stripped ions, are required, or (iii) a good machine vacuum is needed (for example, H^- acceleration). External ion sources are used in high-intensity isotope production cyclotrons but, for example, also in the proposed IBA C400 cyclotron for carbon therapy. Of course, the external ion source is a more complex and more expensive solution, since it requires an injection line with all related equipment such as magnetic or electrostatic beam guiding and focusing elements, vacuum equipment, beam diagnostics, etc.

3.4.1 Different methods of injection

There are a few different ways to inject into a cyclotron.

1. Axial injection: this case is most relevant for small cyclotrons. The beam travels along the vertical symmetry axis of the cyclotron towards the cyclotron centre. In the centre, the beam is bent through 90° degrees from vertical to horizontal into the median plane. This is achieved using an electrostatic or magnetostatic inflector.
2. Horizontal injection: the beam is travelling in the median plane from the outside towards the cyclotron centre. Generally speaking, this type of injection is more complicated than axial injection, owing to the vertical magnetic field exerting a horizontal force on the beam, thereby trying to bend it in the horizontal plane. It has been attempted to cancel this force with electrical forces from an electrode system installed near the median plane [40]. It has also been attempted to tolerate this force and to let the beam make a spiral motion along the hill–valley pole edge towards the cyclotron centre. This is called trochoidal injection and is illustrated in Fig. 45. In the centre, an electrostatic deflector places the particle on the correct equilibrium orbit. Both methods are very difficult and therefore are no longer used.
3. Injection into a separate sector cyclotron: this must be qualified as a special case. Much more space is available in the centre to accommodate magnetic bending and focusing devices. Injection at much higher energies (in the megaelectronvolt range) is possible. This topic is considered as out of the scope of (small) medical accelerators.
4. Injection by stripping: a stripper foil positioned in the centre changes the particle charge state and its local radius of curvature so that the particle aligns itself on the correct equilibrium orbit. This method is mostly applied for separate sector cyclotrons, where the beam is injected horizontally.

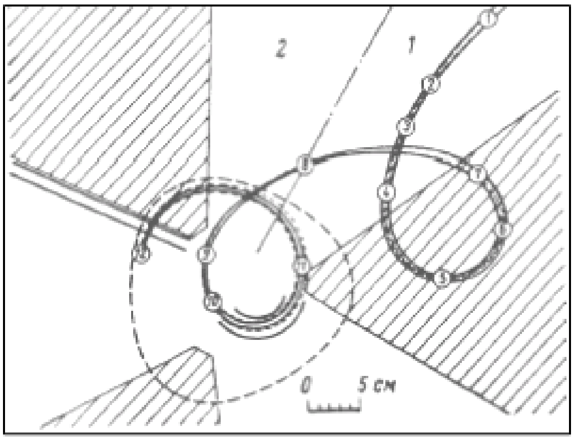


Fig. 45: Horizontal (trochoidal) injection. The beam travels along the hill–valley pole edge. In the centre, an electrostatic device places the beam on the equilibrium orbit. Figure taken from Ref. [28]

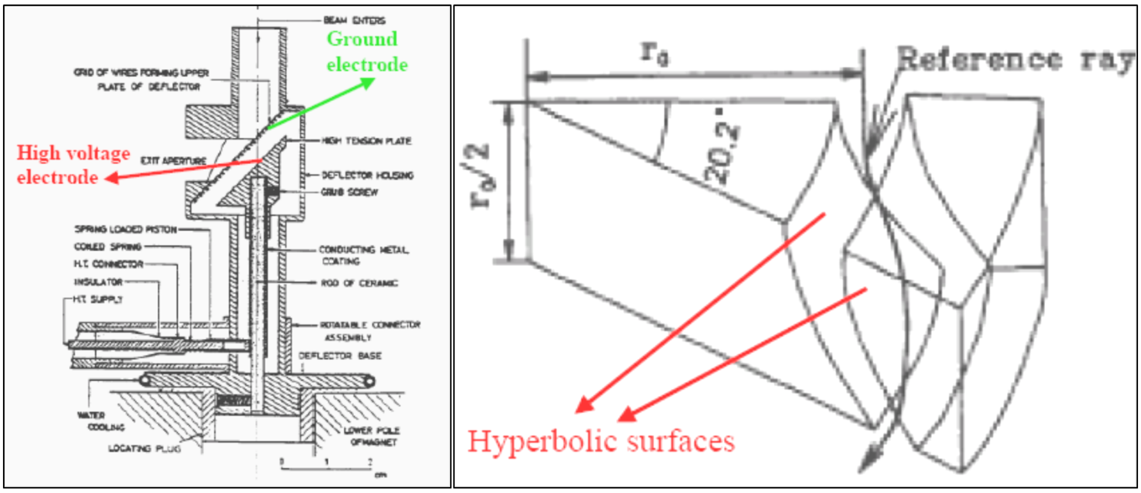


Fig. 46: Left: mirror inflector. Right: hyperboloid inflector. Figure taken from Ref. [28]

3.4.2 *Inflectors for axial injection*

The electrical field between two electrodes bends the beam 90° from vertical to horizontal. The presence of the cyclotron magnetic field creates a complicated 3D orbit; this makes the inflector design difficult. Four different types of electrostatic inflectors are known.

1. The mirror inflector: two planar electrodes are placed at 45° with respect to the vertical beam direction. In the upper electrode is an opening for beam entrance and exit (see Fig. 46). The advantage of the mirror inflector is its relative simplicity. However, because the orbit is not following an equipotential surface, a high electrode voltage (comparable to the injection voltage) is needed. At the entrance, the particle is decelerated and at the exit it is re-accelerated. Furthermore, to obtain a reasonable electrical field distribution between the electrodes, a wire grid is needed across the entrance and exit opening in the upper electrode. Such a grid is very vulnerable and is easily damaged by the beam.
2. The spiral inflector: this is a cylindrical capacitor that is gradually twisted to take into account the spiralling of the trajectory, induced by the vertical cyclotron magnetic field. The design is such

that the electrical field is always perpendicular to the velocity vector of the central particle and the orbit is therefore positioned on an equipotential surface. The electrode voltage can be much lower for a mirror inflector. A simple formula for the electrode voltage is:

$$V = 2 \cdot \frac{E}{q} \cdot \frac{d}{A} , \tag{24}$$

where V is the electrode voltage, E is the injection energy, q is the particle charge, d is the electrode spacing, and A is the electric radius of the inflector (which is almost equal to the inflector height). It can be seen that the ratio between electrode voltage and injection voltage is equal to twice the ratio of the electrode spacing and the height of the inflector. An important advantage of the spiral inflector is that it has two free design parameters that can be used to place the particle on the correct equilibrium orbit. These two parameters are the electrical radius, A, and the so-called tilt parameter, k'. This second parameter represents a gradual rotation of the electrodes around the particle moving direction by which a horizontal electric field component is obtained that is proportional to the horizontal velocity component of the particle. Varying the tilt parameter , k', is, therefore, equivalent (as far as the central trajectory is concerned) to varying the cyclotron magnetic field in the inflector volume. Another advantage of the spiral inflector is its compactness. However, the electrode surfaces are complicated 3D structures, which are difficult to machine. Fortunately, with the wide availability of computer controlled five-axis milling machines, this is not really a problem anymore. Figure 47 shows a 1:1 model of the spiral inflector used in the IBA C30 cyclotron.

3. The hyperboloid inflector: the electrodes are hyperboloids with rotational symmetry around the vertical z-axis (see Fig. 46). As for the spiral inflector, the electrical field is perpendicular to the particle velocity and a relatively low electrode voltage can be used. However, for this inflector, no free design parameters are available. For given particle charge and mass, injection energy and magnetic field, the electrode geometry is fixed and it is more difficult to inject the particle in the correct equilibrium orbit. Furthermore, this inflector is quite large compared with the spiral inflector. However, owing to the rotational symmetry, it is easier to machine.
4. The parabolic inflector: the electrodes are formed by bending sheet metal plates into a parabolic shape. This inflector has the same advantages and disadvantages as the hyperboloid inflector: relatively low voltage and ease of construction, but no free design parameters and relatively large geometry.

3.4.3 Example: axial injection in the IBA C30HC high-intensity cyclotron

Nowadays, the spiral inflector is almost always used for axial injection. Analytical formulae exist for central orbits in a spiral inflector placed in a homogeneous magnetic field [41–44]. However, the field in the cyclotron centre is certainly not uniform, owing to the axial hole needed for axial injection. In practice, the inflector design requires extensive numerical effort, which can be broken down into three main parts: (1) 3D modelling of the electrical fields of the inflector and central region, (2) 3D modelling of the magnetic field in the central region, and (3) orbit tracking in the central region. The complete process is tedious and requires many iterations. First, the central trajectory has to be defined and optimized. There are three main requirements, namely that the injected orbit is nicely on the equilibrium orbit, correctly placed in the median plane and well centred with respect to the inflector electrodes. After an acceptable electrode geometry has been obtained, for which these requirements are fulfilled, the beam optics must be studied. Here the main requirement is that reasonable matching into the cyclotron eigenellipse can be achieved, so that large emittance growth in the cyclotron is avoided [45].

It may be necessary to calculate several inflectors of different height, A, and tilt parameter, k', to optimize this process. At IBA, both the 3D magnetic field computations as well as the 3D electrical field

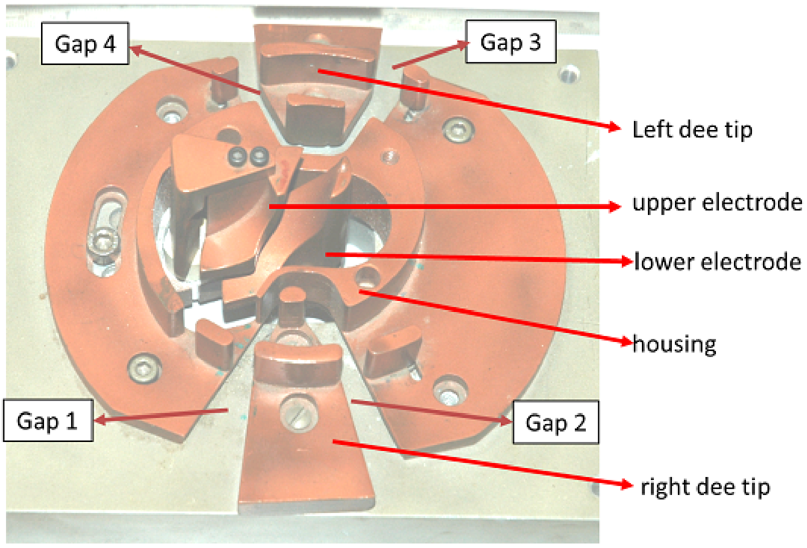


Fig. 47: 1:1 model of the spiral inflector and central region of the IBA Cyclone 30 cyclotron

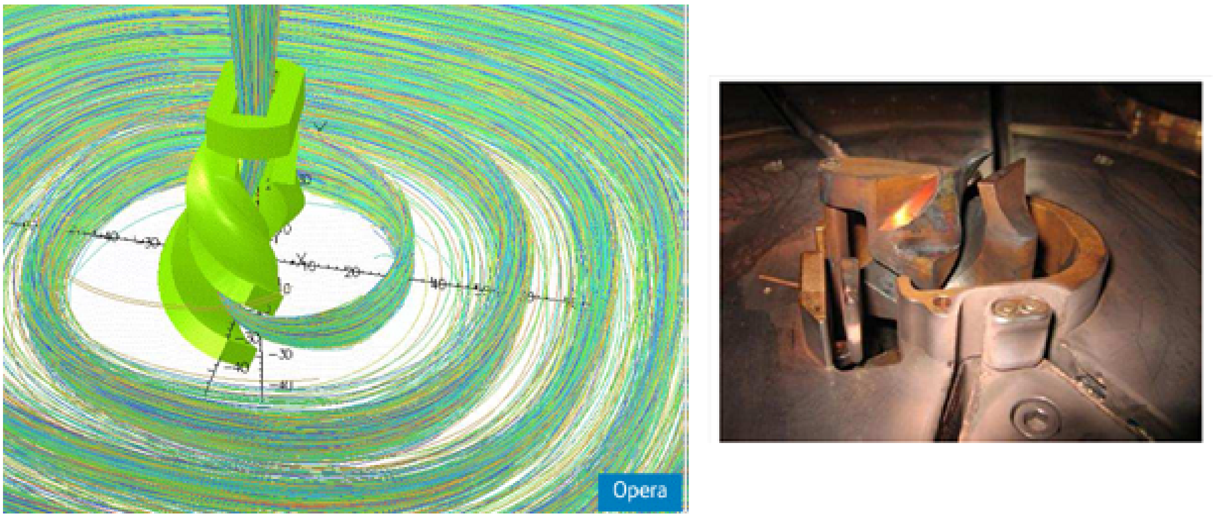


Fig. 48: Left: axially injected 3D orbits calculated with the cyclotron tracking code AOC are imported in the Opera-3d finite-element model of the IBA C30XP cyclotron. Right: central region of the IBA C70XP cyclotron, showing the inflector with an additional electrode at its exit that provides a radial kick, which is needed to centre particles with different q/m ratios on their respective equilibrium orbits.

computations are done using the Opera-3d software package from Vector Fields [7]. Often, the models are completely parameterized, for quick modification and optimization. Figure 48 shows such a model of the central region and the inflector. The inflector model uses the following parameters: the electrode width and spacing (both may vary along the inflector), the tilt parameter, k', and the shape of the central trajectory itself, in terms of a list of points and velocity vectors.

Recently, the IBA C30 has been upgraded to a new high-current version (C30HC) [46]. For this purpose, a new ion source, a new injection line, and a new central region have been installed. A new final amplifier provides 100 kW of RF power as needed for beam acceleration. The new source, as shown in Fig. 49, is the D-pace DC volume cusp source, providing 15 mA of H$^-$ beam within a 4 rms beam

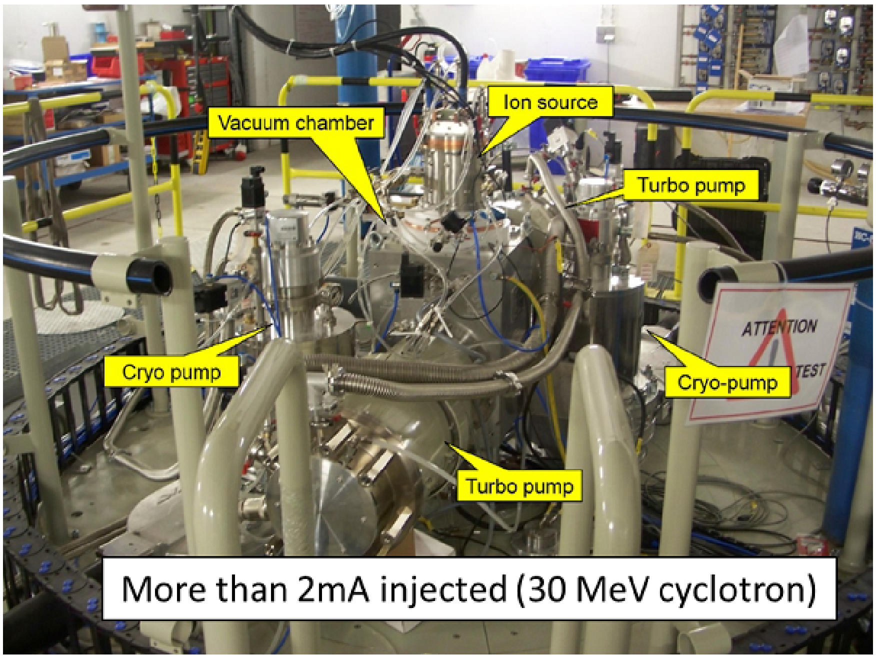


Fig. 49: C30HC injection line installed on top of the cyclotron. The ion source is on top of the vacuum box. This box is pumped by two turbo pumps (front and back).

emittance of 110π mm mrad at an injection energy of 30 keV [47]. This performance approximately doubles the injected beam current as compared with the standard C30 ion source.

The new injection line layout is shown in Fig. 50. The ion source is mounted on top of the vacuum box, which contains an Einzel lens, a buncher, and a Faraday cup. The axial bore of the cyclotron contains a solenoid, two small quadrupoles, and an xy-steering magnet. With this steering magnet and a second pair at the exit of the ion source, the beam at the inflector can be adjusted to the correct position and direction. The design has been optimized to maximize H^- beam injection. The beam line is compact (short) to minimize stripping on the residual gas. Differential pumping is applied with a first turbo pump, which acts directly below the ion source, and a second turbo pump, which pumps the second separated part of the vacuum box and the downstream part of the beam line. Both pumps are magnetically shielded from the cyclotron stray field. The elements in the cyclotron bore are contained at atmospheric pressure around the beam transport tube. This reduces outgassing. The cyclotron bore diameter has been increased to allow sufficient space for these elements. Opera-3d calculations have confirmed that this does not compromise the magnetic field in the median plane. The cyclotron iron is used as return yoke for the solenoid; the quadrupoles have their own return yoke. The beam is focused by the Einzel lens to a small size at the centre of the buncher, as shown in the calculation of the beam envelopes along the beam line (right panel of Fig. 50). In this way, the spread in transit time factor due to the finite beam size is minimized and the bunching efficiency is maximized [48]. Finally, the beam envelope naturally increases at buncher exit, to a maximum value in the solenoid, as permitted by the solenoid bore. This enables the smallest possible beam size to be obtained at the inflector entrance, through the focusing action of the solenoid. The relative positions of the Einzel lens, buncher, and solenoid have been optimized to obtain this condition. The two quadrupoles enable the shape of the beam to be adjusted asymmetrically so that better matching is obtained with respect to the cyclotron acceptance ellipse [45]. For the same reason, the two quadrupoles can be rotated around their axis.

The initial central region was designed almost 30 years ago, when advanced 3D programs for electromagnetic modelling were not yet available. With the better design tools available today, it could

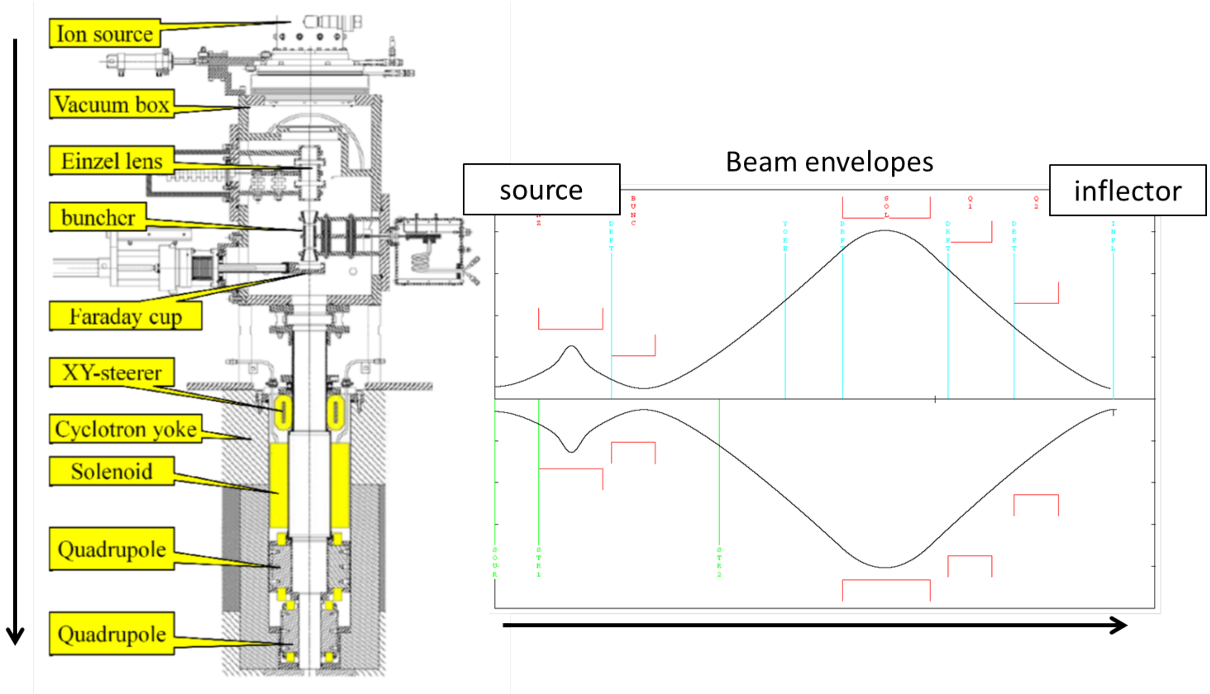


Fig. 50: Left: layout of the C30HC injection line. Right: beam envelopes in this beam line, calculated with TRANSPORT.

be improved and a new inflector and central region was designed. Optimization was conducted for good beam centring in the inflector and in the cyclotron, good vertical focusing, and large beam capture efficiency. The following stepwise approach was used.

1. Construct precise Opera-3d models of the cyclotron magnet and the accelerating structure and obtain the 3D magnetic field in and around the inflector volume.
2. Optimize the azimuthal opening angle of the dees in the centre to maximize the capture efficiency.
3. Determine the accelerated equilibrium orbit by backtracking from high energy towards the centre.
4. Choose an injection point on the accelerated equilibrium orbit, allowing a good position of the inflector relative to the first dee gap.
5. Obtain an initial estimate of the inflector central orbit, passing through the injection point. A special tracking mode is used, simulating the inflector by an electric field that is always perpendicular to the orbit.
6. Construct the real 3D inflector electrodes around the estimated reference orbit using Opera-3d.
7. Verify orbit centring with respect to the inflector electrodes and modify the fringe field parameters if needed.
8. Verify the design by orbit tracking in the real 3D fields and modify the inflector position slight, if necessary, for beam horizontal or vertical centring.

An inflector bending radius of $A = 29.5$ mm was chosen. For this case, a tilt of $k' = -1.0$ was needed to centre the beam. The inflector gap is 8 mm and the aspect ratio equals 2. Figure 48 shows the 3D model of the inflector, developed in Opera-3d. With this new central region and injection line design, a beam current of > 2 mA was obtained at the 1 MeV beam stop in the cyclotron centre.

4 Extraction from cyclotrons

Different solutions for beam extraction are treated. These include extraction by stripping, resonant extraction using a deflector, and the regenerative extraction used in synchro-cyclotrons. The different methods of creating a turn separation are explained. The purpose of different types of extraction device, such as harmonic coils, deflectors, and gradient corrector channels, are outlined. Several illustrations are given, in the form of photographs and drawings.

The topics of cyclotron extraction have already been covered in earlier CAS proceedings in the framework of the general accelerator physics course [28], as well as in the framework of specialized courses [30, 49]. Since then, not so many substantial changes have occurred in the field, especially if one only considers small cyclotrons that are used for applications. For this reason, it was decided to choose an approach where the accent is less on completeness and rigorousness (because this has already been done) but more on explaining and illustrating the main principles that are used in small cyclotrons. Sometimes a more industrial viewpoint is taken. The use of complicated formulae is avoided as much as possible.

Extraction is the process of beam transfer from an internal orbit to a target that is placed outside of the magnetic field. There are three main reasons why extraction is considered as difficult.

1. The magnetic field itself behaves as a kind of trap. When the particles are accelerated into the falling fringe field they will run out of phase with respect to the RF wave; if the phase slip is more than 90°, they will be decelerated instead and move inward. This may be considered as a kind of reflection of the beam on the radial pole edge of the magnet.
2. In a cyclotron, the turn separation is inversely proportional to the radius ($R \approx \sqrt{E}$). Because of this, the turns pile up closely together near the extraction radius. Therefore, it is difficult to deflect the last orbit, without influencing the inner orbits and without important beam losses.
3. During extraction, the beam has to cross the fringe field. This is an area where there are very large gradients and non-linearities in the magnetic field. Special precautions have to be made to avoid substantial beam losses, beam blow-up, or loss of beam quality.

There are a few different ways to solve the problem of extraction.

1. No extraction at all: avoid the problem by using an internal target. This can be done for isotope production, but is a little bit dirty.
2. Extraction by stripping (H^- or H_2^+ cyclotrons): this is often applied in isotope production cyclotrons.
3. Use one (or more) electrostatic deflectors (ESDs) that peel off the last orbit: this is, for example, done in the proton therapy cyclotrons of Varian, IBA, and SHI.
4. Regenerative extraction, as used in synchro-cyclotrons: this is, for example, done in the proton therapy synchro-cyclotrons of Mevion and IBA.
5. Self-extraction: this requires a suitable and precise shaping of the magnetic field. IBA has made one such prototype cyclotron for isotope production.

Cases (3) and (4) require some method to increase the turn separation between the last and penultimate orbits. The different methods will be discussed in some more detail next.

4.1 The use of an internal target

The target is placed between the poles of the cyclotron at a radius where the field is still isochronous. The method was applied quite frequently for the production of radioisotopes, such as $^{103}\mathrm{Pd}$ or $^{201}\mathrm{Tl}$. The method is relatively simple and also non-expensive. The energy can be selected by choosing the correct

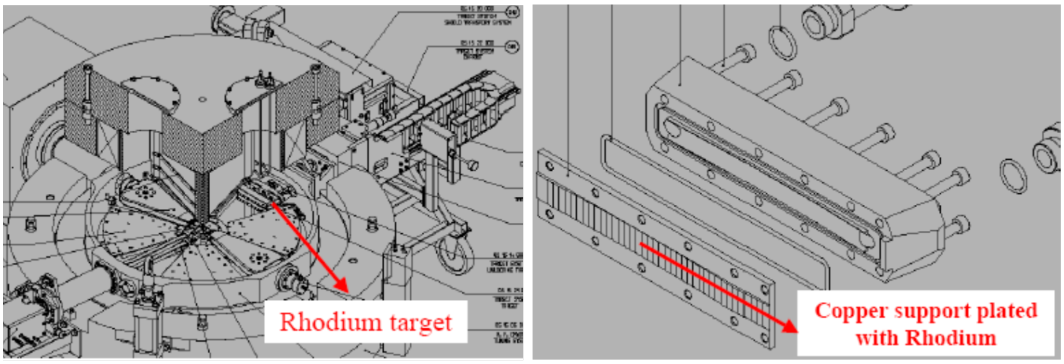


Fig. 51: Left: IBA C18+ cyclotron with an internal rhodium target for the production of palladium. Right: detail of an internal target showing the profiling of the target surface, optimized to maximize the beam spot and minimize beam reflection on the target.

radius of the target in the cyclotron. However, it is a rather dirty way of working because of radioactive contamination of the cyclotron. If the target is perpendicular to the beam, the beam spot is very small and local heating will pose a problem. To avoid this, the target is placed at a small grazing angle with respect to the beam. In this case, however, a certain fraction of the incoming beam will be reflected from the target surface, as a result of multiple scattering. This in turn will activate the cyclotron. The right panel of Fig. 51 shows an IBA C18+ cyclotron with an internal rhodium target, which is used for the production of $^{103}\mathrm{Pd}$. The target can be handled fully remotely. The $^{103}\mathrm{Pd}$ was used for brachytherapy (in the treatment of prostate cancer). Sixteen of these machines have been sold to one customer. The left panel of Fig. 51 shows some detail of the internal Rh/Pd target. The target is heavily water-cooled, so that it can take a beam power of 30 kW (2 mA/15 MeV). The target surface is profiled, which is optimized to maximize the beam spot and minimize the beam reflection.

More strict legal rules concerning the prevention of radioactive contamination will certainly eliminate this method in future.

4.2 Stripping extraction

To extract the beam, the particles pass a thin stripper foil (see the right panel of Fig. 52), by which one or more electrons are removed from the ion. Because of this, there is an instantaneous change of the orbit local radius of curvature. The relation between the local radius before (ρ_i) and after (ρ_f) stripping is given by:

$$\rho_\mathrm{f} = \frac{Z_\mathrm{i}}{Z_\mathrm{f}}\frac{M_\mathrm{f}}{M_\mathrm{i}}\rho_\mathrm{i}\ , \tag{25}$$

where M_i and M_f are the particle mass before and after stripping, respectively. As an example, for H^-, we have $\mathrm{H}^- \Rightarrow \mathrm{H}^+ + 2\mathrm{e}^-$ and the local radius of curvature practically only changes sign ($\rho_\mathrm{f} = -\rho_\mathrm{i}$) because $M_{\mathrm{H}^+} \simeq M_{\mathrm{H}^-}$. As a result, the stripped particle is immediately deflected outward, away from the cyclotron centre. This is illustrated in the left panel of Fig. 52.

Multiple targets can be placed around the machine. This is illustrated in the left panel of Fig. 53. A given target is selected by rotating the corresponding stripper foil into the beam. H^- extraction is applied in many commercial isotope production cyclotrons, fabricated, for example, by IBA (Cyclone 30, C18/9, C10/5), Advanced Cyclotron Systems (TR30, TR13), or General Electric (PETtrace). The right panel of Fig. 53 shows a view on the median plane of a typical IBA isotope production cyclotron. In this cyclotron, eight different target ports are available.

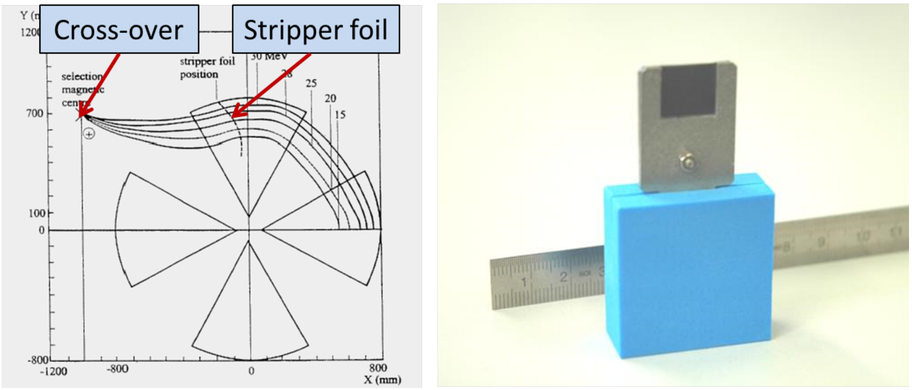


Fig. 52: Left: H$^-$ extraction by stripping. Energy is selected by moving the stripper foil to the correct radius. Right: a simple carbon stripper foil is used to extract the beam (typically 50–200 μg/cm^2).

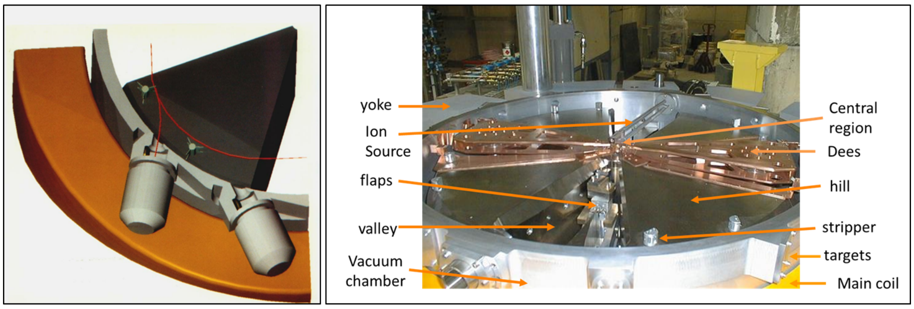


Fig. 53: Left: H$^-$ extraction by stripping. Several targets can be placed around the machine. Right: median plane of a typical IBA isotope production cyclotron, showing the most important subsystems.

The most important features and advantages of stripping extraction are:

1. very simple extraction device;
2. 100% extraction efficiency;
3. variable energy;
4. the ability to place several targets around the machine;
5. simultaneous dual beam extraction;
6. good beam optics.

The energy can easily be varied by moving the radial position of the stripper probe (see Fig. 52). By proper azimuthal positioning of the stripper foil, all orbits come together in the same cross-over point outside the magnetic field. Here, a combination magnet can be placed that deflects the beam into the beam line. Simultaneous dual beam operation is made possible by positioning two stripper foils at an azimuth of 180° with respect to each other. Some fine tuning of the turn-pattern is necessary to distribute the total beam current precisely between the two stripping foils. This may be achieved by fine adjustment of the dee voltage, or by using first-harmonic coils. Since the extracted beam crosses the radial pole edge at an angle that is close to 90°, the large (de-)focusing effects of the fringe field are avoided and the beam quality remains intact. H$^-$ stripping is an ideal solution for low- and medium-energy industrial cyclotrons.

There is a serious limitation of an H$^-$ cyclotron, owing to magnetic stripping that may occur during acceleration. Because of this, the magnetic field cannot be high and to obtain high energy the pole radius of the machine must be increased. A well-known example is the TRIUMF cyclotron [50], which accelerates H^- up to 520 MeV. The average magnetic field is only 0.3 T (in the cyclotron centre), resulting in a magnet diameter of 18 m and a magnet weight of 4000 tonnes. The H^- is also stripped on the vacuum rest gas and, to limit beam losses, good vacuum pumping (expensive) and an external ion source is required (IBA Cyclone-30, ACS TR30). The H^--cyclotron is good for isotope production, but not for proton therapy.

Another example is the acceleration of molecular hydrogen H_2^+ and related extraction by stripping: $\mathrm{H}_2^+ \Rightarrow 2\mathrm{H}^+ + \mathrm{e}^-$. In this case, the radius of curvature does not change sign but is practically divided by two ($\rho_\mathrm{f} = \rho_\mathrm{i}/2$). In this case, extraction is more difficult, because the beam initially remains in the machine, since the particle is deflected inward immediately after stripping. The method, therefore, only works when the flutter is large enough. Note that in this case, the K-value describing the bending power of the cyclotron magnet [5] must be four times higher than K-value for the cyclotron accelerating H^- ions to the same energy.

4.3 Extraction by means of electrostatic deflector devices

4.3.1 Turn separation in a cyclotron

Some qualitative aspects of orbit separation are explained, to illustrate the general effects. In a cyclotron, the position of a particle with a given energy is determined by a betatron oscillation relative to the equilibrium orbit:

$$r(\theta) = r_0(\theta) + x(\theta)\sin(\nu_r\theta + \theta_0)\ . \tag{26}$$

Here r and θ are the polar coordinates of the particle, r_0 is the equilibrium orbit for a given energy, x is the amplitude of the betatron oscillation, ν_r is the radial betatron oscillation frequency, and θ_0 is an arbitrary offset angle. The equilibrium orbit can be ideally centred and shaped (in the case of a perfectly symmetric magnetic field), or it can be displaced with respect to the centre of the cyclotron (when there is a first-harmonic field perturbation in the field). The betatron oscillation is quasi-sinusoidal but the oscillation amplitude may slightly depend on θ, owing to the AVF characteristic of the magnetic field. For the present purpose, this effect is not important. Equation 26 describes the oscillation of a single particle. However, it can also be used to describe a coherent oscillation of the centre of the beam. The latter case is relevant for the study of turn separation. We can evaluate the radius $r(\theta)$ at a fixed azimuth θ_n but for successive turns n. It is easily derived that in this case [49]:

$$\begin{aligned}\Delta r(\theta_n) = \Delta r_0(\theta_n) &+ \Delta x \sin(2\pi n(\nu_r - 1) + \theta_0) \\ &+ 2\pi(\nu_r - 1)x\cos(2\pi n(\nu_r - 1) + \theta_0)\ ,\end{aligned} \tag{27}$$

where Δr is the radial increase between two successive turns. In this equation, there are three different terms. They relate to three different methods that can be used to generate a turn separation. The first term, Δr_0, represents an increase in the radius of the equilibrium orbit and is related to the energy increase ΔE_0 per turn:

$$\frac{\Delta r_0}{r_0} \approx \frac{1}{2}\frac{\Delta E_0}{E_0}\ , \tag{28}$$

where E_0 is the kinetic energy at the radius r_0. Thus, the relative radial increase is only half the relative energy increase. However, for a given cyclotron, the turn separation Δr_0 will double when the dee voltage is doubled.

The second term in Eq. (27) relates to a turn separation due to an increase in the betatron oscillation amplitude between two successive turns. This is, in general, what happens in a resonance. The

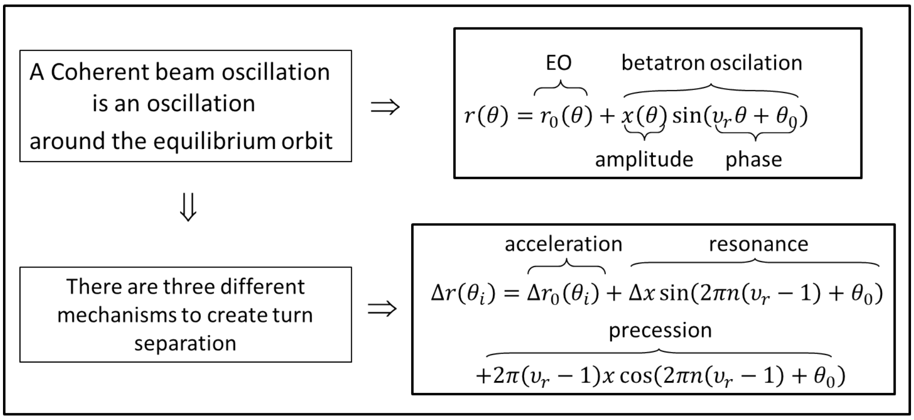

A Coherent beam oscillation is an oscillation around the equilibrium orbit
$\Rightarrow$
EO
betatron oscillation
$r(\theta) = r_0(\theta) + x(\theta)\sin(\upsilon_r\theta + \theta_0)$
amplitude
phase
$\Downarrow$
There are three different mechanisms to create turn separation
$\Rightarrow$
acceleration
resonance
$\Delta r(\theta_i) = \Delta r_0(\theta_i) + \Delta x\sin(2\pi n(\upsilon_r - 1) + \theta_0)$
precession
$+2\pi(\upsilon_r - 1)x\cos(2\pi n(\upsilon_r - 1) + \theta_0)$

Fig. 54: Mechanisms that may contribute to an increase of the turn separation in a cyclotron

resonances that are important for extraction from a (small) cyclotron are the $\nu_r = 1$ resonance and the $2\nu_r = 2$ resonance. The $\nu_r = 1$ resonance is a first-order integer resonance (displacement of the beam) and is driven by a first-harmonic dipole bump. The $2\nu_r = 2$ resonance is a second-order half-integer resonance (exponential growth in the stop band) and is driven by a second-harmonic gradient (quadrupole) bump. The third term in Eq. (27) describes the case where a coherent amplitude already exists, but the turn separation arises from the fact that the phase of the oscillation has advanced between two successive turns. The different mechanisms that create turn separation are illustrated in Fig. 54.

4.3.2 Different extraction methods relying on turn separation

The following classification [51] has been made for the different methods of extraction that rely on an increase in the turn separation.

1. Extraction by acceleration: no means other than acceleration is used to increase the turn separation. This requires a sufficiently high dee voltage and is done, for example, in the IBA/SHI C235 proton therapy cyclotron.
2. Brute force extraction: the beam is extracted during the build-up of the resonance. Owing to this, there is an increase in the amplitude of coherent oscillation in between the last and the penultimate turn. This oscillation is created on or close to the $\nu_r = 1$ resonance by applying a first-harmonic dipole bump. At the same time, the acceleration gives an additional important contribution to the turn separation. The IBA self-extracting cyclotron [38] can probably be best classified in this group.
3. Resonant extraction: here, some coherent beam oscillation is created just before extraction. Two different schemes can be distinguished.
   (a) Precessional extraction: this is more subtle [52–54]. Here, a coherent oscillation is also created with a first-harmonic dipole bump on the $\nu_r = 1$ resonance (or alternatively the oscillation may be created by off-centring the beam at the injection). However, after passing this resonance, the beam is further accelerated into the fringe field of the cyclotron. Here, the value of ν_r drops below 1 (typically to 0.6–0.8). As a result, the betatron phase advance between the last two turns is substantially different from 360° and a turn separation is obtained, which is proportional to the oscillation amplitude and to $\nu_r - 1$. The number of turns in the fringe field should not be too large, to avoid a too large RF phase slip. This extraction method is used in the Varian SC cyclotron for proton therapy.
   (b) Regenerative extraction: this is also subtle. Here the beam is also extracted during resonance build up. The $2\nu_r = 2$ resonance is used. This resonance is driven by a second-harmonic

gradient bump of the field. This means that the second-harmonic bump should show as a function of radius a quadrupole-like dependence (linear increase with radius). This shape is more critical than for the previous methods. If the gradient is large enough, then the resonance will lock the real part of ν_r to a value of one. There is also an imaginary part of ν_r that will cause exponential growth of the betatron oscillation. Regenerative extraction is used in modern superconducting synchro-cyclotrons, such as the IBA S2C2 and the Mevion Monarch.

4.3.3 General layout for resonant extraction

The extraction process for resonant extraction using an electrostatic deflector can generally be subdivided into four steps.

1. Use harmonic coils (placed at a radius where $\nu_r = 1$) to push the beam, create a coherent oscillation, and create the required turn separation. Instead of harmonic coils, the field bump may also be created by properly placed and shaped iron bars. Then obtain turn separation by precession.
2. Use an electrostatic deflector to peel off the last turn and provide an initial radial kick to the beam.
3. Use gradient correctors or focusing channels to guide the beam through the cyclotron fringe field. The primary goal is to reduce the magnetic field locally and to control the field gradients in order to avoid too much optical damage to the beam.
4. Place external focusing elements (quadrupole doublets) as close as possible to the radial pole edge (if necessary, in the cyclotron return yoke), to handle the large beam divergences in the extracted beam.

4.3.4 Harmonic extraction coil

A harmonic extraction coil may be placed at a radius where ν_r is close to one to create a coherent beam oscillation. The principle is simple: if $\nu_r = 1$, the radial kicks that are given to the beam are all in phase and the oscillation amplitude increases linearly with turn number. In reality, this may be more difficult, however. The coil covers a certain radial width (or better, a certain energy range) and ν_r will not be equal to one over the full range. Depending on the number of turns that are seen by the harmonic coil, a situation may very quickly occur where the oscillation amplitude that was already created is lost again because new radial kicks are out of phase with the oscillation built up so far. This occurs when the beam follows the shifted equilibrium orbit adiabatically (too slow acceleration). To avoid this situation, it is important to limit the radial width of the coil. If the shape of the coil follows the local shape of the equilibrium orbit, the particle energy range covered by the coil is small and a non-adiabatic effect may be obtained. The coil should be placed in the pole gap, to use the field amplification produced by the iron. However, this may complicate the design, owing to vertical space limitations. Also, the heat load will be important because water cooling is often not possible. Figure 55 shows a photograph of the harmonic coil that is used in the IBA self-extracting cyclotron.

4.3.5 Electrostatic deflector

The ESD creates an outwardly directed DC electric field between two electrodes. The goal is to give an initial angular deflection to the beam. The inner electrode (called the septum) is placed in between the last and penultimate turn. The septum is at ground potential so that the inner orbits in the cyclotron are not affected. At the entrance, the septum is knife-thin (of the order of 0.1 mm) in order to peel off the last orbit and minimize beam losses on the septum itself. To better distribute the heat due to beam losses, the beginning of the septum is often V-shaped. The septum is water-cooled. The heat loss on the septum

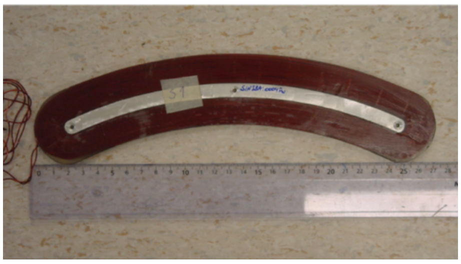


Fig. 55: Harmonic coil used in the IBA self-extracting cyclotron (note that the ruler measures centimeters)

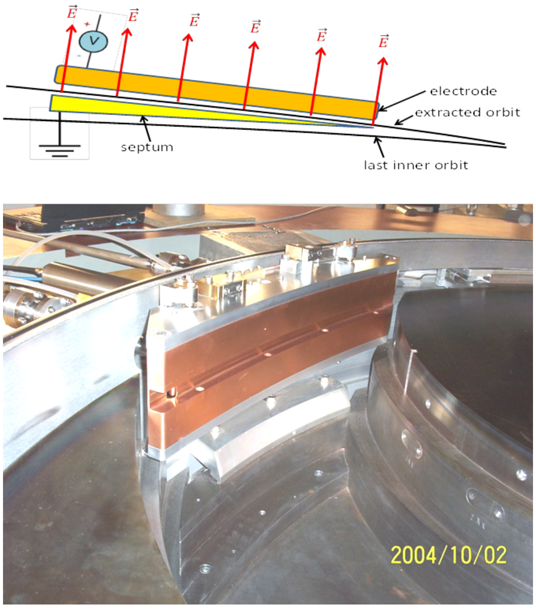


Fig. 56: Top: principle function of an electrostatic deflector. Bottom: electrostatic deflector installed in the IBA C235 cyclotron.

usually determines the maximum current that can be extracted from the cyclotron. The outer electrode is on a negative potential (assuming extraction of positively charged particles). Of course, the shape of the electrodes must follow the shape of the extracted orbit. Figure 56 illustrates the principle and shows the electrostatic deflector that is used in the IBA C235 cyclotron. Figure 57 shows the septum and the electrode.

4.3.6 A gradient corrector channel

The goal of a gradient corrector channel is to guide the beam through the fringe field, to reduce the magnetic field on the extraction path, and to reduce the vertical or increase the radial focusing through the fringe field. Often, more than one magnetic channel is needed along the extraction path. Different types are used:

1. passive channel: made of soft iron bars that are magnetized by the cyclotron magnetic field;
2. active channel: using coils or permanent magnets.

The design always includes an effort to reduce the adverse effect of the channel on the internal orbits. Figure 58 shows a vertical cross-section of the passive focusing channel that was used in the

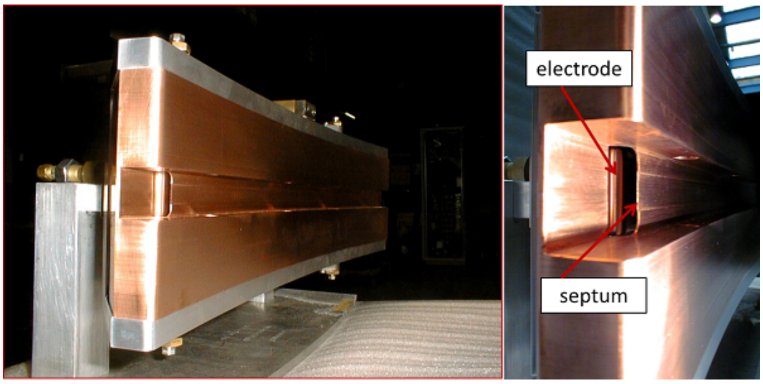


Fig. 57: Close-up of C235 ESD, showing the septum and the electrode

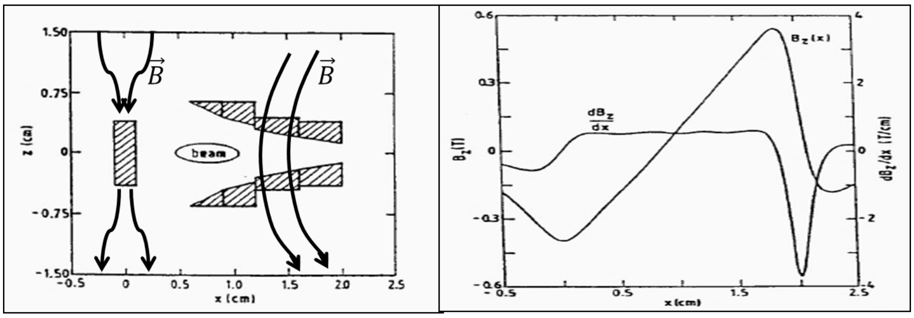


Fig. 58: Left: vertical cross-section of the passive gradient corrector channel used in the small ILEC cyclotron at Technical University Eindhoven. Right: calculated magnetic field and field gradient. The field is increasing with radius, increasing the radial focusing. Figure taken from Ref. [49]

small ILEC cyclotron at Eindhoven University [55]. Here, the poles are shaped to provide a smooth and constant radial gradient at the location of the beam. Near the inner bar, the field decreases because the field lines are sucked into the iron. In between the two outer bars, the field increases because there, the effective pole gap decreases.

4.3.7 Extraction in the IBA C235 cyclotron

Figure 59 shows the extraction scheme that is used in the IBA C235 proton therapy cyclotron. In this cyclotron, turn separation is created by acceleration only. There are no harmonic extraction coils. The only extraction elements are the deflector, the gradient corrector, and the permanent magnet quadrupole doublet that is placed in the return yoke. This cyclotron is a special case because the beam can be accelerated very close to the radial pole edge of the machine. This is achieved by using a pole gap with an elliptical shape. Furthermore, the RF cavities are designed such that there is a strong increase in the dee voltage at large radii so that a larger turn separation is obtained at extraction.

The parameters of the elliptical pole gap are illustrated in the left panel of Fig. 60. It can be seen in the right part of this figure that a good field region is obtained even very close to the radius of the pole, enabling particle acceleration (with isochronous field) very close to the pole radius. Beyond this stable radius, the field decreases very sharply. Because of this feature, only a small kick (provided by the deflector) is needed to extract the beam. The orbit is extracted in about one-quarter of a turn.

