The integration of a hermetic and high-precision calorimeter system into the overall design of the ATLAS detector and its magnet systems has been a task of high complexity where compromises have had to be made, as will be shown in the first part of this section, which describes the basic requirements and features of the calorimeters. As illustrated in the second part, which highlights some aspects of the construction of the most critical element, namely the electromagnetic calorimeter, and of its measured performance in test beam, the impact of the main design choices and of the technology implementations on the performance has been very significant. A few examples of the overall performance expected in the actual configuration of the experiment are presented in Sect. 16.8.2, where it is also compared to the expected performance of the CMS calorimeter system.

16.4.1 General Considerations

16.4.1.1 Performance Requirements

The main performance requirements from the physics on the calorimeter system can be briefly summarised as follows:

- excellent energy and position resolution together with powerful particle identification for electrons and photons within the relevant geometrical acceptance (full azimuthal coverage over $|\eta| < 2.5$) and over the relevant energy range (from a few GeV to several TeV). The electron and photon identification requirements are particularly demanding at the LHC, as already explained in Sect. 16.2.1. These considerations induce requirements of high granularity and low noise on the calorimeters. One has to add to this the operational requirements of speed of response and resistance to radiation (the electromagnetic calorimeters will have to withstand neutron fluences of up to 10^{15} n/cm^2 and ionising radiation doses of up to 200 kGy over 10 years of LHC operation at design luminosity).
- excellent jet energy resolution within the relevant geometrical acceptance, which is similar to that foreseen for the electron and photon measurements (see above). The quality of the jet energy resolution would play an important role in the case of discovery of supersymmetric particles with cascade decays into many hadronic jets [24].
- good jet energy measurements over the coverage required to contain the full transverse energy produced in hard-scattering collisions at the LHC. A calorimetry coverage over $|\eta| < 5$ is necessary to unambiguously ascribe the observation of significant missing transverse energy to non-interacting particles, such as neutrinos from W-boson decay or light neutralinos from supersymmetric particle cascade decays. With adequate calorimetry coverage providing precise measurements of the missing transverse energy, the experiments will be able to reconstruct invariant masses of pairs of hadronically decaying τ-leptons produced for example in the decays of supersymmetric Higgs bosons. They

will also thus be able to identify forward jets produced in vector-boson fusion processes.
- good separation between hadronic showers from QCD jets and those from decays of τ-leptons.
- fast and efficient identification of the processes of interest at the various trigger levels, in particular for the L1 trigger (see Sect. 16.6).

16.4.1.2 General Features of Electromagnetic Calorimetry

The ATLAS EM calorimeter [25] is divided into a barrel part covering approximately $|\eta| < 1.5$ and two end-caps covering $1.4 < |\eta| < 3.2$, and its main parameters are listed in Table 16.10. Its fiducial coverage is without appreciable cracks, except in the transition region between the barrel and end-cap cryostats, where the measurement accuracy is degraded over $1.37 < |\eta| < 1.52$ because of large energy losses in the material in front of the active EM calorimeter, which reaches up to $6\,X_0$. The excellent uniformity of coverage is a direct consequence of the design of this lead/liquid Argon sampling calorimeter with accordion-shaped electrodes and absorbers. The total thickness of the EM calorimeter varies from a minimum of $24\,X_0$ (at $\eta \approx 0$) to a maximum of $35\,X_0$ (at $\eta \approx 2.5$). This depth is sufficient to contain EM showers at the highest energies (a few TeV) and preserve the energy resolution, in particular the constant term which is dominant above a few hundred GeV.

As can be seen from Table 16.10, the ATLAS EM calorimeter has been designed with both excellent lateral and longitudinal granularity, with samplings in depth optimised for energy loss corrections (presampler) and for shower pointing accuracy together with γ/π^0 and electron/jet separation (strips). The intrinsic performance of the EM calorimeter is however significantly affected by the unavoidable amount of material which had to be incorporated in the tracker system (see Fig. 16.10), and also by the cryostats and the solenoid coil in the case of the ATLAS EM calorimeter (see Sect. 16.8.2 for more details).