The beam leaves the cyclotron by crossing the radial pole edge. Here there is a very strong negative radial magnetic field gradient that would completely defocus the beam horizontally. The gradient

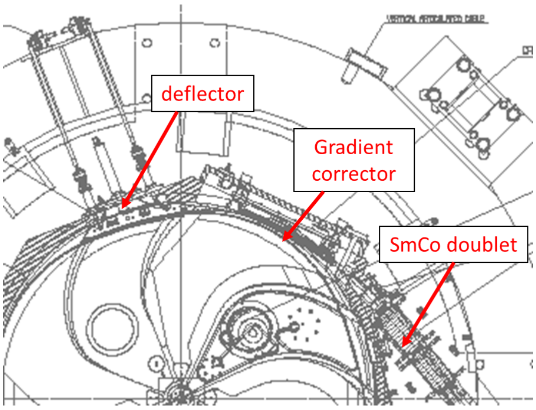


Fig. 59: Extraction scheme used for the IBA C235 proton therapy cyclotron

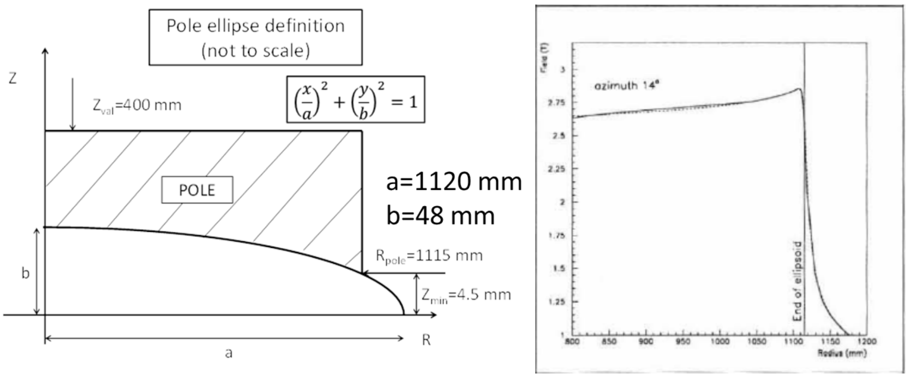


Fig. 60: Left: definition of the elliptical pole gap used in the IBA C235 cyclotron. Right: the magnetic field shows a very sharp cut-off near the radius of the pole.

corrector creates a kind of plateau with descending magnetic field value. Figure 61 shows the installation of this gradient corrector in the C235 cyclotron. It is a passive system made of soft iron that is magnetized by the cyclotron magnetic field and placed between the two main coils, very close to one of the hills of the cyclotron magnet. The gap in the gradient corrector is profiled to obtain a steady magnetic field drop along the direction of the beam but, at the same time, a radial gradient which is well under control. This is shown in the right panel of Fig. 61, where two field profiles are shown in the median plane and perpendicular to the particle orbit, but at two different azimuthal positions along the trajectory (indicated by the blue arrows).

Figure 62 shows the installation of the SmCo permanent magnet doublet in the C235. These magnets are placed immediately at the exit of the vacuum chamber but still in the beam exit penetration of the return yoke. The layout of the permanent magnets is shown in the upper-right panel of Fig. 62. For the first quadrupole, an iron housing is placed around the permanent magnets to shield them from the external cyclotron magnetic field, which is still rather high at this position. The lower-right panel shows the polarity of the individual permanent magnets and the global circulation of the magnetic flux.

4.4 Self-extraction

In a cyclotron, the average magnetic field starts to decrease when approaching the maximum pole radius. This limits the maximum energy that can be achieved in the cyclotron. There are, in fact, two limits.

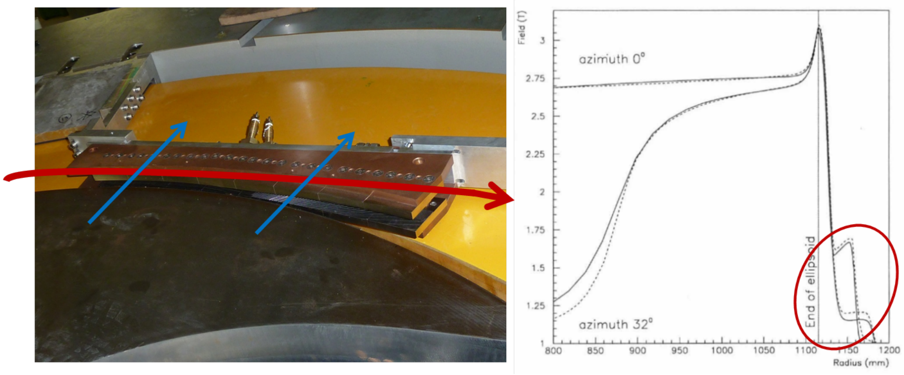


Fig. 61: Left: gradient corrector installed in the C235 cyclotron. Right: magnetic field at two different azimuths, showing the field profile experienced by the extracted beam.

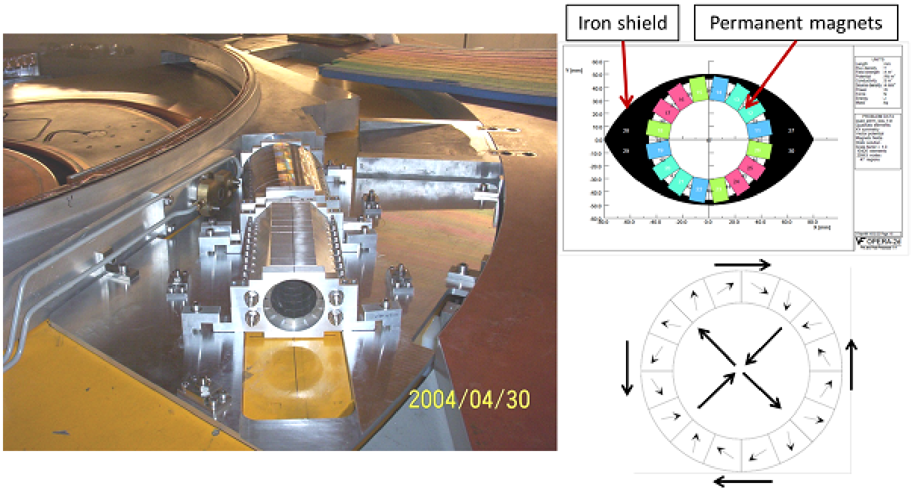


Fig. 62: Left: permanent magnet quadrupole doublet, installed in the IBA C235 cyclotron. Right: layout and polarity of the permanent magnet, producing the high-quality quadrupole field.

1. There is a limit of radial stability. This limit is reached on the equilibrium orbit for which the radial betatron frequency ν_r has fallen to zero ($\nu_r \downarrow 0$). Note that in a rotational symmetric magnetic field this corresponds to the situation where the field index equals -1,
$$n = \frac{r}{B}\frac{\mathrm{d}B}{\mathrm{d}r} = -1\,. \tag{29}$$
This occurs at the radius where the magnetic rigidity $p/q = B \cdot r$ reaches its maximum.
2. There is a limit of acceleration that occurs as a result of the loss of isochronism. This limit is achieved when the RF phase has slipped 90°. Of course, it depends on the dee voltage.

If the vertical pole gap is much larger than the radial gain per turn (as is the case for many cyclotrons that have been built so far), the second limit is achieved earlier than the first. However, when the pole gap is small and, furthermore, when this gap is elliptically shaped (and the dee voltage is not too small), the first limit is achieved first and the beam is self-extracted. However, in this case, the particles will come out at all possible azimuths and there is not really a well-defined coherent extracted beam. To achieve this, an elliptical pole gap is used, which makes enables the realization of very sharp

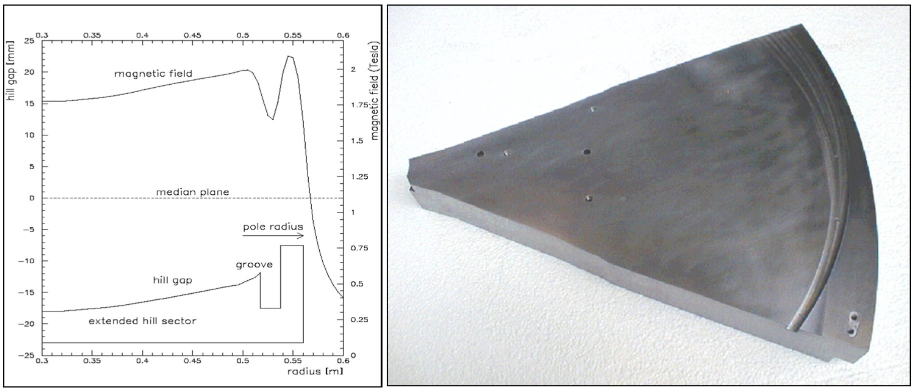


Fig. 63: In the IBA self-extracting cyclotron, a groove is machined in one of the poles (right), to create an extraction channel. This channel provides a magnetic septum at the entrance and, at the same time, gradient correction and focusing. An elliptical pole gap is used, allowing for sharp radial gradients in the magnetic field near the pole edge.

magnetic gradients close to the pole radius. At the same time, a groove or plateau is machined in one of the cyclotron poles to provide an extraction channel. This channel simultaneously serves as magnetic septum and gradient corrector. This is illustrated in Fig. 63.

In principle, the scheme for self-extraction is quite similar to the usual scheme for resonant precessional extraction:

1. harmonic coils create a coherent oscillation;
2. the beam is accelerated into the fringe field where $\nu_r \approx 0.6$;
3. the groove creates a kind of magnetic septum and at the same time provides for a gradient corrector channel;
4. a permanent magnet doublet is placed within the vacuum chamber; this doublet continues the extraction path and focuses the beam in both directions.

The left panel of Fig. 64 shows the interior of the IBA self-extracting cyclotron [38, 56]. The part of the beam that is not well extracted is intercepted by a low-activation water-cooled beam dump. With this machine, a beam intensity of 2 mA has been extracted with an efficiency of 80%. The horizontal beam emittance (2σ) is about 300π mm mrad.

The right panel of Fig. 64 shows the gradient corrector that is used. This is an active channel made of samarium-cobalt permanent magnets. It acts like a quadrupole doublet, but with the first doublet being longer than the second. The quadrupole-like field shape is obtained by using two opposite dipole fields that are radially displaced a few centimetres with respect to each other. Some additional small magnets are placed to minimize the adverse effect of the gradient corrector on the internal orbits.

4.5 Two different extraction schemes in one cyclotron

4.5.1 The IBA C70XP cyclotron

The IBA C70XP cyclotron was discussed in Section 2.5. In this cyclotron, two independent extraction systems are installed:

1. an ESD for the extraction at fixed energy of positively charged particles ($^4\mathrm{He}^{2+}$ and H_2^+);
2. a variable-energy stripper system for negatively charged particles (H^- and D^-).

The stripping extraction is implemented on two opposite poles, allowing simultaneous dual beam extraction. On one side, the two independent extraction modes are linked by the strong geometrical

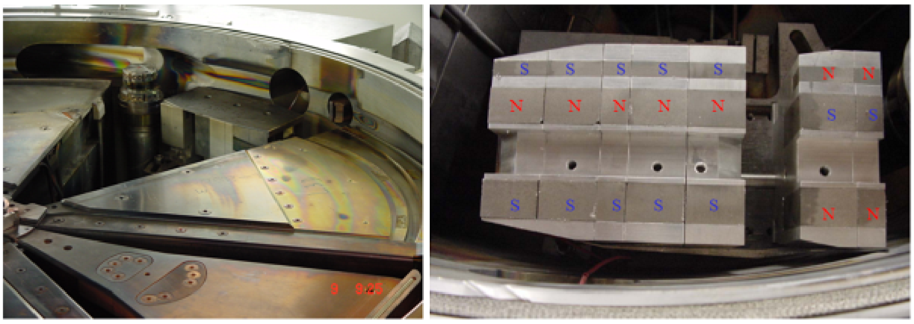


Fig. 64: Left: extraction elements in the IBA self-extracting cyclotron. The harmonic coils are placed underneath the pole covers. Right: lower part of the permanent magnet corrector used in the IBA self-extracting cyclotron. The corrector is placed close to the vacuum chamber visible in the lower part of the photo. The polarity of the magnets is indicated. The polarities are inverted in the upper part of the corrector.

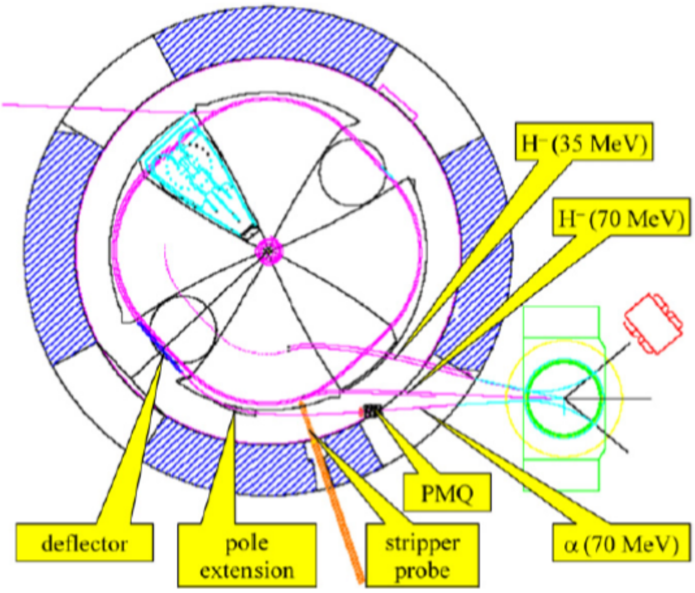


Fig. 65: Dual extraction system (stripping and electrostatic deflector) used in the IBA C70XP-cyclotron: PMQ, permanent magnet quadrupole.

constraint that both beams converge to the centre of one common switching magnet, as illustrated in Fig. 65. This is obtained by having both extractions taking place at the same pole. The position of the stripped beam at the switching magnet is controlled by the position of the stripping probe on the pole. The position of the beam extracted by the ESD is determined by the optimized design and length of the radial pole extension and also depends on the deflector voltage. The radial pole extension locally modifies the falling gradient in the fringe field and facilitates good optical beam transport in this critical region.

Figure 66 shows the ESD. A pre-septum made of tungsten is placed just in front of the actual septum of the ESD. It is water-cooled and protects the septum from overheating. The septum itself consists of two parts: the first part is made of two (upper and lower) tungsten foils that are brazed to their copper supports. This allows for a good heat evacuation. The two tungsten foils can expand independently, avoiding an increase in effective septum thickness due to thermal expansion. The second

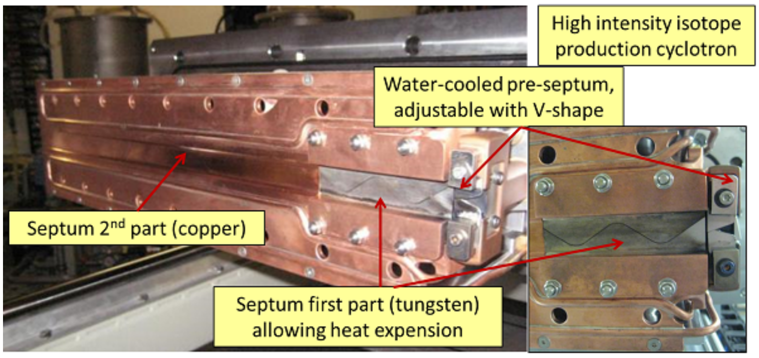
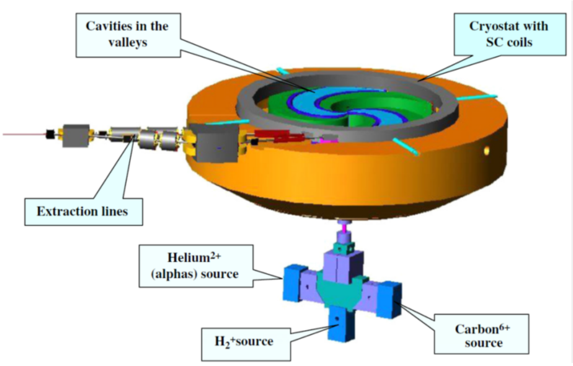

High intensity isotope production cyclotron
Water-cooled pre-septum, adjustable with V-shape
Septum 2nd part (copper)
Septum first part (tungsten) allowing heat expension

Fig. 66: High-power electrostatic deflector used in the IBA C70XP-cyclotron



particles	$^{12}C^{6+}$; $H_2{}^{+}$; $^{4}He^{2+}$
Final energy	
ions	400 MeV/A
protons	265 MeV
Bending limit	K=1600
Weight	700 t
Diameter	6.6 m
Hill field	4.5 Tesla
Valley field	2.45 Tesla
Number of cavities	2
RF frquency	75 MHz; h=4
Vdee	80-160 kV
Number of turns	2000
SC coil	NbTi; Helium cooled
Ischronism of $H_2{}^{+}$	Coil in 2 parts

Fig. 67: C400 superconducting cyclotron and main parameters, proposed by IBA for carbon and proton therapy: SC, superconducting.

part of the septum is made of copper and is machined from a solid copper block. Extraction efficiencies higher than 80% have been obtained with this ESD.

4.5.2 The IBA C400 cyclotron

A design study of the compact superconducting isochronous cyclotron C400 [57, 58]. has been made by IBA in collaboration with the JINR at Dubna. When built, it will be the first cyclotron in the world capable of delivering protons, carbon, and helium ions for cancer treatment. The $^{12}\mathrm{C}^{6+}$ and the $^{4}\mathrm{He}^{2+}$ will be accelerated to 400 MeV$/$A and extracted by an electrostatic deflector. H_2^+ ions will be accelerated to an energy of 265 MeV$/$A and extracted by stripping. The left panel of Fig. 67 shows an artist's impression of the cyclotron. The table on the right lists the main parameters. The magnet yoke has a diameter of 6.6 m and the total magnet weight is about 700 tonnes. The maximum magnetic field in the hills is 4.5 T. Three ion sources are mounted on a switching magnet placed in the injection line below the cyclotron.

Extraction of protons is done with a stripping foil. The minimum proton energy that can be obtained is 320 MeV, for single-turn extraction, and 265 MeV, for two-turn extraction (see Fig. 68). The second solution was chosen in order to be closer to the usually applied energy for the proton beam treatment.

Electrostatic deflection extraction is used for the C and He beams. A single ESD located in a valley is used, with an electric field of $\approx$15 MV$/$m. From the right panel of Fig. 68, it can be seen that the beams from both extraction systems do not exit in the same direction and position: they are combined

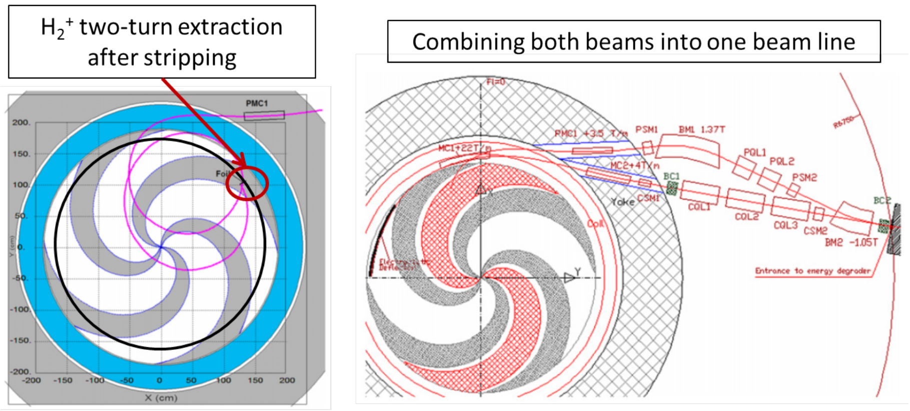


Fig. 68: Left: in the IBA C400 cyclotron, protons are extracted by stripping the accelerated H_2^+ ion by a thin stripping foil. Right: $^{12}C^{6+}$ is extracted with an electrostatic deflector. Both particles, protons and carbon, are combined into one beam line.

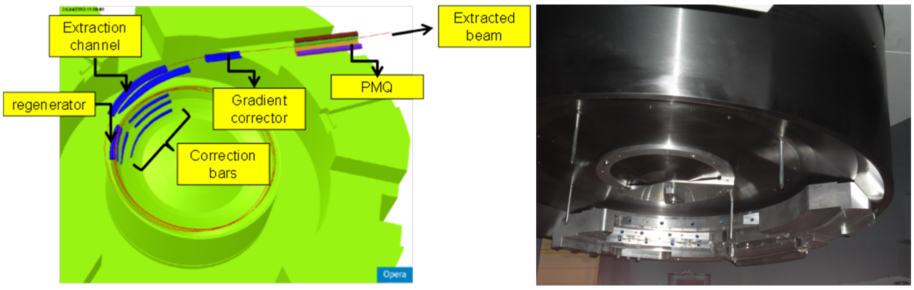


Fig. 69: Passive extraction system of the IBA superconducting synchro-cyclotron S2C2: PMQ, permanent magnet quadrupole.

by two external beam lines onto the common degrader. Downstream, both beams travel in the same beam line.

4.6 Regenerative extraction in the IBA S2C2

In the IBA superconducting synchro-cyclotron (S2C2), extraction with an electrostatic deflector cannot be used, because the turn separation at extraction is far too small. Instead, the regenerative method of extraction based on the $2\nu_r = 2$ resonance is used. The extraction system is fully passive: only soft iron elements are used. The layout is shown in Fig. 69. The regenerator creates a strong magnetic field bump, of which the quadrupole component increases the radial focusing and locks the horizontal betatron frequency ν_r to one. The orbit becomes unstable and is pushed towards the extraction channel by the first-harmonic component of the magnetic field, also produced in the bump. It is essential to avoid the Walkinshaw resonance in this process, as illustrated in Fig. 70. When the extraction sets in, the displacement of the beam towards the extraction channel steadily and exponentially builds up. The magnetic septum of the extraction channel separates the extracted orbit from the last internal orbit. The locking of the betatron frequency ν_r to one can be observed in the left panel of Fig. 70 by the fact that the subsequent azimuth of the nodes of the oscillation is fixed for multiple turns.

A series of iron correction bars is needed to compensate the magnetic field undershoots towards inner radii that are produced by the regenerator and the extraction channel. Farther downstream, between the main coils, a three-bar gradient corrector is placed, which reduces the radial defocusing in the fringe

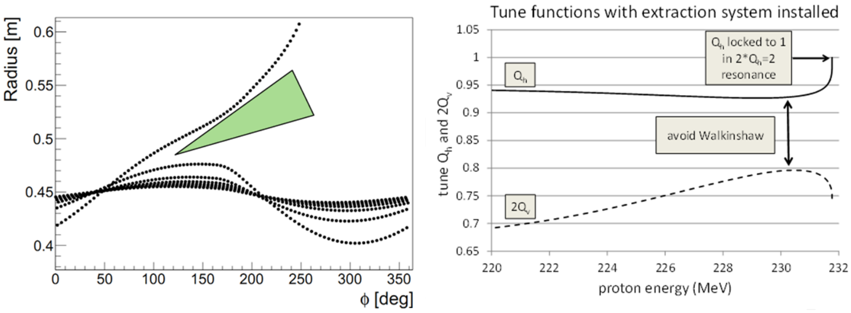


Fig. 70: Regenerative extraction: a strong quadrupole bump increases ν_r and locks it to one. A steady shift of the beam towards the extraction channel is built up. The Walkinshaw resonance ($\nu_r = 2\nu_z$) must be avoided.

field of the pole. Finally a permanent magnet quadrupole is used to further match the extracted beam to the external beam line.

5 Some aspects of cyclotron magnetic design

A good overview of considerations and methods applied for the design of compact cyclotron magnets has already been made in previous topical CAS courses by Jongen and Zaremba [59, 60]. Therefore, after some general considerations, we limit ourselves in this section to some recent examples and some new methods that have been developed.

5.1 Some general considerations

In industrial practice, many design choices of the cyclotron magnet are often already fixed before the start of any design calculations. Such choices are determined by more general considerations, such as:

1. the application of the cyclotron, which will determine the particles to be accelerated and their required maximum kinetic energy;
2. from the maximum particle rigidity, the choice of maximum B-field will determine the pole radius, or vice versa;
3. the choice of the cyclotron type: isochronous or synchro-cyclotron;
4. the coil technology: superconducting or normally conducting;
5. the type of extraction to be applied: stripping extraction, ESD, or regenerative extraction;
6. when choosing, for example, an isochronous solution, the following parameters are often already determined beforehand:
   (a) number of sectors, pole gap, pole angle, valley depth, pole spiral, etc.;
   (b) number of dees, dee voltage, harmonic mode, and RF frequency;
   (c) for injection, the use of an internal or external ion source.

Other choices can be made with some rough back-of-the-envelope calculations: here are a few examples.

1. The maximum magnet rigidity is obtained from the particle maximum kinetic energy as follows:

$$B\rho = \frac{\sqrt{T^2 + 2TE_0}}{300Z} \,, \tag{30}$$

   where Z is the charge ionization number of the particle. The rest-energy, E_0, and the kinetic energy, T, are expressed in MeV and the rigidity $B\rho$ in Tm.

2. For a magnet that is far from magnetic saturation, the relation between the magnetic field in the pole gap and the number of ampere turns in the coil can be approximated as:

$$\oint \vec{H} \cdot \mathrm{d}\vec{l} = (NI)_{\mathrm{tot}} \approx \frac{1}{\mu_0}\left(gB_{\mathrm{gap}}\right) + \frac{1}{\mu_0 \mu_{\mathrm{iron}}}\left(L_{\mathrm{iron}} B_{\mathrm{iron}}\right) . \tag{31}$$

   Here, the integration is made over a closed loop that crosses the gap and closes via the return yoke around the coils of the magnet. Further $(NI)_{\mathrm{tot}}$ is the total number of ampere turns in the coils, g is the pole gap, L_{iron} is the length of the loop in the iron, B_{gap} and B_{iron} are the magnetic fields in the gap and iron, respectively, μ_0 is the magnetic permeability in vacuum and μ_{iron} is the relative magnetic permeability in iron. In practice, for a conventional magnet far from saturation, the contribution of the iron is smaller than a few percent.
3. Assuming a hard-edge approximation for the fields produced from the iron pole sectors, the average field $\bar{B}$ can be estimated as:

$$\bar{B} \approx \alpha B_{\mathrm{h}} + (1-\alpha) B_{\mathrm{v}} . \tag{32}$$

   Here, α is the fraction of the hill angle on one symmetry period and B_{h} and B_{v} are the magnetic field values in the hill and valley, respectively.
4. To estimate the width of the return yoke, the total magnetic flux, Φ, produced between the poles must be roughly equal to the flux guided in the yoke:

$$\Phi \approx 2\pi R_{\mathrm{p}}^2 \bar{B} \approx B_{\mathrm{ret}} A_{\mathrm{ret}} . \tag{33}$$

   Here R_{p} is the (effective) pole radius, B_{ret} is the magnetic field in the return yoke, and A_{ret} is the total surface cross-section of the return yoke.
5. The required increase in average magnetic field, necessary for isochronous operation, may be estimated from the relativistic mass increase of the particles as follows:

$$\bar{B}(r) \approx B_0 \gamma_{\mathrm{rel}} , \tag{34}$$

   where B_0 is the magnetic field in the centre and γ_{rel} is the relativistic gamma.
6. The magnetic flutter may be obtained from the hard-edge approximation, as given in Eq. (13).
7. The betatron frequencies may be related to the flutter and the pole spiral angle, as given in Eqs. (14) and (15).

5.2 Tools for magnetic modelling in Opera

Almost all magnet designs at IBA are done using Opera simulation software from Cobham (VectorFields) [7]. The main packages used are as follows.

5.2.1 Opera-2d

This is the 2D version of the software, which can be used to model magnets with rotational symmetry. It is a perfect choice for an initial design of a synchro-cyclotron magnet (without the extraction system). However, it can also be useful for an initial design of a 3D isochronous cyclotron magnet. In this case, stacking factors can be used to take into account the hill–valley structure of such a magnet. This yields a preliminary idea of the average magnetic field (as a function of radius), the stray field around the cyclotron, and the dimensions of the return yoke.

5.2.2 Opera-3d—modeller interface

This module allows for full 3D simulations. It is easy to use and easy to include fine geometrical details. The 3D finite-element mesh is generated automatically. The most commonly used is the tetrahedral mesh. This mesh is not always very regular, which may result in some noise in the calculated magnetic field.

5.2.3 Opera-3d—preprocessor interface

This module also allows for full 3D simulations, but it is more difficult to use and even more difficult to include fine geometrical details. However, the 3D finite-element mesh is fully created by the user and therefore fully under control. A hexahedral mesh is used and less noisy magnetic fields can be obtained. This module may be useful, for example, for the precise calculation of magnetic forces.

5.3 Design approach using Opera-3d modelling

The cyclotron magnet design approach used at IBA can be characterized with the following features.

1. The 3D models are fully parameterized.
2. Models are automatically generated using macros. Together with the previous feature, this allows some parameters to be changed easily, so that the corresponding new model is quickly obtained.
3. The macros for all different types of IBA cyclotron (such as S2C2, C230, C19/9-family, C30-family, C70-family) have similar structures. This enables the quick creation of new macros for new machines if needed.
4. The magnetic properties of the iron (the BH-curves) are verified with an in-house permeability meter.

The following types of property are parameterized:

1. all important dimensional parameters and pole profiles;
2. main coil settings;
3. material properties, such the different BH-curves for different iron parts;
4. the sizes and other properties of the finite-element mesh;
5. solver tolerances;
6. switches for the selection or de-selection of individual subsystems;
7. switches for filling individual subsystems with air or iron.

The main structure of a cyclotron modelling macro is illustrated in Fig. 71. The first column (blocks in green) shows the main steps, namely: (i) creation of the model in the Opera-3d modeller, (ii) solving the model with the Opera-3d solver, and (iii) analysing the result with the Opera-3d post-processor. The second column (blocks in blue) illustrates how a model is created in the modeller: (i) read from a file all the parameters that are needed to create the model, (ii) create individual subsystems, (iii) assemble these subsystem into one model that represents the full cyclotron, (iv) create and save the database that serves at input for the solver. The third column (yellow blocks) shows some details of some of the blue blocks. Subsystems can be, for example, the yoke (upper, return), pole, or main coils, but also, for example, a passive extraction system, a three-bar gradient corrector, or a permanent magnet quadrupole. Creation of the full cyclotron model includes steps like (i) loading all subsystems, (ii) defining material properties, and (iii) defining boundary conditions or symmetries. The creation of the database requires surface and volume meshing but also enables the inclusion of different cases to be solved, such as, a list of different main coil currents. The last column (in orange) shows some details for the creation of one subsystem, such as (i) creation of the geometry, attaching material labels and mesh-properties to different cells in the subsystem, and saving.

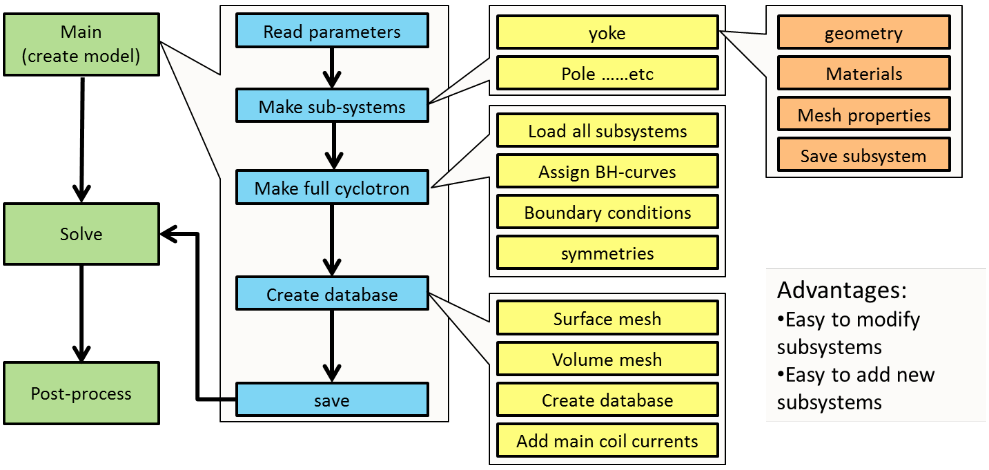


Fig. 71: Logical structure of a typical input file used for the fully parameterized and automated creation of an Opera-3d model of a cyclotron.

5.4 Some examples

In this section, we consider as an example the recent development of the IBA superconducting synchro-cyclotron S2C2. Figure 72 shows a model of the S2C2 made in Opera-2d. The vertical axis is the axis of rotational symmetry. The main coils are shown in red and the magnet iron in deep blue. The iron of the S2C2 magnet is heavily saturated; therefore, additional iron placed around the yoke might introduce a median plane error or increase the vertical forces on the cold mass. The goal of this study was to calculate the vertical asymmetry introduced by the cyclotron support feet. These are shown in light blue. The feet do not obey rotational symmetry (there are four of them), but a stacking factor was used to approximate their average effect. To compensate for the effect of the feet, an iron ring is placed on top of the yoke. Figure 73 shows some results of these simulations. The left panel shows the total vertical force acting on the cold mass during ramp-up of the main coil current. Three cases are shown: (i) for the blue curve, the feet were present in the model but the ring was not; (ii) for the red curve, the ring was present but the feet were not; and (iii) for the green curve, both the feet and ring were present. The dimensions of the ring were optimized to minimize the total vertical force during ramp-up. It can be seen that the total asymmetry due to the feet (about 25000 N at full current) is reduced to about 8000 N by placing the ring. The right panel shows the magnetic median plane error (the radial field in the median plane) as a function of radius for the same three cases at the maximum excitation current. It can be seen that for the same optimized dimensions of the ring, the maximum field error is reduced from 7 G to <1 G. It is noted that in such a synchro-cyclotron, median plane errors must be avoided as much as possible, because of the weak vertical focusing and also because of the large number of turns in such a machine. Based on these calculations, the described method has successfully been used in the actual machine.

The left panel of Fig. 74 shows a model of the S2C2 that was created using the Opera-3d pre-processor. The typical hexahedral mesh is clearly visible in this model. The goal of this model was to calculate precisely the forces and torques on the cold mass as a function of its position and rotation. An understanding of these forces is crucial for the design of the tie rods by which the cold mass is suspended in the cryostat. To have little heat flow from the cold mass to the exterior, the diameter of these tie rods should be as small as possible. Conversely, the diameter should be large enough to withstand the magnetic forces acting on the cold mass. The differential forces on the cold mass due to translations or rotations can be calculated with better precision in the preprocessor. Results of these calculations are summarized in the table on the right of Fig. 74. It was found that all forces and torques vary linearly with displacement or rotation. Furthermore, it can be seen that all coil movements are unstable: a given

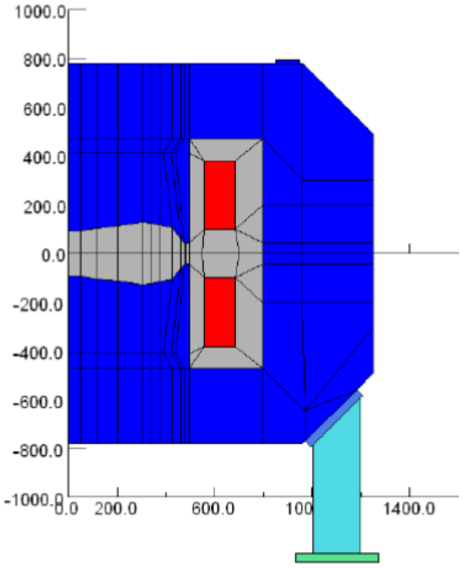
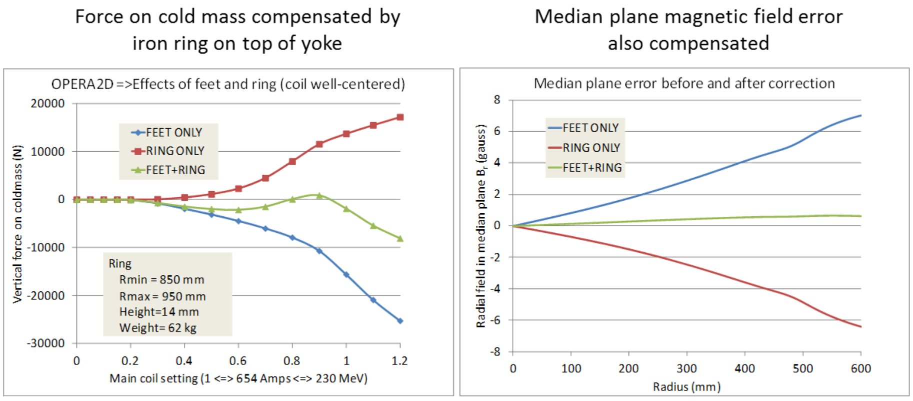


Fig. 72: The Opera-2d model of the S2C2


Force on cold mass compensated by iron ring on top of yoke

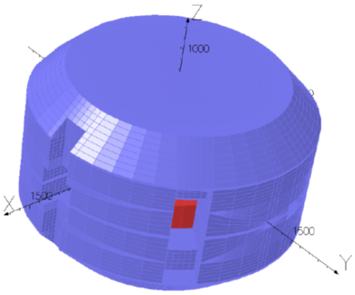
Median plane magnetic field error also compensated

Fig. 73: Opera-2d simulation of the median plane error in the S2C2, as produced by the feet underneath the cyclotron, and compensation by an iron ring placed on top of the yoke.

displacement will create a force in the same direction as the displacement. Horizontal displacements result in a force of about 2 tonnes per mm. For vertical displacements this is about 0.5 tonnes per mm.

The optimization of the magnetic circuit of the S2C2 has been a long process, carried out by three main contributors: IBA, AIMA Developpement, and ASG (the company that made the coil and the cryostat). Many aspects were considered, such as:

1. optimization of the pole-gap profile;
2. definition of the pole radius and the 230 MeV extraction radius;
3. optimization of coil current density and dimensions to assure a margin with respect to the critical surface of the used superconducting material and to allow operation up to 250 MeV;
4. dimensions of the yoke to balance outside stray fields reasonably;
5. dimensions and placement of all horizontal and vertical yoke penetrations (ports are needed for RF system, ion source, vacuum, beam exit, cryo-coolers, 3 horizontal and 2 × 3 vertical tie rods);



FORCES AND TORQUES ACTING ON THE MAIN COIL SYSTEM DUE TO COIL DISPLACEMENTS AND ROTATIONS							
		FORCES			TORQUES		
		dFx ton/mm	dFy ton/mm	dFz ton/mm	dTx Nm/mm	dTy Nm/mm	dTz Nm/mm
coil shift	x-direction	1.99	-0.05	0.00	0	-9	8
	y-direction	-0.05	2.00	0.00	10	2	41
	z-direction	0.00	0.00	0.56	-80	-201	0
coil rotation		dFx ton/deg	dFy ton/deg	dFz ton/deg	dTx Nm/deg	dTy Nm/deg	dTz Nm/deg
	around x-axis	-0.02	0.00	-0.12	91559	-4609	-80
	around y-axis	-0.05	-0.01	-0.30	-4484	91305	79

Fig. 74: Opera-3d preprocessor model of the S2C2, to estimate forces and torques due to displacement acting on the cold mass with respect to the geometrical centre of the cyclotron.

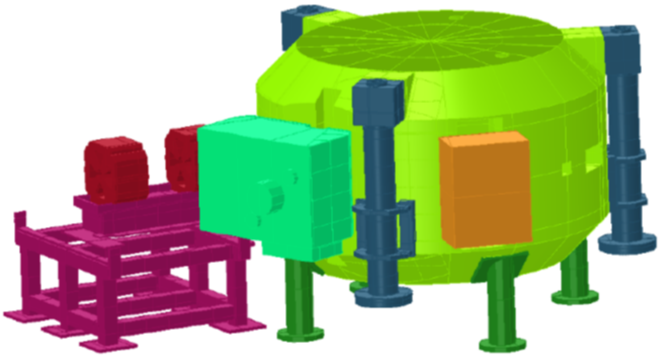


Fig. 75: A complete and detailed Opera-3D model was made of the IBA S2C2

6. optimization of the extraction system;
7. shielding required for external systems, such as the rotating capacitor and the cryogenic coolers;
8. the influence of the external iron systems on the accelerated beam;
9. the influence of the fringe field on the external beam line;
10. median plane errors introduced by the vertical asymmetry in the magnetic design and compensation of these errors;
11. magnetic forces acting on the return yoke, the coils, the extraction system elements, external components, etc.;
12. compensation of first-harmonic field errors;
13. the influence of the cyclotron feet and the yoke lifting system.

For the design of the superconducting coil, the transient behaviour of the magnet with eddy currents and AC losses needs to be studied in detail, along with the quenching behaviour.

To make this optimization, magnetic finite-element models were produced, using Opera-2d, Opera-3d, and CST. In Opera-3d especially, very detailed models were made, as illustrated in Fig. 75. Besides the yoke, poles and main coils, more detailed elements and features were included, such as the yoke penetrations, the extraction system (regenerator and extraction channel with their first-harmonic corrector bars, gradient corrector, and permanent magnet quadrupole), and external systems, such as the cyclotron feet, yoke lifting system, cryo-cooler shields, rotating capacitor shield, and external quadrupoles with support structure.

6 Conclusions

The three main subjects discussed in this report (magnetic field design, beam injection, and beam extraction) are the most difficult problems in cyclotron design. For industrial applications and isotope production, there is generally a need for ever-increasing beam intensity. This implies, at the same time, a minimization of beam losses during both injection and extraction and also an understanding of space charge effects in these processes. For particle therapy applications,the importance lies much more in understanding beam quality, reproducibility, and stability. Except perhaps for Baartman's paper on space charge effects in cyclotrons [61], nothing really new has appeared in theoretical and analytical methods during the last 5–10 years that was helpful in this aspect. However, increasing importance is given to the computational tools that are needed to optimize the design. This concerns 3D finite-element software packages that are used to model the magnetic field as well as electrical field, but also more specialized orbit tracking codes that can be produced in sufficient detail. The latter are usually developed in house.

Acknowledgements

We would like to thank our colleagues Jarno Van de Walle, Vincent Nuttens, Emma Pearson, and Eric Forton for their help in preparing this paper.

References

[1] E.O. Lawrence and M.S. Livingston, *Phys. Rev.* **40**(1) (1932) 9. http://dx.doi.org/10.1103/PhysRev.40.19

[2] C. Oliver *et al.*, Proceedings of the 2013 Cyclotron Conference, Vancouver, Canada, p. 429.

[3] V. Veksler, *J. Phys. USSR* **9**(3) (1945) 153.

[4] E.M. McMillan, *Phys. Rev.* **68**(5-6) (1945) 143L. http://dx.doi.org/10.1103/PhysRev.68.143

[5] J.L. Livingood, *Principles of Cyclic Particle Accelerators* (Van Nostrand, New York, 1961), Chap. 6.

[6] L.H. Thomas, *Phys. Rev.* **54**(8) (1938) 580. http://dx.doi.org/10.1103/PhysRev.54.580

[7] http://operafea.com

[8] K.L. Brown *et al.*, Transport: a computer program for designing charged particle beam transport system, CERN 90/04 (1980), p. 39.

[9] N. Christofilos, Focussing system for ions and electrons, US patent 2736799, priority date March 10, 1950.

[10] E.D. Courant *et al.*, *Phys. Rev.* **88**(5) (1952) 1190. http://dx.doi.org/10.1103/PhysRev.88.1190

[11] H.L. Hagedoorn and N.F. Verster, *Nucl. Instrum. Methods* **18–19** (1962) 201. http://dx.doi.org/10.1016/S0029-554X(62)80032-9

[12] H. Willax, Proposal for a 500 MeV isochronous cyclotron with ring magnet, Proc. Int. Conf. on Sector-Focused Cyclotrons, Geneva, 1963, p. 386.

[13] J.R. Richardson, *AIP Conf. Proc.* **9** (1972) 126. http://dx.doi.org/10.1063/1.2946355

[14] C.B. Bigham *et al.*, *Phys. Canada* **29**(4) (1973) 29.

[15] H.G. Blosser *et al.*, VII Int. Conf. on Cyclotrons and their Applications (1975), p. 584. http://dx.doi.org/10.1007/978-3-0348-5520-4_123

[16] X.Y. Wu, Ph.D. thesis, Michigan State University, 1990.

[17] N.F. Verster and H.L. Hagedoorn, *Nucl. Instrum. Methods* **18–19** (1962) 327. http://dx.doi.org/10.1016/S0029-554X(62)80041-X

[18] M.M. Gordon, *Part. Accel.* **16** (1984) 39.

[19] S. Zaremba *et al.*, Magnetic field design and calculations for the IBA C70 cyclotron, 18th Int. Conf. on Cyclotrons and their Applications (2007), p. 75.

[20] S. Galès, AGOR, a superconducting cyclotron for light and heavy ions, 11th Int. Conf. on Cyclotrons and their Applications (1986), p. 194.

[21] H.W. Schreuder (AGOR team), AGOR: initial beam tests, transport and commissioning, 14th Int. Conf. on Cyclotrons and their Applications (1995), p. 6.

[22] E.L. Kelly, *Nucl. Instr. Meth.* **18–19** (1962) 33. http://dx.doi.org/10.1016/S0029-554X(62)80006-8

[23] N.F. Verster *et al.*, *Nucl. Instr. Meth.* **18–19** (1962) 88.
http://dx.doi.org/10.1016/S0029-554X(62)80014-7

[24] http://cyclotron.lbl.gov/cyclotron-history

[25] W. Kleeven *et al.*, The influence of magnetic field imperfections on the beam quality in an H$^-$ cyclotron, 13th Int. Conf. on Cyclotrons and their Applications (1992), p. 380.

[26] http://www.mevion.com/triniobium-core

[27] W. Kleeven *et al.*, The IBA superconducting synchrocyclotron project S2C2, 20th Int. Conf. on Cyclotrons and their Applications (2013), p. 115.

[28] P. Heikkinen, Injection and extraction for cyclotrons, CAS CERN 94/01 **II** (1994), p. 819.

[29] P. Mandrillon, Injection into cyclotrons, CAS CERN 96/02 (1996), p. 153.

[30] W. Kleeven, Injection and extraction for cyclotrons, CAS CERN 2006/012 (2006), p. 271.

[31] J.-L. Belmont, Ion transport from the source to the first cyclotron orbit, Proc. XXXIII European Cyclotron Progress Meeting [*Nukleonika* **48** supplement 2 (2003) S13].

[32] B.F. Gavin, in *The Physics and Technology of Ion Sources*, 1st ed., Ed. I.G. Brown (Wiley, New York, 1989), p.167.

[33] M. Reiser, *Nucl. Instrum. Methods* **18–19** (1962) 370.
http://dx.doi.org/10.1016/S0029-554X(62)80047-0

[34] N. Hazewindus *et al.*, *Nucl. Instrum. Methods* **118**(1) (1974) 125.
http://dx.doi.org/10.1016/0029-554X(74)90692-2

[35] H. Houtman *et al.*, *Comput. Phys.* **8**(4) (1994) 469. http://dx.doi.org/10.1063/1.168506

[36] M. Reiser and J. Mullendore, *Nucl. Instrum. Methods* **59**(1) (1968) 93.
http://dx.doi.org/10.1016/0029-554X(68)90350-9

[37] E. Liukkonen *et al.*, Design of the central regions for the MSU 500 MeV superconducting cyclotron, Proc. 8th Int. Conf. on Cyclotrons and their Applications [*IEEE Trans. Nucl. Sci.* **NS-26**(2) (1979) 2107]. http://dx.doi.org/10.1109/tns.1979.4329816

[38] W. Kleeven *et al.*, The IBA self-extracting cyclotron project, Proc. 33rd European Cyclotron Progress Meeting [*Nukleonika* **48** supplement 2 (2003) S59].

[39] D. Bohm and L.L. Foldy, *Phys. Rev.* **72**(8) (1947) 649. http://dx.doi.org/10.1103/PhysRev.72.649

[40] R. Beurtly *et al.*, *Nucl. Instrum. Methods* **33**(2) (1965) 338.
http://dx.doi.org/10.1016/0029-554X(65)90069-8

[41] J.-L. Belmont and J.L. Pabot, *IEEE Trans. Nucl. Sci.* **NS-13**(4) (1966) 191.
http://dx.doi.org/10.1109/TNS.1966.4324204

[42] J.-L. Belmont, Axial injection and central region of AVF cyclotrons, Lecture notes of 1986 RCNP KIKUCHI Summer School on Accelerator Technology. Osaka, Research Center for Nuclear Physics, Osaka, p. 79.

[43] L. Root, Ph.D. thesis, University of British Columbia, 1972.

[44] R. Baartman and W. Kleeven, *Part. Accel.* **41** (1993) 41.

[45] W. Kleeven and R. Baartman, *Part. Accel.* **41** (1993) 55.

[46] W. Kleeven *et al.*, *Nucl. Instrum. Methods Phys. Res. B* **269**(24) (2011) 2857.
http://dx.doi.org/10.1016/j.nimb.2011.04.031

[47] http://www.d-pace.com/products_hion.html

[48] W. Kleeven *et al.*, *Nucl. Instrum. Methods Phys. Res. B* **64**(1-4) (1992) 367. http://dx.doi.org/10.1016/0168-583X(92)95496-E

[49] J.I.M. Botman and H.L. Hagedoorn, Extraction from cyclotrons, CAS CERN 96/02 (1996), p. 169.

[50] G. Dutto, The TRIUMF 520 MeV cyclotron, 18 Int. Conf. on Cyclotrons and their Applications, Vancouver 1992, (World Scientific, 1993), p. 138.

[51] W. Joho, Extraction from medium and high energy cyclotrons, Proc. 5 Int. Conf. on Cyclotrons (Butterworths, London, 1971), p. 159.

[52] H.L. Hagedoorn and P. Kramer, *IEEE Trans. Nucl. Sci.* **NS-13**(4) (1966) 64. http://dx.doi.org/10.1109/TNS.1966.4324177

[53] H.L. Hagedoorn and N.F. Verster, The extraction of the beam of the Philips AVF cyclotron, CERN 63-19 (1963), p. 228.

[54] M.M. Gordon and H.H. Blosser , Orbit calculations on the extraction system for the MSU cyclotron, CERN 63-19 (1963), p. 236.

[55] W. Kleeven, Ph.D. thesis, Eindhoven University, 1988.

[56] Y. Jongen *et al.*, High intensity cyclotrons for radioisotope production or the comeback of the positive ions, Proc. 14th Int. Conf. on Cyclotrons and their Applications, Cape Town, Ed. J.C. Cornell (1995), p.115.

[57] Y. Jongen *et al.*, *Nucl. Instrum. Methods Phys. Res A* **624**(1) (2010) 47. http://dx.doi.org/10.1016/j.nima.2010.09.028

[58] Y. Jongen *et al.*, IBA-JINR 400 MeV/u superconducting cyclotron for hadron therapy, Proc. of Cyclotrons, 2010, Lanzhou, China, p. 404.

[59] Y. Jongen and S. Zaremba, Cyclotron magnet calculations, CAS CERN 96/02 (1996), p. 139.

[60] S. Zaremba, Magnets for cyclotrons, CAS CERN 2006/012 (2006), 253.

[61] R. Baartman, Space charge limit in separated turn cyclotrons, 20th Int. Conf. on Cyclotrons and their Applications (2013) p. 305.

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Beam-Transport Systems for Particle Therapy

J.M. Schippers
Paul Scherrer Institut, Villigen, Switzerland

Abstract
The beam transport system between accelerator and patient treatment location in a particle therapy facility is described. After some general layout aspects the major beam handling tasks of this system are discussed. These are energy selection, an optimal transport of the particle beam to the beam delivery device and the gantry, a device that is able to rotate a beam delivery system around the patient, so that the tumour can be irradiated from almost any direction. Also the method of pencil beam scanning is described and how this is implemented within a gantry. Using this method the particle dose is spread over the tumour volume to the prescribed dose distribution.

Keywords
Beam transport; beam optics; degrader; beam analysis; gantry; pencil beam scanning.

1 Introduction

The main purpose of the beam-transport system is to aim the proton beam, with the correct diameter and intensity, at the tumour in the patient and to apply the correct dose distribution. The beam transport from the accelerator to the tumour in the patient consists of the following major sections (see Fig. 1):

– energy setting and energy selection (only for cyclotrons);

– transport system to the treatment room(s), including beam-emittance matching;

– per treatment room—a gantry or a fixed beam line aiming the beam from the correct direction;

– beam-delivery system in the treatment room, by which the dose distribution is actually being applied. These devices are combined in the so called ‘nozzle’ at the exit of the fixed beam line or of the gantry.

In the last few years, there have been many developments by commercial companies, to supply a facility with only one treatment room. Most of the contents of this chapter also apply to these facilities. The only exception is the version in which the beam from a cyclotron is sent to the patient, directly. In that case, the cyclotron is immediately followed by the nozzle components.

The details of the components in the nozzles are discussed in other chapters in these proceedings.

https://doi.org/10.23730/CYRSP-2017-001.241

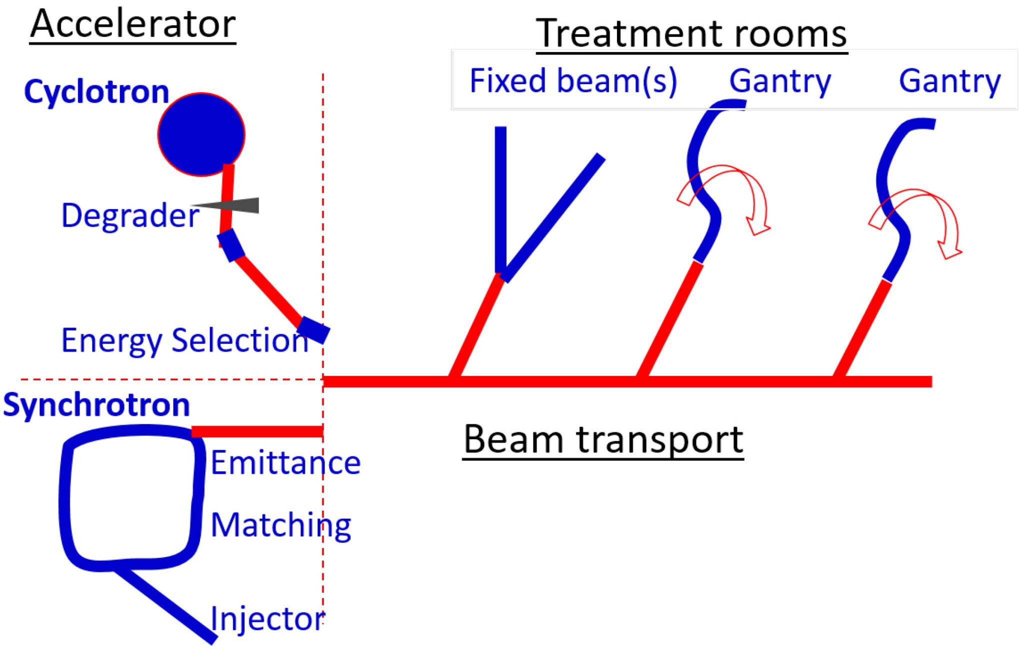


Fig. 1: A schematic overview of the accelerator and beam-transport system in a particle-therapy facility. Both currently used accelerators are indicated. Drawing is not to scale.

There are two major techniques for applying the dose to the patient. After aiming the beam in the desired direction by rotating the gantry (beam-transport system mounted on a rotating structure) to the correct angle, the beam must be spread in the lateral direction since the tumour diameter is larger than the beam cross section. Therefore, one must the match beam shape to the cross section of the tumour, as seen from the incoming beam direction. This is done either by the scattering technique or the scanning technique. In the scattering method, the beam cross section is increased by sending the beam through a system by which the beam diameter is increased by scattering in a foil system (a ‘passive’ technique) to match to the maximum lateral tumour dimension. In the scanning method, a ‘pencil beam’ with a diameter of approximately 1 cm is ‘actively’ scanned in the transverse plane over the tumour cross section. This motion is done in steps, and the applied ‘spot’ dose is varied per step (‘spot scanning’). In a method currently in development, scanning is performed by a continuous shift of the beam along lines in the tumour. During this sweep, the beam intensity is varied to deliver the correct dose along the line (‘continuous scanning’). Until now, the scattering technique has most commonly been used. However, for several years, the scanning technique has been regarded as the optimal technique (i.e. the technique that applies the best possible dose distribution) currently feasible in practice and almost all new facilities are designed to employ this technique. Therefore, in this chapter, the focus will be on the application of scanning techniques.

In this technique, the beam-transport system is controlling the following parameters:

- beam position by scan magnets;
- beam size by focusing;
- beam energy (in a synchrotron or in a degrader following a cyclotron);
- beam intensity by controlling the intensity from the accelerator, or by controlling the beam transmission in the beam-transport system.

2 Beam transport

2.1 Setting of the beam energy

As mentioned in the Introduction, the beam transport between the accelerator and treatment rooms has different functions. In most facilities, these functions are performed in different sections.

In the case of a cyclotron, first, the beam energy is set to the value needed at the treatment site. A choice is made concerning how quickly this system should set the energy. If the beam energy is only set to the value that corresponds to the maximum needed range in the patient (= the deepest part of the tumour as seen from the beam direction), this value does not need to change very quickly (fraction of a minute), since it can be done, for example, during gantry rotation. In that case, the range modulation over the thickness of the tumour is done in the nozzle by using range-shifter plates or a range-modulation wheel, in case scattering is used as the dose-application technique, see Fig. 2.

If a synchrotron is used as the accelerator, each spill is extracted with a specific energy. The waiting time until the next spill is of the order of several seconds, so this can be used, conveniently, to set the maximum energy and perform the modulation in the nozzle. Recent developments in synchrotrons also enable an energy change during a spill [1]. Although not yet implemented, this is a very promising development.

In the case of a cyclotron, the energy modulation can be done at the start of the beam-transport system. This is quite challenging. Apart from the degrader, all following magnets in the beam transport and gantry must be able to make rather fast changes. A high-speed energy change will quickly vary the Bragg-peak position, in depth, and this will limit the time of the total dose application. During modulation of the proton energy, a typical energy step corresponds to a range step of approximately 5 mm in water. This corresponding step of approximately 1% in momentum of the proton beam must be made in a fraction of a second. In that case, the mechanical design of the degrader must be such that the required speed can be reached, but also, the following magnets in the beam line and gantry must be able to change their field accordingly. This requires power supplies that can change the magnet current quickly and use a voltage overshoot to compensate the induction of the magnet.

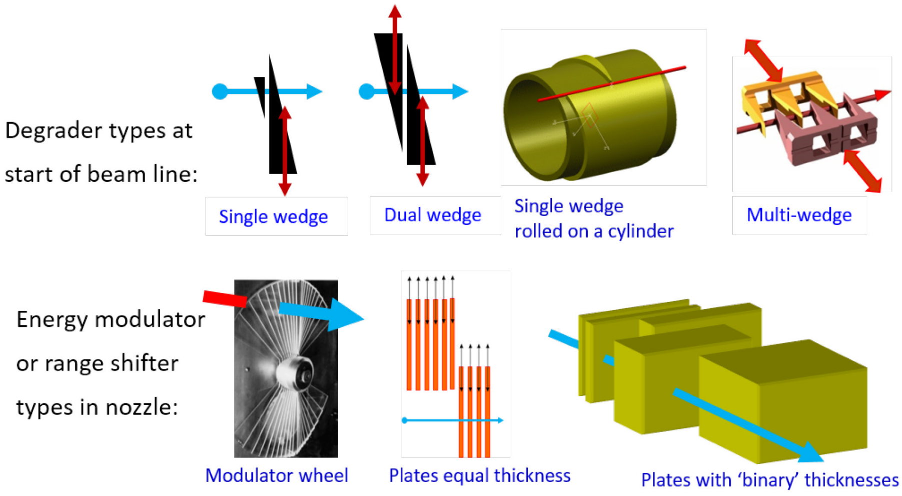

Degrader types at start of beam line:
Single wedge
Dual wedge
Single wedge rolled on a cylinder
Multi-wedge
Energy modulator or range shifter types in nozzle:
Modulator wheel
Plates equal thickness
Plates with ‘binary’ thicknesses

Fig. 2: Top: Several possible degrader systems consisting of wedges that vary the thickness of the material to be crossed by the beam. Bottom: a wheel that modulates the beam energy when it rotates, and two systems of plates that can quickly be put into/out of the beam line for fast energy (range) adjustments. The plates could also be installed as degraders at the beginning of the beam line.

The degrader usually consists of a system of Lucite or graphite wedges, and different mechanical possibilities are shown in Fig. 2. The amount that the wedge(s) has been shifted into the beam line determines the amount of material that is traversed by the beam, and thus, the energy of the ongoing beam. Different mechanical variations of the wedge are in use. The wedge can be rolled along the surface of a cylinder, or a system of multiple wedges can be used. In all cases, the position of the edge(s) must be set quickly and accurately, and at least two wedges are needed to obtain a uniform thickness over the beam diameter.

For a given cyclotron-beam energy, the minimum energy of the particles leaving the degrader depends on the total thickness of the traversed material. Usually, one degrades until a beam energy of approximately 70 MeV is reached. At lower energies, the beam transport could be distorted (become un-sharp) too much by multiple scattering in, for example, the vacuum window at the exit of the beam line and dosimetry devices just before the patient. The maximum energy of the continuous energy variation range is, usually, approximately 20 MeV below the energy from the cyclotron. This is due to the minimum overlap of the wedges, which needs to be the same thickness over the beam cross section.

There will be a spread in beam energy of the particles that leave the degrader. This energy straggling is caused by the statistical variation of the particle tracks in the degrader. This spread increases with the energy loss in the degrader and can be larger than the energy acceptance of the beam-transport system. Therefore, the degrader is followed by an energy selection system (ESS), consisting of a dipole followed by a slit system. In the ESS, a maximum relative (%) beam-momentum spread is selected.

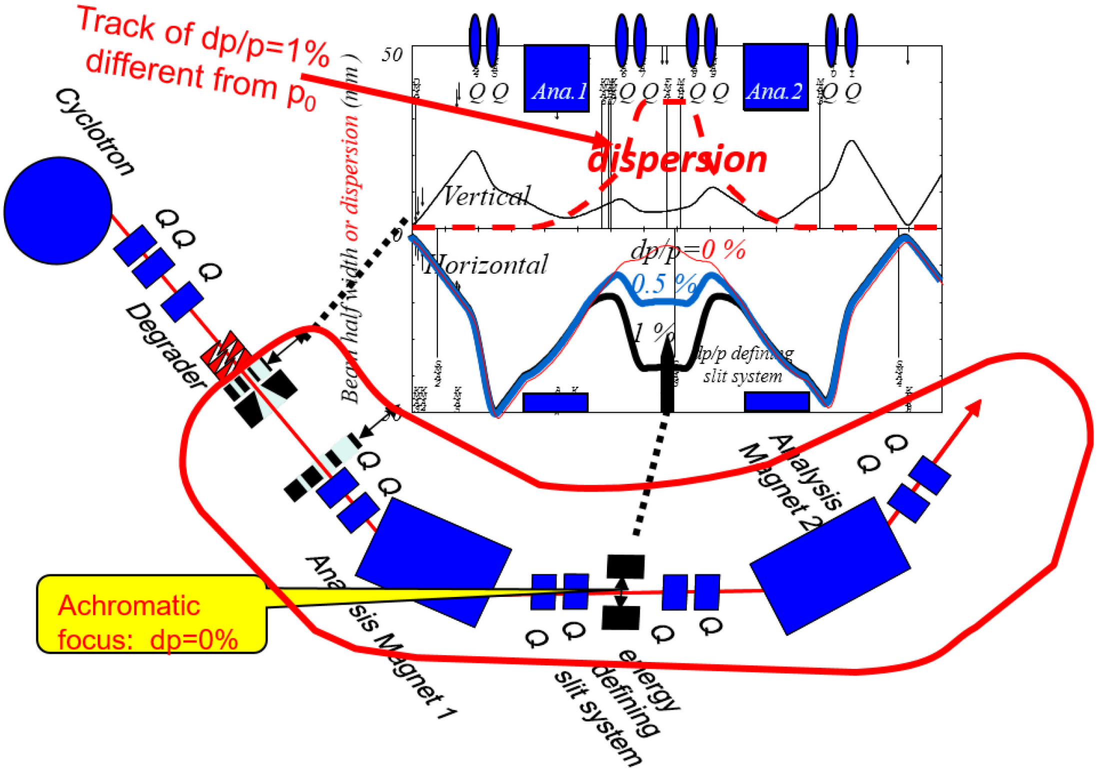


Fig. 3: The beam optics in the ESS. In the top graph, the half beam size is plotted as a function of the position along the beam line. Above the *x*-axis, the vertical dimension is plotted as a solid line and the dispersion as a dashed line. Below the *x*-axis, the horizontal beam size is plotted for 0%, 0.5% and 1% momentum spread. The black arrows indicate the momentum acceptance slit aperture.

As shown in Fig. 3, the beam optics in the ESS are usually made such that there is a large dispersion at the slit position (‘dispersion’ is the track of a particle with 1% momentum deviation). But, to get a well resolved energy-position dependence at one’s position, one also needs a monochromatic focus at one’s position. This is the case when there would be a small beam width if the beam were without energy spread (the 0% dp/p line in Fig. 3). This combination of dispersion and monochromatic focusing results in a beam profile in which there is a unique relation between horizontal position and energy. Then, the total width is mostly determined by the energy spread in the beam, as shown in Fig. 4.

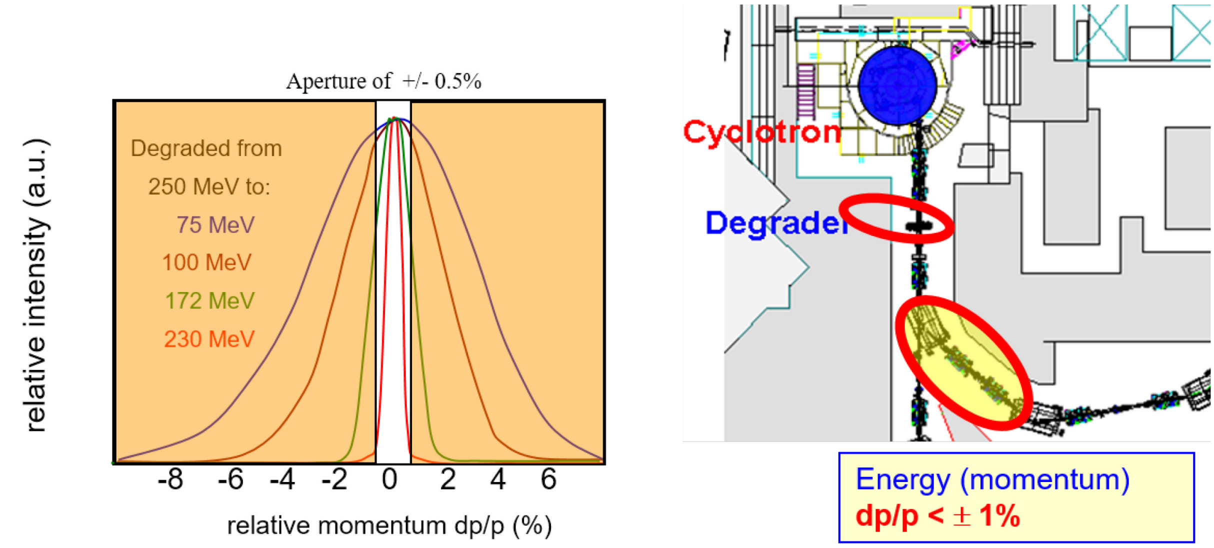


Fig. 4: Left: the momentum spread, which almost equals the beam spread in the horizontal plane, at the momentum-selecting slit. The aperture used, of e.g. 0.5%, has been indicated. Right: the location of the degrader and the ESS in the PSI proton-therapy facility.

Due to the momentum selection at the slit, the beam energy behind the ESS is mainly determined by the magnetic field of the bending magnet. The wrong magnetic field will send the incorrect energy through the slit system, yielding the wrong range within the patient. The relation is

$$\frac{\mathrm{d}R}{R} = 3.2\frac{\mathrm{d}B}{B}, \tag{1}$$

so that an error of 1% in the magnetic field B would lead to a range error of 6.4 mm, at a (water equivalent) depth of $R = 20$ cm in the patient. Therefore, the field setting must be reproducibly correct within 10^{-4}. The absolute degrader setting is less critical ($<10^{-2}$), due to the width of the energy straggling. A small setting error will only lead to a shift of the profile at the slit (see Fig. 4), yielding a slightly reduced transmission through the ESS.

Using the aperture in the slit system, one thus selects a certain momentum spread, typically, approximately, 0.5–1%. This is also the momentum spread at the patient. One should realize that this momentum spread is smaller than in the case of energy degrading (to modulate the range) being performed in the nozzle, since, in that case, no energy selection is done after crossing the range-shifter plates or the modulation wheel. The smaller momentum spread after an ESS will give sharper Bragg peaks in the patient. As shown in Fig. 5, especially for low energies, this has consequences for the sharpness of the Bragg peak, which must be taken into account in treatment planning.

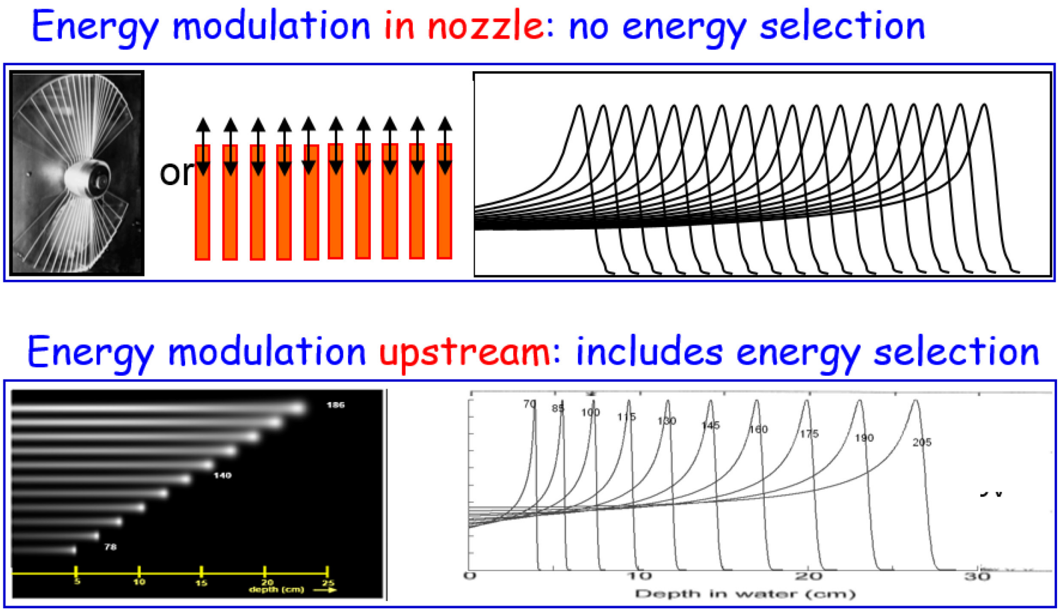


Fig. 5: The shape (sharpness) of the Bragg peak in the patient for different energy-reducing methods. Top: methods used in the nozzle, where no energy selection is done. Bottom: a degrader, followed by an ESS, will give sharp Bragg peaks at low energies.

The degrader not only decreases the beam energy; it also increases the emittance due to multiple scattering. In order to minimize this effect, but while keeping sufficient stopping power, a material of low atomic number *Z* is used. However, the emittance increase is considerable, and the degrader must be followed by a set of apertures that limit the emittance to the acceptance of the following beam line. Since the emittance also increases with decreasing energy out of the degrader, the fraction of the beam that is accepted in the beam transport, behind these collimators and the ESS, is also decreasing, the lower the energy. Figure 6 [2], shows the order of magnitude of this transmission loss at the 250 MeV beam from the cyclotron at PSI.

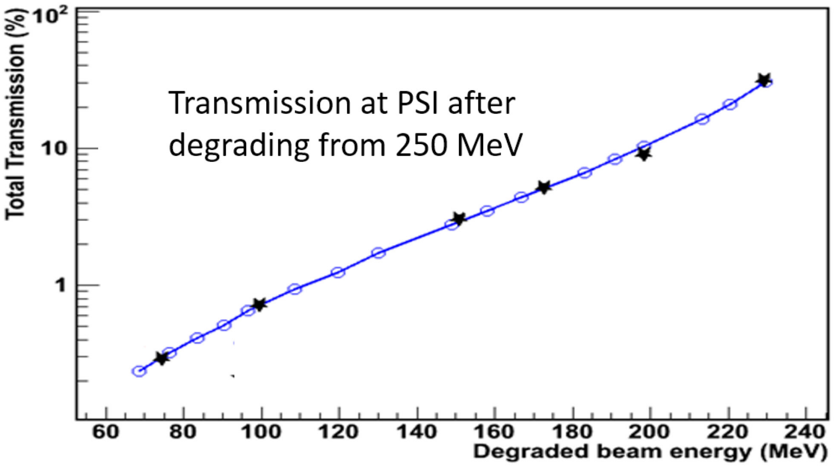


Fig. 6: The transmission of the proton beam in the ESS and beam-transport system, as a function of the energy behind the degrader, for a beam of 250 MeV from the cyclotron [2].

The beam intensity at the patient would be strongly dependent on beam energy, due to this energy-dependent transmission. This would complicate the treatment and could have consequences for safety, due a limitation of reaction times. Therefore, an intensity compensation is necessary to obtain a beam intensity that is not so dependent on energy. This can be done by adjusting the intensity in the cyclotron, but one could also design a beam-line setting with energy-dependent controlled beam losses at dedicated collimators in the beam-transport system.

Of course, all beam losses will create activation. Therefore, it is important to concentrate the beam losses at well-known locations. These can be shielded if necessary, and also, one can select materials that will not contain long-lived radioactive isotopes. In a cyclotron facility, the degrader, acceptance defining collimators, and the ESS are usually the only such locations in the beam-transport system. The losses due to intensity compensation are relatively low.

2.2 Optics of the beam transport

In a synchrotron, the beam energy is set in the ring by acceleration until the desired energy has been reached. A change to other energies thus requires a new spill, which takes a few seconds. Therefore, in synchrotron facilities, energy modulation is usually done in the nozzle. Therefore, scattering is still the method most commonly used in synchrotron facilities. However, as discussed in the chapter on synchrotrons (these proceedings), there have been very interesting recent developments, aimed at beam-energy reduction during a spill.

Another difference with cyclotrons is the extracted emittance. The emittance of the beam from a cyclotron is slightly asymmetric (horizontally versus vertically), but this asymmetry is overruled by the emittance increase in the degrader, which is symmetric. The horizontal emittance of a synchrotron beam can be a factor of ten smaller than in the vertical direction. This would give a gantry-angle dependence in the dose-application process. Therefore, emittance matching is necessary if a gantry is used. This can be done by a system of rotating quadrupoles that rotate the emittance with the gantry (see chapter on this topic), or by passing the beam through a scatter foil to make a symmetric emittance.

The further optics of the beam transport should have a layout that is as stable as possible, thus providing a high reproducibility of the beam-transport characteristics. Since the emittance of a degraded beam from a cyclotron is much larger than that of a synchrotron, the aperture of most magnets is usually larger in the case of a cyclotron. However, apart from the degrader with an ESS and the emittance matching, the beam optics are rather similar for both types of machines, as already indicated in Fig. 1. A beam transport with several intermediate images is very stable and reproducible and easy to verify by profile monitors at these locations. Collimators, at one or more of these locations, also convert beam alignment errors into an intensity reduction, which is usually less of a problem.

One should be able to rely on the reproducibility of the beam-line setting. Therefore, the temperature of the vaults and the cooling water should not fluctuate. Also, one should always use the same procedures for, e.g., field ramping and beam-line settings for other energies. During or in-between patient treatments there is no possibility to tune (fiddle with) the beam or to try a new setting.

Beam-diagnostic tools, such as profile monitors, should only be used when developing new beam-line settings or in quality assurance (QA) programs. They should be removed from the beam trajectory and this position should be continuously verified. A non-interceptive beam-diagnostic tool is an advantage and can be used during treatments.

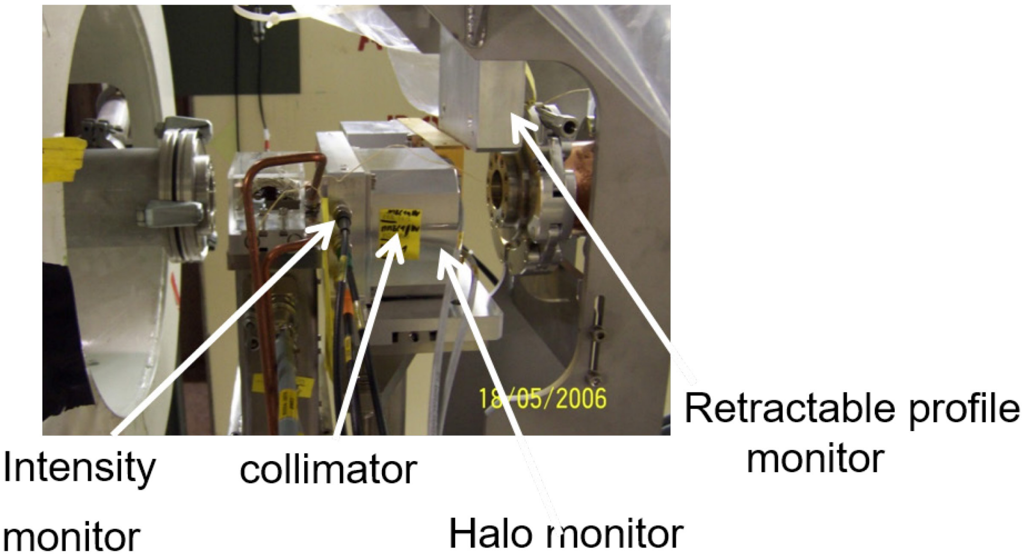


Fig. 7: The gantry coupling point at PSI. The beam direction is from right to left. In an air gap of 20 cm, a collimator and beam-diagnostic tools have been mounted. The non-interceptive intensity and halo monitors continuously provide the control system information, including during patient treatment.

For example, at PSI such beam-diagnostic tools have been mounted at the coupling points in all treatment rooms and are always active [3]. Here, the vacuum of the beam line is separated from the vacuum of the gantry by an air gap of approximately 20 cm, as shown in Fig. 7. A retractable profile monitor must be retracted during treatments. However, the ionization of the air in the gap behind this monitor is used to measure the beam intensity as well as the presence and orientation of the beam halo. There are no foils in the beam. Electric fields are guiding the charge to foils next to the beam path.

3 Gantry

In ion-therapy facilities, normally, use is made of different fixed beams, using beam directions in the horizontal direction, in the vertical direction, and/or from an inclined direction. In combination with a slightly flexible patient-positioning system, beam directions around the fixed directions can be used. But, in order to direct the particle beam from all different directions to the tumour, other methods are in use. A gantry offers most flexibility [4].

When using a gantry, the beam-transport system ends at a coupling point to the gantry. Downstream of this point, the beam-transport system is mounted on a mechanical structure, which can rotate 360 degrees, or a bit more than 180 degrees, around the patient. At the coupling point, the beam should be symmetric, both in shape and in divergence, so that there is no dose-application dependence on the gantry angle. In a gantry, the beam must be bent several times to get the correct beam, which is perpendicular to the incoming beam direction. The magnetic rigidity of the protons implies the use of strong (>1.5 T) and large magnets. A distance of at least 2 m is needed in order to provide the necessary distance for a scattering system and the length to spread the scattered beam. For pencil-beam scanning, a similar amount of space is needed for the scanning magnets and the length to get sufficient lateral displacement of the scanned beam. Also, space is needed for dose monitors, range-shifter and scanning magnets, or energy modulator and collimation systems between the exit of the last magnet and the patient. Therefore, the typical diameter of these 100 ton gantries is 10–11 m, both for scattered beams and for scanned beams (see, for example, Fig. 8).

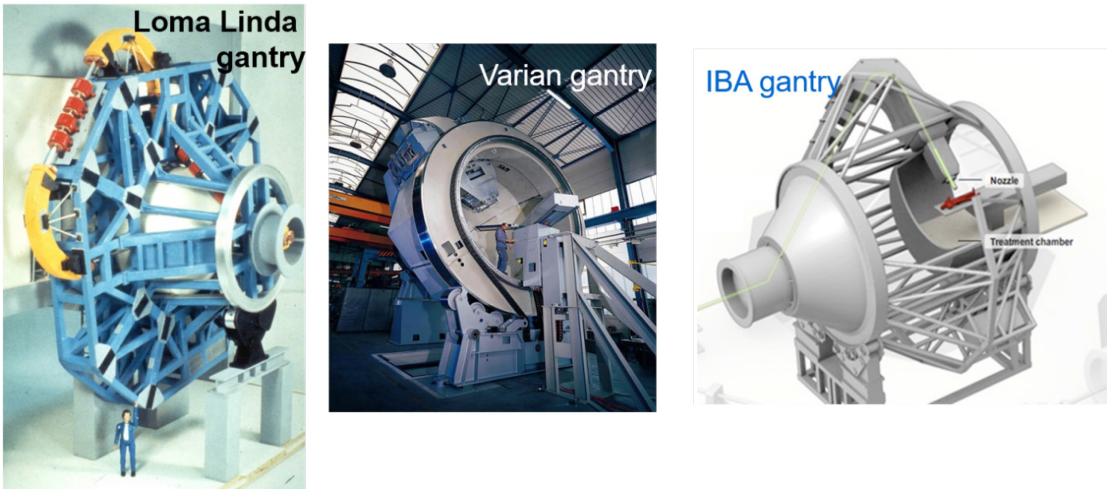


Fig. 8: Left—the ‘corkscrew’ gantry at Loma Linda; middle—the Varian gantry designed for scanning beams during its construction phase; and right—the IBA gantry.

Due to the almost three times larger magnetic rigidity of carbon ions (approximately 6.8 Tm), the radius of curvature of particle trajectories is more than 3 m with conventional magnets. The huge difficulties and costs associated with a carbon gantry have resulted in only one existing gantry for carbon ions (13 m diameter, 570 tons), which has been installed at the HIT facility in Heidelberg, see Fig. 9 [5]. However, a gantry with superconducting magnets has been designed and built for the extension of the Japanese ion-therapy facility HIMAC [6].

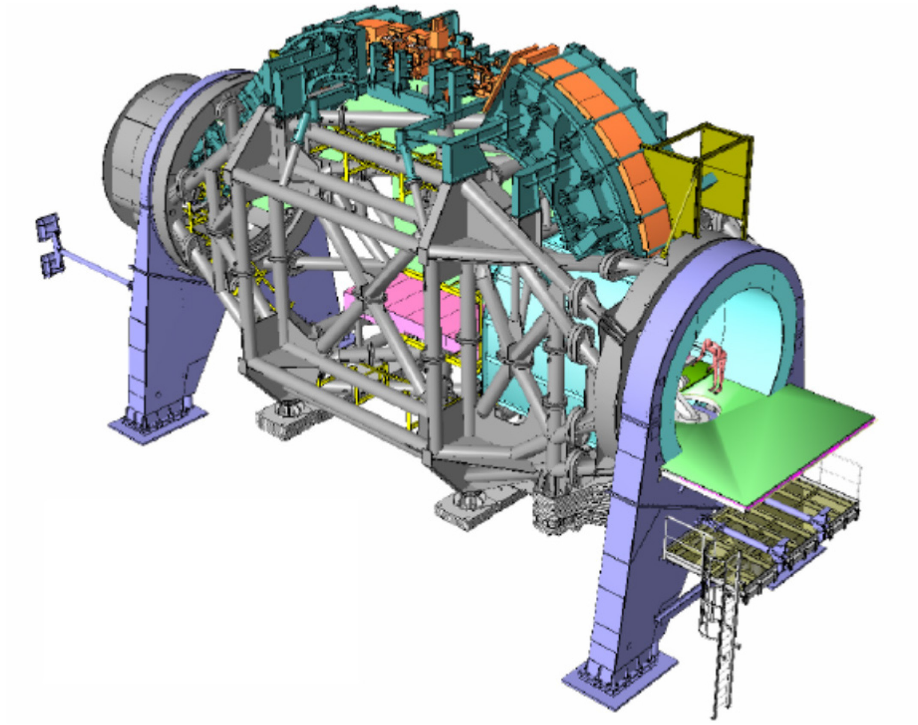


Fig. 9: The gantry at HIT in Heidelberg for carbon therapy

The first three gantries in particle therapy originated from a design by the Harvard group [7] and were installed at the facility in Loma Linda. The beam-transport-system layout is referred to as a *corkscrew* and consists of a double-achromatic-bending system, resulting in a rather short gantry design, as shown in Fig. 8. The nozzle has been designed for a scattered beam. But, with the increasing demand for pencil-beam scanning, dedicated nozzle designs have been made to allow (also or only) scanning. In the case of scanning, the scanning magnets have usually been mounted at the exit of the last bending magnet (‘downstream scanning’). Their location acts as the virtual source of the pencil beams.

At PSI, a compact proton gantry (‘Gantry1’) has been in use since 1996 [8], which is optimized for pencil-beam scanning. In this gantry, a scanning magnet has been mounted before (‘upstream scanning’) the last bending magnet (90 degrees), which bends the beam towards the isocentre. The beam optics has been designed such that a deflection of the pencil beam by the scanning magnet causes a parallel shift of the beam at the isocentre. In this way, the virtual source of the pencil beams is located at infinity and an orthogonal pencil-beam arrangement is obtained. The other two orthogonal displacements are performed by inserting range-shifter plates in the nozzle to shift the Bragg peak in depth and by shifting the table in the direction orthogonal to the magnetic scanning direction. The maximum speed of the dose application is limited by the necessary slow motion of the patient table.

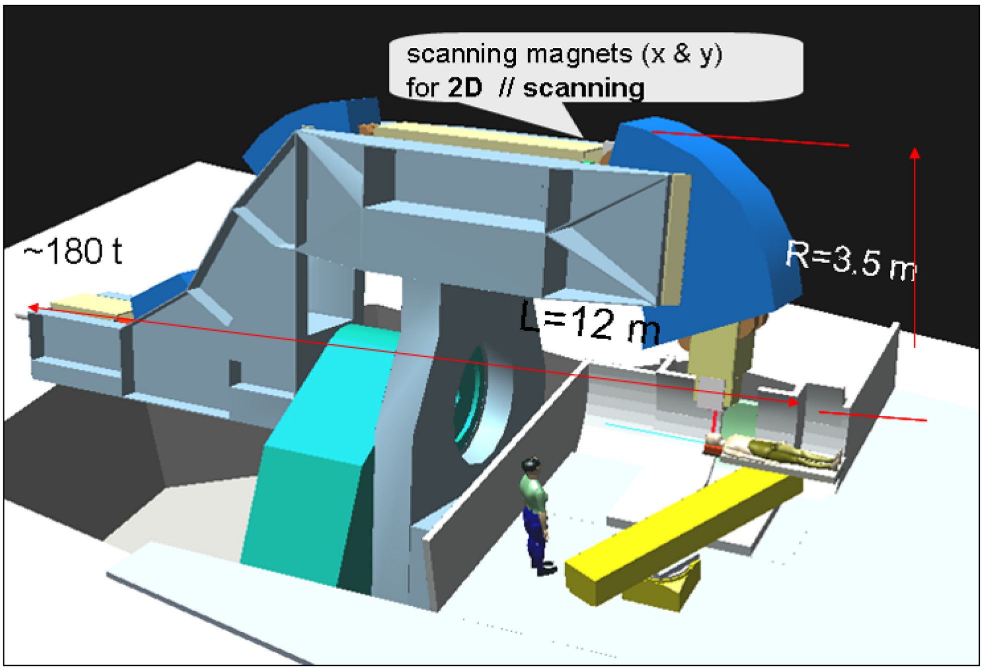


Fig. 10: Design of Gantry 2 at PSI. This gantry is designed for parallel beam scanning in two dimensions and fast energy scanning.

In order to have fast (i.e. magnetic) scanning in two directions, ‘Gantry2’ has been designed and built at PSI, see Fig. 10 [9]. This gantry allows for double magnetic scanning in a field of 12×20 cm^2, with parallel beam displacements in both directions. To allow scanning over 12 cm in the direction orthogonal to the bending plane of the last (also 90 degrees) bending magnet, a relatively large gap of this bending magnet is needed. Since this gap limits the field size in this direction, a table shift of <12 cm is necessary once or twice at a certain gantry angle. The magnets of Gantry 2 are laminated, so that their magnetic fields can change with the energy of the incoming beam. With this combination of laminated magnets in the gantry and in the preceding beam line following a degrader, a range change of ~5 mm in water is achieved in ~80 ms.

A gantry design employing a beam-focusing concept normally used in FFAG (Fixed Field Alternating Gradient) accelerators, allows a very large momentum acceptance of ±15% [10]. The design of the magnet system shows very-tightly packed focusing and defocusing magnets, with gradients up to 70 T/m to be achieved with superconducting magnets. This has not been built yet, however, but it has triggered several groups to work, also, on various other designs of gantries with superconducting magnets [11].

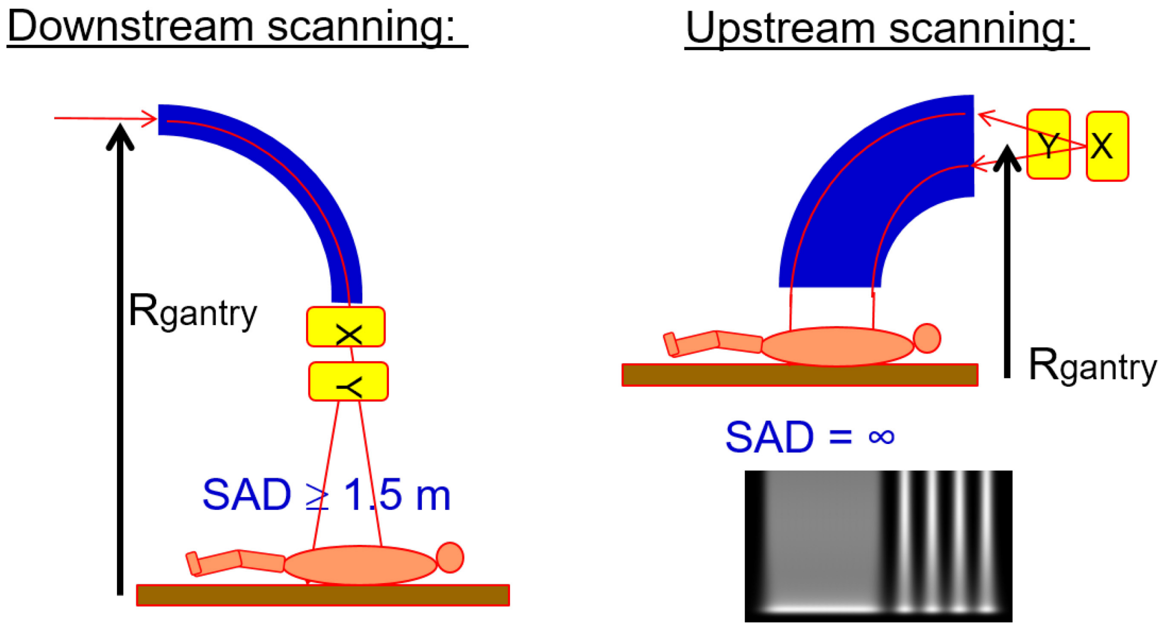


Fig 11: Schematic overview of the two most used scanning magnet configurations: before ('upstream') or behind ('downstream') the last bending magnet. SAD= source–axis distance.

So, when implementing the scanning technique, one can make a choice between upstream scanning, downstream scanning (see Fig. 11), or a combination of the two [12]. The advantage of downstream scanning is the simpler layout of the large final bending magnet. However, due to the needed space behind the last bending magnet, the gantry radius will not be small.

In the case of upstream scanning, a large aperture and more complicated optics are needed to obtain appropriate 2D pencil-beam motions in combination with sufficient focusing of the pencil beams. The gantry radius can be made rather small, since only approximately 1 m of space is needed for dosimetry equipment, etc. Upstream scanning also offers the possibility to get a parallel displacement of the pencil beam, which has several advantages. This is done via a so-called point-to-parallel imaging, by means of the inclined angles of the entrance and exit of the last bending magnet.

4 Conclusions

The beam transport between accelerator and gantry has different purposes. The most important is to aim the proton beam with the correct diameter and intensity at the tumour in the patient and to apply the correct dose distribution. This task is achieved by performing several actions, usually performed in clearly distinguishable beam-line sections. For example, these could be a section for setting the energy (in the case of a cyclotron), a section to perform the transport of the beam to the treatment room(s), and a final section that aims the beam at the tumour from the correct direction.

An accelerator facility for particle therapy implements a variety of technical measures to ensure an accurate and reproducible dose delivery to patients. In comparison to an accelerator for physics research, a medical irradiation facility differs mostly in relation to reliability and simplicity of the equipment and operational procedures. Certain measurements need to be performed frequently, so that absolute dose values are delivered to the patients with an accuracy of a few percent. The necessary safety measures related to the irradiation treatment impose a stringent discipline, and procedure-following regulations, on the operation and implementation of changes or upgrades.

Acknowledgement

I would like to thank the numerous colleagues at other particle-therapy institutes and at several companies for the countless discussions at conferences and workshops and the detailed information given at site visits.

References

[1] Y. Iwata *et al.*, Multiple-energy operation with quasi-dc extension of flattops at HIMAC, MOPEA008, Proc. IPAC'10, Kyoto, Japan, 2010, p. 79.
http://epaper.kek.jp/IPAC10/index.htm

[2] M.J. van Goethem *et al.*, *Phys. Med. Biol.* **54** (2009) 5831.
http://dx.doi.org/10.1088/0031-9155/54/19/011

[3] R. Dölling et al., Beam diagnostics for the proton therapy facility proscan, Proc. AccApp'07, 2007, p. 152.

[4] J.B. Flanz, Proc. PAC95, Dallas, TX, USA, May 1–5, 1995, p. 2004.
http://accelconf.web.cern.ch/AccelConf/p95/

[5] Th. Haberer *et al.*, *Radiother. Oncol.* **73** S186 (2004).
http://dx.doi.org/10.1016/S0167-8140(04)80046-X

[6] K. Noda *et al.*, New heavy-ion cancer treatment facility at HIMAC, TUPP125. Proc. EPAC08, 2008, p. 1818.

[7] A.M. Koehler, Proc. 5th PTCOG Meeting: Int. Workshop on Biomedical Accelerators, Lawrence Berkeley Laboratory, Berkeley, CA, 1987, p. 147.

[8] E. Pedroni *et al.*, *Med. Phys.* **22(1)** (1995) 37. http://dx.doi.org/10.1118/1.597522

[9] E. Pedroni *et al.*, *Z. Med. Phys.* **14(1)** (2004) 25. http://dx.doi.org/10.1078/0939-3889-00194

[10] D. Trbojevic *et al.*, Proc. PAC07, Albuquerque, New Mexico, USA, June 25-29, 2007, p. 3199. http://epaper.kek.jp/p07/INDEX.HTML

[11] W. Wan *et al.*, Alternating gradient canted cosine theta superconducting magnets for future compact proton gantries, Phys. Rev. ST—Acc. and Beams 18, 103501 (2015)

[12] H. Vrenken *et al., Nucl. Instr. Meth.* **A 426(2–3)** (1999) 618. http://dx.doi.org/10.1016/S0168-9002(99)00039-X

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Beam Dynamics in Synchrotrons

B.J. Holzer
CERN, Geneva, Switzerland

Abstract

This paper gives an overview of particle dynamics in synchrotrons and storage rings. Both the transverse and the longitudinal plane are described in a linear approximation. The main emphasis is on giving an introduction to the basic concepts and allowing the reader to deduce the main parameters of a machine, based on some simple scaling laws.

Keywords

Accelerator physics; synchrotron; storage ring; transverse dynamics; longitudinal dynamics.

1 Introduction

We would like to start this little overview with some kind of definition of a synchrotron, in an attempt to achieve the impossible task of summarizing in a few lines the key issues associated with such a machine. And we would like to ask you, the esteemed reader, to come back to this point at the end of the story and let us know whether or not the definition is a valuable one.

So, a synchrotron is a type of circular accelerator that needs:

- a magnetic bending field to keep the particles on a closed, more or less circular orbit;
- a mechanism to lock this B-field to the changing particle energy and thus keep the particles on, or close to, this design orbit over the complete energy range of the machine;
- focusing forces that follow the energy gain of the beam to keep the particles together, and that ultimately lead to a well-defined beam size;
- a Radio Frequency (RF) structure to accelerate the particles and create the necessary energy gain per turn via longitudinal electric fields;
- a mechanism to synchronize the RF frequency to the timing of the circulating particles and to provide a longitudinal focusing (phase-focusing) effect that keeps the particles longitudinally bunched.

So much for the definition.

This seems to deserve a remark to reassure the reader: these machines exist, they are very robust, they deliver stable particle beams, and, most importantly, they can be built.

Two examples will act as proof of this strong statement: the ADA (Annelli de Accumulatione) (Fig. 1), as far as we know, the very first particle collider and certainly one of the smallest synchrotrons, built in Frascati by Bruno Touschek in 1944; and the LHC [1], at present the largest storage ring ever built, running at the highest achievable particle energies at CERN (Figs. 2 and 3).

2 Introduction to transverse beam dynamics

The transverse beam dynamics of charged particles in an accelerator describes the movement of single particles under the influence of the external transverse bending and focusing fields. It includes the detailed arrangement (for example, their positions in the machine and their strength) of the accelerator magnets used to obtain well-defined, predictable parameters of the stored particle beam, and it describes

https://doi.org/10.23730/CYRSP-2017-001.253

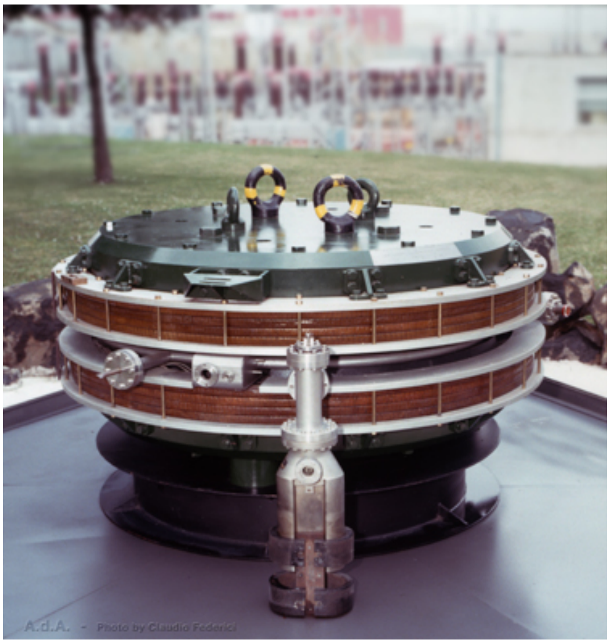


Fig. 1: ADA, the first electron positron collider ring

Fig. 2: A view of the tunnel of the LHC proton–proton collider at CERN, Geneva

methods to optimize the trajectories of single particles, as well as the dimensions of the beam considered as an ensemble of many particles. A detailed treatment of this field in full mathematical detail, including sophisticated lattice optimizations such as the right choice of the basic lattice cells and the design of dispersion suppressors or chromaticity compensation schemes, is beyond of the scope of this basic overview. For further reading and for more detailed descriptions, we therefore refer to the more detailed explanations in [1–4].