16.4.1.3 General Features of Hadronic Calorimetry

Figure 16.13 shows the total number of absorption lengths contained in the ATLAS hadronic calorimetry and in front of the muon system as a function of pseudorapidity. Good containment of jets of typically 1 TeV energy requires about $11\,\lambda$ in the full calorimeter, a target which has been achieved over most of the pseudorapidity range.

For the central part of the hadronic calorimetry, which covers the range $0 < |\eta| < 1.7$, the sampling medium consists of scintillator tiles and the absorber medium is steel. The tile calorimeter is composed of three parts, one central barrel and two extended barrels. The choice of this technology provides maximum radial depth for the least cost for ATLAS. The hadronic calorimetry is extended to

Table 16.10 Main parameters of the ATLAS calorimeter system

	Barrel		End-cap	
EM calorimeter				
Number of layers and $\|\eta\|$ coverage				
Presampler	1	$\|\eta\| < 1.52$	1	$1.5 < \|\eta\| < 1.8$
Calorimeter	3	$\|\eta\| < 1.35$	2	$1.375 < \|\eta\| < 1.5$
	2	$1.35 < \|\eta\| < 1.475$	3	$1.5 < \|\eta\| < 2.5$
			2	$2.5 < \|\eta\| < 3.2$
Granularity $\Delta\eta \times \Delta\phi$ versus $\|\eta\|$				
Presampler	0.025×0.1	$\|\eta\| < 1.52$	0.025×0.1	$1.5 < \|\eta\| < 1.8$
Calorimeter (strip layer)	$0.025/8 \times 0.1$	$\|\eta\| < 1.40$	0.050×0.1	$1.375 < \|\eta\| < 1.425$
	0.025×0.025	$1.40 < \|\eta\| < 1.475$	0.025×0.1	$1.425 < \|\eta\| < 1.5$
			$0.025/8 \times 0.1$	$1.5 < \|\eta\| < 1.8$
			$0.025/6 \times 0.1$	$1.8 < \|\eta\| < 2.0$
			$0.025/4 \times 0.1$	$2.0 < \|\eta\| < 2.4$
			0.025×0.1	$2.4 < \|\eta\| < 2.5$
			0.1×0.1	$2.5 < \|\eta\| < 3.2$
Calorimeter (middle layer)	0.025×0.025	$\|\eta\| < 1.40$	0.050×0.025	$1.375 < \|\eta\| < 1.425$
	0.075×0.025	$1.40 < \|\eta\| < 1.475$	0.025×0.025	$1.425 < \|\eta\| < 2.5$
			0.1×0.1	$2.5 < \|\eta\| < 3.2$
Calorimeter (back layer)	0.050×0.025	$\|\eta\| < 1.35$	0.050×0.025	$1.5 < \|\eta\| < 2.5$
Number of readout channels				
Presampler	7808		1536 (both sides)	
Calorimeter	101,760		62,208 (both sides)	
LAr hadronic end-cap				
$\|\eta\|$ coverage			$1.5 < \|\eta\| < 3.2$	
Number of layers			4	
Granularity $\Delta\eta \times \Delta\phi$			0.1×0.1	$1.5 < \|\eta\| < 2.5$
			0.2×0.2	$2.5 < \|\eta\| < 3.2$
Readout channels			5632 (both sides)	
LAr forward calorimeter				
$\|\eta\|$ coverage			$3.1 < \|\eta\| < 4.9$	
Number of layers			3	

(continued)

Table 16.10 (continued)