2.1 Geometry of the ring

In general, magnetic fields are used in circular accelerators to provide the bending force and to focus the particle beam. In principle, the use of electrostatic fields would be possible as well, but at high momenta (i.e., if the particle velocity is close to the speed of light), magnetic fields are much more efficient. The force acting on the particles, the Lorentz force, is given by

$$\mathbf{F} = q \cdot (\mathbf{E} + \mathbf{v} \times \mathbf{B}). \tag{1}$$

Fig. 3: The LHC proton–proton collider

For high-energy particle beams, the velocity v is close to the speed of light and so represents a nice amplification factor whenever we apply a magnetic field. As a consequence, it is much more convenient to use magnetic fields for bending and focusing the particles.

Therefore, neglecting electric fields for the moment, we write the Lorentz force and the centrifugal force on the particle on its circular path as

$$F_{\text{Lorentz}} = e \cdot v \cdot B, \tag{2}$$

$$F_{\text{centrifugal}} = \frac{\gamma m_0 v^2}{\rho}. \tag{3}$$

Assuming an idealized homogeneous dipole magnet along the particle orbit, having pure vertical field lines, we define the condition for a perfect circular orbit as equality between these two forces. This yields the following condition for the idealized ring:

$$\frac{p}{e} = B \cdot \rho, \tag{4}$$

where we are referring to protons and have accordingly set $q = e$. This condition relates the so-called beam rigidity $B\rho$ to the momentum of a particle that can be carried in the storage ring, and it ultimately defines, for a given magnetic field of the dipole magnets, the size of the storage ring.

In reality, instead of having a continuous dipole field the storage ring will be built with several dipole magnets, powered in series to define the geometry of the ring. For a single magnet, the trajectory of a particle is shown schematically in Fig. 4. In the free space outside the dipole magnet, the particle trajectory follows a straight line. As soon as the particle enters the magnet, it is bent onto a circular path until it leaves the magnet at the other side.

The overall effect of the main bending (or 'dipole') magnets in the ring is to define a more or less circular path, which we call the 'design orbit'. By definition, this design orbit has to be a closed loop, and so the main dipole magnets in the ring have to define a full bending angle of exactly 2π. If α denotes the bending angle of a single magnet, then

$$\alpha = \frac{\mathrm{d}s}{\rho} = \frac{B\,\mathrm{d}s}{B \cdot \rho}. \tag{5}$$

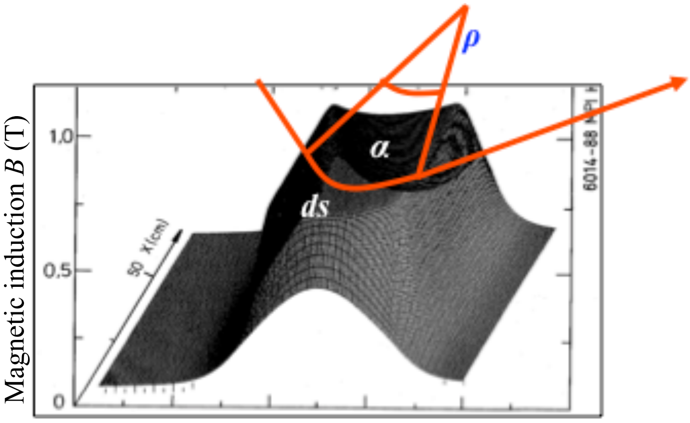


Fig. 4: Field map of a storage ring dipole magnet, and schematic path of a particle

We therefore require that

$$\frac{\int B\,\mathrm{d}s}{B \cdot \rho} = 2\pi. \tag{6}$$

Thus, a storage ring or synchrotron is not a 'ring' in the true sense of the word but more a polygon, where 'poly' means the discrete number of dipole magnets installed in the 'ring'.

In the case of the LHC, the dipole field has been pushed to the highest achievable values: 1232 superconducting dipole magnets, each 15 m long, define the geometry of the ring and, via Eq. (6), the maximum momentum for the stored proton beam. Using the equation given above, for a maximum momentum $p = 7$ TeV/c, we obtain a required magnetic field of

$$B = \frac{2\pi \cdot 7000 \cdot 10^9\ \mathrm{eV}}{1232 \cdot 15\ \mathrm{m} \cdot 2.99792 \cdot 10^8\ \mathrm{m\,s^{-1}}}, \tag{7}$$

or

$$B = 8.33\ \mathrm{T}, \tag{8}$$

to bend the beams. For convenience, we have expressed the particle momentum in units of GeV/c here. Figure 5 shows a photograph of one of the LHC dipole magnets, built with superconducting NbTi filaments, which are operated at a temperature $T = 1.9$ K.

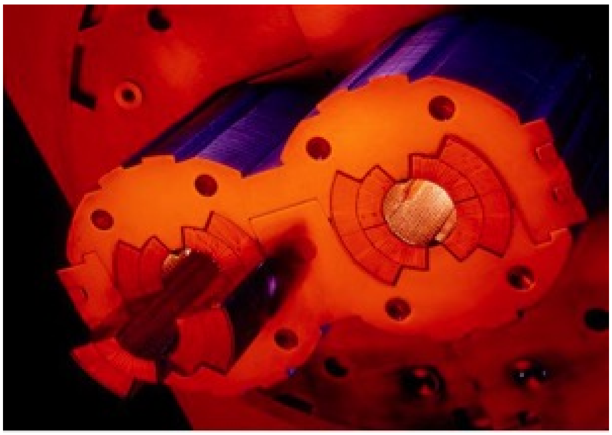


Fig. 5: Superconducting dipole magnet in the LHC storage ring

2.2 Focusing properties

In addition to the main bending magnets that guide the beam onto a closed orbit, focusing fields are needed to keep the particles close together. In modern storage rings and light sources, we have to keep more than 10^{12} particles in the machine, distributed over a number of bunches, and these particles have to be focused to keep their trajectories close to the design orbit. Furthermore, these particles are stored in the machine for many hours, and a carefully designed focusing structure is needed to maintain the necessary beam size at different locations in the ring and guarantee stability of the transverse motion.

Following classical mechanics, linear restoring forces are used, just as in the case of a harmonic pendulum. Quadrupole magnets provide the corresponding field property: they create a magnetic field that depends linearly on the amplitude of the particle, i.e., the distance of the particle from the design orbit:

$$B_x = g \cdot y, \qquad B_y = g \cdot x. \tag{9}$$

The constant g is called the gradient of the magnetic field and characterizes the focusing strength of the quadrupole lens in both transverse planes. For convenience, it is normalized (like the dipole field) to the particle momentum. This normalized gradient is denoted by k and defined as

$$k = \frac{g}{p/e} = \frac{g}{B\rho}. \tag{10}$$

The technical layout of such a quadrupole is depicted in Fig. 6. As in the case of the dipoles, the LHC quadrupole magnets were built using superconducting technology.

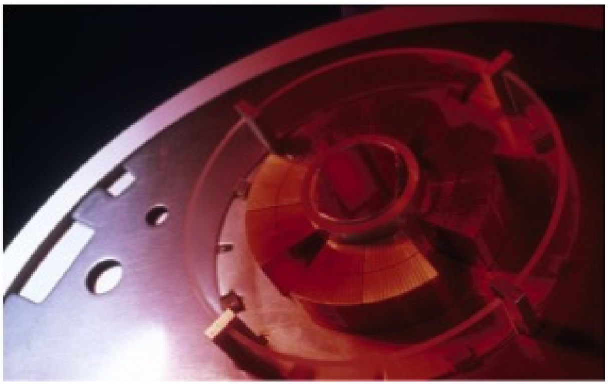


Fig. 6: Superconducting quadrupole magnet in the LHC storage ring

Now that we have defined the two basic building blocks of a storage ring, we need to arrange them in a so-called magnet lattice and optimize the field strengths in such a way as to obtain the required beam parameters. An example of how such a magnet lattice looks like is given in Fig. 7. This photograph shows the dipole (orange) and quadrupole (red) magnets in the TSR storage ring in Heidelberg [5]. Eight dipoles are used to bend the beam into a 'circle', and the quadrupole lenses between them provide the focusing to keep the particles within the aperture limits of the vacuum chamber.

A general design principle of modern synchrotrons and storage rings should be pointed out here. In general, these machines are built following a so-called separate-function scheme: every magnet is designed and optimized for a certain task, such as bending, focusing, or chromatic correction. We separate the magnets in the design according to the job they are supposed to do; only in rare cases a combined-function scheme is chosen nowadays, where different magnet properties are combined in one piece of hardware. To express this principle mathematically, we use the general Taylor expansion of the normalized magnetic field,

$$\frac{B(x)}{p/e} = \frac{1}{\rho} + k \cdot x + \frac{1}{2!}mx^2 + \frac{1}{3!}nx^3 + \cdots . \tag{11}$$

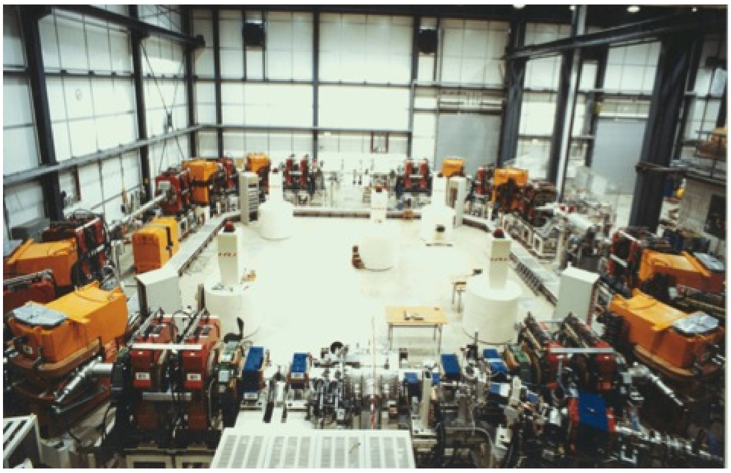


Fig. 7: The TSR storage ring, Heidelberg, as a typical example of a separate-function strong-focusing storage ring [5].

Following the arguments above, for the moment we take into account only constant (dipole) or linear (quadrupole) terms. The higher-order contributions to the field will be treated later as (hopefully) small perturbations.

Under these assumptions, we can derive — in a linear approximation — the equation of motion of the transverse particle movement. To derive the equation of motion, we start with a general expression for the radial acceleration as known from classical mechanics (see, e.g., [6]):

$$a_r = \frac{\mathrm{d}^2\rho}{\mathrm{d}t^2} - \rho\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)^2 . \tag{12}$$

The first term refers to an explicit change in the bending radius, and the second to the centrifugal acceleration. Referring to our coordinate system, and replacing the ideal radius ρ by $\rho + x$ for the general case (Fig. 8), we obtain for the balance between the radial force and the counter-acting Lorentz force the relation

$$F = m\frac{\mathrm{d}^2}{\mathrm{d}t^2}(x+\rho) - \frac{mv^2}{x+\rho} = evB. \tag{13}$$

On the right-hand side of the equation, we take only linear terms of the magnetic field into account,

$$B_y = B_0 + x\frac{\mathrm{d}B_y}{\mathrm{d}x}, \tag{14}$$

and for convenience we replace the independent variable t by the coordinate s,

$$x' = \frac{\mathrm{d}x}{\mathrm{d}s} = \frac{\mathrm{d}x}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}s}, \tag{15}$$

to obtain an expression for the particle trajectories under the influence of the focusing properties of the quadrupole and dipole fields in the ring, described by a differential equation. This equation is derived in its full beauty elsewhere [6], so we shall just state it here:

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

where k is the normalized gradient introduced above and the $1/\rho^2$ term represents the so-called weak focusing, which is a property of the bending magnets.

The particles will now follow the ‘circular’ path defined by the dipole fields, and in addition will undergo harmonic oscillations in both transverse planes. The situation is shown schematically in Fig. 8. An ideal particle will follow the design orbit represented by the circle in the diagram. Any other particle will perform transverse oscillations under the influence of the external focusing fields, and the amplitude of these oscillations will ultimately define the beam size.

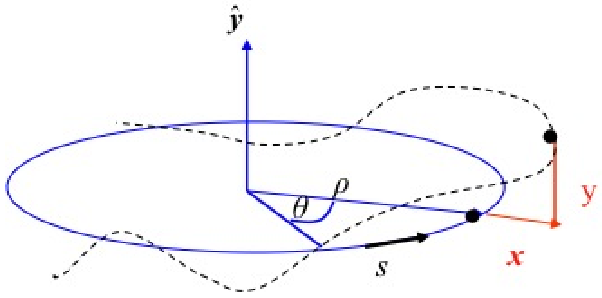


Fig. 8: Coordinate system used in particle beam dynamics; the longitudinal coordinate s moves around the ring with the particle considered.

Unlike the case of a classical harmonic oscillator, however, the equations of motion in the horizontal and vertical planes differ somewhat. Assuming a horizontal focusing magnet, the equation of motion is as shown in Eq. (16). In the vertical plane, on the other hand, because of the orientation of the field lines and thus by Maxwell’s equations, the forces instead have a defocusing effect. Also, the weak focusing term disappears in general:

$$y'' - y \cdot k = 0. \tag{17}$$

The principal problem arising from the different directions of the Lorentz force in the two transverse planes of a quadrupole field is sketched in Fig. 9. So, we have to explicitly introduce quadrupole lenses that focus the beam in the horizontal and vertical directions in some alternating order, and it is the task of the machine designer to find an adequate solution to this problem and to define a magnet pattern that will provide an overall focusing effect in both transverse planes.

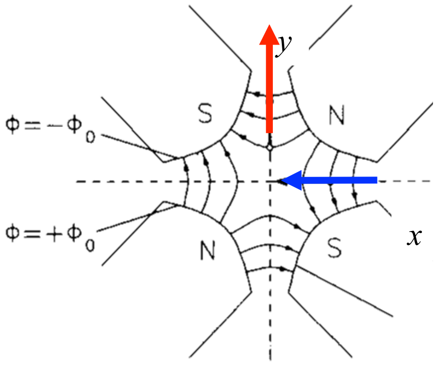


Fig. 9: Field configuration in a quadrupole magnet and the direction of the focusing and defocusing forces in the horizontal and vertical planes.

Following closely the example of the classical harmonic oscillator, we can write down the solutions of the above equations of motion. For simplicity, we focus on the horizontal plane; a ‘focusing’ magnet is therefore focusing in this horizontal plane and at the same time defocusing in the vertical plane. Starting with the initial conditions for the particle amplitude x_0 and angle x'_0 in front of the magnet element, we obtain the following relations for the trajectory inside the magnet:

$$x(s) = x_0 \cdot \cos\left(\sqrt{|K|}\, s\right) + x'_0 \cdot \frac{1}{\sqrt{|K|}} \sin\left(\sqrt{|K|}\, s\right), \tag{18}$$

$$x'(s) = -x_0 \cdot \sqrt{|K|} \sin\left(\sqrt{|K|}\, s\right) + x'_0 \cdot \cos\left(\sqrt{|K|}\, s\right). \tag{19}$$

Here the parameter K combines the quadrupole gradient and the weak focusing effect: $K = k - 1/\rho^2$. Usually, these two equations are combined into a more elegant and convenient matrix form,

$$\begin{pmatrix} x \\ x' \end{pmatrix}_s = \mathbf{M}_{\text{foc}} \begin{pmatrix} x \\ x' \end{pmatrix}_0 , \tag{20}$$

where the matrix $\mathbf{M}_{\text{foc}}$ contains all the relevant information about the magnet element:

$$\mathbf{M}_{\text{foc}} = \begin{pmatrix} \cos(\sqrt{|K|}\, s) & \frac{1}{\sqrt{|K|}} \sin(\sqrt{|K|}\, s) \\ -\sqrt{|K|} \sin(\sqrt{|K|}\, s) & \cos(\sqrt{|K|}\, s) \end{pmatrix} . \tag{21}$$

The situation is visualized schematically in Fig. 10.

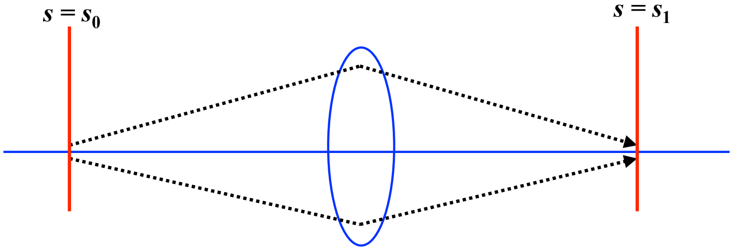


Fig. 10: Schematic illustration of the principle of the effect of a focusing quadrupole magnet

In the case of a defocusing magnet, we obtain analogously that

$$\begin{pmatrix} x \\ x' \end{pmatrix}_s = \mathbf{M}_{\text{defoc}} \begin{pmatrix} x \\ x' \end{pmatrix}_0 , \tag{22}$$

with

$$\mathbf{M}_{\text{defoc}} = \begin{pmatrix} \cosh(\sqrt{|K|}\, s) & \frac{1}{\sqrt{|K|}} \sinh(\sqrt{|K|}\, s) \\ \sqrt{|K|} \sinh(\sqrt{|K|}\, s) & \cosh(\sqrt{|K|}\, s) \end{pmatrix} ; \tag{23}$$

see Fig. 11.

For completeness, we also include the case of a field-free drift. With $K = 0$, we obtain

$$\mathbf{M}_{\text{drift}} = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} . \tag{24}$$

This matrix formalism allows us to combine the elements of a storage ring in an elegant way, and so it is straightforward to calculate particle trajectories. As an example, we consider the simple case of an alternating focusing and defocusing lattice, a so-called FODO lattice [2]; see Fig. 12.

As we know the properties of each and every element in the accelerator, we can construct the corresponding matrices and calculate step by step the amplitude and angle of a single-particle trajectory

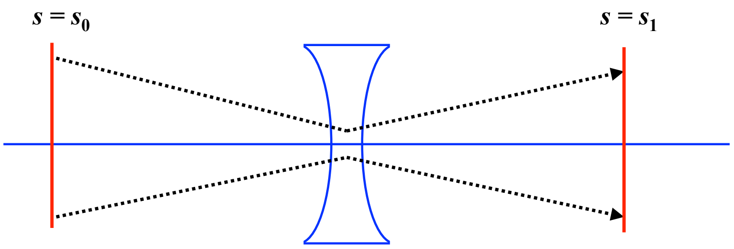


Fig. 11: Schematic illustration of the principle of the effect of a defocusing quadrupole magnet

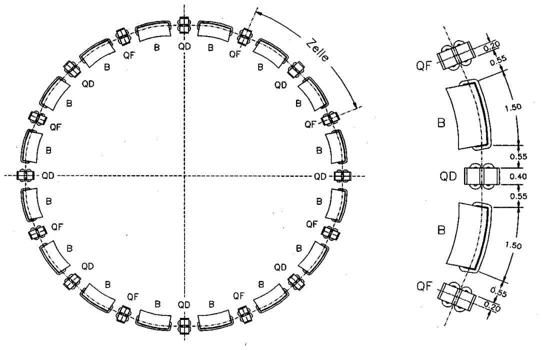


Fig. 12: A simple periodic chain of bending magnets and focusing/defocusing quadrupoles forming the basic structure of a storage ring [2].

around the ring. Even more conveniently, we can multiply out the different matrices and, given initial conditions x_0 and x'_0, obtain directly the trajectory at any location in the ring:

$$\mathbf{M}_{\text{total}} = \mathbf{M}_{\text{foc}} \cdot \mathbf{M}_{\text{drift}} \cdot \mathbf{M}_{\text{dipole}} \cdot \mathbf{M}_{\text{drift}} \cdot \mathbf{M}_{\text{defoc}} \cdots . \tag{25}$$

The trajectory thus obtained is shown schematically in Fig. 13.

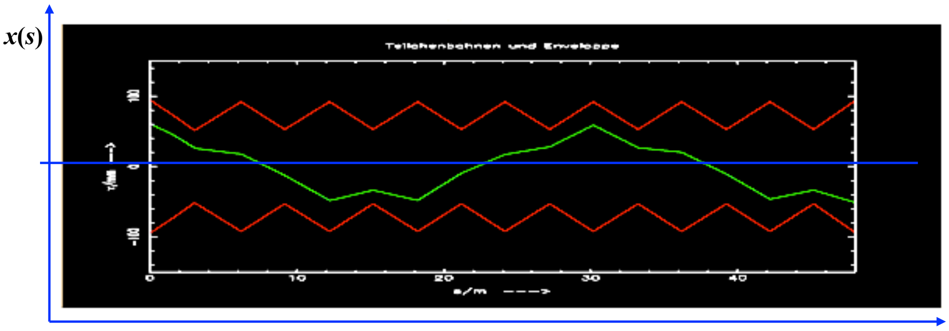


Fig. 13: Calculated particle trajectory in a simple storage ring

We emphasize the following facts in this context.

- At each moment, which means inside each lattice element, the trajectory is a part of a harmonic oscillation.

– However, because of the different restoring or defocusing forces, the solution will look different at each location.
– In the linear approximation that we have made use of in this context, all particles experience the same external fields, and their trajectories will differ only because of their different initial conditions.
– There seems to be an overall oscillation in both transverse planes while the particle is travelling around the ring. Its amplitude stays well within the boundaries set by the vacuum chamber, and its frequency in the example of Fig. 13 is roughly 1.4 transverse oscillations per revolution, which corresponds to the eigenfrequency of the particle under the influence of the external fields.

Coming closer to a real, existing machine, we show in Fig. 14 an orbit measured during one of the first injections into the LHC storage ring. The horizontal oscillations are plotted in the upper half of the figure and the vertical oscillations in the lower half, on a scale of ±10 mm. Each histogram bar indicates the value recorded by a beam position monitor at a certain location in the ring, and the orbit oscillations are clearly visible. By counting (or, better, fitting) the number of oscillations in both transverse planes, we obtain values of

$$Q_x = 64.31, \qquad Q_y = 59.32. \tag{26}$$

These values, which describe the eigenfrequencies of the particles, are called the horizontal and vertical *tunes*, respectively. Knowing the revolution frequency, we can easily calculate the transverse oscillation frequencies, which for this type of machine usually lie in the range of some hundreds of kilohertz.

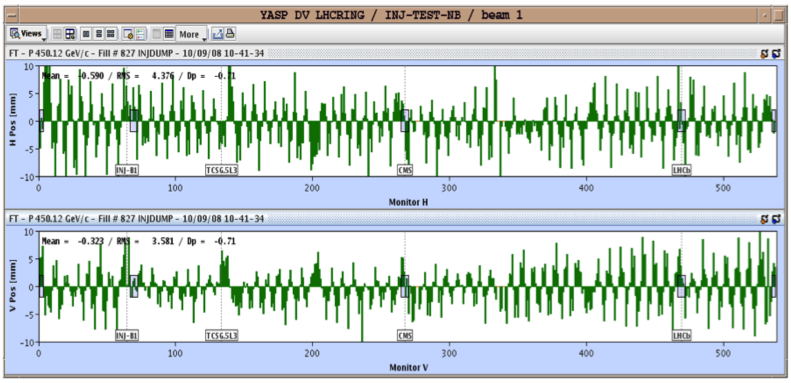

YASP DV LHCRING / INJ-TEST-NB / beam 1
Views
More
FT - P 450.12 GeV/c - Fill # 827 INJDUMP - 10/09/08 10-41-34
Mean = -0.590 / RMS = 4.376 / Dp = -0.71
H Pos [mm]
10
5
0
-5
-10
INJ-B1
TCSG.5L3
CMS
LHCb
0
100
200
300
400
500
Monitor H
FT - P 450.12 GeV/c - Fill # 827 INJDUMP - 10/09/08 10-41-34
Mean = -0.323 / RMS = 3.581 / Dp = -0.71
V Pos [mm]
10
5
0
-5
-10
INJ-B1
TCSG.5L3
CMS
LHCb
0
100
200
300
400
500
Monitor V

Fig. 14: Measured orbit in LHC during the commissioning of the machine

As the tune characterizes the particle oscillations under the influence of all external fields, it is one of the most important parameters of a storage ring. Therefore it is usually displayed and controlled at all times by the control system of such a machine. As an example, Fig. 15 shows the tune diagram of the HERA proton ring [7]; this was obtained via a Fourier analysis of a spectrum measured from the signal of the complete particle ensemble. The peaks indicate the two tunes in the horizontal and vertical planes of the machine, and in a sufficiently linear machine a fairly narrow spectrum is obtained.

Briefly referring back to Fig. 13, the question is what the trajectory of the particle will look like in the second turn, or the third, or after an arbitrary number of turns. Now, as we are dealing with a circular machine, the amplitude and angle x and x' at the end of the first turn will be the initial conditions for the second turn, and so on. After many turns, the overlapping trajectories begin to form a pattern, such as that in Fig. 16, which indeed looks like a beam that here and there has a larger and a smaller size but still remains well defined in its amplitude by the external focusing forces.

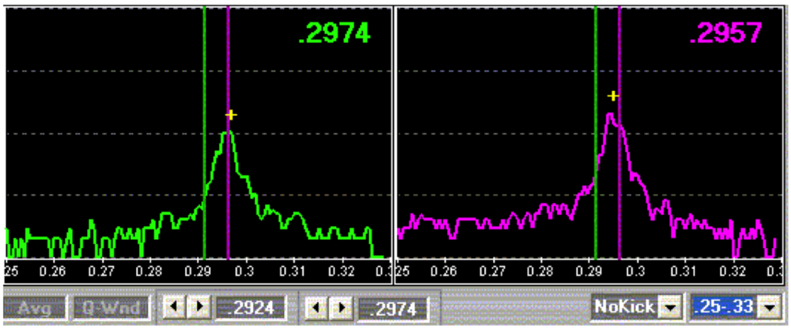


Fig. 15: Tune signal of a proton storage ring (HERA-p)

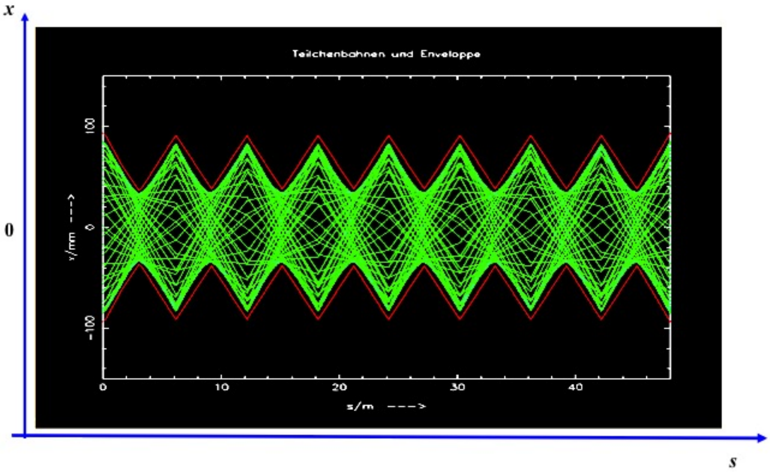


Fig. 16: Many single-particle trajectories together form a pattern that corresponds to the beam size in the ring.

3 The Twiss parameters α, β, and γ

As explained above, repeating the calculations that lead to the orbit of the first turn will result in a large number of single-particle trajectories that overlap in some way and form the beam envelope. Figure 16 shows the result for 50 turns. Clearly, as soon as we are talking about many turns or many particles, the use of the single-trajectory approach is quite limited and we need a description of the beam as an ensemble of many particles. Fortunately, in the case of periodic conditions in the accelerator, there is another way to describe the particle trajectories and, in many cases, it is more convenient than the above-mentioned formalism. It is important to note that, in a circular accelerator, the focusing elements are necessarily periodic in the orbit coordinate s after one revolution. Furthermore, storage ring lattices have an internal periodicity in most cases: they are often constructed, at least partly, from sequences in which identical magnetic structures, the lattice cells, are repeated several times in the ring and lead to periodically repeated focusing properties. In this case, the equation of motion can now be written in a slightly different form:

$$x''(s) - k(s) \cdot x(s) = 0, \qquad (27)$$

where, for simplicity, we refer to a pure quadrupole magnet and so the $1/\rho^2$ term does not appear. The main issue, however, is that unlike the treatment above, the focusing parameters (or restoring forces) are no longer constant but are functions of the coordinate s. However, they are periodic in the sense that, at least after one full turn, they repeat themselves, i.e., $k(s+L) = k(s)$, leading to the so-called Hill differential equation. Following Floquet's theorem [3], the solution of this equation can be written in its general form as

$$x(s) = \sqrt{\varepsilon}\sqrt{\beta(s)}\cos(\psi(s) + \phi), \tag{28}$$

where ψ is the phase of the oscillation, ϕ is its initial condition, and ε is a characteristic parameter of a single particle or, if we are considering a complete beam, of the ensemble of particles. Taking the derivative with respect to s, we get

$$x'(s) = -\frac{\sqrt{\varepsilon}}{\sqrt{\beta(s)}}\sin(\psi(s) - \phi) + \cos(\psi(s) + \phi). \tag{29}$$

The position and angle of the transverse oscillation of a particle at a point s are given by the value of a special amplitude function, the β-function, at that location, and ε and ϕ are constants of any particular trajectory. The β-function depends in a rather complicated manner on the overall focusing properties of the storage ring. It cannot be calculated directly by an analytical approach, but instead has to be either determined numerically or deduced from properties of the single-element matrices described above (see, e.g., [4]). In any case, like the lattice itself, it has to fulfil the periodicity condition

$$\beta(s + L) = \beta(s). \tag{30}$$

Inserting the solution (28) into the Hill equation and rearranging slightly, we get

$$\psi(s) = \int_0^s \frac{\mathrm{d}s}{\beta(s)}, \tag{31}$$

which describes the phase advance of the oscillation. It should be emphasized that ψ depends on the amplitude of oscillation of the particle. At locations where β reaches large values, i.e., the beam has a large transverse dimension, the corresponding phase advance is small; conversely, at locations where we create a small β in the lattice, we obtain a large phase advance. In the context of Fig. 13, we introduced the tune as the number of oscillations per turn, which is nothing other than the overall phase advance of the transverse oscillation per revolution in units of 2π. So, by integrating Eq. (31) around the ring, we get the expression

$$Q = \frac{1}{2\pi}\oint \frac{\mathrm{d}s}{\beta(s)}. \tag{32}$$

The practical significance of the β-function is shown in Figs. 16 and 17. Whereas in Fig. 16 the single-particle trajectories are plotted turn by turn, Fig. 17 shows schematically a section through the transverse shape of the beam and indicates the beam size inside the vacuum chamber. The hyperbolic profile of the pole shoes of the quadrupole lens is sketched as a yellow dashed line, and the envelope of the overlapping trajectories, given by $\hat{x} = \sqrt{\varepsilon\beta(s)}$, is marked in red and used to define the beam size in the sense of a Gaussian density distribution.

3.1 β, ε, and the phase space ellipses

Although the β-function is a somewhat abstract parameter that results from all focusing and defocusing elements in the ring, the integration constant ε has a well-defined physical interpretation. Given the solution of Hill's equation

$$x(s) = \sqrt{\varepsilon}\sqrt{\beta(s)}\cos(\psi(s) + \phi) \tag{33}$$

and its derivative

$$x'(s) = -\frac{\sqrt{\varepsilon}}{\sqrt{\beta(s)}}\sin(\psi(s) - \phi) + \cos(\psi(s) + \phi), \tag{34}$$

we can transform Eq. (33) to

$$\cos(\psi(s)) = \frac{x(s)}{\sqrt{\varepsilon\beta(s)}} \tag{35}$$

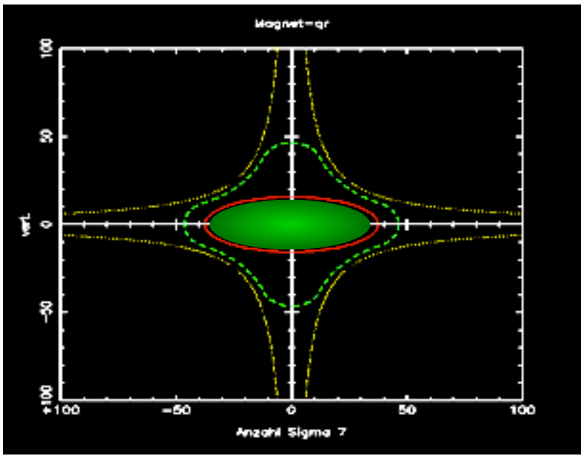


Fig. 17: Transverse beam shape inside a quadrupole magnet

and insert the expression into Eq. (34) to get an expression for the integration constant ε:

$$\varepsilon = \gamma(s)x^2(s) + 2\alpha x(s)x'(s) + \beta(s)x'^2(s). \tag{36}$$

Here we have followed the usual convention in the literature and introduced the two parameters

$$\alpha(s) = -\frac{1}{2}\beta'(s) \tag{37}$$

and

$$\gamma(s) = \frac{1+\alpha^2(s)}{\beta(s)}. \tag{38}$$

We obtain for ε a parametric representation of an ellipse in the (x, x') 'phase space', which, according to Liouville's theorem, is a constant of motion, as long as only conservative forces are considered. The mathematical integration constant thus gains physical meaning. In fact, ε describes the space occupied by the particle in the transverse (x, x') phase space (simplified here to a two-dimensional space). More specifically, the area in the (x, x') space that is covered by the particle is given by

$$A = \pi \cdot \varepsilon, \tag{39}$$

and, as long as we consider only conservative forces acting on the particle, this area is constant according to Liouville's theorem. Here we take these facts as given, but we should point out that, as a direct consequence, the so-called emittance ε cannot be influenced by any external fields; it is a property of the beam, and we have to take it as given and handle it with care.

To be more precise, and following the usual textbook treatment of accelerators, we can draw the ellipse of the particle's transverse motion in phase space; see, for example, Fig. 18. Although the shape and orientation are determined by the optics function β and its derivative, $\alpha = -\frac{1}{2}\beta'$, and so change as a function of the position s, the area covered in phase space is constant.

In Fig. 18, expressions for the dependence of the beam size and divergence and, as a consequence, the shape and orientation of the phase space ellipse are included. For the sake of simplicity, we shall not derive these expressions here but instead refer to [4].

Referring again to the single-particle trajectory discussed above (see Fig. 13), but now plotting for a given position s in the ring the coordinates x and x' turn by turn, we obtain the phase space coordinates of the particle as shown in Fig. 18 (marked as dots in the figure). These coordinates follow

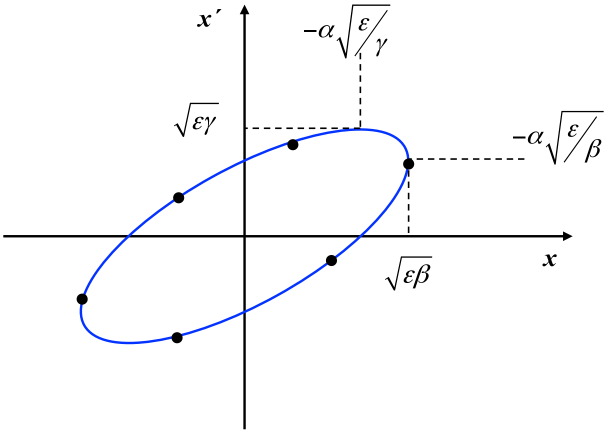

x′
$-\alpha\sqrt{\varepsilon/\gamma}$
$\sqrt{\varepsilon\gamma}$
$-\alpha\sqrt{\varepsilon/\beta}$
$\sqrt{\varepsilon\beta}$
x

Fig. 18: Ellipse in (x, x') phase space

the form of an ellipse, whose shape and orientation are defined by the optical parameters at the reference position s in the ring. Each point in Fig. 18 represents the transverse coordinates for a certain turn at that position in the ring, and the particle performs from one turn to the next a number of revolutions in phase space that corresponds to its tune. We have already emphasized that, as long as only conservative forces are considered (i.e., no interaction between the particles in a bunch, no collisions with remaining gas molecules, no radiation effects, etc.), the size of the ellipse in (x, x') space is constant and can be considered as a quality factor of a single particle. Large areas in (x, x') space mean large amplitudes and angles of transverse particle motion, and we would consider this as meaning a low particle 'quality'.

Let us now talk a little more about the beam as an ensemble of many (typically 10^{11}) particles. Referring to Eq. (28), at a given position in the ring the beam size is defined by the emittance ε and the function β. Thus, at a certain moment in time, the cosine term in Eq. (28) will be equal to one and the amplitude of the trajectory will reach its maximum value. Now, if we consider a particle at one standard deviation (sigma) of the transverse density distribution, then by using the emittance of this reference particle we can calculate the size of the complete beam, in the sense that the complete area (within one sigma) of all particles in the (x, x') phase space is surrounded (and thus defined) by our one-sigma candidate. Thus the value $\sqrt{\varepsilon \cdot \beta(s)}$ defines the one-sigma beam size in the transverse plane.

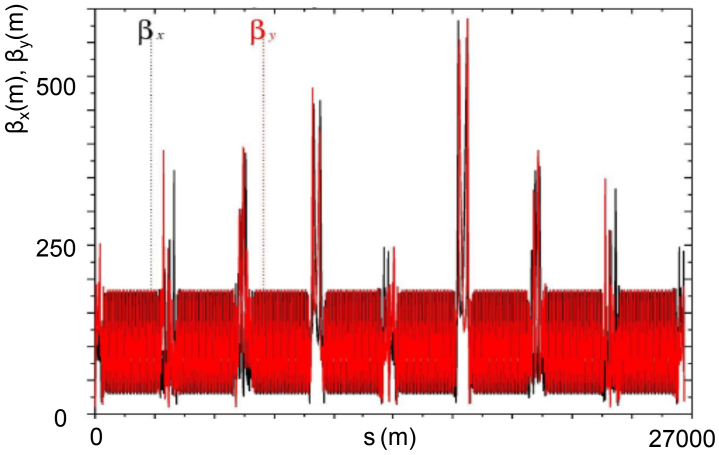


Fig. 19: LHC beam optics

As an example, we shall use the values for the LHC proton beam (Fig. 19). In the periodic pattern of the arc, the β-function is equal to 180 m and the emittance ε at the flat-top energy is roughly 5×10^{-10} rad m. The resulting typical beam size is therefore 0.3 mm. Now, clearly, we would not design a vacuum aperture for the machine based on a one-sigma beam size; typically, an aperture requirement corresponding to 12σ is a good rule to guarantee a sufficient beam lifetime, allowing for tolerances arising from magnet misalignments, optics errors, orbit fluctuations, and operational flexibility. In Fig. 20, part of the LHC vacuum chamber is shown, including the beam screen used to protect the cold bore from synchrotron radiation; this corresponds to a minimum beam size of 18σ.

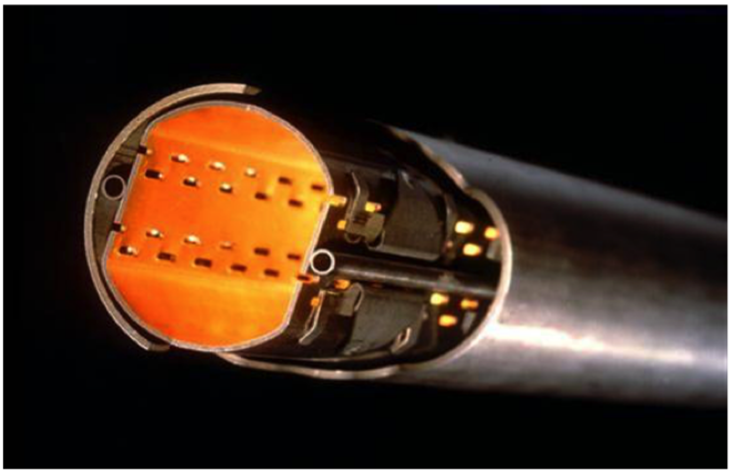


Fig. 20: The LHC vacuum chamber with a beam screen to shield the bore of the superconducting magnet from synchrotron radiation.

4 Errors in field and gradient

Up to now, we have treated the beam and the equation of motion as a monoenergetic problem. Unfortunately, in the case of a realistic beam, we have to deal with a considerable distribution of the particles with respect to energy or momentum. A typical value is

$$\frac{\Delta p}{p} \approx 1.0 \cdot 10^{-3}. \tag{40}$$

This momentum spread leads to several effects concerning the bending of the dipole magnets and the focusing strength of the quadrupoles. It turns out that the equation of motion, which has been a homogeneous differential equation until now, acquires a non-vanishing term on the right-hand side.

4.1 Dispersive effects

Replacing the ideal momentum p in Eq. (10) by $p_0 + \Delta p$, we obtain instead of Eq. (16)

$$x'' + x \cdot \left(\frac{1}{\rho^2} + k\right) = \frac{\Delta p}{p} \cdot \frac{1}{\rho}. \tag{41}$$

The general solution of our now inhomogeneous differential equation is therefore the sum of the solution of the homogenous equation of motion and a particular solution of the inhomogeneous equation:

$$x(s) = x_{\mathrm{h}}(s) + x_{\mathrm{i}}(s). \tag{42}$$

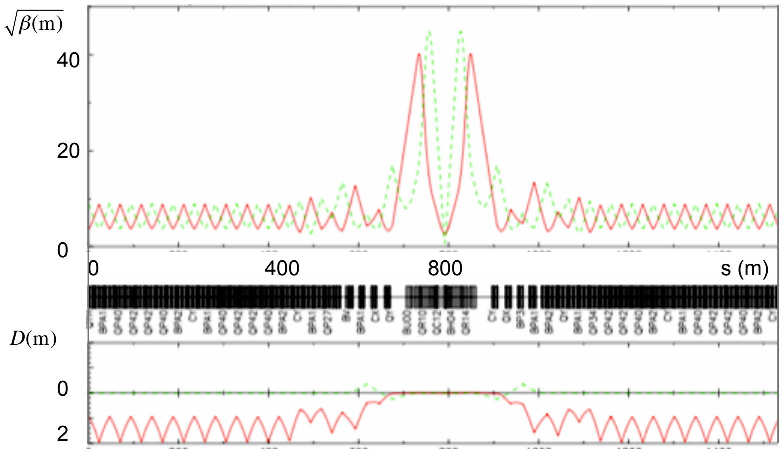


Fig. 21: β-function (upper part) and dispersion (lower part) of a typical high-energy collider ring

Here x_{h} is the solution that we have discussed up to now and x_{i} is an additional contribution that still has to be determined. For convenience, we usually normalize this second term and define a special function, the so-called dispersion:

$$D(s) = \frac{x_{\mathrm{i}}(s)}{\Delta p/p}. \tag{43}$$

This describes the dependence of the additional amplitude of the transverse oscillation on the momentum error of the particle. In other words, it fulfils the condition

$$x_{\mathrm{i}}''(s) + K(s) \cdot x_{\mathrm{i}}(s) = \frac{1}{\rho} \cdot \frac{\Delta p}{p}. \tag{44}$$

The dispersion function is defined by the magnet lattice and is usually calculated by optics programs in the context of the calculation of the usual optical parameters; it is of equal importance. Analytically, it can be determined for single elements via the expression

$$D(s) = S(s) \cdot \int \frac{1}{\rho(\bar{s})} C(\bar{s})\, \mathrm{d}\bar{s} - C(s) \cdot \int \frac{1}{\rho(\bar{s})} S(\bar{s})\, \mathrm{d}\bar{s}, \tag{45}$$

where $S(s)$ and $C(s)$ correspond to the sine-like and cosine-like elements of the single-element matrices or of the corresponding product matrix if there are several elements considered in the lattice.

Although all this sounds somewhat theoretical, we would like to stress that typical values for the beam size and dispersive effect in the case of a high-energy storage ring are

$$x_{\beta} \approx 1\text{–}2\ \mathrm{mm}, \qquad D(s) \approx 1\text{–}2\ \mathrm{m}. \tag{46}$$

Thus, for a typical momentum spread of $\Delta p/p = 1 \cdot 10^{-3}$, we obtain an additional contribution to the beam size from the dispersion function that is of the same order as that from the betatron oscillations, x_{β}. An example of a high-energy beam optics system including the dispersion function is shown in Fig. 21. It should be pointed out that the dispersion describes the special orbit that an ideal particle would have in the absence of betatron oscillations ($x_{\beta} = x'_{\beta} = 0$) for a momentum deviation of $\Delta p/p = 1$. Nevertheless, it describes 'just another particle orbit' and so it is subject to the focusing forces of the lattice elements, as seen in the figure.

4.2 Chromaticity

Whereas dispersion is a problem that describes the non-ideal bending effect of dipoles in the case of a momentum error (or spread) in the particles, the careful reader will not be surprised to learn that a similar

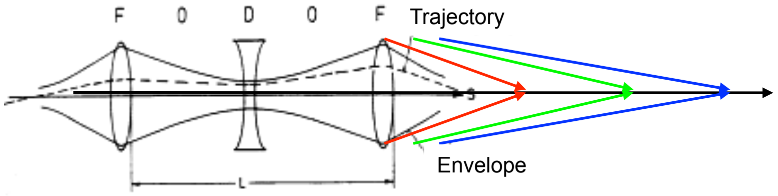


Fig. 22: Schematic view of the chromaticity effect in a quadrupole lens

effect exists for the quadrupole focusing. We call this *chromaticity*. The chromaticity Q' describes an optical error of a quadrupole lens in an accelerator: for a given magnetic field, i.e., gradient of the quadrupole magnet, particles with a smaller momentum will feel a stronger focusing force, and particles with a larger momentum will feel a weaker force. The situation is shown schematically in Fig. 22. As a consequence, the tune of an individual particle will change, and the chromaticity Q' relates the resulting tune shift to the relative momentum error of the particle. By definition, we write

$$\Delta Q = Q' \cdot \frac{\Delta p}{p}. \tag{47}$$

Q' is a consequence of the focusing properties of the quadrupole magnets and is thus given by the characteristics of the lattice. For small momentum errors $\Delta p/p_0$, the focusing parameter k can be written as

$$k(p) = \frac{g}{p/e} = \frac{ge}{p_0 + \Delta p}, \tag{48}$$

where g denotes the gradient of the quadrupole lens, p_0 denotes the design momentum, and the term Δp refers to the momentum error. If Δp is small, as we have assumed, we can write in a first-order approximation

$$k(p) \approx \frac{ge}{p_0}\left(1 - \frac{\Delta p}{p_0}\right). \tag{49}$$

This describes a quadrupole error

$$\Delta k = -k_0 \cdot \frac{\Delta p}{p}. \tag{50}$$

The negative sign indicates that a positive momentum deviation leads to a weaker focusing strength and, accordingly, to a negative tune shift:

$$\Delta Q = -\frac{1}{4\pi}\int \Delta k\, \beta(s)\, \mathrm{d}s, \tag{51}$$

$$\Delta Q = -\frac{1}{4\pi}\frac{\Delta p}{p}\int k_0 \beta(s)\, \mathrm{d}s. \tag{52}$$

By definition, the chromaticity Q' of a lattice is therefore given by

$$Q' = -\frac{1}{4\pi}\int k(s)\beta(s)\, \mathrm{d}s. \tag{53}$$

Now, unfortunately, although the dispersion created in the dipole magnets requires nothing more than some more aperture in the vacuum chamber, the chromaticity of the quadrupoles has an influence on the tune of the particles and so can lead to dangerous resonance conditions. Particles with a particular momentum error will be pushed into resonances and be lost within a very short time. A look at the tune spectrum visualizes the problem. Whereas an ideal situation leads to a well-compensated chromaticity

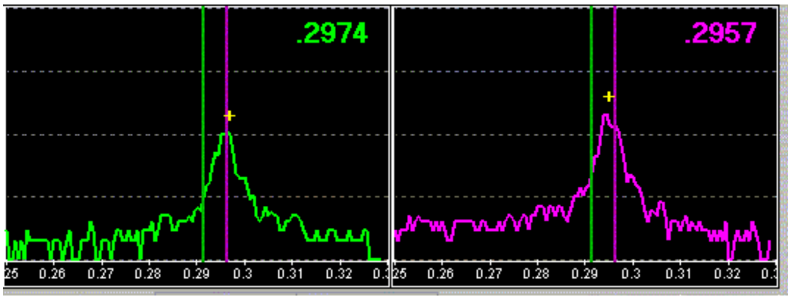


Fig. 23: Tune spectrum of a proton beam with a well-corrected chromaticity $Q' \approx 1$

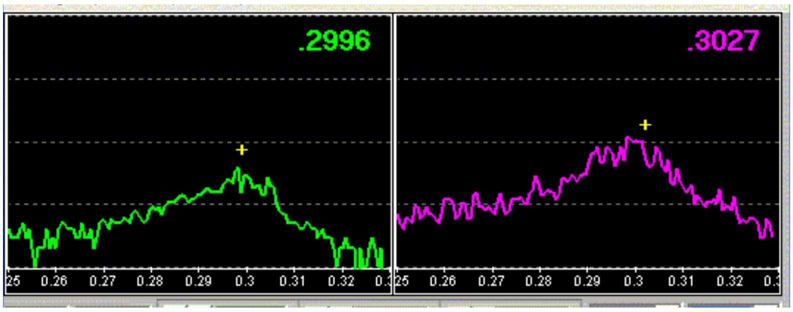


Fig. 24: Tune spectrum of a proton beam with a poorly matched chromaticity $Q' \approx 20$

and the particles oscillate with basically the same frequency (Fig. 23), a non-corrected chromaticity (Q' = 20 units in the case of Fig. 24) broadens the tune spectrum and a number of particles are pushed towards dangerous resonance lines.

In large storage rings and synchrotrons in particular, this problem is crucial and represents one of the major factors that limit machine performance: because of the strong focusing of the quadrupoles and the large size, the chromaticity can reach considerable values. A chromaticity correction scheme is therefore indispensable. The trick is performed in three steps.

- We sort the particles in the horizontal plane according to their momentum. This is done whenever we have a non-vanishing dispersion, for example close to the focusing quadrupoles in the arc, where both the dispersion and the β-function reach high values and the particle trajectories are determined by the well-known relation $x_\mathrm{d}(s) = D(s) \cdot \Delta p/p$.
- At these places, we create magnetic fields that have a position-dependent focusing strength, in other words, fields that represent a position-dependent gradient. Sextupole magnets have exactly this property: if g' describes the strength of the sextupole field, we get

$$B_x = g' \cdot xy \tag{54}$$

for the horizontal field component and

$$B_y = g' \frac{1}{2} \cdot (x^2 - y^2) \tag{55}$$

for the vertical component. The resulting gradient in both planes is obtained as

$$\frac{\mathrm{d}B_x}{\mathrm{d}y} = \frac{\mathrm{dB_y}}{\mathrm{d}x} = g' \cdot x. \tag{56}$$

- We now only have to adjust the strengths of two sextupole families (one to compensate the horizontal and another to compensate the vertical chromaticity) to get an overall correction in both planes.

In a little more detail, and referring again to normalized gradients, we can write

$$k_{\mathrm{sext}} = \frac{e}{p} g' \cdot x = m \cdot x, \tag{57}$$

which leads for a given particle amplitude

$$x_{\mathrm{d}} = D \cdot \frac{\Delta p}{p} \tag{58}$$

to the normalized focusing strength (of the sextupole magnet)

$$k_{\mathrm{sext}} = m \cdot D \frac{\Delta p}{p}. \tag{59}$$

The combined effect of the so-called natural chromaticity created by the quadrupole lenses (Eq. (53)) and the compensation by the sextupoles leads to an overall chromaticity

$$Q' = -\frac{1}{4\pi} \oint (K(s) - m(s) \cdot D(s)) \beta(s) \, \mathrm{d}s \tag{60}$$

and needs to be compensated to zero in both transverse planes.

To summarize and make things as crystal clear as possible, the focusing properties of the magnet lattice lead to restoring forces in both transverse planes. The transverse motion of a particle is therefore a quasi-harmonic oscillation as the particle moves through the synchrotron, and the tune describes the frequency of these oscillations. As we cannot assume that all particles have exactly the same momentum, we have to take into account the effect of the momentum spread in the beam: the restoring forces are a function of the momentum of each individual particle and so the tune of each particle is different. We have to correct for this effect, and we do so by applying sextupole fields in regions where a non-vanishing dispersion distributes the off-momentum particles in the horizontal plane.
As easy as that!

5 Longitudinal beam dynamics

5.1 Introduction

Following the tradition in most textbooks, we treat particle dynamics in the longitudinal direction in a separate section of this paper. And, before going into the technical and mathematical details of the treatment, we would like to describe the motivation behind this decision.

Technically, magnetic fields are most appropriate for guiding the beam and creating the transverse focusing forces that are needed to keep the particles on a stable orbit. As we have shown in the discussion of Eq. (1), the electric fields that could be used as well are far weaker. However, the Lorentz force resulting from a magnetic field is always perpendicular to the particle's velocity vector. And, as an unfortunate consequence, we are obliged to make use of the weak electric fields as soon as we talk about particle acceleration. Technically speaking, we must replace our beloved magnets by electric fields, created in devices that we call RF resonators or cavities, where RF waves are built up to act on the beam, with the field vector pointing in the direction of the particle motion.

As a direct consequence, the resulting forces in the longitudinal plane are much weaker and so the longitudinal oscillation frequencies are much smaller. Thus, in addition to the technically quite different approaches to transverse and longitudinal dynamics, there is a physics argument that allows us to treat the two aspects independently. The resulting frequencies are very different, and there is hardly any crosstalk (or coupling) between the transverse and longitudinal oscillations of the beam particles. An extreme example might again be the LHC: whereas the frequency of the betatron oscillations is of the order of 1 MHz, the frequency of the longitudinal movements is about 23 Hz. Nevertheless, for completeness, we should mention that in a higher order of approximation coupling between the two modes is indeed possible and needs to be avoided. But these topics are beyond the scope of this paper, and the curious reader is referred to Ref. [8].

The longitudinal movement of the particles in a storage ring — and, closely related to this, the acceleration of the beam — is strongly related to the problem of synchronization between the particles and the accelerating system. This synchronization may be established via the basic hardware and design of the machine (as in a Wideroe structure) or via the orbit (as in a cyclotron), or it may be a fundamental feature of the ring in a more sophisticated way, which leads to the name 'synchrotron' for the specific type of machine concerned. We shall treat these different aspects in more detail. But, before we do so, we would like to start on a very fundamental basis and at the same time go back a little in history.

5.2 Electrostatic machines

The most prominent example of an electrostatic machine, besides the Cockcroft–Walton generator [9–11], is the Van de Graaff accelerator [12]. A sketch of the principle is shown in Fig. 25. Using a moderate DC high voltage, charges are sprayed onto a moving belt or a chain with insulated links and transported to a kind of Faraday screen, where the charges accumulate, leading to considerably high voltages. By their design concept, these machines deliver an excellent energy resolution, as basically each and every particle sees the same accelerating potential (which is no longer the case when we have to consider RF accelerating structures). A short overview of these machines can be found in Ref. [13].

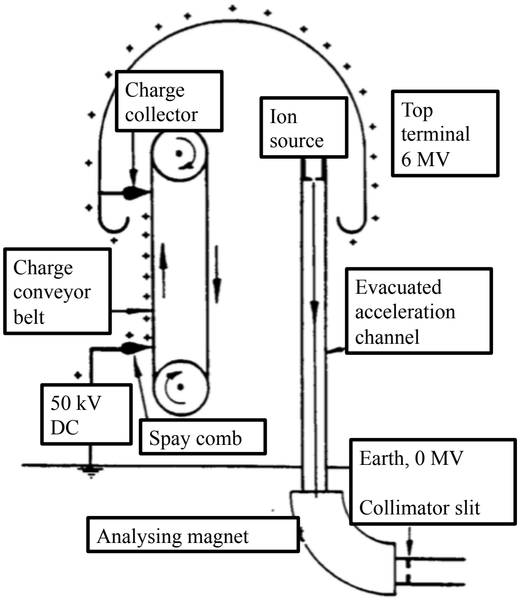


Fig. 25: Technical principle of a Van de Graaff accelerator [13]

The kinetic energy of the particle beam is given by integration of the electric field E in the direction z of the particle motion, and is measured as usual in units of electronvolts (eV):

$$\mathrm{d}W = e \cdot E_z \,\mathrm{d}s \Rightarrow W = e \int E_z \,\mathrm{d}s = [\mathrm{eV}]. \tag{61}$$

An example of such a machine is shown in Fig. 26. This ‘tandem’ Van de Graaff accelerator uses a stripper foil in the middle of the structure. Thus, in the first half, it accelerates negative ions that have been produced in a Cs-loaded source. After stripping of the electrons, the same voltage is applied in the second part of the structure to gain another step in energy. All in all, the particle energy is determined by the high voltage that can be created or, more precisely, by the potential difference the particles pass through, and so it is finally always limited by discharge effects in the electric field.

Fig. 26: Tandem Van de Graaff accelerator at the Max Planck Institute, Heidelberg

We emphasize here that special synchronization (i.e., timing between the particles and accelerating field) is not needed, as a DC voltage is used for acceleration. The disadvantage, however, is the fact that the DC voltage applied can only be used once and the energy gain of the particles is limited by the technically achievable voltage (or, more precisely, the feasible E-field), which defines the potential difference.

5.3 Radio frequency accelerators: The Wiederoe-type linac

The basic limitation imposed by the maximum achievable voltage in electrostatic accelerators can be overcome by applying AC voltages. However, a more complicated design is needed to prevent the particles from being decelerated during the negative half-wave of the RF system. In 1928, Wiederoe presented the layout of such a machine (and built it) for the first time. Figure 27 shows the principle.

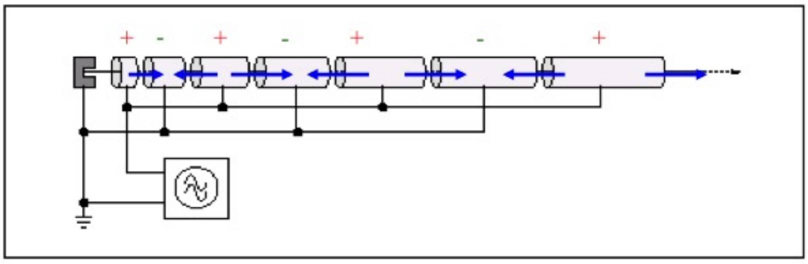


Fig. 27: Schematic view of a Wiederoe drift-tube linac

The arrows in the figure show the direction of the electric field at a given moment in time (i.e., the positive half-wave of the AC voltage), and the corresponding polarity applied to the electrodes is indicated by the $+$ and $-$ signs. In principle, arbitrarily high beam energies can be achieved by applying the same voltage to a high enough number of electrodes, provided that the particle beam is shielded from the electric field during the negative RF half-wave. Accordingly, the electrodes are designed as drift tubes (thus the expression ‘drift tube linac’) whose length is chosen according to the particle velocity and RF period. The design principle is shown schematically in Fig. 28.

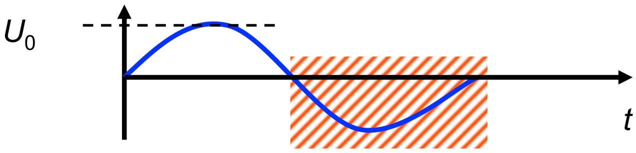


Fig. 28: The length of the drift tube has to be optimized to shield the particles from the negative RF half-wave

The area shaded red in the figure corresponds to the time during which the particle has to be shielded from the decelerating field direction, and is defined by the RF period: $t_{\text{shield}} = \tau_{\text{rf}}/2$. Accordingly, the length of the drift tube has to be

$$l_i = v_i \cdot \frac{\tau_{\text{rf}}}{2}. \tag{62}$$

If the kinetic energy (in the classical regime) is given by

$$E_i = \frac{1}{2} m v^2, \tag{63}$$

we obtain an equation for the drift tube length that, for a given accelerating voltage U_0 and charge q, depends on the RF frequency ν_{rf} and the number of the accelerating step i:

$$l_i = \frac{1}{\nu_{\text{rf}}} \sqrt{\frac{i q U_0 \cdot \sin \psi_{\text{s}}}{2m}}. \tag{64}$$

The parameter ψ_{s} describes the so-called synchronous phase and can be chosen to be 0° in this case to obtain the highest acceleration performance.

One of the best-known examples of such an accelerator is running at GSI in Darmstadt, and is used as a universal tool for the acceleration of (almost) any heavy-ion species. The internal structure is shown in Fig. 29, including the drift tubes and the surrounding vessel.

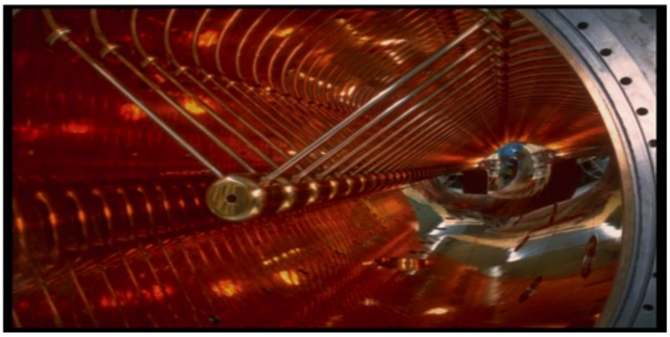


Fig. 29: The UNILAC drift tube accelerator at GSI

Unlike the case for DC accelerators, the timing between the particles and the RF field is suddenly of major importance: synchronization has to be obtained between the particle's time of arrival at the resonator and the accelerating RF field, which—as shown above—is represented by the drift tube length, and so in a certain sense is built into the hardware of the system.

6 Longitudinal particle dynamics in synchrotrons

The design of a synchrotron follows this approach, except that the drift tube of the RF structure, where the particles are shielded from the decelerating field, is replaced by the machine itself. If we define the longitudinal aspects of a synchrotron as a circular accelerator with:

- a design orbit of constant radius, defined by the arrangement and strength of a number of dipole magnets;
- an RF system, located at a distinct place in the ring and powered at an RF frequency that is equal to the revolution frequency of the particles or an integer multiple (so-called harmonic) of it;

we are already quite close to reality. The rest is some mathematics.

For a description of the particle dynamics, we refer to a synchronous particle of ideal energy, phase, and energy gain per turn. As we shall see, the synchronization between the RF system and the particle beam is of major importance in this type of machine and has to be explicitly included in the design. To understand the principle, we have to refer briefly to the transverse dynamics of a particle with a momentum error and the resulting dispersive effects (Fig. 30).

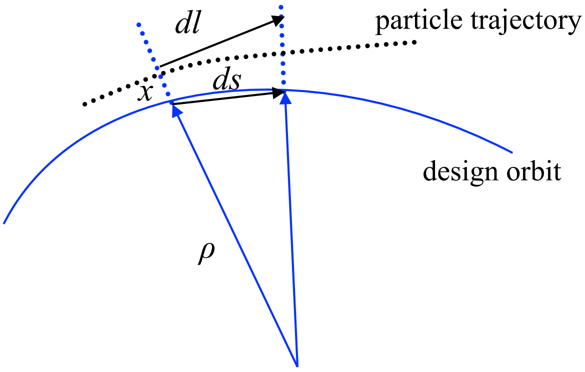


Fig. 30: Particle orbits in a synchrotron for an ideal and an off-momentum particle

Whereas the ideal particle will run on the design orbit defined by the dipole magnets and will proceed a distance $\mathrm{d}s$, a non-ideal particle will run on a displaced orbit (displaced to the outer side of the ring in the example of Fig. 30) and will travel a corresponding distance $\mathrm{d}l$:

$$\frac{\mathrm{d}l}{\mathrm{d}s} = \frac{\rho + x}{\rho}. \tag{65}$$

Solving for $\mathrm{d}l$, we obtain

$$\mathrm{d}l = 1 + \frac{x}{\rho(s)}\mathrm{d}s, \tag{66}$$

and by integrating around the machine we get the orbit length of the non-ideal particle, which depends on the radial displacement x:

$$l_{\Delta E} = \int \mathrm{d}l = \int \left(1 + \frac{x_{\Delta E}}{\rho(s)}\right) \mathrm{d}s, \tag{67}$$

where we assume that the radial displacement $x_{\Delta E}$ is caused by a momentum error and the dispersion function of the magnet lattice:

$$x_{\Delta E}(s) = D(s) \cdot \frac{\Delta p}{p}. \tag{68}$$

We obtain an expression for the difference in orbit length between the ideal and the dispersive particle, which is determined by the size of the relative momentum error and the dispersion function of the storage ring:

$$\delta l_{\Delta E} = \frac{\Delta p}{p} \int \frac{D(s)}{\rho(s)} \mathrm{d}s. \tag{69}$$

The ratio between the relative orbit difference and the relative momentum error is called the *momentum compaction factor* α_p and is determined by the integral of the dispersion function around the ring and the bending radius of the dipole magnets:

$$\frac{\delta l_{\Delta E}}{L} = \alpha_p \frac{\Delta p}{p}, \tag{70}$$

where

$$\alpha_p = \frac{1}{L}\int \frac{D(s)}{\rho(s)}\,\mathrm{d}s. \tag{71}$$

Although the expression 'momentum compaction factor' might be an unfortunate choice and we would like instead to call it the lengthening factor, its physical meaning is as important as it is clear: it describes the lengthening of the orbit for particles that have a given momentum deviation with respect to the ideal particle. And it is also clear that in a circular accelerator this orbit-lengthening effect cannot be avoided, because of the dispersion function.

For some initial estimates, we assume equal bending radii in all dipoles, so $1/\rho = \mathrm{const}$ and we replace the integral of the dispersion around the ring by a sum over the average dispersion in the dipole magnets (outside the dipoles the term $1/\rho = 0$, so this assumption is justified for a rough estimate):

$$\int_{\text{dipoles}} D(s)\,\mathrm{d}s \approx l_{\Sigma(\text{dipoles})}\langle D\rangle_{\text{dipole}}. \tag{72}$$

We get a nice, simple expression for the momentum compaction factor that depends only on the ratio of the average dispersion to the geometric radius R of the machine:

$$\alpha_p = \frac{1}{L} l_{\Sigma(\text{dipoles})}\langle D\rangle_{\text{dipole}}\frac{1}{\rho} = \frac{1}{L}2\pi\rho\langle D\rangle\frac{1}{\rho}, \tag{73}$$

$$\alpha_p = \frac{2\pi}{L}\langle D\rangle \approx \frac{\langle D\rangle}{R}. \tag{74}$$

For a quick estimate, α_p is given by the ratio of the average dispersion to the geometric radius of the ring. Assuming, finally, that the particles are moving at the speed of light, i.e., $v \approx c = \mathrm{const}$, the relative error in time is given by the relative change in the orbit length and thus by the momentum compaction factor and the relative momentum error:

$$\frac{\delta t}{t} = \frac{\delta l_\epsilon}{L} = \alpha_p \frac{\Delta p}{p}. \tag{75}$$

So the secret of the longitudinal motion is already disclosed, even if you might not have realized yet: the dispersive effect in a synchrotron or, in other words, the orbit lengthening for off-momentum particles, described by α_p, is the fundamental feature of the principle of operation of a synchrotron: it relates the time of arrival in the RF structure to the momentum error of the particle.

6.1 Dispersive effects in synchrotrons

Well ... we have seen this topic come up already, but we would like to look at the issue from the point of view of timing, i.e., from the point of view of the synchronization between the particles and the RF system. And we have to enlarge our point of view and include the case of non-relativistic particles (a rigorous treatment can be found, for example, in Refs. [14] and [15]). So our problem of synchronization needs a more careful treatment, which must include the fact that the particles are travelling at a speed that might be considerably lower than the speed of light. The parameter of interest, however, is still the ratio between the relative momentum error and the relative frequency deviation of a particle:

$$\frac{\mathrm{d}f_\mathrm{r}}{f_\mathrm{r}} = \eta\frac{\mathrm{d}p}{p}. \tag{76}$$

Here we have introduced the parameter η to address this issue, and we have added explicitly a subscript 'r' to denote the revolution frequency of the particle. We shall derive an expression for this η-parameter in the next few lines. But we would like to point out now that η combines the effect of the changing velocity of the particle *and* the relativistic increase in mass with changing energy. And thus it is this strange η that is the key parameter for anything about timing in a synchrotron.

But step by step ... Given the revolution frequency as a function of machine circumference and speed, the revolution frequency around the ring is

$$f_{\mathrm{r}} = \frac{\beta c}{2\pi R}, \tag{77}$$

where β is the relativistic parameter v/c and R stands for the geometric radius of the machine (defined by the length of the design orbit, which is $2\pi R$). Via the logarithmic derivative, we obtain the obvious relation

$$\frac{\mathrm{d}f_{\mathrm{r}}}{f_{\mathrm{r}}} = \frac{\mathrm{d}\beta}{\beta} - \frac{\mathrm{d}R}{R}. \tag{78}$$

Now, from Eq. (70) we know that the relative change in radius, i.e., the second term in the expression, is given by the momentum compaction factor α_p:

$$\frac{\mathrm{d}R}{R} = \alpha_p \frac{\mathrm{d}p}{p}; \tag{79}$$

and, as the momentum is related to the particle energy,

$$p = mv = \beta\gamma \frac{E_0}{c}, \tag{80}$$

we can write the following for the relative momentum change:

$$\frac{\mathrm{d}p}{p} = \frac{\mathrm{d}\beta}{\beta} + \frac{\mathrm{d}(1-\beta^2)^{-1/2}}{(1-\beta^2)^{-1/2}} = \gamma^2 \frac{\mathrm{d}\beta}{\beta}. \tag{81}$$

Introducing the two equations (79) and (81) into Eq. (78), we finally obtain the required relation between the frequency offset and the momentum error,

$$\frac{\mathrm{d}f_{\mathrm{r}}}{f_{\mathrm{r}}} = \left(\frac{1}{\gamma^2} - \alpha\right) \frac{\mathrm{d}p}{p}. \tag{82}$$

Accordingly, the η-parameter defined above is given as

$$\eta = \frac{1}{\gamma^2} - \alpha_p. \tag{83}$$

This combines the effect of a momentum deviation on the orbit size (described by α_p above) and the effect of the speed of the particle, which increases with increasing momentum, until we reach the ultra-relativistic regime and $v \approx c = \mathrm{const}$.

This relation is indeed extremely interesting, as it tells us *what to do when we start to accelerate particles* in a synchrotron. The easiest situation occurs when we are dealing with ultrarelativistic particles. In this case the last equation reduces to the simplified situation described in Eq. (75): if γ is high, the first term, $1/\gamma^2$, tends to zero and the change in the revolution frequency is defined by the momentum compaction factor α_p. For small energies, however, things get more complicated.

As an important remark, we state that the change in revolution frequency depends on the particle energy γ and may possibly change sign during acceleration. Particles become faster at the beginning of the process and arrive earlier at the location of the cavity (classical regime), whereas particles that travel at $v \approx c$ will not get any faster but instead become more massive and, being pushed to a dispersive orbit, will arrive later at the cavity (relativistic regime). The boundary between the two regimes is defined by the case where no dependence of the frequency on $\mathrm{d}p/p$ is obtained, namely, $\eta = 0$, and the corresponding energy is called the *transition energy*:

$$\gamma_{\mathrm{tr}} = \frac{1}{\sqrt{\alpha_p}}. \tag{84}$$

In general, we design machines in such a way as to avoid the crossing of this transition energy. As it involves changes in the RF phase unless the particles lose the longitudinal focusing created by the sinusoidal RF function, the bunch profile will be diluted and become lost. Qualitatively, the longitudinal focusing effect and the problem of the gamma transition are explained in Fig. 31.

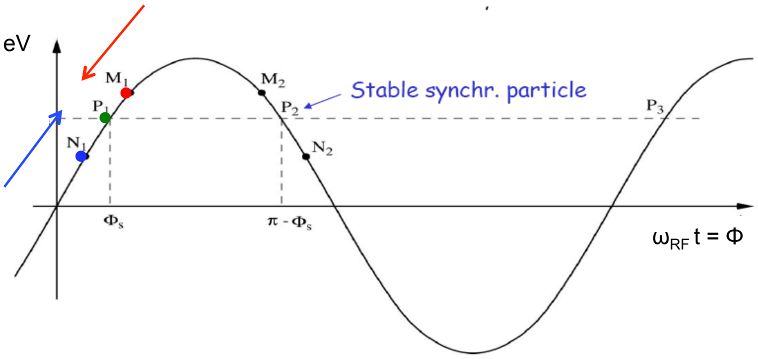


Fig. 31: Qualitative picture of the phase-focusing principle below the transition

6.2 The classical regime

Assume that an ideal particle is passing through the cavity at a certain ideal position in time (or phase), as indicated by the green spot in Fig. 31. It will see a certain accelerating voltage and, correspondingly, receive an energy increase. We call this phase the synchronous phase. A particle that has a smaller energy than the ideal value will travel at a lower speed and will arrive later after the next turn, and thus at a larger phase, and it will see a stronger accelerating voltage. It will therefore compensate the lack in energy and, step by step, come closer to the ideal particle. Just the opposite happens to a particle that has a positive energy offset. As it is faster than the synchronous particle, it will arrive at the cavity earlier and see a smaller voltage, and will again approach the ideal particle step by step. In both cases a net focusing effect is obtained, which is due to the relation between momentum and speed and the right choice of the synchronous phase. This focusing effect leads to stable longitudinal oscillations of the particles, keeping them close together or, more precisely, close to the synchronous particle, and so it forms a so-called bunch of particles in the longitudinal direction.

Here we have to pause for a moment and contemplate the situation a little: it is evident that perfect synchronization can be obtained in the case where the revolution frequency f_{r} is equal to the RF frequency f_{rf}. But it is also evident that we can again obtain a synchronous condition if f_{rf} is an integer multiple of f_{r}:

$$f_{\mathrm{rf}} = hf_{\mathrm{r}}. \tag{85}$$

We call the integer h the 'harmonic number', and it defines the number of synchronous 'locations' on the closed orbit. This is clear enough: at each of these h locations, the longitudinal-focusing principle is equally valid. Therefore we obtain h so-called 'buckets' in the machine that can be occupied by particle bunches.

For highly relativistic particles, the same effect exists but the origin of the focusing effect is now the relativistic increase in mass with energy. As visualized in Fig. 32, the high-energy particle (marked in blue) will, because of its higher mass, move on a longer orbit and, as its speed is constant ($v \approx c$), it will arrive later at the cavity location. As a consequence, the synchronous phase has to be chosen depending on whether we are running the machine below or above the transition. Synchrotrons that have to pass through the transition will have to apply a phase jump to keep the particles bunched.

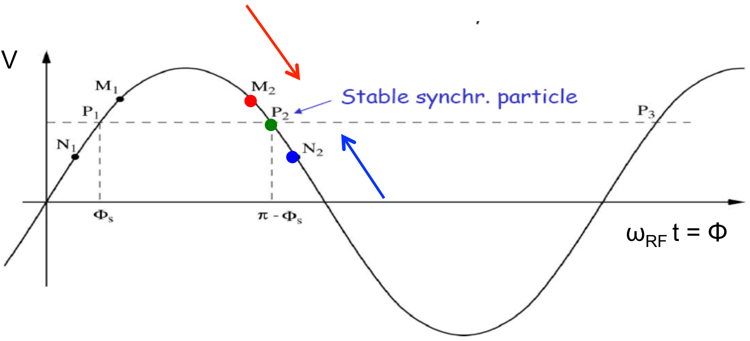


Fig. 32: Qualitative picture of the phase-focusing principle above the transition

In this context, it is worth taking a look at the acceleration mechanism itself. The particle momentum is defined via the beam rigidity by the dipole field B:

$$p = eB\rho. \tag{86}$$

As a consequence, a change in the particle momentum is reflected by an appropriate change in the B-field:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = e\rho\dot{B}. \tag{87}$$

The momentum increase per turn is therefore given by

$$(\Delta p)_{\text{turn}}\,\mathrm{d}t = e\rho\dot{B}T_{\text{r}} = \frac{2\pi e\rho R\dot{B}}{v} \tag{88}$$

and, referring to the energy change rather than the change in momentum, we obtain using

$$E^2 = E_0^2 + p^2c^2 \rightarrow \Delta E = v\,\Delta p \tag{89}$$

the change in energy per turn, which is clearly related to the accelerating voltage and the synchronous phase ϕ_{s} of the particles:

$$\Delta E_{\text{turn}} = \Delta W_{\text{turn}} = 2\pi e\rho R\dot{B} = e\hat{V}\sin\phi_{\text{s}}. \tag{90}$$

The following remarks might be worth making.

- The dipole field changes the orbit, and this leads to a change in the time (or phase) of arrival at the RF cavities and so to an accelerating effect on the whole beam.
- The energy gain depends on the rate of change of the dipole field.
- The number of stable synchronous particles is equal to the harmonic number h, which is ultimately the number of RF wavelengths that fit into the machine circumference. Thus we get h synchronous particles that are equally spaced around the circumference.

- All synchronous particles satisfy the relation $p = eB\rho$. They have the nominal energy and follow the nominal trajectory.
- As long as the particles are not fully relativistic, their revolution frequency changes, and so the RF frequency, which is a multiple of f_r, must also change to stay in synchronization during the complete acceleration process.

6.3 Frequency change during acceleration

One last comment might be useful before we can go to our workshop, take a jigsaw and hammer, and start building a machine. As soon as we start to accelerate a particle in our ring, two things will happen according to Eq. (78): the velocity will increase (which justifies the term ‘accelerator’) and, at the same time, the relativistic mass will increase via $m = \gamma \cdot m_0$. Because of the first effect, we expect a change in revolution frequency, and because of the second this velocity change will reduce and the mass effect will take over.

Now, the relation between the revolution frequency and the RF frequency is defined by the harmonic number and depends on the size of the ring and the magnetic dipole field:

$$f_\mathrm{r} = \frac{f_\mathrm{rf}}{h} = \mathrm{function}(B, R_\mathrm{s}). \tag{91}$$

Hence, using the beam rigidity relation and the average dipole field to define the radius of the ideal particle, we obtain the following for an average magnetic field $\langle B(t) \rangle$:

$$\frac{f_\mathrm{rf}(t)}{h} = \frac{v(t)}{2\pi R_\mathrm{s}} = \frac{1}{2\pi}\frac{e}{m}\langle B(t)\rangle \tag{92}$$

and

$$\frac{f_\mathrm{rf}(t)}{h} = \frac{1}{2\pi}\frac{ec^2}{E_\mathrm{s}(t)}\frac{r}{R_\mathrm{s}}B(t). \tag{93}$$

I hope it is becoming clearer that the ‘independent parameter’ that drives the particle acceleration in a synchrotron is the magnetic dipole field (even if we have to admit that a certain number of RF cavities is also useful for doing the job). Using the relativistic overall energy

$$E^2 = p^2c^2 + (m_0c^2)^2, \tag{94}$$

we finally obtain an expression for the RF frequency as a function of the changing external dipole field:

$$\frac{f_\mathrm{rf}(t)}{h} = \frac{c}{2\pi R_\mathrm{s}}\left\{\frac{B(t)^2}{(m_0c^2/ecr)^2 + B(t)^2}\right\}^{1/2}. \tag{95}$$

So all we have to do for successful operation is to put the form of our ramping B-field into Eq. (95) and adjust the frequency control of our RF system accordingly. And the machine will run automatically and be ‘synchronized’—which is where the name ultimately comes from.

At high energies, or, more accurately, when

$$B > \frac{m_0c^2}{ecr}, \tag{96}$$

the velocity increase becomes increasingly small (we get closer and closer to the speed of light), the second term becomes negligible, and the situation simplifies a quite a bit:

$$\frac{f_\mathrm{rf}(t)}{h} = \frac{c}{2\pi R_\mathrm{s}} = \mathrm{const}. \tag{97}$$

In the case of electron synchrotrons, because of the small mass of electrons and, as a consequence, the high values of γ, it is evident that the condition for Eq. (97) is nearly always fulfilled and that that relation can be applied right from the beginning: in these machines, the revolution frequency does not change by any considerable amount during acceleration and can be considered as constant. For proton and heavy-ion beams, however, even up to LHC energies (7 TeV in the case of protons), the effect has to be taken into account up to the flat-top energy and so proton and heavy-ion synchrotrons need more sophisticated RF control.

7 Synchrotron motion

Once more unto the breach, dear friends ... [16]

In the following, we shall again contemplate the longitudinal motion a little. However, we shall try to put things on a mathematically more solid basis. As shown qualitatively in Fig. 32, we expect a longitudinal oscillation in phase and energy under the influence of the focusing mechanism explained above. The relation between the relative frequency deviation and relative momentum error has been derived in (Eq. (82)):

$$\frac{\mathrm{d}f_{\mathrm{r}}}{f_{\mathrm{r}}} = \left(\frac{1}{\gamma^2} - \alpha_p\right)\frac{\mathrm{d}p}{p}, \tag{98}$$

which translates into a difference in revolution time

$$\frac{\mathrm{d}T}{T_0} = \left(\alpha_p - \frac{1}{\gamma^2}\right)\frac{\mathrm{d}p}{p} \tag{99}$$

and leads to a difference in phase on arrival at the cavity

$$\begin{aligned}
\Delta\psi &= 2\pi\frac{\Delta T}{T_{\mathrm{rf}}} = \omega_{\mathrm{rf}}\cdot\Delta T \\
&= \frac{h\cdot 2\pi}{\beta^2}\left(\alpha_p - \frac{1}{\gamma^2}\right)\frac{\mathrm{d}E}{E} \\
&= h\cdot\omega_0\cdot\Delta T = h2\pi\frac{\Delta T}{T_0} \\
&= h\cdot 2\pi\left(\alpha_p - \frac{1}{\gamma^2}\right)\frac{\mathrm{d}p}{p} \\
&= \frac{h\cdot 2\pi}{\beta^2}\left(\alpha_p - \frac{1}{\gamma^2}\right)\frac{\mathrm{d}E}{E}.
\end{aligned}$$

As before, the revolution frequency f_{r} and the RF frequency ω_{rf} are related to each other via the harmonic number h. Hence the difference in energy and the offset in phase are connected to each other through the momentum compaction factor or, more accurately, the parameter η.

Differentiating the last expression with respect to time gives the rate of change of the phase offset per turn:

$$\Delta\dot{\psi} = \frac{\Delta\psi}{T_0} = \frac{h2\pi}{\beta^2 T_0}\left(\alpha_p - \frac{1}{\gamma^2}\right)\frac{\mathrm{d}E}{E}. \tag{100}$$

This expression tells us about the rate of change of the phase of a particle as a function of its changing energy.

On the other hand, the difference in energy gain of an arbitrary particle that has a phase distance of $\Delta\psi$ from the ideal particle is given by the voltage and phase of the RF system (a trivial statement but worth mentioning):

$$\Delta E = e\cdot U_0(\sin(\psi_{\mathrm{s}} + \Delta\psi) - \sin\psi_{\mathrm{s}}). \tag{101}$$

As before, we describe the phase of the ideal ('synchronous') particle by ψ_s and the phase difference by $\Delta\psi$. For small amplitudes $\Delta\psi$ of the phase oscillations, we can simplify the treatment by assuming

$$\sin(\psi_\mathrm{s} + \Delta\psi) - \sin\psi_\mathrm{s} = \sin\psi_\mathrm{s}\cos\Delta\psi - \cos\psi_\mathrm{s}\sin\Delta\psi - \sin\psi_\mathrm{s}. \tag{102}$$

For small amplitudes $\Delta\psi$, we can make the approximation

$$\sin\Delta\psi \approx \Delta\psi, \qquad \cos\Delta\psi \approx 1, \tag{103}$$

and obtain the following for the rate of energy change per turn:

$$\Delta\dot{E} = e \cdot \frac{U_0}{T_0}\Delta\psi\cos\psi_\mathrm{s}. \tag{104}$$

A second differentiation with respect to time delivers

$$\Delta\ddot{E} = e \cdot \frac{U_0}{T_0}\Delta\dot{\psi}\cos\psi_\mathrm{s}. \tag{105}$$

Combining Eqs. (100) and (105), we finally get a differential equation for the longitudinal motion under the influence of the phase-focusing mechanism:

$$\Delta\ddot{E} = e \cdot \frac{U_0}{T_0}\frac{2\pi h}{\beta^2 T_0}\left(\alpha_p - \frac{1}{\gamma^2}\right)\frac{\mathrm{d}E}{E}\cos\psi_\mathrm{s}. \tag{106}$$

For a given energy, the parameters in front of the right-hand side are constant and describe the longitudinal, or 'synchrotron', oscillation frequency. Therefore, using

$$\Omega = \omega_0 \cdot \sqrt{\frac{-eU_0 h\cos\psi_\mathrm{s}}{2\pi\beta^2 E}\left(\alpha_p - \frac{1}{\gamma^2}\right)}, \tag{107}$$

we get the equation of motion in the approximation of small amplitudes:

$$\Delta\ddot{E} + \Omega^2\,\Delta E = 0. \tag{108}$$

This describes a harmonic oscillation in (E–ψ) phase space of the difference in energy of a particle from the ideal (i.e., synchronous) particle under the influence of the phase-focusing effect of our sinusoidal RF function.

As already discussed qualitatively, the expression in Eq. (107) leads to real solutions if the argument of the square root is a positive number. Two possible situations therefore have to be considered: below the gamma transition the η-parameter is positive, and above it it is negative. The synchronous phase (which is the argument of the cosine function) therefore has to be chosen to get an overall positive value under the square root:

$$\begin{array}{lll} \gamma < \gamma_\mathrm{tr} & \eta > 0, & 0 < \psi_\mathrm{s} < \pi/2, \\ \gamma > \gamma_\mathrm{tr} & \eta < 0, & \pi/2 < \psi_\mathrm{s} < \pi. \end{array}$$

This, finally, is the mathematical description of the classical and relativistic regimes that defines stable conditions of synchrotron motion, as explained qualitatively in Figs. 31 and 32. Figure 33 shows as an example the superconducting RF system of the LHC, and the basic parameters of this system, including the synchrotron frequency, are listed in Table 1.

Figure 34, finally, shows the longitudinal focusing effect that was observed during the commissioning phase of the LHC. On the left-hand side, beam had been injected into the storage ring while the RF system still was switched off. The bunch, nicely formed by the RF voltage of the pre-accelerator, is visible for only a few turns and the bunch profile decays rapidly, as no longitudinal focusing is active. The right-hand side shows the situation with the RF system activated and the phase adjusted. The injected particles stay nicely bunched and the acceleration process can start.

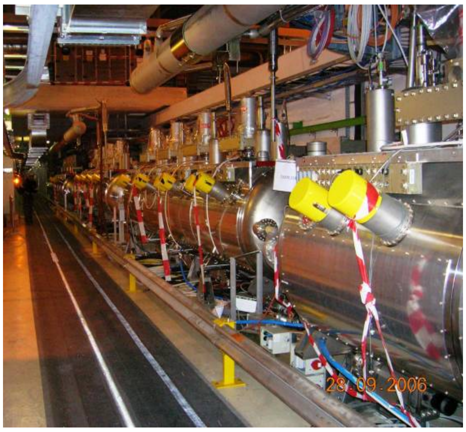


Fig. 33: LHC RF system

Table 1: Parameters of the LHC RF system

Bunch length (4σ)	1.06 ns
Energy spread (2σ)	0.22×10^{-3}
Number of stored bunches	2808
RF frequency	400 MHz
Harmonic number h	335 640
RF voltage per beam	16 MV
Energy gain per turn	485 keV
Synchrotron frequency	23 Hz

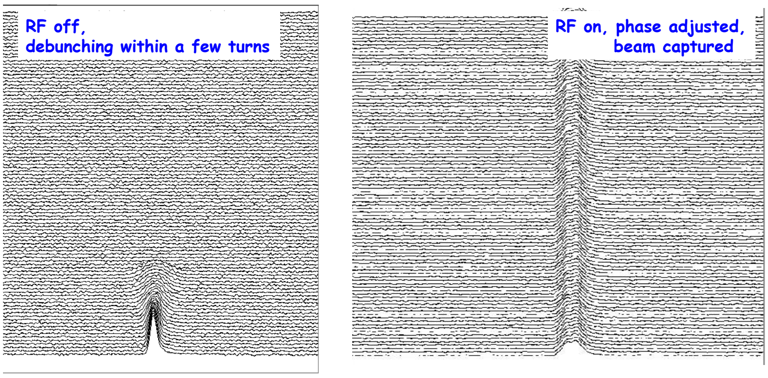


Fig. 34: Left: injection into the LHC while the RF is switched off: the bunches decay in a small number of turns. Right: the RF is switched on and phased: the bunches remain long, focused, and stable.

References

[1] LHC design report, edited by O.S. Brüning et al., CERN-2004-003 (CERN, Geneva, 2004), http://dx.doi.org/10.5170/CERN-2004-003-V-1, http://dx.doi.org/10.5170/CERN-2004-003-V-2, http://dx.doi.org/10.5170/CERN-2004-003-V-3.

[2] K. Wille, *The Physics of Particle Accelerators* (Oxford University Press, Oxford, 2000).

[3] J. Rossbach and P. Schmueser, in Proceedings of the CAS–CERN Accelerator School, 5th General Accelerator Physics Course, Jyväskylä, Finland, 7–18 September 1992, edited by S. Turner, CERN-1994-001 (CERN, Geneva, 1994), pp. 17–88. http://dx.doi.org/10.5170/CERN-1994-001.17

[4] B. Holzer, in Proceedings of the CAS–CERN Accelerator School on Superconductivity for Accelerators, Erice, Italy, 24 April–4 May 2013, edited by R. Bailey, CERN-2014-005 (CERN, Geneva, 2014), pp. 21–40. http://dx.doi.org/10.5170/CERN-2014-005.21.

[5] E. Jaeschke *et al.*, The Heidelberg test storage ring for heavy ions TSR, Proc. EPAC, Rome, 1988.

[6] H. Goldstein, *Klassische Mechanik* (Akademische Verlaggesellschaft Wiesbaden, 1981).

[7] HERA Design Team, HERA: a proposal for a large electron proton colliding beam facility at DESY, DESY-HERA-81/10 (1981).

[8] A. Piwinski, in Proceedings of the CAS–CERN Accelerator School, Advanced Accelerator Physics Course, Santa Margherita di Pula, Italy, 31 Jan–5 Feb 1985, edited by S. Turner, CERN-1987-003-V-1 (CERN, Geneva, 1987), pp. 187–202. http://dx.doi.org/10.5170/CERN-1987-003-V-1.187.

[9] J.D. Cockcroft and E.T.S. Walton, *Proc. R. Soc. (Lond.)* **A136**(830) (1932) 619. http://dx.doi.org/10.1098/rspa.1932.0107

[10] J.D. Cockcroft and E.T.S. Walton, *Proc. R. Soc. (Lond.)* **A137**(831) (1932) 229. http://dx.doi.org/10.1098/rspa.1932.0133

[11] J.D. Cockcroft and E.T.S. Walton, *Proc. R. Soc. (Lond.)* **A144**(853) (1934) 704. http://dx.doi.org/10.1098/rspa.1934.0078

[12] R.J. Van de Graaff, *Phys. Rev.* **38** (1931) 1919.

[13] P.J. Bryant, in Proceedings of the CAS–CERN Accelerator School, 5th General Accelerator Physics Course, Jyväskylä, Finland, 7–18 September 1992, edited by S. Turner, CERN-1994-001 (CERN, Geneva, 1994), pp. 1–16. http://dx.doi.org/10.5170/CERN-1994-001.1.

[14] F. Tecker, Longitudinal beam dynamics, Proc. CAS–CERN Accelerator School, Advanced Accelerator Physics Course, Trondheim, Norway, 18–29 Aug 2013, edited by W. Herr, CERN-2014-009 (CERN, Geneva, 2014), pp. 1–21. http://dx.doi.org/10.5170/CERN-2014-009.1.

[15] J. LeDuff, in Proceedings of the CAS–CERN Accelerator School, 5th General Accelerator Physics Course, Jyväskylä, Finland, 7–18 September 1992, edited by S. Turner, CERN-1994-001 (CERN, Geneva, 1994), pp. 289–311. http://dx.doi.org/10.5170/CERN-1994-001.289.

[16] W. Shakespeare, *Henry V*.

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Medical Physics Commissioning

D. Meer
Paul Scherrer Institut, Villigen, Switzerland

Abstract
Medical commissioning is an important step in bringing a particle gantry into clinical operation for tumour treatments. This involves the parametrization and characterization of all relevant systems including beam delivery, patient table, imaging systems and connection to all required software components. This article is limited to the necessary tasks for the beam delivery system of a pencil beam scanning system. Usually, commissioning starts with the characterization of the unscanned beam and calibration of the beam energy. The next steps are the parametrization of the scanning system, the commissioning of the beam position monitoring system and characterization of the spot size, which all require precision better than 1 mm. The commissioning effort for these tasks also depends on the gantry topology. Finally, calibration of the dose measurement system ensures that any dose distribution can be delivered with an absolute precision better than 1%.

Keywords
Proton therapy; pencil beam scanning; commissioning; calibration; medical physics.

1 Introduction

Proton therapy systems are complex and commissioning is an important step in bringing such a facility into clinical operation. Medical physics commissioning is the intermediate step between technical commissioning and medical acceptance. Technical commissioning includes the installation and functional testing of system components, e.g., the beam line magnets or the rotating mechanical gantry structure.

Within *medical physics commissioning*, the main focus is on integral system tests and parametrization of the machine characteristics. The correct interaction between different system components is verified and important machine parameters are measured. This task not only consists of checking the beam scanning system and monitoring system, but also includes checking the mechanical systems (gantry rotation), the patient positioning systems, the different imaging systems and the transformation of coordinate system between them, as well as the software connection to the treatment planning software (TPS) and oncology information system. Finally, a properly configured safety system ensures a safe operation. After this important step, the system must be able to precisely deliver a predefined dose distribution with a given dose to a target.

What follows is medical acceptance of the whole treatment unit when operated by a responsible end-user. The system specifications are validated by numerous tests, performed by medical physicists. Many of these tests have end-to-end character and successful completion is a prerequisite before clinical operation can be started.

There is no strict separation between technical and medical commissioning, and responsibilities between producer and end-user might shift from one installation to another. It also makes a difference whether an established commercial product is installed in a clinical environment or a system is developed for clinical application at a research institute. In this report we look at the latter case, mainly reporting on the experience gained in the commission of the second gantry at the Paul Scherrer Institut (PSI). Gantry 2 [1] is an active scanning system and we therefore only concentrate on pencil beam scanning

https://doi.org/10.23730/CYRSP-2017-001.285

(PBS) rather than passive scattering. We mainly focus on the dose delivery system, although medical commissioning comprises much more than this, as mentioned before. In the first section, we discuss the commissioning of the unscanned beam. The following section explains the lateral beam spreading with the scanning system and the last section describes the commissioning tasks for the dose-monitoring system.

2 Commissioning the unscanned beam

Almost all particle therapy systems use beam transfer lines to transport the particles from the source (the accelerator) to the treatment rooms. Dipole and quadrupole magnets along the beam line help to direct and focus the particle beam to the isocentre. Additionally, a collimator, mechanical slits or scattering foils help to create the required phase space at the treatment location. Today, the vast majority of PBS systems do not use local energy modulation right in front of the patient but rely on variable beam energy from the accelerators or on upstream energy degrading systems.

2.1 Beam line tuning

To cover the full clinical range in depth, at least 50 different beam energies must be transported from the accelerator to the isocentre, and a beam line usually contains more than 30 devices. To optimize the beam line settings, well proven tools like TRANSPORT [2] or Mad-X [3] are available.

From a particular solution for the magnet settings for a given energy, beam line settings for other energies can be obtained by scaling the magnet currents with the particle momentum. This approach also gives smooth behaviour of the beam characteristics over the entire energy range. However, a linear model, as used in TRANSPORT, cannot describe the entire beam line, which also contains vacuum windows, collimators and drift distances in air or energy-absorbing materials. In order to obtain a full beam line description and a qualitative number for the beam line transmission, these models must be combined with Mote Carlo methods. Several beam profile monitors along the beam line can help to validate the predictions from such models.

To complete the beam tuning, some manual optimization with steering magnets is necessary in order to centre the beam. It is of particular importance that the beam is well centred on the mechanical rotation axis of the gantry. This procedure can be simplified if a position monitor on the still horizontal but rotating part of the gantry is available and data from this monitor can be read while the beam is turned on and the gantry is rotating. If a circle is visible on this position monitor, the beam centring can still be improved as shown in Fig. 1. Steering magnets on the gantry also help to centre the beam. If necessary, gantry angle dependent solutions must be found.

2.2 Energy calibration

The range of the particles in matter (water) is one of the crucial parameters for TPS and needs to be adjusted with an absolute precision better than ± 0.5 mm. Nevertheless, the calculations of beam tuning tools are based on beam momentum and, due to several factors, the comparison of the converted particle range with the measured range at the beam line often shows not negligible differences. Therefore, energy-to-range calibration is necessary and is performed with help of Bragg peak measurements. The range of a particle beam is defined as the 80% point of the maximum (R_{80}) in the distal fall-off of the Bragg peak. By this definition, the range is insensitive to different momentum spread [4].

The preferred tools for measuring range in water are so-called range scanners. They consist of two dose-measurement chambers (plain parallel ionization chambers (ICs)): one is fixed at the entrance to the measurement phantom and the second one moves at depth. The ratio between the two chambers is calculated at different depths to obtain the depth–dose curve. Commercial measurement tools such as the PEAKFINDER Water Column from PTW-Freiburg, Germany, are available for such measurements. The typically requested precision for range measurements is better than 0.2 mm, which corresponds

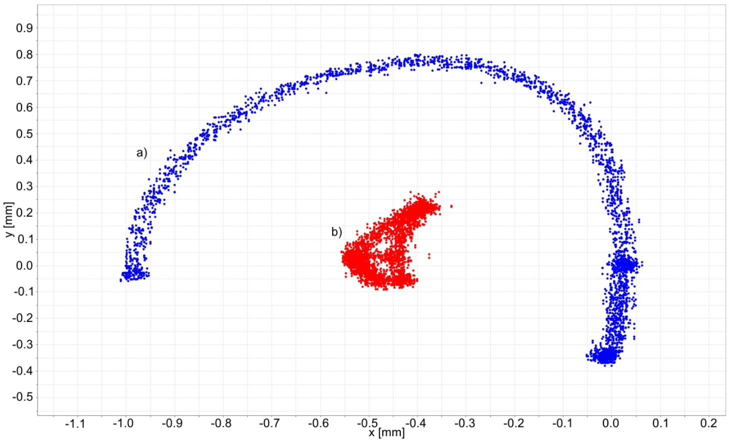


Fig. 1: Beam centring at the beginning on the rotating part of the gantry. The beam position is recorded continuously while the gantry rotates. The initial beam positions (a) and beam positions after proper centring (b) are shown. Maximum deviation is about $350\ \mu$m after optimization.