	Barrel		End-cap	
Granularity $\Delta x \times \Delta y$ [cm]			FCal1: 3.0 × 2.6	$3.15 < \|\eta\| < 4.30$
			FCal1: ∼ four times finer	$3.10 < \|\eta\| < 3.15$,
				$4.30 < \|\eta\| < 4.83$
			FCal2: 3.3 × 4.2	$3.24 < \|\eta\| < 4.50$
			FCal2: ∼ four times finer	$3.20 < \|\eta\| < 3.24$,
				$4.50 < \|\eta\| < 4.81$
			FCal3: 5.4 × 4.7	$3.32 < \|\eta\| < 4.60$
			FCal3: ∼ four times finer	$3.29 < \|\eta\| < 3.32$,
				$4.60 < \|\eta\| < 4.75$
Readout channels			3524 (both sides)	
Scintillator tile calorimeter				
	Barrel		Extended barrel	
$\|\eta\|$ coverage	$\|\eta\| < 1.0$		$0.8 < \|\eta\| < 1.7$	
Number of layers	3		3	
Granularity $\Delta\eta \times \Delta\phi$	0.1 × 0.1		0.1 × 0.1	
Last layer	0.2 × 0.1		0.2 × 0.1	
Readout channels	5760		4092 (both sides)	

larger pseudorapidities by a copper/liquid-argon calorimeter system, which covers the range $1.5 < |\eta| < 3.2$, and by the forward calorimeters, a set of copper-tungsten/liquid-argon detectors at larger pseudorapidities. The hadronic calorimetry thus reaches one of its main design goals, namely coverage over $|\eta| < 4.9$.

The ATLAS forward calorimeters are fully integrated into the cryostat housing the end-cap calorimeters, which reduces the neutron fluence in the muon system and, with careful design, affects very little the neutron fluence in the tracker volume. The main role of these calorimeters is to keep the tails in the measurement of missing transverse energy at a low level and to tag jets in the forward direction rather than to accurately measure their energy, so their geometry has been simplified and their readout costs have been minimised. The forward calorimeters are based on copper (front) and tungsten (back) absorber bodies and absorber rods, the latter being parallel to the beam and slotted into precisely machined holes. The gaps in these holes are filled with LAr and operated at an electric field of about 1 kV/mm.

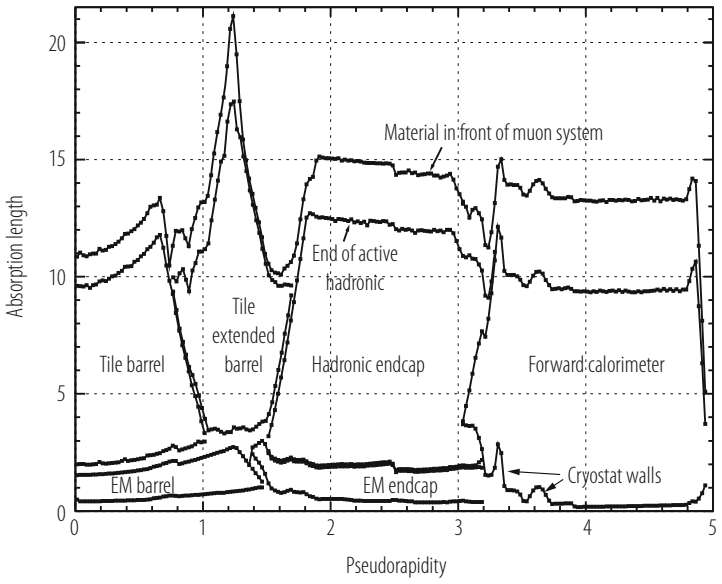

Fig. 16.13 Distribution of amount of material (in absorption lengths) for the ATLAS calorimetry (and in front of the muon system) as a function of η

16.4.2 Construction Experience and Measured Performance in Test Beam

As has been described above, the ATLAS calorimeters comprise a variety of technologies, each with its own challenges and pitfalls, and only a few of the most prominent examples of lessons learned during construction can be given in this review.