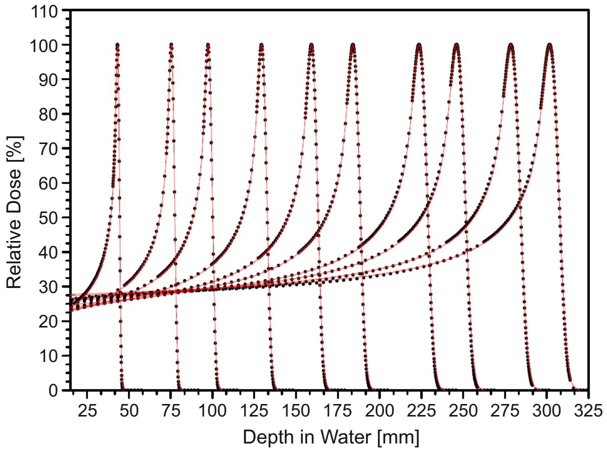


Fig. 2: Measured integral depth–dose curves for 10 different beam energies. The adaptive measurement granularity provides high resolution and efficient data acquisition.

approximately to a resolution of 0.1 MeV. At PSI, we have developed a water scanner which only consists of the movable chamber. We have access to the last dose monitor on the gantry which is used for data normalization.

For determination of the R_{80}, it is sufficient to only measure the region around the Bragg peak with high resolution. However, the integral depth–dose profile is an important input for the TPS and must also be measured with high precision. As shown in Fig. 2, an adaptive granularity can help to accelerate measurements.

2.3 Beam line supervision

In systems with variable beam energy, the beam line settings are changed with high frequency and online supervision helps to verify the correct settings. Of particular importance are the dipoles of the energy-selection system, collimators and slits. The correct settings for the dipoles can be verified by an independent measurement of the magnetic field with Hall probes. The signal of the Hall probe is proportional to the B field

$$V_{\mathrm{HallProbe}} \propto B \propto pc = \sqrt{E_{\mathrm{kin}}^2 + 2E_{\mathrm{kin}}E_0}\ , \tag{1}$$

and can be parametrized as a function of beam energy E,

$$V_{\mathrm{HallProbe}} = a_1\sqrt{(E-a_2)^2 + 2E_0(E-a_2)} + a_3 \tag{2}$$

with three parameters a_1, a_2 and a_3. At PSI, the stability of the electronics limits the energy resolution of the beam line supervision to about 1 MeV, corresponding to approximately 2 mm range uncertainty in water.

3 Calibration of the scanning system

Once the unscanned beam has been tuned, commissioning of the scanning system can be started. First, the impact of the gantry topology on the calibration of the scanner magnets is discussed.

3.1 Up-stream and down-stream scanning

The location of the scanner magnets on a PBS gantry has significant implications for both gantry design and commissioning. In a conventional design, the scanner magnets are the last devices in the beam line as shown in Fig. 3(b). In this so-called *down-stream scanning*, the beam direction is divergent while the beam is scanned laterally. The scanner magnets are placed at least 2 m away from the isocentre since a small source-to-axis distance (SAD) is unfavourable, leading to an increased skin dose, for example. To first order there is a linear correlation between the spot position at the isocentre and the scanner magnet current. Another advantage is that the spot shape is unaffected for different scan positions. On the other hand, the calibration of the scanning system is sensitive to the longitudinal alignment of the device for the position measurement in the case of a finite SAD.

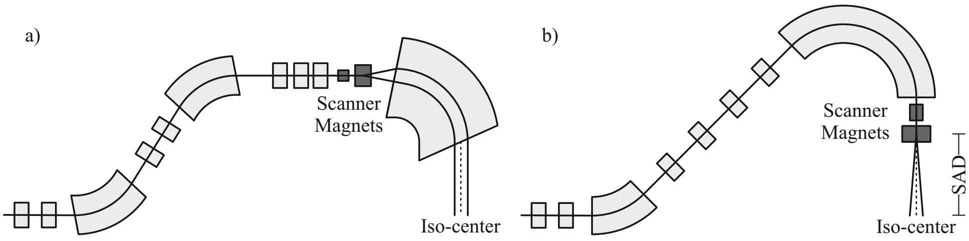


Fig. 3: Two possible gantry topologies with up-stream (a) and down-stream (b) scanning

Up-stream scanning is the other gantry topology in which the scanner magnets are installed before the last bending dipole as shown in Fig. 3(a). In this configuration, a parallel beam (infinite SAD) can be obtained by properly optimizing the beam optics and the last bending dipole. A parallel scanned beam at the isocentre simplifies treatment planning and measurements for the quality assurance program. However, there are also additional technical challenges: beam-energy dependent inhomogeneities and fringe fields of the dipole can affect spot shape at the isocentre and beam focus needs to be dynamically adjusted when scanning the beam in the divergent plane. Consequently, higher-order corrections for position-to-current conversion need to be considered as well. PSI Gantry 2 [5] is one of the few upstream scanning systems, and the calibration of that is discussed in the following section.

3.2 Scanner magnet calibration

The important prerequisite for accurate calibration of the scanner magnets is the availability of an appropriate device to measure the beam position at the isocentre for the full scanning range and different gantry angles. We decided to place a copy of the nozzle strip chamber on a rotatable support at the isocentre (see Fig. 4 (right)). As a consequence of a parallel beam, the scan range at the isocentre is the same as in the nozzle, and we could therefore use a spare position monitor from the nozzle. This has the additional advantage that the standard read-out from the therapy control system (TCS) can be used, and data from the isocentre measurement can be integrated easily into the TCS logging file. The rotational support can turned remotely and integrates a beam dump which greatly simplifies the measurement at different gantry angles.

An alternative setup is to use a measurement device that can be attached to the gantry nozzle. In that case, the deformation of the supporting structure for different gantry angles needs to be considered.

The result of the sweeper calibration is two functions which provide the scanner magnet current

$$I_{\mathrm{S}}^{x} = f^{x}(E,x,y)\ ,\ I_{\mathrm{S}}^{y} = f^{y}(E,x,y) \tag{3}$$

as a function of beam energy and lateral position at the isocentre for both scanner magnets. These functions can be look-up tables with additional interpolation or an analytic function. We decided to use polynomial functions that would be less sensitive to measurement errors and to give a smooth position dependency. The functions are calculated for every 10 MeV and interpolation is used for intermediate energies. Unfortunately, it was not possible to find one single polynomial expression for the full scanning range because each corner of the scan range showed substantial and specific deviation from the linear position to current correlation. To better consider the peculiarity of each corner, the global polynomial model was expanded with a local solution for each quadrant:

$$\begin{aligned} I_{\mathrm{S}}^{x}(x,y) = {} & \overbrace{g_1x + g_2y + g_3x^2 + g_4y^2 + g_5y^3 + g_6y^4}^{\text{Global function}} + \\ & \sum_{q=1}^{4} \underbrace{l_1^q xy + l_2^q x^2y + l_3^q xy^2 + l_4^q x^3y + l_5^q x^2y^2 + l_6^q xy^3 + l_7^q x^4y + l_8^q x^3y^2 + l_9^q x^2y^3 + l_{10}^q xy^4}_{\text{Local function for each quadrant}} \end{aligned} \tag{4}$$

and an analogue expression for I_{S}^{y}.

The ansatz with only mixed terms in x and y for the local solution ensures that the local contribution is vanishing when $x \to 0$ or $y \to 0$, and this provides a smooth function of position. The position to current calibration with this model requires 46 parameters for one beam energy. This is still much less data than a look-up table to deliver beam spots with a precision in the order of 100 μm.

3.3 Position projection to the isocentre

During dose delivery with active scanning, the position of the pencil beam is continuously monitored with a position monitor. Segmented ICs are the monitors of choice for minimizing the additional amount of material in the beam path. However, they measure the beam in the nozzle approximately 1 m before the isocentre and therefore the measured position needs to be projected to the isocentre. This requires profound knowledge of the beam angle for the full scanning range and all beam energies. An uncertainty in the beam angle of approximately 1 mrad already leads to a position measurement error of 1 mm (see Fig. 4 (left-hand side)). The same measurement setup as for the scanner magnet calibration can be used.

3.4 Spot size measurements

In addition to the depth–dose curve, the spot size is the other important input parameter for the TPS. The TPS can work with a single spot size for a given energy as long as spot size variations are negligible over the entire scan range. To assess this assumption a two-dimensional spot shape measurement needs to be performed at the isocentre over the full scan and proton energy range. At PSI, we are using a system

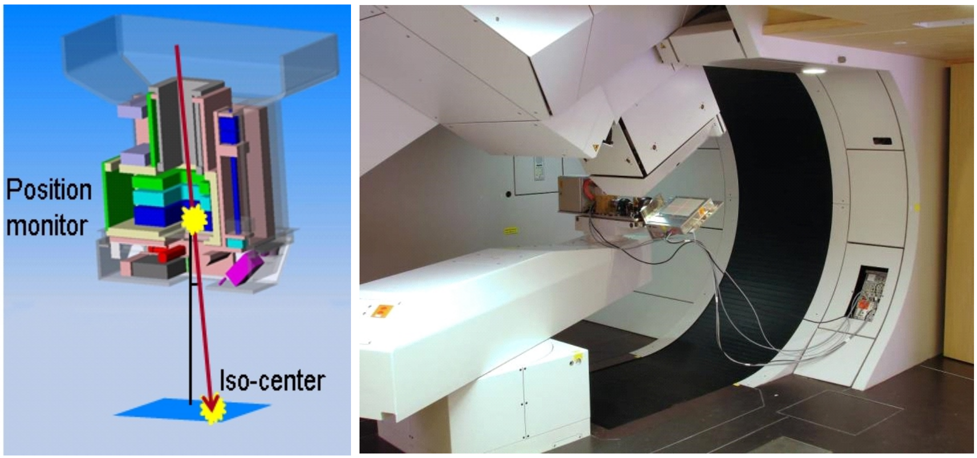


Fig. 4: Precise data on the beam angle is required to predict the position at the isocentre from the measured position in the nozzle chamber (left-hand side). At PSI, a second strip chamber was positioned at the isocentre to measure the correlation (right-hand side). The same setup is also used for the calibration of the scanner magnets.

which is based on a CCD camera looking at a scintillating foil, which provides high spatial resolution (see Fig. 5 (right-hand side)). Similar commercial equipment is available, such as the Lynx detector from IBA Dosimetry in Schwarzenbruck (Germany), for example. The spot size can be extracted from a delivered spot pattern as shown in Fig. 5 (left-hand side). Multiple spots in one single data acquisition increase the measurement efficiency as long as the spots are well separated and the tails of the spots are not overlapping. A two-dimensional Gaussian is fitted for each spot to determine the spot size, and the required parameters for the TPS are extracted, depending on the beam model in the TPS.

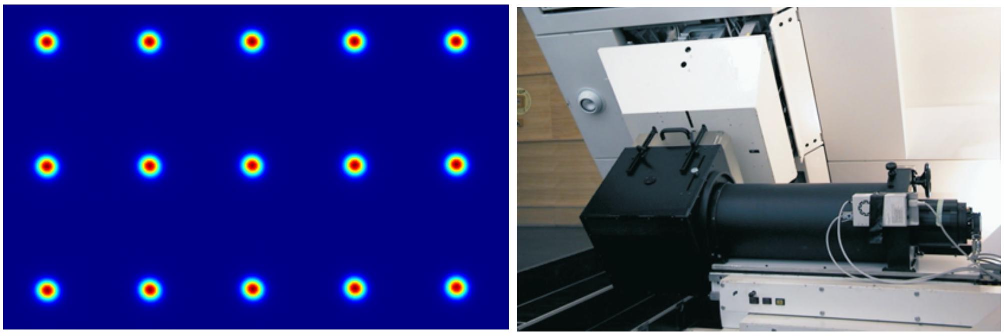


Fig. 5: Typical spot pattern for spot size determination for a beam energy of 150 MeV (left-hand side). The device for this measurement is based on a scintillating screen and a CCD camera (right-hand side).

4 Dose monitor calibration

For a correct dose distribution, the exact dose per spot determination is as essential as the spatial location of the spot, and the number of delivered protons must be detected with high accuracy. ICs with a minimal amount of material are again the preferred measurement devices because the dose monitor stays permanently in the beam path. The collected charge from the IC

$$Q_{\mathrm{IC}} = G_{\mathrm{IC}} \cdot S(E) \cdot \frac{1}{k_{Tp}} \cdot q_e \cdot n_p \tag{5}$$

is proportional to the number of protons n_p and the stopping power $S(E)$ for a given energy. The factor k_{Tp} corrects for temperature and pressure other than standard conditions. The theoretical calculated value for the IC gain G_{IC} based on chamber geometry has large uncertainties and therefore experimental determination is required. At PSI, we use a Faraday cup for this calibration. The charge measured with the Faraday cup is $Q_{\mathrm{FC}} = q_e n_p$ and can be used for direct gain determination

$$G_{\mathrm{IC}} = \frac{Q_{\mathrm{IC}} \cdot k_{Tp}}{S(E) \cdot Q_{\mathrm{FC}}} \, . \tag{6}$$

There is also an alternative calibration procedure based on a measurement with a small IC [6].

In general, dose distributions are calculated with pencil beam models which predict the dose per incident proton [7]. The IC calibration is needed to convert the number of protons from a pencil beam to the expected IC signal. The validity of the IC calibration is verified by measuring the dose delivered to a water phantom with a certified thimble IC following a code of practice [8]. At PSI, we found an agreement of 2–3% between the dose calculated with the beam model and absolute dosimetry measurements. An additional empirical correction factor reduces this difference to less than 0.5%.

5 Summary and conclusion

It is far beyond the scope of this article to give a complete description of medical commissioning. Important topics such as calibration of the imaging systems or the patient table were not touched. The goal of this text is to illustrate the methodology for some of the major tasks.

After successful medical commissioning, the system is validated by acceptance tests with similar characteristics and is then ready for patient treatment. After clinical operation has started, the performance of the system is continuously monitored by quality assurance (QA) tests. These tests have different repetition periods ranging from daily to yearly tests. The careful execution of a QA program is necessary to guarantee clinical operation with constant high quality over many years.

Acknowledgements

The author would like to thank all members of the Center for Proton Therapy at PSI who contributed to this work.

References

[1] E. Pedroni *et al.*, *Z. Med. Phys.* **14** (2004) 25. http://dx.doi.org/10.1078/0939-3889-00194

[2] K.L. Brown, *et al.*, Transport, a Computer Program for Designing Charged Particle Beam Transport Systems, CERN 80-04 (1980). http://dx.doi.org/10.5170/CERN-1980-004

[3] W. Herr and F. Schmidt, in Proceedings of the CAS-CERN Accelerator School: Intermediate Course on Accelerator Physics, Zeuthen, Germany, 15–26 september 2003, CERN-2006-002 (CERN, Geneva, 2004), pp. 505-528, http://dx.doi.org/10.5170/CERN-2006-002.505

[4] W.C. Hsi *et al.*, *Med. Phys.* **36** (2009) 2297. http://dx.doi.org/10.1118/1.3132422

[5] E. Pedroni, *et al.*, *Eur. Phys. J. Plus* **126** (2011) 66. http://dx.doi.org/10.1140/epjp/i2011-11066-0

[6] C. Gomà *et al.*, *Phys. Med. Bio.* **59** (2014) 4961. http://dx.doi.org/10.1088/0031-9155/59/17/4961

[7] E. Pedroni, *et al.*, *Phys. Med. Biol.* **50** (2004) 4961. http://dx.doi.org/10.1088/0031-9155/50/3/011

[8] P. Andreo *et al.*, Absorbed Dose Determination in External Beam Radiotherapy, International Atomic Energy Agency, Technical Reports Series No. 398 (2000)

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

(The) Future (of) Synchrotrons for Particle Therapy

J. Flanz
Massachusetts General Hospital, Boston, MA, USA; Harvard Medical School, Cambridge, MA, USA

Abstract
The field of particle therapy is quickly growing and yet its more widespread adoption is limited by size, cost, and the need for adaptation to more conformal treatment techniques. In order to realize the benefits of this modality the equipment used to generate and deliver the beam is evolving. The accelerator is one of the key components of this equipment, and its future will be dictated by its ability to accommodate clinical requirements. This lecture is intended to provide an introduction to these requirements and identify how synchrotrons are designed to deliver the desired beams, as well as what limitations exist, and expectations for the future of synchrotrons.

Keywords
Particle therapy; synchrotron; dose; extraction; beam parameters; clinical parameters; pencil; crayon; scanning.

1 Introduction

This lecture is intended to offer a perspective on the subject of synchrotrons for particle therapy, about where the field is now and where it needs to go. These questions are answered by reference to an analysis of the requirements of particle therapy.

We start with our present view of the near future. The following are some of the recent themes that have been driving the development of particle therapy.

- *Beam scanning* (‘pencil’ or ‘crayon’, PBS). The method of choice for spreading the beam is beam scanning: more particularly, using magnetic fields to move the beam across the target, thus ‘painting’ the desired area. The size of the ‘brush’ is the beam size, which is strongly related to the properties of the largely unperturbed beam emerging from the accelerator (and the subsequent focusing systems). The depth of penetration of the beam is primarily determined by the beam energy.
- *Image-guided radiation therapy* (IGRT). The beam position is determined by the use of imaging technology of some sort. For moving targets, the beam properties may require adjustment by feedback during the motion. With protons, in particular, it is possible to image anatomy and directly determine the effective stopping power along the path to the target. Proton radiography and tomography depend upon the ability of the beam to penetrate the patient, and thus require an appropriate beam energy.
- *Adaptive radiotherapy*. Imaging techniques and treatment planning must evolve to a point where a target today that has a different geometry from yesterday (or a minute ago) can be effectively treated. The treatment parameters need to be modified almost on-the-fly. This has implications not only for beam delivery but also for quality assurance.
- *End of range*. Currently, there is some uncertainty in the range of the particles in the patient. This uncertainty results from errors in conversion from X-ray-based imaging and from organ motion or redistribution. Such range information can potentially be obtained more accurately

https://doi.org/10.23730/CYRSP-2017-001.293

using particle-based imaging or other on-line detection methods, which would then require adjustment of the delivered beam energy during delivery.

- *Ions*. It has been suggested that the treatment of a single tumour could benefit from the use of multiple particles with different values of linear energy transfer, delivered during a single irradiation.
- *Effective cost*. It is a continuing concern that the capital investment is higher for particle facilities than for some other modalities. The basis of that conclusion may be from inappropriate comparisons. In any case, the goal must be to achieve a cost balance in terms of capital investment, patient throughput, and treatment efficacy and accuracy so as to be competitive with other modalities.

Consideration of the above goals of particle therapy, as well as the specific clinical requirements placed on the beam parameters, should be factored into the requirements for the accelerator.

2 Flow of requirements

In any discussion of the future of synchrotrons, the above goals should be kept in mind. As part of this, one must clearly define the clinical beam requirements and determine which ones are related to the accelerator design. For some parameters, the characteristics of the accelerator are critical to the beam delivery process, and for other parameters they are almost irrelevant. One must design the accelerator to achieve all the desired clinical goals, not (as in the past) take an existing accelerator and figure out how to apply it to some clinical goals.

The key goals of radiotherapy are:

- to deliver the required dose;
- to deliver that dose with a prescribed dose distribution;
- to deliver that dose in the right place.

The beam delivery system, which is in between the accelerator and the patient, will play a role in how the safe delivery of clinical beam parameters are related to the accelerator parameters. As an example of this, Table 1 shows a few possible parameters and the flow of values from the clinical values to the beam parameters and then to the accelerator parameters that are involved for the case of a beam scanning delivery system.

Table 1: Sample of flow from clinical values to accelerator parameters

Clinical **parameter**	Sample **clinical value**	Beam **parameter**	Accelerator **parameter**
Dose rate	1 Gy/L min	~100 × 10^9 protons/min	Beam current
Range	32 cm (in water)	226.2 MeV protons	Beam energy
Scanned-beam penumbra	80% to 20% fall-off = 3.4 mm (in air)	3 mm sigma ($e^{-1/2}$ for a Gaussian beam)	Beam size, beam emittance

As implied by Table 1 and the above text, there is a flow from the clinical requirements and safety requirements to the accelerator requirements. This is depicted in Fig. 1. Starting from any position in the chart other than the top will likely result in compromised treatment parameters.

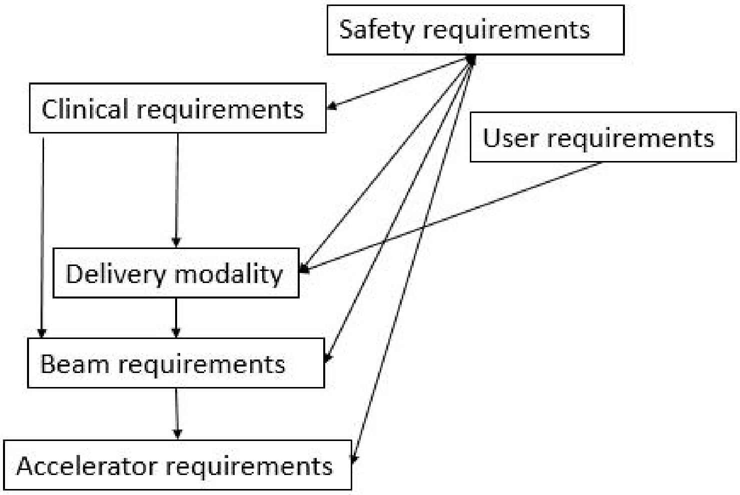


Fig. 1: Flow of requirements

3 Beam delivery modalities

This lecture is not intended to describe the details of beam delivery modalities for therapy; however, to obtain some basis for the accelerator requirements, it is useful to understand some aspects of this technology.

3.1 Beam scattering

A beam-scattering system uses the effects of multiple scattering when a beam passes through a material to spread the beam from the unperturbed ‘pencil’ to a beam size consistent with the target size. In a double scattering system (a sample of which is shown in Fig. 2), two scatterers with a special profile are used to create a spread-out beam with a uniform transverse distribution as shown in Fig. 3 (lower curve). The beam is also spread longitudinally, to obtain a Spread Out Bragg Peak (SOBP) by selectively degrading the beam energy in the correct proportions to create a flat, spread-out longitudinal distribution as shown in Fig. 3 (upper curve). The properties of the input beam must be tailored to these requirements. However, in this case, because of scattering effects, practically speaking only the initial beam energy is relevant to the beam delivery unless the SOBP requires current modulation to adjust the relative amplitudes of the Bragg peak. The tolerance of the beam position is another factor that needs to be controlled in this beam delivery scenario.

3.2 Beam scanning

A beam-scanning system uses magnetic deflection, as shown in Fig. 4, to move the unperturbed beam across the target cross-section, thus spreading out the beam. The unperturbed beam is characterized by a Gaussian profile as in Fig. 5(b), the lower curve. The Gaussian profile is integrated as the beam is moved across the target. Longitudinally the beam profile is that of a Bragg peak as in Fig. 5(a), the upper curve. The energy of the beam is varied, thus adjusting the beam range in such a way as to obtain the desired longitudinal dose distribution.

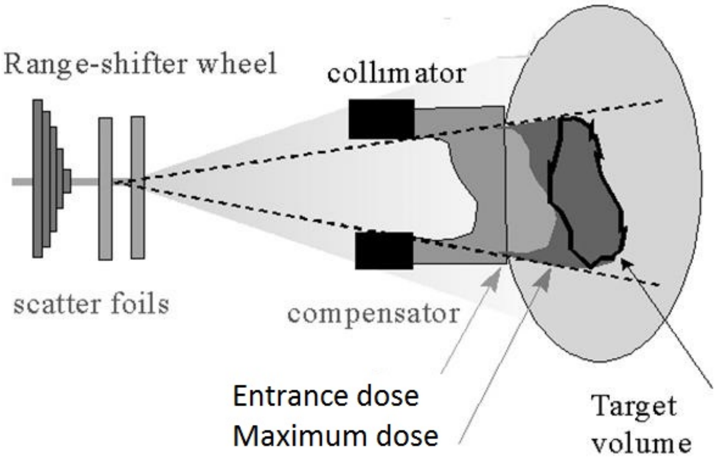


Fig. 2: Components of a scattering nozzle

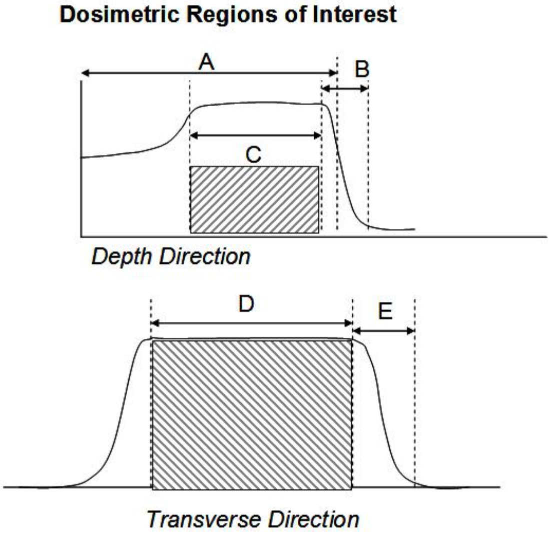


Fig. 3: Dosimetric quantities for scattered beams: depth dose (top), transverse dose (bottom)

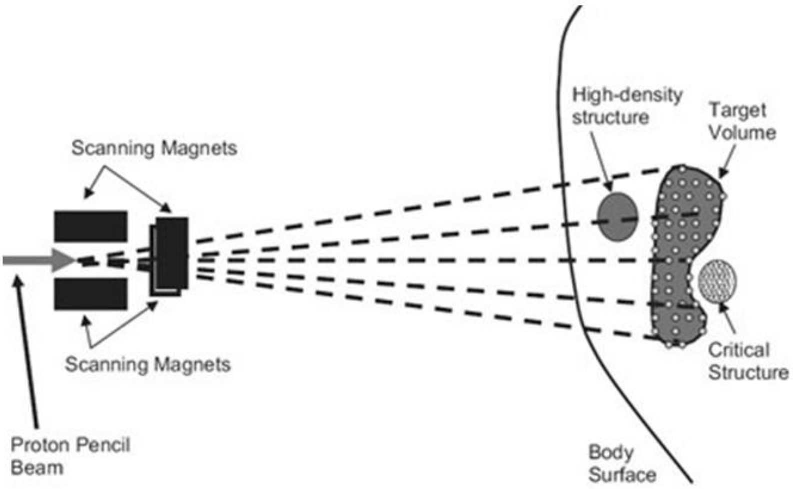


Fig. 4: Components of a scanning nozzle

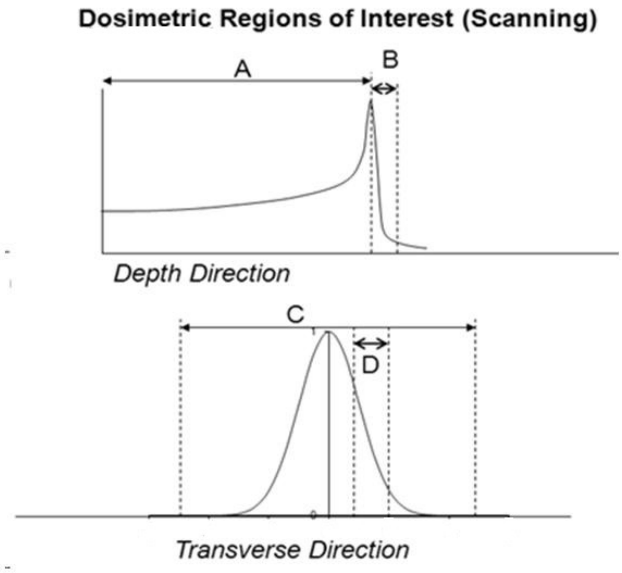


Fig. 5: Dosimetric quantities for scanned beams: depth dose (top), transverse dose (bottom)

4 Dose and dose rate

The correspondence between an accelerator parameter and a clinical parameter is typically not simple; however, crude estimates may enable an order-of-magnitude value to be determined and used for preliminary accelerator design.

4.1 Conversion from beam parameters to clinical parameters

Here is an example of this conversion. We start with the assumption of an accelerator beam of 150 MeV protons and a beam current of 1 nA. This beam is assumed to be incident (via a scanning system) directly (without modification or loss) on a target. Therefore:

- power = joules/second = energy × current
  - e.g., 150 MeV × 1 nA = 0.15 W
- dose = joules/kg ≡ gray (Gy)
  - dose = (power × seconds)/kg
  - e.g., 150 MeV × 1 nA × 60 s = 9 J (for one minute)
- water: 1 kg/1000 cm^3 = 1 kg/L
- dose = 9 J/1 kg (in a litre) = 9 Gy
  - 150 MeV, 1 nA = 9 Gy in 1 L in 1 minute
- But not all energy goes into the target (see Bragg peak) ⇒ 3–6 Gy in 1 litre in 1 minute
- 1 nA in 60 s ⇒ 60×10^{-9} C ⇒ 3.7×10^{11} protons for 3 Gy
- Therefore, for 1 Gy in 1 litre we need ~120 gigaprotons (1.2×10^{11})
  - 120 Gp/min ⇒ ~0.3 nA (averaged over a minute, but synchrotrons are cyclic…)

This gives an indication of the number of protons needed to treat a target, depending upon the dose (in Gy) that is prescribed. This number of protons must be extracted from an accelerator in the desired time interval. This lecture is about synchrotrons, so all examples will be relevant to these devices.

4.2 Applicability to synchrotrons

A synchrotron is a closed loop of magnetic components in which particles are stored, accelerated, and then extracted (details are discussed in other lectures in this course). During the time the particles are stored and accelerated, they must all live nicely with each other. However, they are all charged, and thus they repel each other. This effect is called the space charge force and is represented in Fig. 6, which indicates the strength of the repulsive/defocusing force across the distribution.

However, as the particles move, they are a medium carrying a current and since parallel currents attract, this attractive force partially cancels the repulsive force, by an amount that depends on the magnitude of the current. As the particles move faster, the deleterious space charge effects are reduced. Thus the worst-case situation occurs during low-energy injection and the number of charged particles that can be stored in the ring depends upon the injection energy of the particles.

Figure 7 shows a graph of the number of protons that can be stored in various medical synchrotrons (the limitation arising from the space charge forces only). Note from the above that 1 Gy/min in a litre ⇒ 120 gigaprotons/min ⇒ <4 Gp/acceleration cycle (assuming a 2 s cycle), where 4 Gp = 4 × 10^9 protons.

Thus a connection between a prescription and an accelerator constraint is obtained.

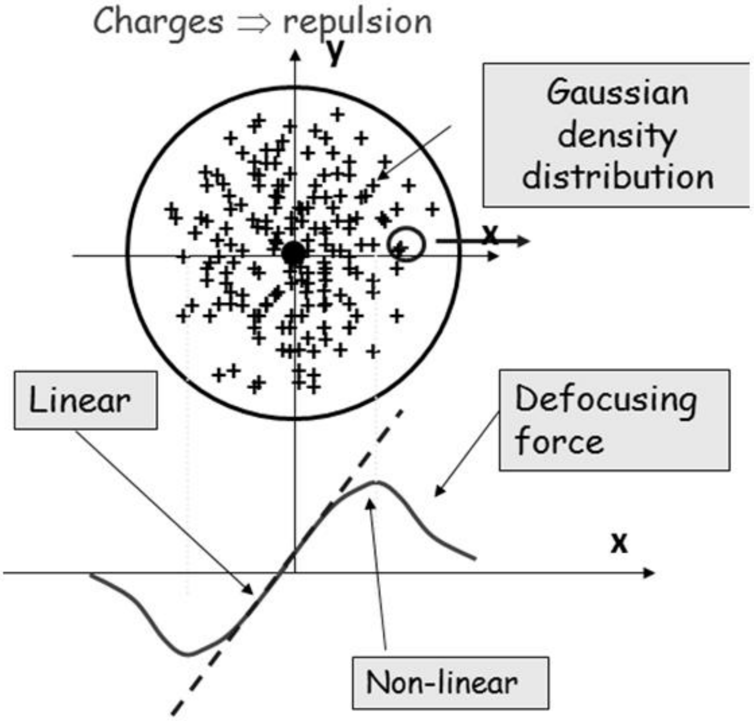


Fig. 6: Repulsive forces in a charged beam

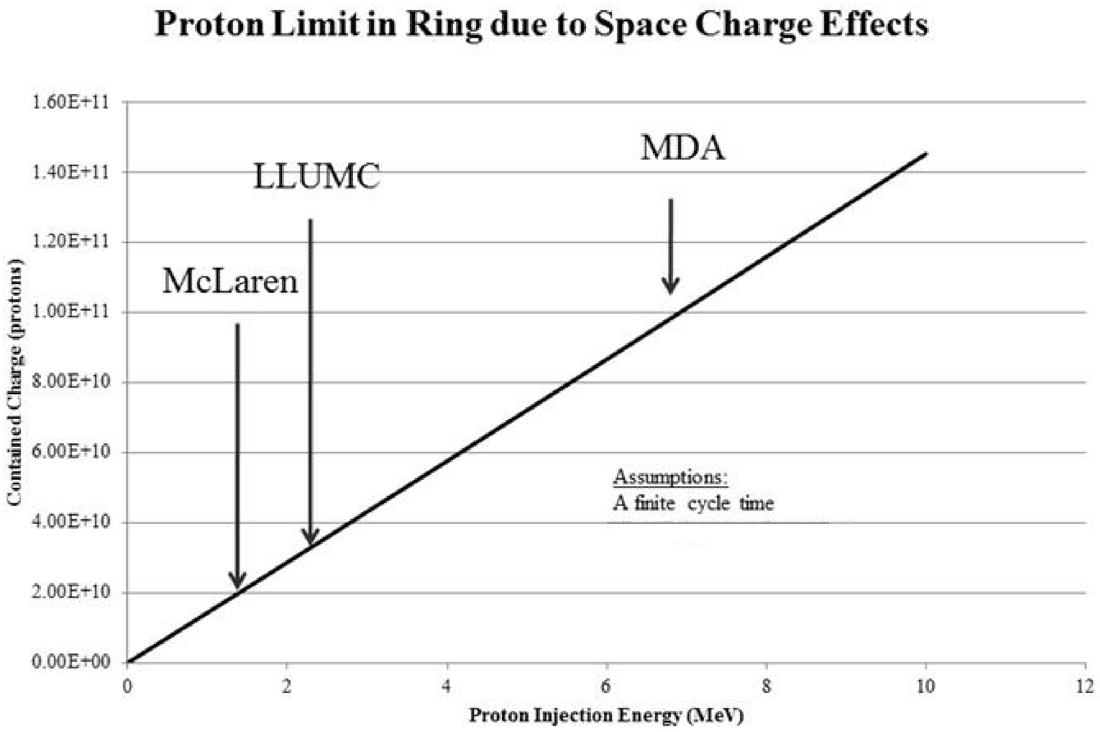


Fig. 7: Space charge limits in proton storage rings

4.3 Beam current issues

The situation may change depending upon the details of the treatment. A few examples follow.

- The beam-spreading modality plays a significant role. If scattering is used, the beam current incident on the beam delivery system may need to be on the order of nanaoamps, whereas for scanning, only tenths of a nanoamp may be required.
- It may be desired to reduce the number of fractions required to deliver the total dose, in which case the dose per fraction would be increased, thus increasing the desired dose rate (so that the treatment time per fraction does not increase).
- Considerations of target motion inside the patient may affect the time constraints on the beam delivery.
- The instrumentation and the beam analysis time will affect the dose rate that can be safely applied.

4.4 Tolerances

Depending upon the situation, clinical tolerances can have an impact on the machine performance. Here is one example. Assume that it is desired to deliver 40 gigaprotons to a target (the total dose in a particular field). Clinical tolerance dictates a 2% accuracy in the dose delivery, which results in a tolerance of 8×10^8 protons. Assume further that this target has transverse dimensions of 10 cm × 10 cm and that the beam spot size used is 5 mm (1 sigma). Thus it will take roughly 20 × 20 beam spots to cover the target. These 400 beam spots will each have to be delivered with a tolerance of 2×10^6 protons, indicating the level of control required for the beam. Note further that if it takes 100 μs to respond when measuring and reacting to the beam delivery (it could take longer), this tolerance translates to a requirement of not delivering more than 2×10^6 protons in 100 μs, or a maximum current of 3.2 nA.

4.5 Extraction effects

Extracting the beam from a synchrotron is a semi-stochastic process. There are a variety of methods of extracting the beam. One can imagine a pail of water in which water is stored and in which a spigot is inserted into the bottom with a valve to control when, and at what rate, the water is extracted. The water would be expected to come out smoothly, but if it were filled with air bubbles, the result would be different. Figure 8 shows an example of the time dependence of an uncorrected resonant extraction. One class of extraction method is to divide the transverse space inside the synchrotron into a stable and an unstable region. This can be done (in a simplified view) by introducing a non-linear magnetic field, which is weaker than the focusing forces in the ring below a certain radius, but larger at a higher radius (this is a simplified explanation). An example of the kind of time dependence of the extracted beam that can be achieved is shown in the curves in Fig. 9.

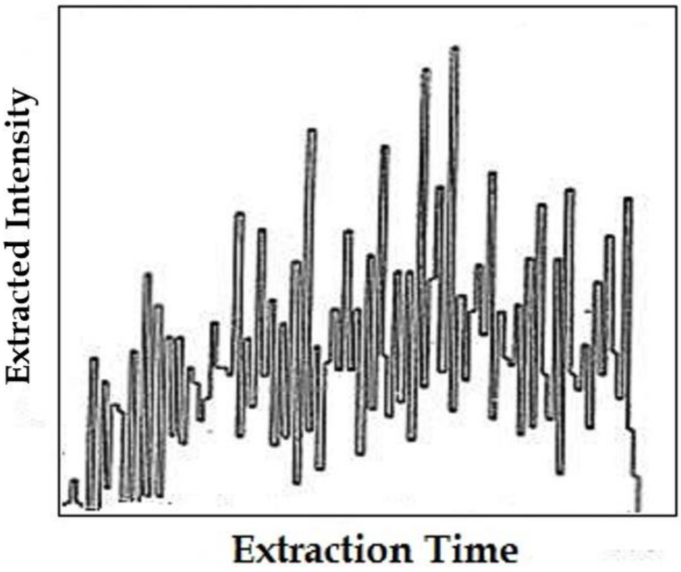


Fig. 8: Example of uncorrected time dependence of extracted beam

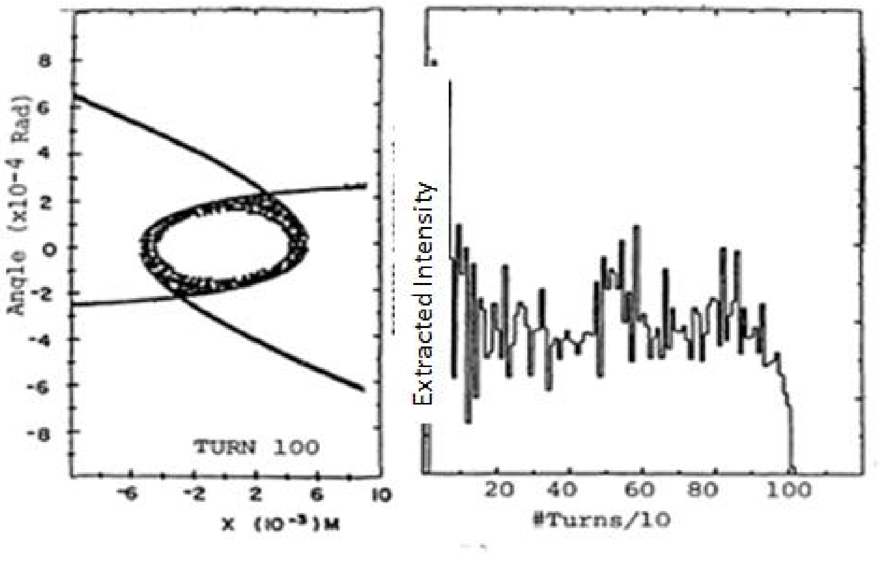


Fig. 9: Left: phase space of beam, with separatrix for chromatic extraction. Right: example of time dependence of chromatic extracted beam.

In Fig. 9, the curve on the left shows the phase space of the beam, with the intersecting parabolas representing the phase space boundaries between stable (inside) and unstable (outside) regions. The time dependence of the intensity of a beam extracted from a stretcher ring is shown in Fig. 9 on the right; here, special optics were used in the ring to couple the transverse dimensions to the energy in order to smooth the extraction [1]. Figure 10 shows an example of a scope trace of a beam extracted from a synchrotron used for medical treatment which employs a sophisticated feedback extraction correction system. Issues related to the time it takes to turn off the beam (e.g., when the desired dose is reached) and the range of controllable intensity play a role in determining the appropriateness of an accelerator design.

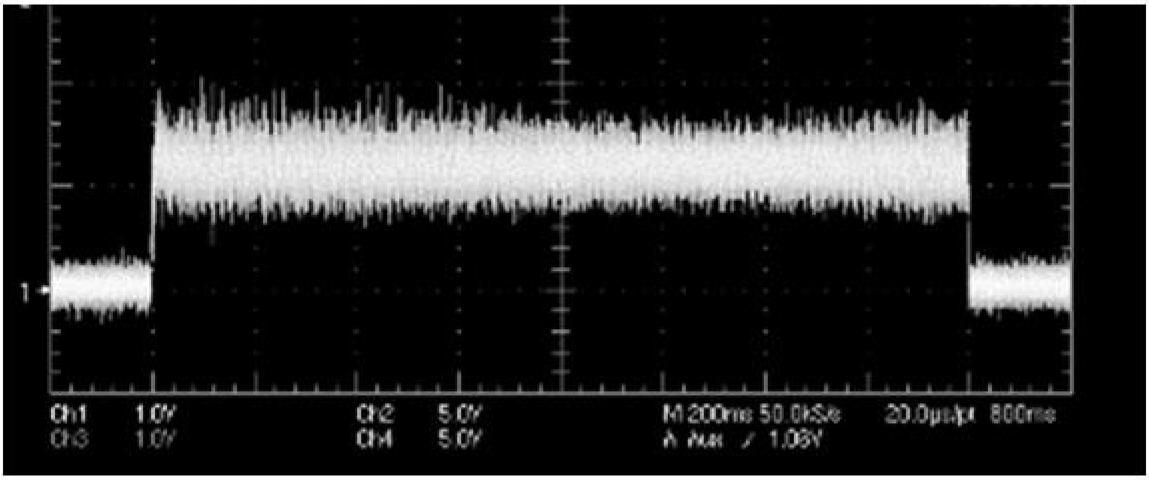


Fig. 10. Example of corrected time dependence of RF-excited extracted beam

5 Beam size and shape

It is desired to deliver a dose to the target but to allow all surrounding tissue to remain unharmed. The degree of conformity is related to the shape of the beam incident on the target. As noted earlier, the beam may have a Gaussian shape, but it may also have other shapes. A Gaussian shape is particularly well suited for scanned beams since Gaussians can combine well and produce a uniform or otherwise conformal pattern. Figure 11 shows how multiple Gaussians can be combined to form a flat top. The right-hand part of the figure shows that as the Gaussians are separated, the combined dose eventually shows the beam structure.

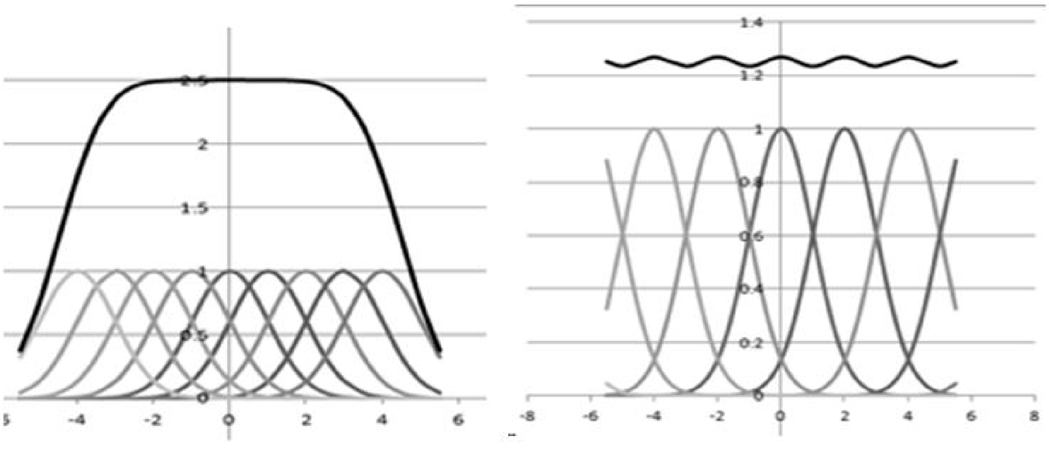


Fig. 11: Left: multiple Gaussians spaced such that resultant summed intensity is uniform. Right: multiple Gaussians spaced further apart than optimal, showing the summed intensity structure.

5.1 Penumbra

Note that the dose 'falls off' at the sides. The steepest fall-off is at the edge of the Gaussian (without external collimators). The steepness of this fall-off determines how much dose will be deposited into healthy tissue, as shown in Fig. 12. The fall-off region is generally called the penumbra, after the use of this term in scattered-beam physics. If it is desired to ensure that the target dose is within 2.5% of the nominal dose, then the edge of the target must be contained in the top left part of the distribution in the figure (upper box). Suppose, for example, that an important structure occurs in the lower box, to the right; assume, for example, that its edge is 5 mm from the edge of the target. Suppose also that the physician has determined that this critical structure cannot receive more than 50% of the target dose. Using the equation of a Gaussian,

$$e^{-x^2/2\sigma^2},$$

one can calculate that the Gaussian shape cannot have an r.m.s. width (one sigma, σ) larger than 7 mm.

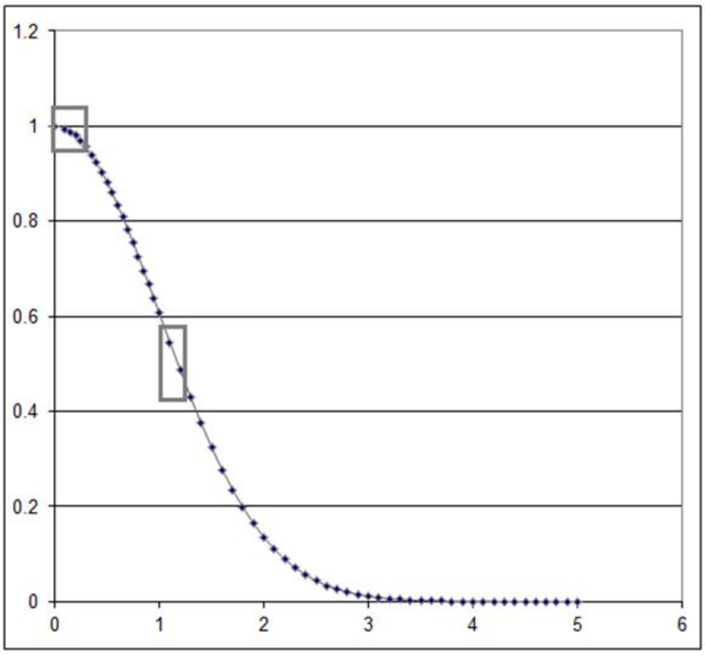


Fig. 12. Example of relative intensities of a Gaussian for a given spacing

If the shape of the beam is *not* Gaussian, much of the above discussion becomes invalid and specific calculations are required to determine the adequacy of the beam for treatment. Note that in some synchrotron extraction schemes the beam has a sharp edge in the plane of the extraction (this may occur if there is a septum in the synchrotron).

5.2 Effects of emittance

The beam size requirements identified above, coupled with the method by which the beam is delivered (e.g., whether or not it is delivered through a beamline), can place a constraint on the beam emittance. Assuming that it is desired to have a specific beam size at the target and that the final magnetic element is at a fixed distance from the target, a larger-emittance beam will have to be much larger than a smaller-emittance beam at the location of that last magnetic element. If the last magnetic element is on a gantry, then the size and weight of the gantry will be affected by this beam size constraint. Figure 13 shows the power and weight requirements for a gantry dipole located 3 m away from the patient isocentre as a function of the beam size desired. The upper curve is for a beam with an emittance of 25 mm mrad and the lower curve is for a beam with one of 5 mm mrad. Thus one can appreciate the ramifications of larger- versus smaller-emittance beams (independently of how they are achieved).

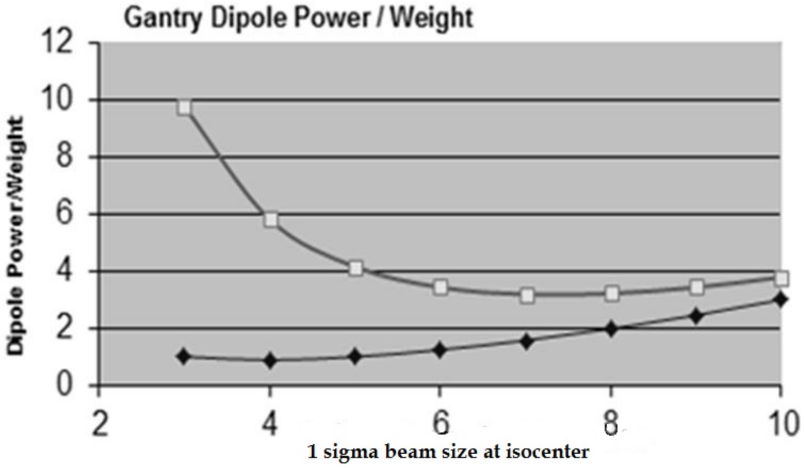


Fig. 13: Power and weight of a gantry dipole for a larger-emittance (upper curve) and smaller-emittance (lower curve) beam as a function of the desired beam size at the target.

6 Timing

There is a possibility of a strong interaction between the time distribution of the beam extracted from the accelerator and the beam delivery method. In the past few years, beam delivery for particle therapy has evolved into beam scanning as this enables the most conformal dose distribution possible. Therefore we will focus on this aspect of beam timing.

6.1 Beam delivery and accelerator timing

There are essentially two styles of scanning beam delivery, one that may be called 'dose driven' and another that may be called 'time driven'. The first integrates the dose at a given location before moving on to the next location. The second assumes that the beam current has the desired stability and uses time as the variable to identify the dose deposited at a given location. If, for example, one moves the beam continuously, the amount of dose deposited along the beam path in any given time interval is determined by the beam current, the scanning speed, or both. If these are not precisely correct, the dose deposited will not be precisely correct. Alternatively, one can deposit a dose at a specific location (spot) and wait until the dose desired there has been delivered (dose-driven method) before moving onto the next location, and therefore the time dependence of the beam (during the time in which the dose is being delivered) is not as relevant, with the following exceptions.

In dose-driven beam delivery, one must stop the beam when the desired dose has been reached. There is, however, a time required for the beam instrumentation to measure and analyse this dose, and a time required for the beam to be turned off. Thus there is a time lag between the time the system decides to turn off the beam and the time the beam is actually turned off. Therefore, either one can anticipate that the beam to be delivered in this time frame will be known and begin the turn-off process earlier, or one can lower the beam current to ensure that the dose that will be delivered in this time frame will not be significant. In addition, in both time-driven and dose-driven continuous scanning, the beam turn-off time will occur while the beam is moving, and one must account for the locations that receive a dose during the turn-off time.

The above considerations highlight the importance of knowing the beam extraction stability and the turn-off time, as discussed above in Section 4.4. The smoothness and controllability of the extracted beam can determine which method of beam scanning can be applied. The smoothness can also determine the anticipated time required to turn off the beam. If the extracted beam current is unstable, one must prepare for the highest beam current and the dose rate must be reduced.

6.2 Organ motion

In some cases, the target moves. One hopes to reduce the dose delivered to healthy tissue, and this poses some challenges when the beam delivery has to be done in a time-dependent way. If one could deliver the dose simultaneously to the entire 3D volume of the target (as is almost the case with scattered-beam delivery), one would only need to consider the location of the moving target as a function of time. One could, for example, deliver dose to the entire volume within which the target was moving; then healthy tissue would be irradiated, but the target would receive the desired dose in the shortest possible time. If the beam is gated on only when the target is in the beam path, then the macroscopic timing capability of the accelerator and the time frame of the motion have to be taken into consideration in determining the length of the treatment. Figure 14 shows an example of this situation, which depends upon the time cycle of the accelerator [2]. A large improvement in the efficiency of a synchrotron for beam treatment was achieved with the development of a variable-cycle synchrotron whose injection and extraction can be synchronized with the target motion.

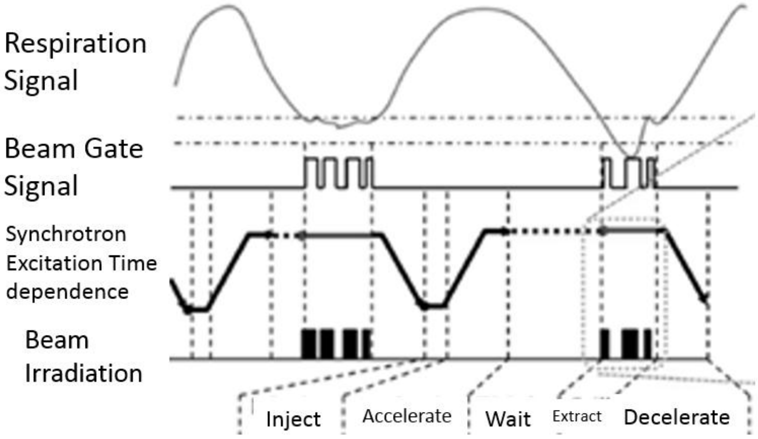


Fig. 14: Time dependence of synchrotron operation

An additional effect arises when the beam delivery is done using scanning, which has a 3D position–time dependence. One must consider the timing of the beam delivery, the timing of the target motion, and the timing of the accelerator in order to achieve an appropriate dose distribution in a reasonable time frame.

Some types of beam cycle for synchrotrons include rapid cycling with fast, one-turn extraction or very short pulses of a periodic (e.g., 30 Hz) beam, where each pulse can be at a different energy. Alternatively, one can use slow extraction, pulling out particles from the accelerator as needed, at the beam current needed until they are used up. If there are, for example, 10^{10} particles in the synchrotron, then it may take 1 s to use up those particles at the maximum current identified earlier. If more are needed, one has to wait for the time it takes to inject and accelerate another bunch. Also, if a different energy is needed, one must wait for another acceleration cycle, unless one is able to extract at different energy levels during one extraction cycle as shown in Fig. 15 [3]. Note that a breathing cycle can take about 3 s.

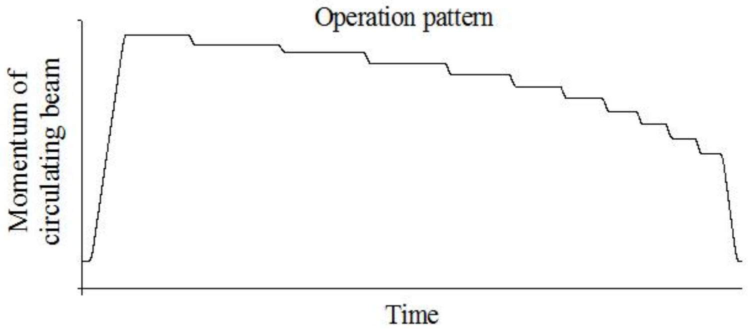


Fig. 15: Time dependence of beam energy during extraction in a multienergy extraction process

7 Cost

If particle therapy is to become more widely used, the cost of particle therapy systems needs to be reduced. One must reduce both the building and the equipment costs associated with a facility. Today a variety of synchrotrons are being used in medical treatment, some of which are shown in Fig. 16.

Smaller synchrotrons for ions heavier than protons, using superconducting magnets, are being investigated. Proton synchrotrons have already been reduced in size by ProTom and Hitachi (to name two), with diameters on the order of 5 m. Some groups are also attempting to reduce the size and cost of gantry structures. One facility for proton therapy is already being constructed in an existing conventional radiotherapy clinic at Massachusetts General Hospital, with no new building being built for the machine (Fig. 17).

Synchrotrons for heavier ions (e.g., carbon) are much larger, and today resemble a particle physics laboratory accelerator; however, smaller systems utilizing superconducting technologies are under investigation.

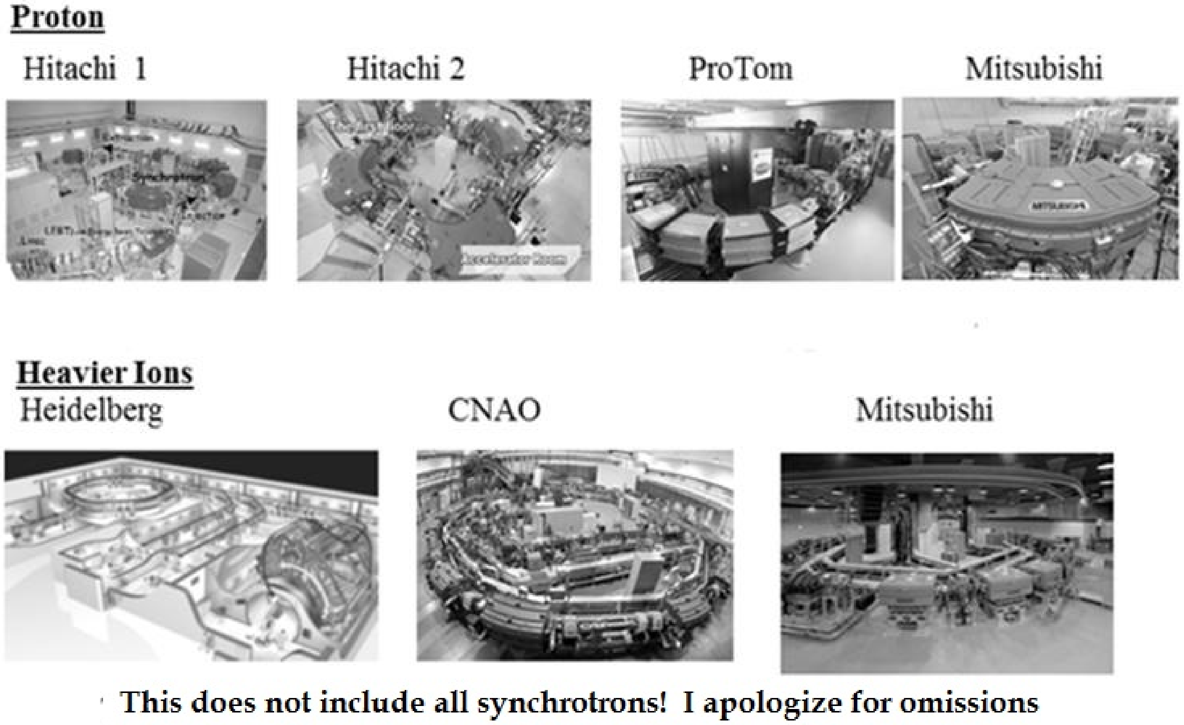


Fig. 16: Some particle therapy synchrotrons currently in operation

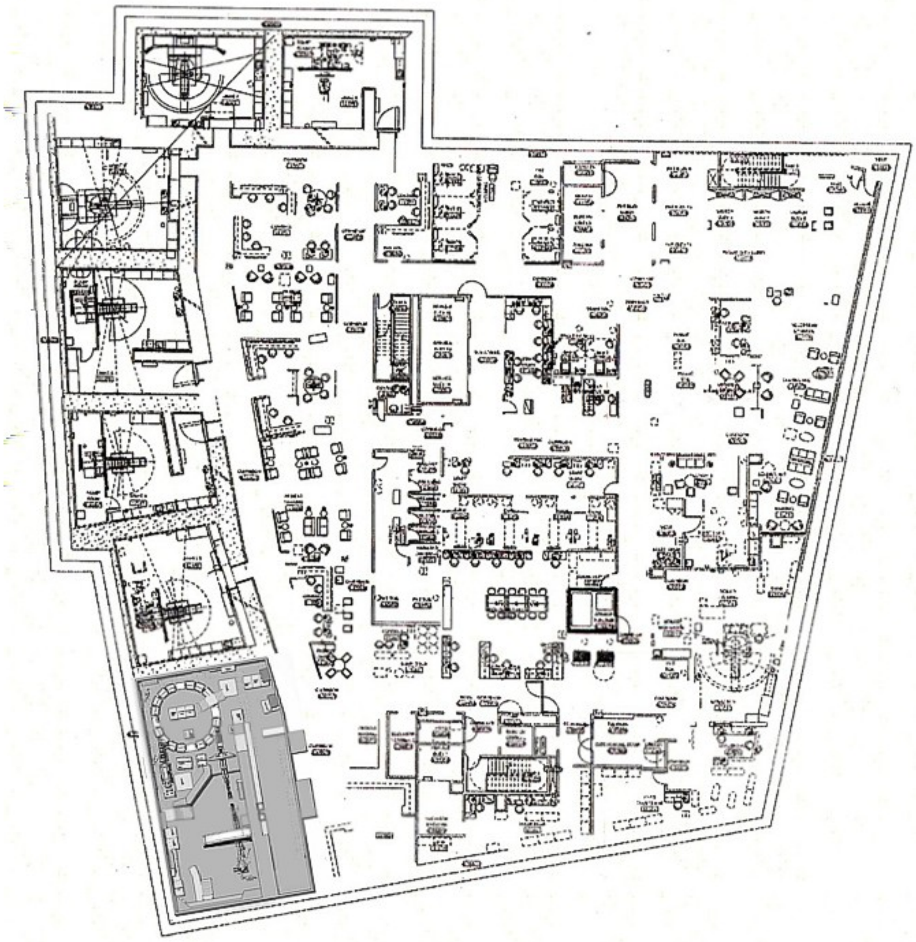


Fig. 17: Plan of radiation oncology floor at Massachusetts General Hospital, showing a proton therapy room integrated with linacs (space shaded grey).

8 The future of synchrotrons

Many developments of the synchrotron have been necessary in order to develop it into an efficient accelerator for application in medical particle therapy. The design of the synchrotron must correspond to the important treatment parameters. Development work is continuing with the aim of obtaining adequate intensity storage while attempting to minimize the cost of higher-energy injection and smoothly controllable extraction, with a flexible accelerator cycle that can change the beam energy extracted during a single cycle. Thus far, however, all of these capabilities have not been used in any one synchrotron for treatment. For proton synchrotrons, the performance has to continue to improve and the cost has to continue to decrease. There will be increasing demands for faster treatment without compromising accuracy. For heavier particles, the use of superconducting technology can reduce the size, but the cost and the rapidity of change of the magnet excitation will be affected.

There are a number of factors for future consideration (some of which may compete):

– cost:
  - size vs. superconductivity;
  - injector energy;
– intensity:
  - injector energy vs. cost;

- energy:
  - therapeutic energy vs. imaging vs. low (shallow-range) energies;
- energy change speed:
  - effects of superconductivity;
  - beam storage stability;
- turn-off time:
  - instrumentation time;
  - analysis time;
  - extraction control;
- irradiation time:
  - full-volume irradiation in a short time.

The future development of synchrotrons must be directed towards meeting the demands of optimal, safe delivery of particle therapy at a cost that is competitive with conventional radiotherapy systems. It is no longer adequate to identify an accelerator and then ask how it can be used for particle therapy; this has been done. One now has to optimize the delivery of particle therapy, including the beam parameters, timing, size, and cost. Thus far, there is no single standout technology that can accomplish all this, but the technology to achieve operation with the desired parameters does exist and the field is ripe for new insights and developments.

References

[1] J.B. Flanz and C.P. Sargent, *IEEE Trans. Nucl. Sci.* **NS-23**(5) (1985) 2444.
http://dx.doi.org/10.1109/tns.1985.4333941

[2] M. Umezawa *et al.*, *Hitachi Rev.* **64**(8) (2015) 506.

[3] Y. Iwata *et al.*, Multiple-energy operation with quasi-DC extension of flattops at HIMAC, Proc. IPAC’10, Kyoto, Japan, 2010, p. 79, paper MOPEA008.

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015, edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Therapy Control and Patient Safety for Proton Therapy

M. Grossmann
Paul Scherrer Institut, Villigen, Switzerland

Abstract
This contribution describes general concepts for control and safety systems in proton therapy. These concepts are illustrated by concrete examples implemented in the Proscan facility at Paul Scherrer Institut (PSI).

Keywords
Proton therapy; control systems; real time; safety.

1 Introduction

Protons have successfully been used for many years in the treatment of cancer. Their physical properties allow a better conformation of the dose to a target volume, which (depending on the tumour location) can give protons an advantage over conventional radiation therapy with photons.

When preparing a radiation therapy a therapy plan is created that designates the region in the patient to be treated and the dose to be applied. This plan is then translated into a steering file containing the machine information for the treatment facility. It is the task of the therapy control system to make sure that the treatment follows this prescription. The patient safety system supervises the treatment, detects dangerous situations and if necessary interrupts the treatment.

This article describes general concepts for proton therapy control and safety systems, illustrating them using examples from the proton therapy facility at Paul Scherrer Institute (PSI).

2 Proton therapy at PSI

PSI operates the Proscan facility for proton therapy [1]. The layout of the facility is shown in Fig. 1. The proton beam is produced by the superconducting 250 MeV cyclotron COMET. A degrader at the exit of the cyclotron reduces the beam energy according to the therapeutic requirements. The beam is then sent into four treatment rooms.

The OPTIS treatment room is dedicated to the treatment of ocular tumours (mostly melanoma). It uses a conventional scattering technique to apply the required dose. With this technique patients have been treated at PSI since 1984. The installation is an upgrade of the original facility and started operation in 2010.

Gantry 1 went into operation in 1996 as the world's first spot-scanning gantry that applies the dose by magnetic scanning of a focused pencil beam. It aims at the treatment of deep-seated tumours, many of them some kind of brain tumour.

With Gantry 2, PSI continued the technological development of magnetic pencil beam scanning [2]. It is based on the experience with Gantry 1 and is optimized for fast and advanced scanning modes, allowing the extension of treatment indications. Patient operations started in 2013.

The most recent treatment room Gantry 3 is currently under construction. Unlike the other rooms which were built by PSI, it is a commercial device based on Varian Medical System's ProBeam system. Treatment of patients is scheduled to start in 2017.

https://doi.org/10.23730/CYRSP-2017-001.309

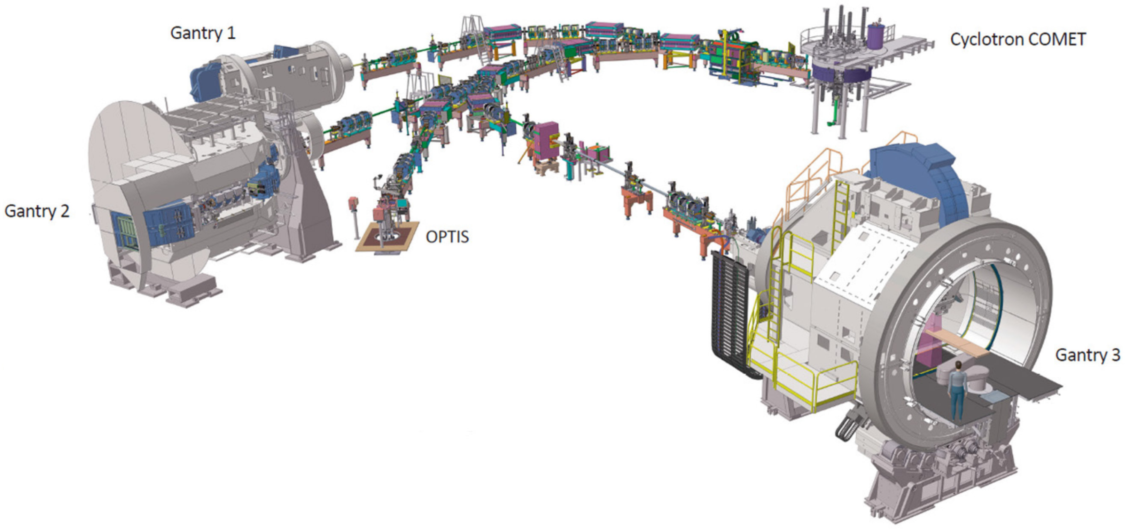


Fig. 1: Layout of the Proscan facility beamlines including the new Gantry 3 area. The existing treatment areas are Gantry 1, OPTIS 2 and Gantry 2.

3 Control systems

3.1 Machine control system

The *machine control system (MCS)* groups together all the subsystems necessary to control the accelerator and the beamline. Its purpose is to deliver a beam with modifiable beam energy, momentum spread and intensity to the treatment area.

The MCS can be based on a framework such as EPICS[1]. Access to the equipment under control is provided through *Versa Module Eurocards*, so-called *VME modules*, housed in several crates distributed over the Proscan facility. Each VME crate is equipped with a single-board computer called *input output controller (IOC)*. These IOCs can be accessed over the network using the EPICS concept of *channels* by other applications or user interfaces.

The MCS provides a *graphical user interface (GUI)* running on workstations located both in the central control room and at each treatment room. Write access from these workstations is restricted during therapy.

Communication between a treatment area and the MCS is organized through the *beam allocator gateway (BALL)*. Besides providing an interface to the MCS, BALL restricts write access to MCS-controlled devices in order to prevent interference from different areas during therapy.

3.2 Machine protection

The purpose of the machine protection system (at PSI called *run permit system (RPS)*) is to protect the beamlines and the accelerator from damage by the proton beam. It defines limits for the setting of devices in the beamlines, e.g. an upper and a lower limit for the current in magnets along the beam line. If the current setting is outside of these limits, the beam cannot be switched on. If the beam is on when a violation of the supervised limits occurs, it is switched off. The RPS has its own independent interface to a set of final elements for this purpose.

[1] EPICS is a software framework for experimental physics and industrial control systems.

The limits used by the RPS depend on the treatment area that requests the beam. The RPS is informed by the MCS which treatment area is requesting the beam and then sets the limits accordingly.

3.3 The concept of mastership

During treatment of a patient only one area can have full control over the facility, and no other area is able to interfere. Based on the planned therapy flow, one area requests exclusive access, called *mastership*, to the beamline from BALL. If available, BALL grants mastership, giving exclusive access to both the shared and local beamline sections all the way upstream to the cyclotron. This includes control of the degrader (energy selection) and the kicker magnet (beam on/off control). The requesting area becomes 'master' and the patient treatment scheduled in this area can be performed.

3.4 Therapy control system

The *therapy control system (TCS)* consists of two components, the *therapy delivery system (TDS)* and the *therapy verification system (TVS).*

TDS and TVS run on IOCs in separate VME crates. These crates contain (besides the IOCs) the necessary modules to interface with the actuators and sensors used to perform the treatments, and with the other control system parts (see Fig. 2).

Proceeding spot by spot, the TDS initiates the setting of the actuators to control the beam position (the two sweeper magnets, the patient table, the gantry motors and the beamline) and sets the required dose for the spot to be applied. When all preparations are successfully completed, the command to switch on the beam is issued.

The TVS checks that the spot sequence proceeds as planned. It verifies the correct setting of the actuators by means of dedicated sensors (e.g. Hall probes to check the field of the sweeper magnets; position encoders to check the positions of the patient table and gantry body). In addition, it performs a number of checks before, during and after the spot application (e.g. verification of the dose applied and of the spot position).

In case of a failure, the TDS and TVS can interrupt the treatment by generating an interlock. Both TDS and TVS use individual steering files which contain definitions of the dose and of the nominal values for actuators and sensors for each spot.

The user interface to the TCS consists of two applications: the *GUI server*, communicating with TDS and TVS over the network, and the *GUI client*, communicating with the GUI server over the network. There may be more than one instance of the GUI client running at a time. The GUI server guarantees that only one GUI client has write access to the TCS.

The proton dose is monitored by ionization chambers. The primary (dose-defining) monitor is located just upstream of the patient. Additional monitors in the treatment room and along the beamline supervise the correct function of the primary monitor and serve for additional safety functions (e.g. interrupting the treatment when beam intensity becomes too high).

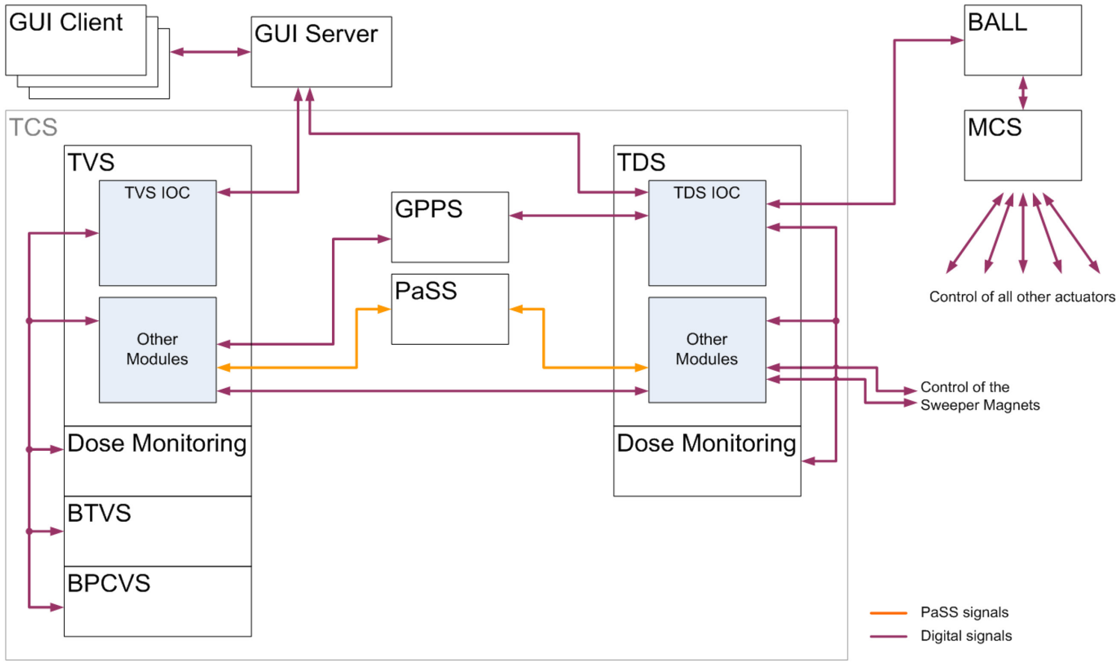


Fig. 2: Simplified scheme of the therapy control system (TCS) with its interface to associated control systems and the GUI server and clients. PaSS signals are failsafe connections (three-wire current loop), digital signals are I/O signals like TTL, or serial, parallel and network connections.

4 Patient safety system

4.1 Principle

The purpose of the *patient safety system (PaSS)* is to protect the patient from radiation hazards by minimizing the risk for an uncontrolled irradiation.

The PaSS follows the principles of any safety system:

- *sensors* measure critical parameters;
- a *logic* evaluates the data provided by the sensors and decides if the ongoing (therapy) process should be interrupted;
- *final elements* reach the safe state (no beam, stop of mechanical motion).

The PaSS is implemented with logic hard coded in electronic chips and using only point-to-point connections, avoiding the use of communication buses. Reflecting the architecture of the Proscan facility, the PaSS has local and central parts (schematically shown in Fig. 3):

- the local PaSS for a specific treatment area, with its own sensors, logic and final elements;
- the central PaSS for those parts of the facility that are used by all treatment areas, with sensors, logic and final elements that must be shared between treatment areas.

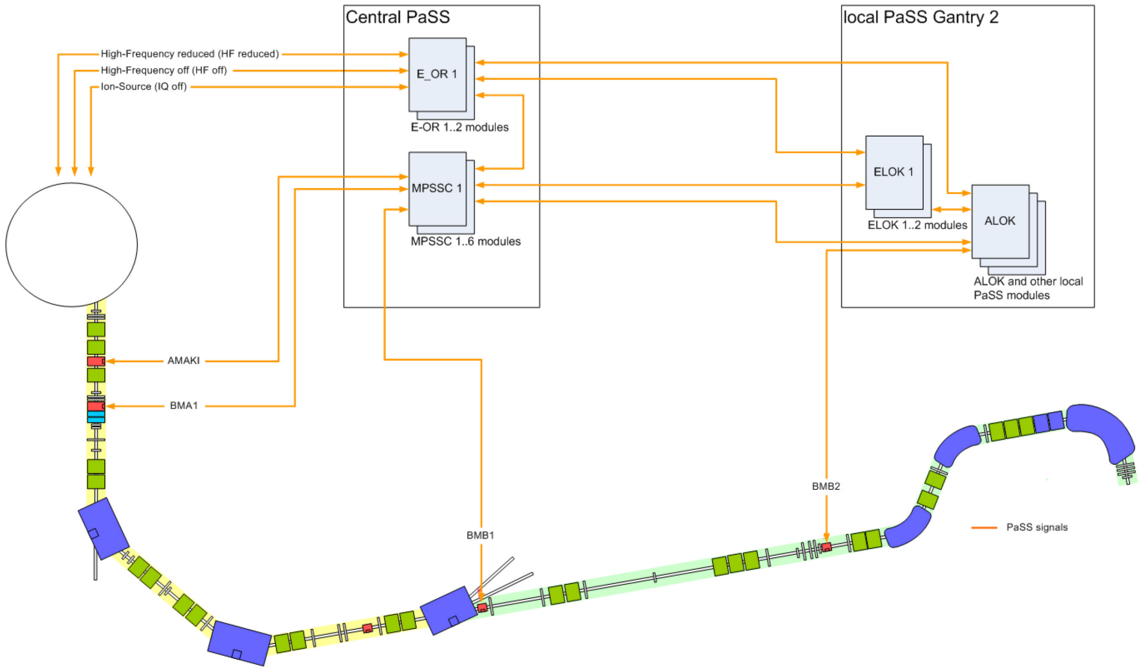


Fig. 3: General architecture of the patient safety system with local and central components, each containing several modules. This figure is largely simplified: the single wires shown in the figure are in reality multiple hardware, point-to-point connections. Inputs to the PaSS (from sensors directly and from control systems) are not shown.

4.2 Final elements

Several final elements allow us to switch off the beam to prevent uncontrolled irradiation. Both redundancy and diversity are required. At PSI the following elements are used (see also Fig. 3 for their location).

4.2.1 Kicker magnet AMAKI

The kicker magnet AMAKI is the fastest device dedicated to turning the beam on and off. If the magnet is energized, it deflects the beam into a beam dump. AMAKI was primarily designed as a steering element for spot scanning. Due to its fast reaction time and its ability to handle many on/off cycles in an extremely reliable way, it also has the function of a safety element. A magnetic switch is used to supervise the magnetic field of the AMAKI. Reaction time for AMAKI is 300 µs.

4.2.2 High-frequency generator

The high-frequency (HF) generator of the cyclotron can be used as final element in two stages: the power can be reduced to 80% (HF_RED)—this already stops acceleration of protons completely but prevents cooling down of the generator, which can lead to beam instabilities. In a second stage the generator can be completely switched off (HF_OFF). Reaction time for both stages is 400 µs.

4.2.3 Ion source

Powering off the ion source at the centre of the cyclotron prevents protons entering the accelerator (IQ_OFF). Reaction time for the ion source is 20 ms.

4.2.4 Local fast mechanical blocker

At the entrance point of the proton beam into the treatment room a fast mechanical beam blocker (BMB2) is installed. It is a block of 10 cm copper which can be moved in or out of the beamline by pressurized air and a mechanical spring. In case of failure of the pressurized air it will fall into the beam-blocking position by gravity. Reaction time for the local fast mechanical blocker is 60 ms.

4.2.5 Local slow mechanical blocker

Located in the local part of the beam line, but still upstream of the treatment room, is a slow mechanical blocker (BMB1) made of graphite. It can be moved in or out of the beam line by pressurized air. In case of failure of the pressurized air it will fall into the beam-blocking position by gravity. Reaction time for the local slow mechanical blocker is 1 s.

4.2.6 Central slow mechanical blocker

Located near the exit of the cyclotron is a slow mechanical blocker (BMA1) made of graphite. It can be moved in or out of the beam line by pressurized air. In case of failure of the pressurized air it will fall into the beam-blocking position by gravity. Reaction time for the central slow mechanical blocker is 1 s.

4.3 The beam switch-off functions

The beam *switch-off* can have two different functions.

- A *control function*: the TCS controlling the dose application switches the beam off when the specified dose has been applied.
- A *safety function*: a potential radiation hazard is detected and the PaSS switches the beam off. Depending on the risk associated with the radiation hazard and the source of the hazard, three different interlock levels have been defined. They are labelled ALOK, ATOT and ETOT (see Fig. 4)

4.3.1 Beam switch-off control function

The beam *switch-off* control function is used during regular operation to terminate the spot once the nominal dose has been reached. At PSI this is realized by energizing the kicker magnet AMAKI, the *primary final element.*

Since in spot scanning the dead time between the single dose element applications needs to be kept below a few milliseconds, the regular beam *switch-off* control function does not trigger any mechanical beam blockers to be closed.

4.3.2 ALOK beam switch-off safety function

If a problem with any equipment of the local beamline occurs which might cause a radiation hazard, an ALOK is raised. An example is an unexpected beam position.

In the presence of an ALOK, the safe state, i.e. beam off, is achieved by the activation of the following final elements:

- the fast local beam blocker is closed (BMB2);
- the kicker magnet is energized (AMAKI).

The correct reaction to an ALOK is supervised. If one or several monitoring functions do not respond within the defined timeout, the ALOK is escalated to the next interlock level ATOT. Reasons which can lead to an escalation are:

- AMAKI status not changed after timeout;

- BMB2 not closed after timeout;
- nozzle dose monitors measure beam after timeout.

4.3.3 ATOT beam switch-off safety function

If a problem originating from a device in the shared beamline occurs and could cause a radiation hazard or if an ALOK was escalated, an ATOT is raised. This happens for example when, after a beam *switch-off* command from the TDS, the beam current does not reach zero within the specified time limit.

In the presence of an ATOT, the following final elements are activated:

- the local slow beam blocker is closed (BMB1);
- the central slow beam blocker is closed (BMA1);
- the cyclotron HF generator power is reduced to 80% (HF_RED).

The correct reaction to an ATOT is supervised. If one or several monitoring functions do not respond within the defined timeout, the ATOT is escalated to the next interlock level ETOT. Reasons which can lead to an escalation are:

- BMA1 not closed after timeout;
- BMB1 not closed after timeout;
- HF power not reduced after timeout;
- beam line dose monitors measure beam after timeout.

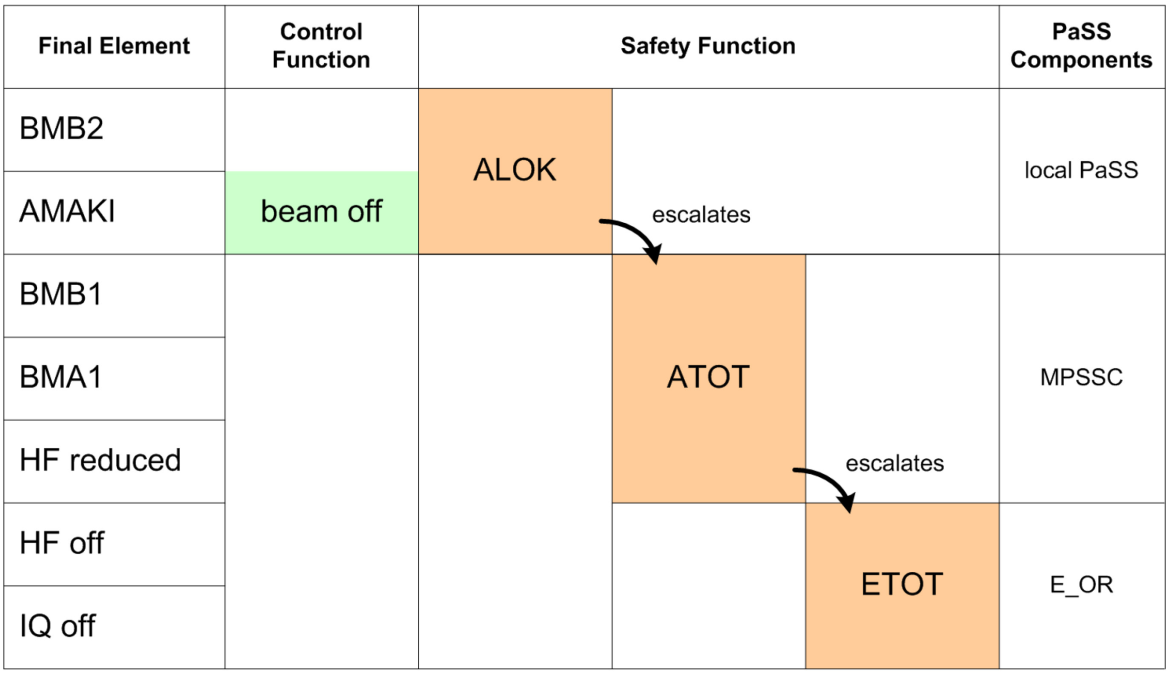
4.3.4 ETOT beam switch-off safety function

In case of a high-risk radiation hazard, the ETOT (emergency interlock) signal is raised. This happens when:

- a person operates a mechanical emergency-stop button; or
- beam is detected in a treatment area which is not master; or
- an ATOT was escalated.

In the presence of an ETOT, the following final elements are activated:

- the cyclotron HF generator is powered off completely (HF_OFF);
- the cyclotron ion-source power supply is powered off (IQ_OFF).



Final Element	Control Function	Safety Function			PaSS Components
BMB2		ALOK			local PaSS
AMAKI	beam off		escalates		
BMB1			ATOT		MPSSC
BMA1					
HF reduced				escalates	
HF off				ETOT	E_OR
IQ off					

Fig. 4: Summary of the final elements used by the regular beam-off command and by each of the safety functions, ALOK, ATOT and ETOT. The column to the right shows which of the PaSS components is responsible for controlling the respective final element.

4.4 Shared and redundant access to the common final element subsystems

With the exception of BMB2, access to the final elements is shared between all treatment areas. This access is implemented by the two components of the central PaSS:

- the *main patient safety switch and controller (MPSSC)*; and
- the *emergency-or module (E-OR*).

The MPSSC implements the following functions:

- area reservation and mastership validation:
  - only one single slow local beam blocker BMB1 can be open at a time and this only for the master area;
  - the mastership setting defined by BALL is cross-checked with direct hardware signals from the TCS of the treatment areas;
- ALOK safety function and AMAKI control:
  - the ALOK function is enabled for the master area exclusively;
  - access to the regular *beam on/off* command is restricted to the master area exclusively;
- ATOT safety function:
  - the ATOT safety function is restricted to the master area exclusively.

Redundantly, the E-OR module implements the ETOT safety functions:

- HF_OFF;
- IQ_OFF.

The E-OR allows all treatment areas to trigger the ETOT function at all times.

4.5 Implementation

The PaSS logic is implemented on several XILINX Spartan 2 FPGAs[2] (Fig. 5). These FPGAs are placed on industry-pack (IP) modules. Each module provides 10 channels (input and output). Hytec 8003 VME carrier boards can hold up to four IP modules. The carrier boards provide power to the IP modules and an interface to perform resets, set modes and read out the status of the IP modules. A transition board on the back side of the VME crate provides connections from the sensors and to the final elements.

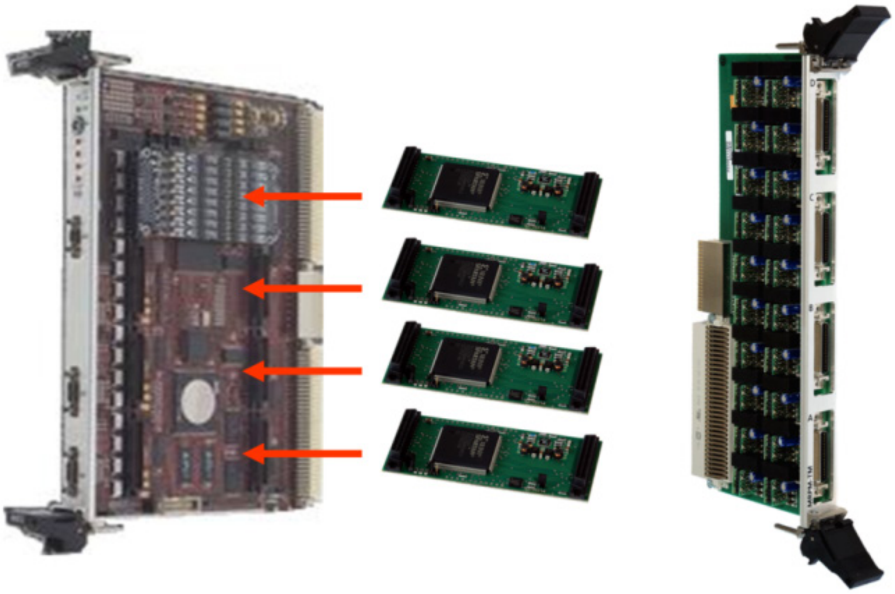


Fig. 5: Hardware implementation of the PSI patient safety system: four IP modules are mounted on a VME carrier board; the transition module on the right provides connections to sensors and final elements.

5 Integration of the commercial Gantry 3

Gantry 3 [3] is the latest extension of the PSI proton therapy facility. As already mentioned in the introduction the design is not based on PSI technology but the gantry is built by Varian Medical Systems.

Obviously Gantry 3 is delivered with its own control system. The challenge for the control system integration is to merge two quite different worlds: on one hand Varian's ProBeam scanning system, on the other hand PSI's existing systems, most of them built in-house. The Varian scanning system needs access to the cyclotron, the beamlines and the safety elements to control the beam and to switch it off in case of an interlock.

The chosen architecture leaves most of the existing systems untouched, with newly developed interfaces to connect them; see Fig. 6. The Varian system controls the scanning system (scanning magnets and dose monitors), the gantry and patient table positioning and the part of the beamline on the gantry itself. PSI systems control the upstream part of the beamline, the cyclotron and the central components of the PSI safety system.

The PSI part of the interface consists of two components, the *TCS* and the *PaS*S adapters.

[2] A field-programmable gate array (FPGA) is an integrated circuit designed to be configured by a customer or a designer

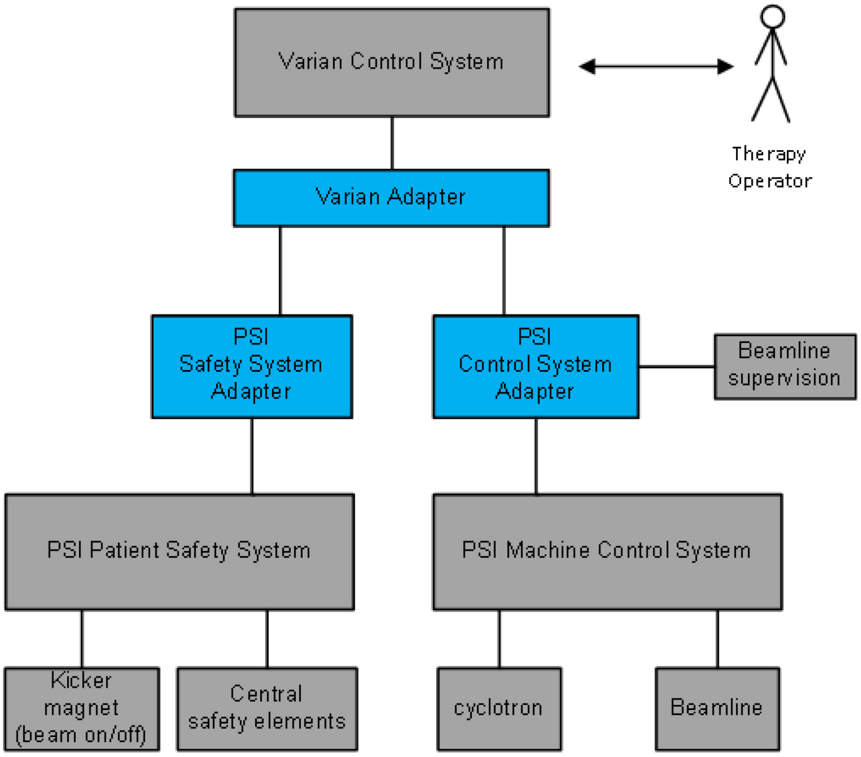

Varian Control System
Therapy Operator
Varian Adapter
PSI Safety System Adapter
PSI Control System Adapter
Beamline supervision
PSI Patient Safety System
PSI Machine Control System
Kicker magnet (beam on/off)
Central safety elements
cyclotron
Beamline

Fig. 6: Integration of Gantry 3 into PSI's control and safety systems is provided by the 'adapter' interfaces

5.1 TCS adapter

The TCS adapter communicates with the Varian system over a network connection. It serves as a gateway to PSI's machine control system. Typical commands are requests for treatment room reservation, beam energy and intensity. It also supervises the correct setting of the energy-selection system. It is implemented as a VME system with a Motorola single-board computer running VxWorks.

5.2 PaSS adapter

The PaSS adapter constitutes the interface between the Varian (local) and PSI (central) systems. While re-using most of the technology from the existing treatment rooms it was decided to program the logic on a state-of-the-art platform [4]. The choice fell for the IFC1210 controller, developed jointly by PSI and IOxOS [5]. It features a user-programmable Virtex 6 FPGA chip. It contains the safety system logic which after system startup is totally autonomous. On board are a further two Power PC CPUs running SMP Linux. They provide a standardized EPICS communication interface. This is used by the GUI, which provides access to the safety logic and automated actions like logging and statistics.

A generic platform, the so-called *signal converter box (SCB)*, supports the interconnection of IO signals from and to all subsystems. Data exchange between the SCB and the logic controller is handled over a high-speed communication link. It was developed jointly by PSI and Supercomputing Systems (SCS) [6].

For the development of the safety logic the same process as for the existing areas is followed. It comprises extensive verification and validation steps to ensure the correctness and integrity of the logic.

Acknowledgements

The development of PSI's control and safety systems and equally the whole proton therapy facility was great team work. The author would like to thank the members of the Center for Proton Therapy and the many others at PSI who contributed to this effort.

References

[1] J.M. Schippers *et al.*, *Nucl. Instrum. Methods* B **261** (2007) 773. http://dx.doi.org/10.1016/j.nimb.2007.04.052

[2] E. Pedroni *et al.*, *Z. Med. Phys.* **14** (2004) 25, http://dx.doi.org/10.1078/0939-3889-00194

[3] A. Koschik *et al.*, PSI Gantry 3: Integration of a new gantry into an existing roton therapy facility, Proc. IPAC16, 2016. http://jacow.org/ipac2016/papers/tupoy014.pdf

[4] P. Fernandez *et al.*, Reusable patient safety system framework for the proton therapy center at PSI, Proc. ICALEPCS2015, Pre-Press Release, 2015.

[5] IOxOS Technologies, IFC_1210 – Intelligent FPGA controller P2020 VME64x single board computer, rev. A0, data sheet, 2011, http://www.ioxos.ch

[6] Supercomputing Systems, http://www.scs.ch

Proceedings of the CAS-CERN Accelerator School: Accelerators for Medical Applications, Vösendorf, Austria, 26 May–5 June 2015,
edited by R. Bailey, CERN Yellow Reports: School Proceedings, Vol. 1/2017, CERN-2017-004-SP (CERN, Geneva, 2017)

Fixed-Field Alternating-Gradient Accelerators

S.L. Sheehy
University of Oxford, UK

Abstract

These notes provide an overview of Fixed-Field Alternating-Gradient (FFAG) accelerators for medical applications. We begin with a review of the basic principles of this type of accelerator, including the scaling and non-scaling types, highlighting beam dynamics issues that are of relevance to hadron accelerators. The potential of FFAG accelerators in the field of hadron therapy is discussed in detail, including an overview of existing medical FFAG designs. The options for FFAG treatment gantries are also considered.

Keywords

FFAG; medical; accelerator; hadron therapy.

1 Preface

These notes are broken down into two main parts. The first gives a general introduction to Fixed-Field Alternating-Gradient (FFAG) accelerators (often referred to simply as FFAGs) including an overview of the basic transverse dynamics of both the so-called ‘scaling’ and the ‘non-scaling’ variety, and a brief comparison with other machines. In the second part, I discuss the motivations for considering FFAGs for medical applications, focusing on charged-particle therapy (hadron therapy). I then introduce a number of FFAG designs specifically aimed at medical applications, including developments of FFAG gantries for medical use.

The FFAG accelerator is a class of circular accelerator that combines properties of both the cyclotron and the synchrotron. It uses a magnetic field which is constant in time, hence the ‘fixed-field’, together with an increased focusing strength achieved using the ‘alternating-gradient’ principle [1]. The RF acceleration scheme is usually variable-frequency, but in some specific instances a fixed-frequency system is possible.

Many accelerator physicists have difficulty when first encountering FFAGs, as they may have learned about only one example of an FFAG machine and therefore approach the topic with some preconceptions. Many questions arise, such as ‘aren’t the magnets very large?’, ‘have any of them actually been built?’, and ‘isn’t the FFAG only good for large muon beams?’ The difficulty with these questions is that the modern FFAG is not really a single type of machine. This would be like assuming that a synchrotron can only be a light source, rather than a hadron collider or a medical accelerator. For the record, the answers to the questions are ‘sometimes’, ‘yes’, and ‘not only’. Of course, for synchrotrons it is easy to see that the application can change the layout, design, and properties of an accelerator. This is perhaps even more true in the case of the FFAG.

Starting with the idea that FFAGs are just accelerators which have both a fixed field and alternating-gradient focusing produces a large spectrum of designs. Some designs may have purely linear (quadrupole) magnets and fixed-frequency RF acceleration, whereas others have large-aperture magnets which produce a complex variation of the field with radius and a variable-frequency RF system. Most FFAGs have a very large dynamic aperture. This flexibility of FFAG design has only emerged in roughly the last 15 years and the field continues to be a rich source of novel developments.

From these lecture notes you should be able to answer questions like those above, and describe the various types of FFAG accelerators and some of the design principles behind them. I hope you will also learn of the potential and some of the challenges in utilizing these machines in medical applications.

https://doi.org/10.23730/CYRSP-2017-001.321

2 Fixed-field alternating-gradient accelerators

2.1 A historical perspective

The concept of an FFAG accelerator is not new. This type of accelerator was invented in the 1950s and 1960s at the same time as the synchrotron was being developed. Much of the early work in developing FFAGs was carried out at the Midwestern Universities Research Association (MURA), but only electron FFAG accelerators were constructed at the time [2]. It is interesting also to note that in Soviet terminology the FFAG is known as a 'ring phasotron' [3]. It was not until the 1990s that interest in this type of accelerator re-emerged in Japan, with a particular focus on what such machines could offer as hadron accelerators. For a little more on the history, see Ref. [4] and the references therein.

A particular area of recent interest in the field of FFAGs is their potential for high-intensity operation, because of their high repetition rate, large acceptance, simpler and cheaper power supplies, and flexibility of the RF acceleration system. High intensity may be required for some medical applications (such as radioisotope production), but for hadron therapy the beam currents are typically low. For this reason, we shall not discuss high intensities here but instead focus on the advantages in terms of high repetition rate, flexibility, and simple magnet power supplies. The impact of these qualities will be discussed in Section 4.

2.2 The original or 'scaling' FFAG

In 1943 Marcus Oliphant described the idea of the synchrotron as follows:

> Particles should be constrained to move in a circle of constant radius thus enabling the use of an annular ring of magnetic field ... which would be varied in such a way that the radius of curvature remains constant as the particles gain energy through successive accelerations.

He intended that the magnetic field should be varied temporally and the beam should always follow the same annulus [5]. However, in principle there is no reason why the annulus may not change radius and the field vary spatially rather than temporally. This is the fundamental idea behind the FFAG. A large variation of the field with radius will constrain the change in radius of the orbits; this can lead to a larger field increase with radius and more compact orbits than in a cyclotron. This is the original type of FFAG, which we now call 'scaling'.

To introduce the scaling FFAG, it is useful to review the four main different types of circular accelerator, which we can classify by the magnetic field they use to guide the particles [3]. These are the following.

1. Fixed-field constant-gradient accelerators, including conventional cyclotrons, synchrocyclotrons, and microtrons.
2. Pulsed-field constant-gradient accelerators, which includes weak-focusing synchrotrons and betatrons.
3. Pulsed-field alternating-gradient accelerators, which are the well-known AG synchrotrons.
4. Fixed-field alternating-gradient accelerator, otherwise known as FFAGs.

The fourth variety, FFAG accelerators, were proposed independently in the early 1950s by Ohkawa in Japan [6], Symon *et al.* in the United States [7], and Kolomensky in Russia [8]. Symon *et al.* proposed:

> A type of circular accelerator with magnetic guide fields which are constant in time, and which can accommodate stable orbits at all energies from injection to output energy.

This relies on introducing sectors with a reversed magnetic field into a cyclotron-like machine, producing strong focusing throughout the energy range. The field may rise rapidly with radius such that the orbits are relatively compact over a large energy range. It is possible to accelerate both light particles (electrons) and heavier particles (hadrons) to relativistic energies with this method. The time-independent magnetic field means that the repetition rate can be much higher than in a pulsed-field alternating-gradient machine (i.e., synchrotron), as the RF modulation can be on a much shorter time-scale than the modulation of a magnetic field. The consequences of this for medical applications will be discussed later.

The field is arranged in such a way that the increase in gradient with momentum results in the beam experiencing the same focusing independent of radius. This means that the betatron tunes are constant for all orbits. This constant focusing (or constant betatron tune) is ensured if two conditions are met. First, the field index k must be constant, where we can define k in terms of the bending radius ρ, the vertical magnetic field B_y, and its derivative in the horizontal direction x:

$$k = -\frac{\rho}{B_y}\frac{\partial B_y}{\partial x}. \tag{1}$$

Therefore we require

$$\left.\frac{\partial k}{\partial p}\right|_{\theta=\mathrm{const.}} = 0. \tag{2}$$

The second requirement is that the shape of the particle orbits remains constant as the size of the orbits 'scales' with energy, such that each higher-energy orbit is a geometrically similar enlargement of the lower-energy orbits as described by the following equation, derived by Kolomensky [3]:

$$\left.\frac{\partial}{\partial p}\left(\frac{\rho_0}{\rho}\right)\right|_{\theta=\mathrm{const.}} = 0. \tag{3}$$

If the field meets these two conditions, the FFAG is referred to as being of the 'scaling' variety.

To satisfy these requirements, we use a magnetic field that increases with radius. The particular shape of the field is given by the r^k law of the following equation, which describes the increase in field with radius r with respect to a reference radius r_0, where the field increase is characterized by the field index, k:

$$B_y = B_0\left(\frac{r}{r_0}\right)^k. \tag{4}$$

The field is shown in Fig. 1. It should be clear that for a given value of the field index k, the field at a smaller radius not is only lower but also has a lower gradient. As the momentum increases and particles move to higher radius, the value of the confining field increases along with the field gradient. Of course, this is the field for only one of the two alternating-gradient types, the 'F' or focusing type. The field for the 'D' (defocusing) type has the opposite sign, as expected.

The fact that the different sectors in the scaling FFAG have reverse-polarity magnetic fields means that the length of the orbit and the mean radius of the machine are necessarily larger than if there were no reverse fields, as shown in Fig. 2. This is necessary to ensure stability in both planes and is the main disadvantage of the scaling FFAG. This can be partly or fully overcome, however, as the maximum magnetic-field value can be higher than that in a synchrotron with a time-varying field, which can help to constrain the machine size. In addition, there are no separate-function magnets, so provided sufficient space is left for injection, extraction, accelerating cavities, vacuum ports, and so forth, the machine may be made compact as the main magnets perform all the required functions simultaneously. The FFAG layout in Fig. 2 is called a 'radial-sector' layout as the faces of the magnets lie along radial lines from the machine centre.

There is in fact a second type of FFAG which does away with the need for the reverse-polarity field, known as the 'spiral' FFAG. This is formed from a succession of hills and valleys of field distributed in the azimuthal direction, where the magnets have a spiral angle with respect to the beam as shown in Fig. 3. In this case the beam does not enter the magnet exactly perpendicular to the face of the

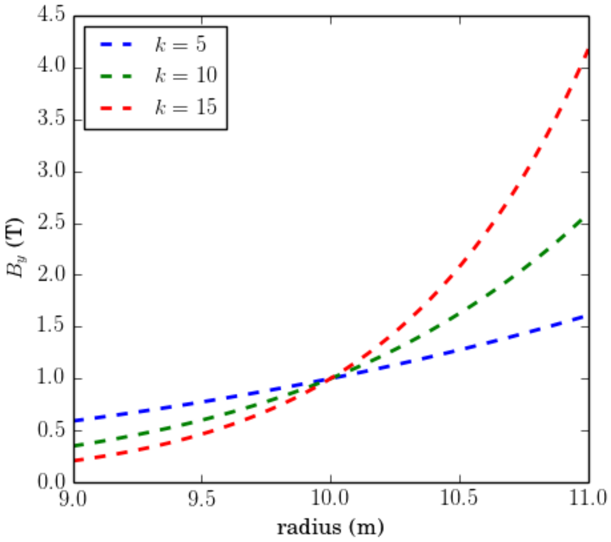


Fig. 1: An example of the characteristic scaling law for different values of the field index k, with $r_0 = 10$ m and $B_0 = 1.0$ T.

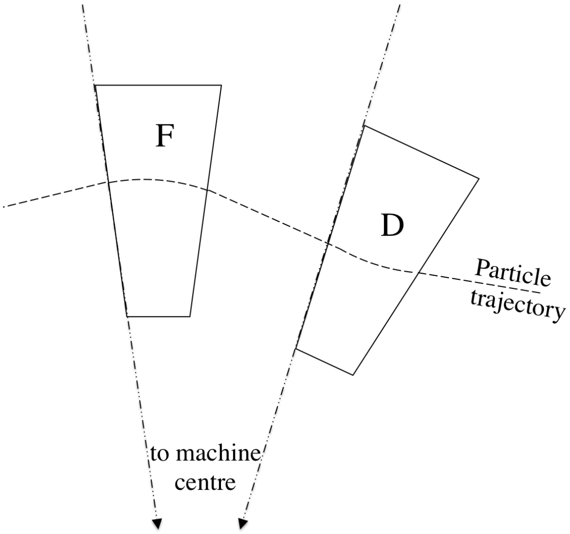


Fig. 2: Radial-sector FFAG layout with reverse bend

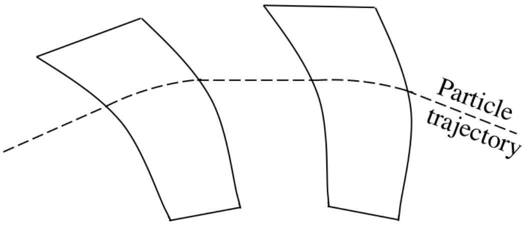


Fig. 3: Spiral FFAG layout, removing the need for reverse bends

magnet, and thus edge focusing results. This removes the need for the opposing reverse bending magnet, while maintaining strong focusing. In the spiral FFAG, the focusing can also be made independent of momentum.

The field in a spiral FFAG in general has the form

$$B(r,\theta) = B_0 \left(\frac{r}{r_0}\right)^k F(\vartheta), \tag{5}$$

where we call ϑ a generalized angle. This is related to the usual azimuth by

$$\vartheta = \theta - \tan\zeta \ln \frac{r}{r_0}. \tag{6}$$

The value of ζ is the angle between the field spiral and the radial direction, and $\tan\zeta$ is constant. As before, B_0 and r_0 are constants. The function $F(\vartheta)$ is an arbitrary periodic function with period $\vartheta_0 = 2\pi/N$, and N is the number of periods. In the radial-sector scaling FFAG that we discussed earlier, all the orbits enlarge and remain geometrically similar, but in the spiral FFAG they also turn about the centre. This can be seen in an example spiral FFAG design for the RACCAM project in Section 4.

In terms of beam dynamics, it is useful to compare the scaling FFAG with a synchrotron. Modern synchrotrons employ the principle of alternating-gradient or 'strong' focusing [9, 10], in which alternating focusing and defocusing magnets lead to much stronger focusing forces in the transverse plane than in constant-gradient weak-focusing synchrotrons. This alternating-gradient focusing is also employed in the FFAG. The transverse beam dynamics in the FFAG is therefore much the same as in the synchrotron, at least for a single orbit or energy, in the sense that we may discuss beta functions, dispersion, and so forth. The difference is that in this case the field is highly nonlinear and these transverse optics functions may vary with radius.

3 The non-scaling FFAG

The non-scaling FFAG allows the strict scaling laws applied in the original scaling FFAG to be relaxed. The idea of violating the strict scaling law of the FFAG occurred to Kent Terwilliger and Lawrence W. Jones in the 1950s [11], but such a machine was never pursued. Two of the main disadvantages of the original FFAG are the highly nonlinear magnetic field required and the large aperture of the magnets and RF arising from the shift of the orbit with energy, which can be up to the order of 0.5–1.0 m. The non-scaling FFAG arose from the question "what if we violate the scaling law?" Or, more specifically, "What if we take a line tangent to the scaling law in Fig. 1, such that the field is linear with radius?" This radical idea led to the linear non-scaling FFAG and was proposed in the 1990s [12].

3.1 Linear non-scaling FFAG

The linear non-scaling FFAG is so called because it uses only up to linear focusing elements, that is, quadrupole and dipole fields. When only quadrupoles and no higher-order multipoles are used for focusing, the beam shifts outward with acceleration because of dispersion and sees a reduced level of focusing. This is really like considering a synchrotron where we do not ramp the magnets with time. In the scaling FFAG, we got around this by varying the gradient with the momentum and by making the beam pipe wider to allow for the orbits moving. But in the linear non-scaling FFAG we ignore the scaling law and any focusing issues for now, which allows us to increase the gradient and reduce the dispersion function even further to reduce the shift of the orbit with momentum. To achieve this, a linear non-scaling FFAG lattice may use normal bending with a defocusing 'D' quadrupole and reverse bending with a focusing 'F' quadrupole, and may (or may not, depending on the design) change at high momentum to use the 'D' quadrupole for reverse bending and the 'F' for normal bending, as described in Ref. [4].

One must then ask what happens to the beam dynamics in such an accelerator. One consequence is that the orbits no longer 'scale', so they are no longer geometrically similar at different energies.

The orbits can be made much more compact than in the scaling FFAG. However, the most dramatic difference is that the betatron tunes are no longer constant with energy. This may seem surprising if you have worked on (almost) any other type of accelerator, as the betatron tunes are usually designed to be kept constant. In the linear non-scaling FFAG, they vary dramatically throughout the acceleration cycle, crossing not just high-order betatron resonances but also integer resonances.

In theory, if the acceleration is fast enough, the beam may be able to cross betatron resonances before they have time to build up, and therefore any amplitude growth effects may be mitigated. How fast this crossing needs to be depends on imperfections and alignment errors in the machine, and clearly necessitates a fast acceleration rate. In fact, the linear non-scaling FFAG was proposed in the context of muon acceleration, where very fast acceleration before the muons decay is an absolute requirement. The many questions surrounding the dynamics of such a machine led to the construction of the first non-scaling FFAG accelerator, known as EMMA.

3.2 The EMMA non-scaling FFAG

EMMA is the Electron Machine for Many Applications, a 10–20 MeV electron accelerator constructed at Daresbury Laboratory to demonstrate the technology of the linear non-scaling FFAG accelerator. The accelerator is shown in Fig. 4, and measures just over 5 m in diameter. Despite the name, EMMA is in fact a demonstrator machine; the primary purpose of its commissioning in 2010–2011 was to investigate some of the key dynamics issues and technical issues involved in realizing the technology. This being demonstrated, it aimed to show that the non-scaling FFAG was a viable technology for use in ‘many applications’ ranging from medical use and muon acceleration to high-power proton accelerators.

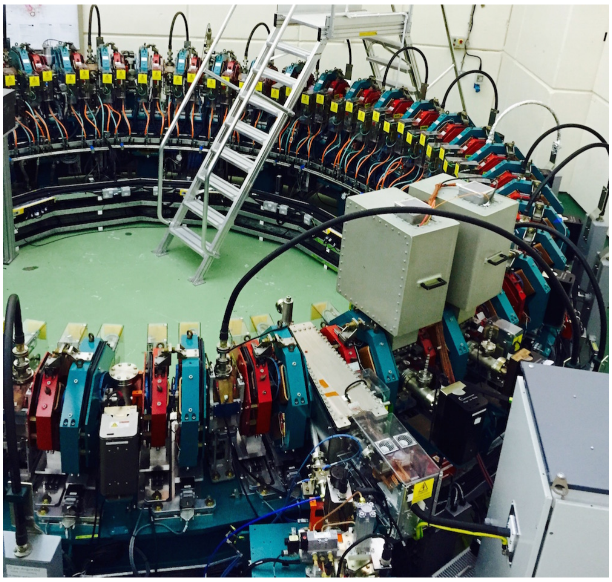


Fig. 4: The EMMA non-scaling FFAG at the STFC Daresbury Laboratory, UK

A few key points of interest, of relevance to medical accelerators, were learned from the EMMA experiment.[1] Most importantly, EMMA demonstrated the successful operation of a linear non-scaling FFAG, bringing this technology into reality. Next, EMMA demonstrated that fast crossing of integer betatron resonances is possible if the beam is accelerated quickly enough. However, EMMA accelerates in just 10 turns from injection to extraction and it is not practicable to accelerate a hadron beam so quickly over the energies required for medical use. We shall return to this point in our discussion of hadron FFAG design studies. More information about EMMA can be found in Refs. [13, 14].

[1] EMMA also demonstrated a number of other interesting accelerator physics concepts, including a novel acceleration mechanism known as ‘serpentine acceleration’.

4 FFAGs for hadron therapy

In this section we turn to addressing the specific idea of FFAG accelerators for medical applications, in particular hadron therapy. State-of-the-art hadron therapy (proton and light-ion therapy) centres must provide beams for patient treatment with unprecedented precision, flexibility, and reliability within a hospital environment. This can pose challenges for present accelerator technology, in terms of both beam requirements and considerations of machine size and cost. Existing technological options may meet present requirements, but one should ask whether they are the optimal solution for the future of this treatment modality. Can we do better?

In accelerator terms, the beam energies required for therapy are relatively low compared with the multi-TeV energy range required for high-energy physics experiments. Proton energies up to 250 MeV are required for treatment, and up to 330 MeV if online proton radiography is incorporated. For heavier ions, full-body-treatment-energy C^{6+} is equivalent in magnetic rigidity to protons with an energy of around 1.2 GeV. These energies can be achieved with a number of different types of accelerator, including synchrotrons, cyclotrons, and linear accelerators. Because of the limited space available in a clinical environment, a circular accelerator is usually chosen as the main accelerator for a facility.[2] There are a number of possible choices of circular accelerator. Most existing or planned facilities use the established technology of either the synchrotron or the cyclotron.

Cyclotrons have a limited energy range, as it is difficult to maintain vertical focusing at high energy and the magnets become increasingly unwieldy as the energy and thus machine size grow. They are also limited by their fixed extraction energy. For hadron therapy, this means in practice that the beam energy must be adjusted between the accelerator and the patient using an energy degrader. Although a cyclotron can deliver ample dose to the patient, the use of energy degraders and passive scattering systems has led to concerns about activation and radiation protection [16]. State-of-the-art systems ought to use active scanning systems and active energy variation to deliver the best-quality treatment.

Synchrotrons operate with a pulsed beam; however, most synchrotrons have a slow cycle rate, on the order of 1 Hz. Rapid-cycling synchrotrons have been proposed for this application, with rates up to around 50 Hz. Their main advantage is their flexibility, particularly in terms of energy reach and easy variable-energy extraction. However, the pulsed nature of these machines makes scanning rather slow. This problem has been partly overcome by using a stable slow extraction of the beam, although the energy variability is still limited by the repetition rate of the synchrotron. In addition, synchrotrons are generally much larger than cyclotrons.

This discussion leads us to ask where the FFAG accelerator might enter the picture. With a fixed magnetic field, the repetition or cycle rate can be much higher than in a synchrotron, up to the range of kHz, limited only by the speed of the RF system. With variable-energy extraction, this high repetition rate could enable slice-to-slice energy variation without limiting the dose rate for full 3D conformal irradiation, or even 3D tumour motion tracking during treatment. Such a machine is not limited in energy range by its dynamics like a cyclotron, and so can easily reach the energies required for therapy. If fast extraction can be achieved at any energy, one may even conceive of a machine which operates with different ion species on a pulse-by-pulse basis, using some pulses for imaging and some for treatment while using the same acceleration and beam delivery system. Of course, this requires that the extraction line, beamlines, gantry, diagnostics, and quality assurance are able to handle such flexibility. In essence, the FFAG has the potential to remove some of the limitations of existing technologies and provide a flexible, relatively compact alternative.

4.1 The KEK 150 MeV FFAG

The first prototype FFAG for medical applications was a 150 MeV proton accelerator at KEK in Japan. This was one of the first hadron FFAGs ever to be built and is a radial-sector scaling FFAG using a

[2] However, recent developments such as the ‘Cyclinac’ design are being studied for this application [15].

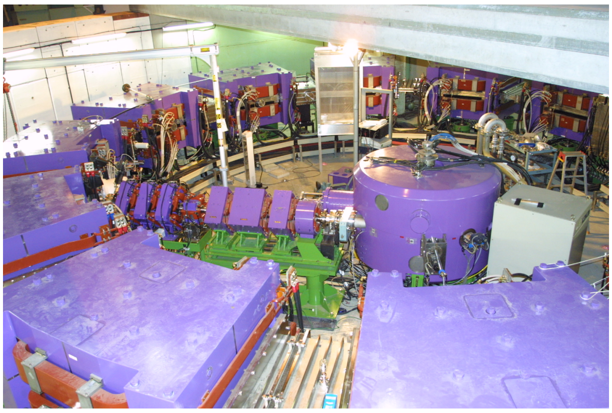


Fig. 5: The KEK 150 MeV scaling FFAG, Japan

12-sector DFD triplet lattice structure. The orbits vary from an average radius of 4.47 m at 12 MeV to 5.2 m at 150 MeV at the centre of the F magnet. This machine was as much about proof of technology as it was about suitability for medical applications. A number of key innovations were made which addressed some of the technology challenges of the scaling FFAG, including wide-aperture magnetic-alloy RF cavities for acceleration [17] and the development of a 'return-yoke-free' main magnet triplet, which eased beam injection and extraction by allowing traversal through the side of the main magnets. A very similar machine for studies of accelerator-driven systems is also in operation at Kyoto University Research Reactor Institute, and the original KEK machine is now in use at Kyushu University. The machine is shown in Fig. 5 along with the cyclotron injector and transport line.

4.2 The RACCAM project

As discussed earlier, one issue with the radial-sector scaling FFAG is that it has a relatively large circumference because of the reverse bends, but the spiral FFAG does not have this issue. The RACCAM project, which ran from 2006 to 2008 in France, was a university and industry collaboration to design and prototype a spiral scaling FFAG for medical applications. It combined some key benefits, including variable-energy fast extraction and multiport extraction to different beamlines. The machine was designed to accelerate beams over a variable energy range from as low as 5.55 MeV to a maximum of 180 MeV with a repetition rate of at least 100 Hz. The design principles of the machine can be found in Ref. [18] and the layout and orbits for varying energies can be seen in Fig. 6.

As part of the RACCAM project, a suitable spiral FFAG magnet was designed in iteration with beam dynamics studies, and a prototype magnet was fabricated and measured. Details of the design can be found in Ref. [19]. Along with the machine study and the technical design, a lot of work was undertaken in an attempt to optimize the facility itself. The collaboration carried out a study on how best to utilize multiple treatment rooms taking advantage of a machine with multiple extraction ports [20].

4.3 Linear non-scaling FFAG designs

Around the time the linear non-scaling FFAG concept was invented, a number of designs to apply this concept to hadron therapy arose. One example which was studied in detail was designed by Keil *et al.* [21] and consists of three concentric rings, to produce full-energy protons and carbon ions for hadron therapy.

This is the simplest type of non-scaling FFAG, the linear version, consisting of F and D quadrupoles, providing alternating-gradient (F/D) strong focusing. In this lattice, the use of F and D doublet magnets is proposed, in which a radial offset between the F and D quadrupoles provides the dipole bending field, removing the need for separate dipole magnets, as in the EMMA design. The lat-

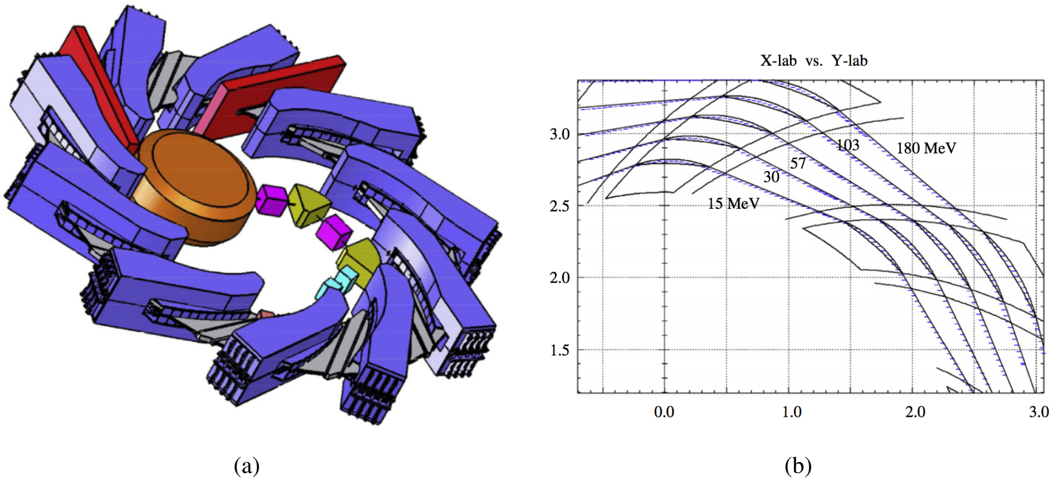


Fig. 6: (a) Machine layout and (b) orbits in the RACCAM FFAG

tice of the second ring, which covers the 31–250 MeV proton energy range (and the equivalent for carbon ions), comprises 48 FD doublet cells and the ring radius is around 6.9 m. The layout of the lattice cell can be seen in Fig. 7 and the optics in Fig. 8.

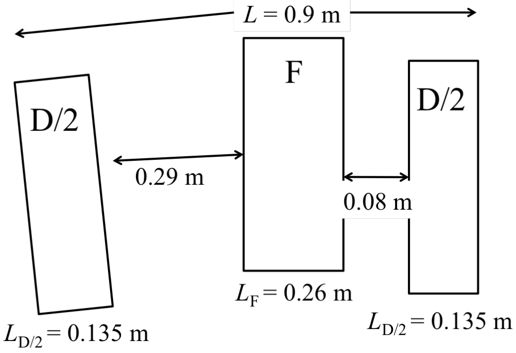


Fig. 7: Layout of one cell of the second (middle) ring from centre to centre of the D magnets in two adjacent cells

As we now know, the betatron tune in a linear non-scaling FFAG changes during acceleration; the variation of the tune in this lattice is shown in Fig. 9. This changing tune will cross integer and half-integer betatron resonances, where any errors in the lattice will result in small kicks to the beam which will build up with each subsequent turn, potentially damaging the beam quality. Acceleration in this case occurs in roughly 1000 turns, which is far slower than the 10-turn acceleration in EMMA.

Detailed studies were carried out on the sensitivity of such a design to alignment errors, and a comparison was made with an alternative design by Trbojevic [22]. The results showed that alignment tolerances in such a machine may be as tight as 10 microns, an extremely stringent requirement. One may improve things by optimizing the lattice for the application so as to cross fewer resonances, but it is likely that resonance crossing will remain an issue [23, 24]. The topic of resonance crossing in non-scaling FFAGs is an ongoing area of active research [25]. In addition, one only needs to note the density of such a lattice to realize that injection and extraction and the lack of long straight sections may be problematic in an operational machine. However, the advantages of the linear non-scaling FFAG may be realized if the machine is made non-symmetric with long straight insertions. The principle may also be used to optimize the treatment gantry rather than the accelerator itself, an option which we shall discuss later.

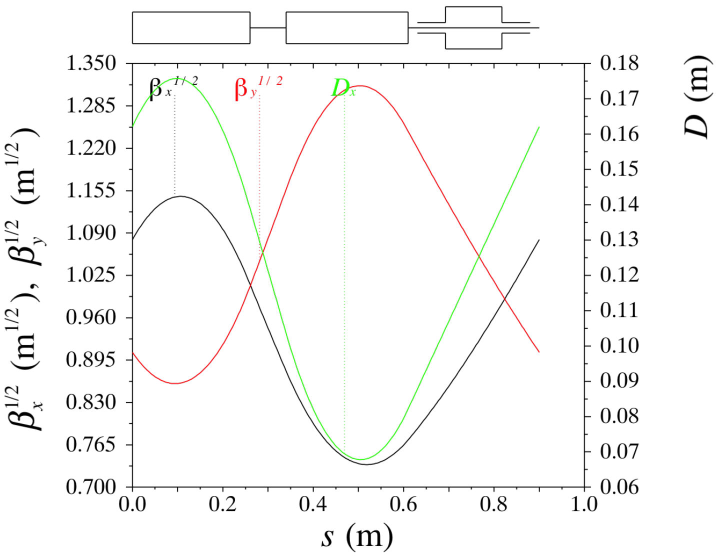


Fig. 8: Beta functions and dispersion in the Keil *et al.* non-scaling FFAG for hadron therapy

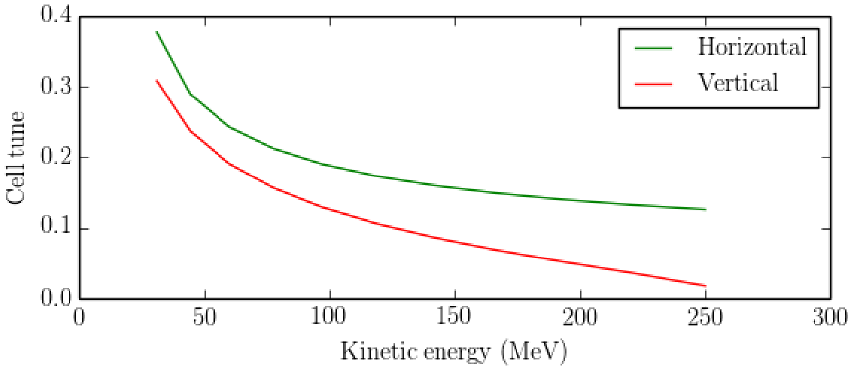


Fig. 9: Variation of betatron tune with acceleration. The cell tune is shown; the machine tune can be found by multiplying by the number of cells, which is 48 in this ring, resulting in the tunes crossing around 10 integers in each plane.

4.4 PAMELA: Particle Accelerator for MEdicaL Applications

As part of the CONFORM (COnstruction of a Non-scaling FFAG for Oncology, Research and Medicine) project, which included the construction of the EMMA non-scaling FFAG, a design study for a hadron non-scaling FFAG for hadron therapy was undertaken from 2007 to 2011. Known as the PAMELA project, this was a design study aimed at applying the concept of a non-scaling FFAG to hadron therapy. The design principle was based on two rings to cover the full energy range of both protons and carbon ions up to an equivalent magnetic rigidity of 6.7 T m, that is, 440 MeV/nucleon for C^{6+} ions (Fig. 10). The machine was designed to operate with a very fast repetition rate of 1 kHz.

The starting point for the study was a linear non-scaling FFAG. However, the concern about beam deterioration from resonance crossing discussed in the previous section was significant enough to motivate studies of alternative designs of FFAGs for medical use. This resulted in a non-linear FFAG design. In each ring, the design strategy resulted in a variation of the total betatron tunes with acceleration which was well within half an integer (i.e., $\Delta\nu_{\text{total}} = n_{\text{cell}}\,\Delta\nu_{\text{cell}} < 0.5$). It also maintained some advantageous non-scaling properties such as small orbit excursion and simpler magnets which are easy to align.

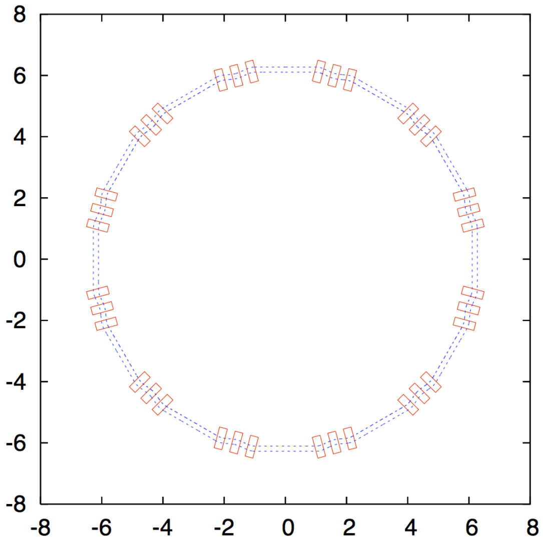


Fig. 10: Layout of the PAMELA proton ring lattice showing the low-energy (inner) and high-energy (outer) orbits. The axis units are metres.

The steps of the design were as follows:

1. The designers started with a scaling law based on a radial-sector FFAG with an FDF triplet lattice.
2. The second stable region of Hill's equation was used, with a horizontal phase advance per cell greater than 180°. This allowed a larger field index to be used, resulting in a reduction in the orbit excursion by a factor of roughly five compared with the first stability region [26].
3. The field was decomposed into its multipole components and the expansion was truncated at some low order, usually octupole or decapole, to simplify the magnets.
4. The shape of the magnets was changed to rectangular rather than sector-shaped, and aligned on a straight line rather than along an arc to simplify construction.
5. The field profile was finally optimized such that the variation in betatron tune was minimized throughout acceleration.

The resulting machine was studied up to engineering design level, including a novel superconducting magnet design [27]; studies of RF cavities; engineering design of the cryostats, injectors, injection and extraction magnets, and beamlines; and a first look at a gantry system [28]. The layout of the injectors and two main rings is shown in Fig. 11.

4.5 NORMA: NORmal conducting Medical Accelerator

The NORMA study followed on from the PAMELA design to address some of the issues that arose during that study and to iterate the design. For example, it was felt that the PAMELA machine might have been closer to implementation if normal-conducting magnets had been used instead of the novel superconducting design. The NORMA design also extended the energy range of a single ring to provide beams up to 350 MeV which could travel through a patient to enable proton radiography. Although the design was for protons only (unlike PAMELA, which was for both protons and carbon ions), it provided detailed studies and optimization based on dynamic aperture [29].

The cell structure (Fig. 12) in this design was very similar to that of PAMELA, using an FDF triplet. The NORMA study also made further innovations by introducing two long straight sections almost 5 m long for injection, extraction, and acceleration by creating a racetrack-shaped machine, shown in Fig. 13.

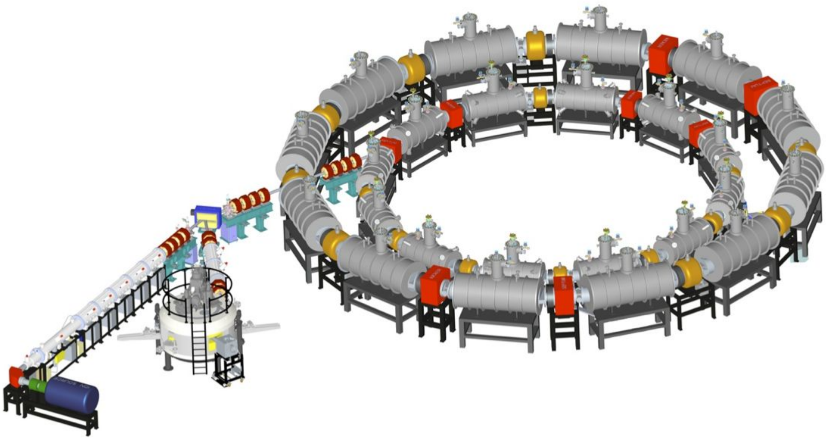


Fig. 11: Layout of the PAMELA facility

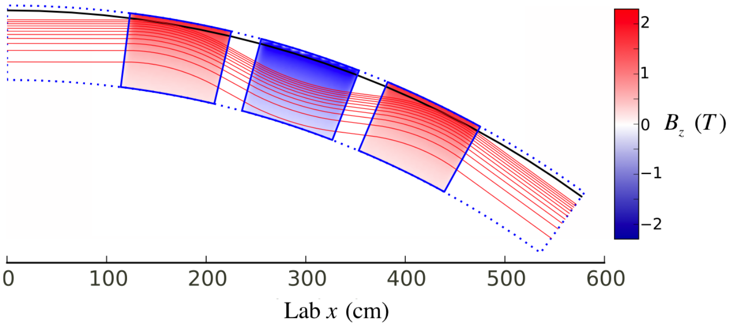


Fig. 12: Layout of the NORMA cell, showing the magnetic field throughout the cell and trajectories of low-energy (inner) to high-energy (outer) orbits.

4.6 Arbitrary-edge-angle non-scaling FFAG designs

Contemporary with the PAMELA design, an alternative approach by Johnstone and Koscielniak used wedge-shaped quadrupoles to achieve control over the betatron tunes with acceleration [30]. In this case both the path length through the quadrupoles and the contribution of edge focusing were designed to be a function of the beam momentum. However, over the energy ranges considered, the orbit shift was relatively large, of the order of 1 m. More information on these designs can be found in Refs. [31, 32].

5 FFAG gantry designs

One of the most promising application areas for the FFAG is in the beam delivery system from accelerator to patient. Existing treatment gantries are large, expensive, and slow to provide variable-energy beams to the patient, as one must wait for the magnetic field to be adjusted before a different energy can be accepted through the transport line. In particular, for carbon ions, the existing gantry at the Heidelberg Ion Therapy centre is both large and heavy, weighing in at 630 t, where 135 t of that is the magnets and the remainder is the supporting structure to allow rotation around the patient.

An FFAG gantry would be able to have a large energy acceptance with a single magnet setting and no limit on rapid variation of energy. This means that if the machine could provide energy variability at a rate of 1 kHz, the beam delivery system and gantry would be able to transport this to the patient treatment room.

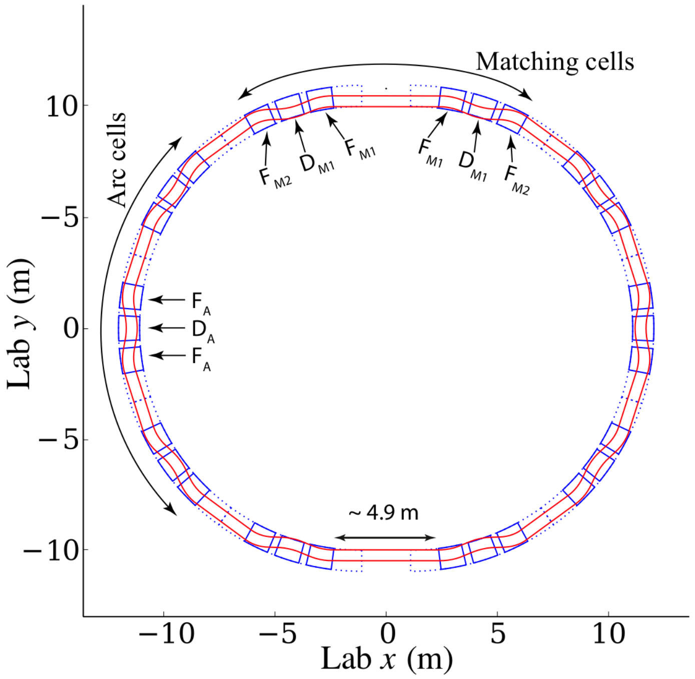


Fig. 13: Layout of the NORMA racetrack FFAG

A transport line and gantry design based on similar optics was developed in the context of the PAMELA project [33–35]. However, in the PAMELA-type design the aperture is still relatively large. Noting that the linear non-scaling FFAG can have an extremely compact aperture over a large momentum range has motivated significant development work on the use of this technique for gantries. Resonance crossing and injection or extraction are obviously not issues in single-pass gantries. The major advantages that could be realized using this technology are reductions in the size, cost, and complexity of treatment gantries.

A design exists for a superconducting non-scaling FFAG gantry for carbon ions [36–38] which may reduce the weight of the magnetic components of such a design from 135 t to around 2 t. There are also designs for compact proton FFAG gantries utilizing either compact magnets or even permanent magnets [39, 40].

6 Summary

I hope you have gained from these notes a basic understanding of FFAG accelerators and their potential for use in medical applications. Here we have focused on hadron therapy applications but, of course, in the high-intensity regime FFAGs also hold promise for other medical applications such as radioisotope production and boron neutron capture therapy. Listed throughout are many references which will provide the reader with a starting point for learning more about the application of FFAG accelerators.

Acknowledgements

The author would like to thank members of the FFAG community who have contributed valuable references, images, and discussions. In particular, I would like to thank S. Tygier, F. Méot, S. Machida, and D. Kelliher.

References

[1] E. Courant and H. Snyder, *Ann. Phys. (N. Y).* **3**(1) (1958) 1. http://dx.doi.org/10.1016/0003-4916(58)90012-5

[2] L. Jones, F. Mills, A. Sessler, K. Symon, and D. Young, *Innovation Was Not Enough* (World Scientific, Singapore, 2009).

[3] A.A. Kolomensky and A.N. Lebedev, *Theory of Cyclic Accelerators* (North-Holland, Amsterdam, 1966), Vol. 1.

[4] S. Machida, Fixed field alternating gradient accelerators, Proc. CERN Accelerator School, Granada, Spain, 2012, https://cas.web.cern.ch/cas/Granada-2012/Lectures/GranadaLectures/Machida.pdf.

[5] M. Oliphant, The acceleration of particles to very high energies, Classified memo submitted to DSIR, University of Birmingham Archive (1943).

[6] T. Ohkawa, Proc. Annual Meeting of JPS, 1953.

[7] K.R. Symon, D.W. Kerst, L.W. Jones, L.J. Laslett, and K.M. Terwilliger, *Phys. Rev.* **103**(6) (1956) 1837. http://dx.doi.org/10.1103/PhysRev.103.1837

[8] A.A. Kolomensky, *ZhETF* **33298** (1957) 1371.

[9] E.D. Courant, M.S. Livingston, and H.S. Snyder, *Phys. Rev.* **88**(5) (1952) 1190. http://dx.doi.org/10.1103/PhysRev.88.1190

[10] E.D. Courant and H.S. Snyder, *Ann. Phys.* **3**(1) (1958) 1. http://dx.doi.org/10.1016/0003-4916(58)90012-5

[11] L.W. Jones, Kent M. Terwilliger: graduate school at Berkeley and early years at Michigan, 1949–1959, Kent M. Terwilliger Memorial Symposium, 1989 (AIP Conference Proceedings No. 237, 1991), p. 1.

[12] C. Johnstone, W. Wan, and A. Garren, Fixed field circular accelerator designs, Proc. Particle Accelerator Conference, New York, 1999, **5** p. 3069. http://dx.doi.org/10.1109/pac.1999.792155

[13] S. Machida *et al.*, *Nature Phys.* **8**(3) (2012) 243. http://dx.doi.org/10.1038/nphys2179

[14] R. Barlow *et al.*, *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.* **624**(1) (2010) 1. http://dx.doi.org/10.1016/j.nima.2010.08.109

[15] U. Amaldi *et al.*, Cyclinacs: fast-cycling accelerators for hadrontherapy (2009), http://arxiv.org/pdf/0902.3533.

[16] E. Pedroni, Latest developments in proton therapy, Proc. EPAC'00, Vienna, 2000, p. 240, http://www.jacow.org.

[17] Y. Yonemura, A. Takagi, M. Yoshii, Y. Mori, M. Aiba, K. Okabe, and N. Ikeda, *Nucl. Instrum. Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip* **576**(2–3) (2007) 294. http://dx.doi.org/10.1016/j.nima.2006.11.072

[18] S. Antoine *et al.*, *Nucl. Instrum. Methods A* **602**(2) (2009) 293. http://dx.doi.org/10.1016/j.nima.2009.01.025

[19] T. Planche *et al.*, *Nucl. Instrum. Methods A* **604**(3) (2009) 435. http://dx.doi.org/10.1016/j.nima.2009.02.026

[20] F. Meot, *Phys. Procedia* **66** (2015) 361. http://dx.doi.org/10.1016/j.phpro.2015.05.045

[21] E. Keil, A.M. Sessler, and D. Trbojevic, *Phys. Rev. ST Accel. Beams*, vol. **10**(5), p. 054701, May 2007. http://dx.doi.org/10.1103/PhysRevSTAB.10.054701

[22] D. Trbojevic, Small proton therapy accelerator by non-scaling FFAG, Presented at FFAG'08 Workshop, Manchester, UK, 2008, http://www.cockcroft.ac.uk/events/FFAG08/programme.htm.

[23] S.L. Sheehy and D.J. Kelliher, *Int. J. Mod. Phys. A* **26**(10–11) (2011) 1842. http://dx.doi.org/10.1142/S0217751X11053237

[24] S.L. Sheehy, D.Phil. thesis, University of Oxford, 2010.

[25] K. Moriya *et al.*, *Phys. Rev. ST Accel. Beams* **18**(3) (2015) 034001. http://dx.doi.org/10.1103/PhysRevSTAB.18.034001

[26] S. Machida, *Phys. Rev. Lett.* **103**(16) (2009) 164801. http://dx.doi.org/10.1103/PhysRevLett.103.16480

[27] H. Witte, *et al.*, *IEEE Trans. Appl. Supercond.* **22**(2) (2012) 2100110. http://dx.doi.org/10.1109/TASC.2012.2186135

[28] K.J. Peach *et al.*, *Phys. Rev. ST Accel. Beams* **16**(3) (2013) 030101. http://dx.doi.org/10.1103/PhysRevSTAB.16.030101

[29] J.M. Garland, R.B. Appleby, H. Owen, and S. Tygier, *Phys. Rev. ST Accel. Beams* **18**(9) (2015) 094701. http://dx.doi.org/10.1103/PhysRevSTAB.18.094701

[30] C. Johnstone and S. Koscielniak, New nonscaling FFAG for medical applications, Proc. IEEE Particle Accelerator Conf., 2007. http://dx.doi.org/10.1109/pac.2007.4440631

[31] C. Johnstone *et al.*, Nonscaling FFAG variants for HEP and medical applications, Proc. Particle Accelerator Conf., Vancouver, 2009, no. TU6PFP080, p. 1478, http://www.jacow.org.

[32] C. Johnstone *et al.*, *Int. J. Mod. Phys. A*, vol. **26**(10–11) (2011) p. 1690. http://dx.doi.org/10.1142/S0217751X11053110

[33] S. Machida and R. Fenning, *Phys. Rev. ST Accel. Beams* **13**(8) (2010) 084001. http://dx.doi.org/10.1103/PhysRevSTAB.13.084001

[34] R. Fenning, Ph.D. thesis, Brunel University, Uxbridge, UK, 2011.

[35] R. Fenning, S. Machida, D. Kelliher, A. Khan, and R. Edgecock, JINST **7** (2012) P05011.

[36] D. Trbojevic, R. Gupta, B. Parker, E. Keil, and A.M. Sessler, Superconducting non-scaling FFAG gantry for carbon/proton cancer therapy, Proc. IEEE Particle Accelerator Conference, 2007, p. 3199. http://dx.doi.org/10.1109/pac.2007.4440714

[37] D. Trbojevic *et al.*, Superconducting non-scaling FFAG gantry for carbon/proton cancer therapy, Brookhaven National Laboratory, BNL-77556-2007-CP (2007), http://www.bnl.gov/isd/documents/35754.pdf.

[38] D. Trbojevic, B. Parker, E. Keil, and A.M. Sessler, *Phys. Rev. ST Accel. Beams* **10**(5) (2007) 053503. http://dx.doi.org/10.1103/PhysRevSTAB.10.053503

[39] D. Trbojevic, FFAG gantry design, presented at FFAG'14 workshop, Brookhaven National Laboratory, 2014.

[40] D. Trbojevic, Update on the innovative carbon/proton non-scaling FFAG isocentric gantries for the cancer therapy, Proc. IPAC'10, Kyoto, 2010, p. 124, http://www.jacow.org

Participants

ABELLEIRA FERNANDEZ, J.L.	CERN, Geneva, CH
ADJEI, D.	Institute of Optoelectronics, Warsaw, PL
ALARAUDANJOKI, J.	University of Jyvaskyla, Jyvaskyla, FI
ALMALKI, M.	GSI, Frankfurt, DE
ASHANIN, I.	NRNU MEPhI, Moscow, RU
BELLINZONA, V.	University of Pavia, Pavia, IT
BENEDETTI, S.	CERN, EPFL, TERA Foundation, Geneva, CH
BENNA, M.	Varian Medical Systems, Troisdorf, DE
BODENDORFER, M.	CERN, Geneva, CH
BRACHT, S.	GSI, Darmstadt, DE
DELA CRUZ, J.	Canadian Light Source Inc., Saskatoon, CA
DIMOV, V.	CERN, Geneva, CH
DOS SANTOS AUGUSTO, R.M.	CERN, Geneva, CH
DUECK, J.	Paul Scherrer Institut, Villigen, CH
EBERHARDT, M.	Varian Medical Systems, Troisdorf, DE
EICHINGER, M.	MedAustron, Wiener Neustadt, AT
ESPOSITO, M.	ADAM SA, Geneva, CH
ESTEBAN MULLER, J.	CERN, EPFL, Geneva, CH
FUERTINGER, M.	MedAustron, Wiener Neustadt, AT
GARLASCHE, M.	CERN, Geneva, CH
GINER NAVARRO, J.	CERN, Geneva, CH
GONSALVES, B.	CERN, Geneva, CH
HUGGINS, A.	Varian Medical Systems Particle Therapy, Troisdorf, DE
IAKOVENKO, V.	Kiev Institute for Nuclear Research, Kiev, UA
JAATINEN, J.	University of Jyvaskyla, Jyvaskyla, FI
JANSSEN, X.	VDL ETG Research, Eindhoven, NL
JASELSKYTE, E.	Hospital of Lithuania, Kaunas, LT
JUNTONG, N.	Synchrotron Light Research Institute, Muang, TH
JUNUZOVIC, J.	MedAustron, Wiener Neustadt, AT
KANAPELKA, M.	JIPNR, Minsk, RU
KLUCHEVSKAYA, Y.	NRNU MEPhI, Moscow, RU
KOUBEK, B.	CERN, Geneva, CH
KOZLOWSKA, W.	CERN, Geneva, CH
KRANTZ, C.	Eppelheim, DE
KRONBERGER, M.	MedAustron, Wiener Neustadt, AT
KURFUERST, C.	MedAustron, Wiener Neustadt, AT
LANGEGGER, R.	MedAustron, Wiener Neustadt, AT
LANTE, V.	CNAO, Pavia, IT
LASHEEN, A.	CERN, Geneva, CH
LI, H.	Uppsala University, Uppsala, SE
LOPEZ, R.	CERN, Geneva, CH
MIZIC-BAJRIC, S.	MedAustron, Wiener Neustadt, AT
MORA VALLEJO, L.	CERN, Geneva, CH
NADIG, P.	MedAustron, Wiener Neustadt, AT
OLIVER, C.	CIEMAT, Madrid, ES
PATTARANUTAPORN, P.	Mahidol University, Bangkok, TH
PELLETIER, S.	MedAustron, Wiener Neustadt, AT
PEREZ, J.M.	CIEMAT, Madrid, ES

PERRELET, D.	CERN, Geneva, CH
PERUSKO, D.	Cosylab D.D., Ljubljana, SL
PIOLI, S.	University of Rome "La Sapienza", Rome, IT
PITMAN, S.	The Cockcroft Institute, Warrington, GB
PIVI, M.	MedAustron, Wiener Neustadt, AT
REPOVZ, M.	MedAustron, Wiener Neustadt, AT
RIZZOGLIO, V.	Paul Scherrer Institut, Villigen, CH
ROSSI, A.	CERN, Geneva, CH
ROTHE, S.	CERN, Geneva, CH
ROY, G.	CERN, Geneva, CH
SABATO, L.	University of Sannio, Benevento, IT
SANCHES ARIAS, J.	MedAustron, Wiener Neustadt, AT
SCHMITZER, C.	MedAustron, Wiener Neustadt, AT
SCHOEMERS, C.	HIT, Heidelberg, DE
SCHUH, S.	CERN, Geneva, CH
SCHWARZ, S.	MedAustron, Wiener Neustadt, AT
SCHWINDLING, J.	CEA Saclay, Gif sur Yvette Cedex, FR
SEDLACKOVA, K.	Slovak University of Technology, Bratislava, SK
SEIDEL, S.	Helmholtz Zentrum Berlin, Berlin, DE
SENZACQUA, M.	University of Rome "La Sapienza", Rome, IT
STADLBAUER, T.	MedAustron, Wiener Neustadt, AT
STROHMEIER, M.	Marburg Ion Beam Therapy Center, Marburg, DE
SUKHIKH, E.	Tomsk Polytechnik University, Tomsk, RU
TRIEBL, O.	MedAustron, Wiener Neustadt, AT
WALLNER, J.	MedAustron, Wiener Neustadt, AT
WASTL, A.	MedAustron, Wiener Neustadt, AT
WU, J.	Zion Biotech Technology Inc., Hsinchu City, CN
XU, Y.	Peking University, Beijing, CN