The biggest challenge has clearly been the construction of the electromagnetic calorimeters. The technology chosen for the ATLAS EM calorimeter, although based on a well established technique had a number of innovative features, which resulted in some major production issues:

- the most difficult part of the project, by far, has been the fabrication in industry of large electrodes of about 2 m length containing about 1000 resistive pads each. This problem was overcome through the careful monitoring of the production on-site by experts from the collaboration.
- a total of about 20,000 m^2 of honeycomb spacers have been used to maintain the flexible electrodes in the centre of the gap between absorbers. To avoid major problems with the high-voltage behaviour of assembled modules, a rigorous and careful cleaning procedure for all parts, especially the honeycomb, had to be implemented.

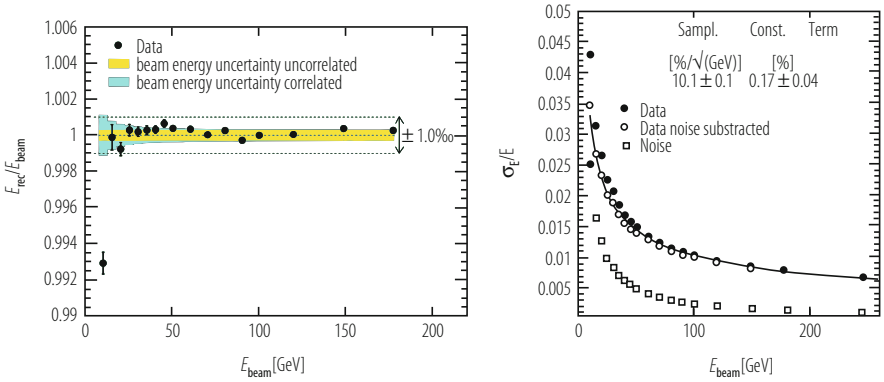

Fig. 16.14 Linearity of response (left) and energy resolution (right) obtained for a production module of the ATLAS barrel EM calorimeter as a function of the incident electron beam energy

- radiation-tolerant electronics had to be produced for all components in the cavern. This comprises all the front-end electronics boards housed near the signal feed-throughs.

The ATLAS collaboration has performed an extensive programme of test-beam measurements to calibrate and characterise the EM calorimeter modules [26]. The original plans called for a test-beam calibration of about 20% of the modules. In the end, a smaller fraction of 15% of the ATLAS EM modules underwent detailed test-beam measurements, and a few recent results from these stand-alone calibration campaigns are presented here.

Figure 16.14 shows that a linearity of response of ±1 per mil has been obtained over an electron energy range from 20 to 180 GeV for an ATLAS barrel LAr EM module. To achieve this, while preserving the energy resolution (also shown in Fig. 16.14), requires a thorough understanding of the material in front of the active calorimeter and a careful evaluation of the weights and corrections to be applied to the raw cluster energy. The uniformity of response across the whole module has also been measured and found to contribute an r.m.s. of 0.4% to the global constant term, which is within the specifications set to the LAr EM calorimeter (see Sect. 16.8.2 for a more detailed discussion of the various contributions to the constant term for the EM calorimeters).

16.5 Muon Spectrometer System

Muons are a very robust, clean and unambiguous signature of much of the physics that ATLAS has been designed to study. The ability to trigger and to reconstruct muons at the highest luminosities of the LHC has been incorporated into the design of the experiment from the very beginning [29]. In fact, the concepts chosen for

measuring muon momenta have shaped the experiment more than any other physics consideration (see also Sect. 16.2.1).

As discussed already in Sect. 16.2.2, the choice of magnet was motivated by the method which would be used for the measurement of muons with momenta up to $\sim$TeV scales. ATLAS has thus opted for a high-resolution, stand-alone measurement independently of the rest of the sub-detectors, resulting in a very large volume, with low material density, over which the muon measurement takes place. The ATLAS toroidal magnetic field provides a momentum resolution which is essentially independent of pseudorapidity up to a value of 2.7.

This section reviews the main features of the muon spectrometer system and discusses a few of the challenges encountered. A few examples of the overall performance expected in the actual configuration of the experiment are presented in Sect. 16.8.3, where it is also compared to the expected performance of the CMS muon system.

16.5.1 General Considerations

The physics signatures that give rise to muons are numerous and varied. At the highest momenta, they include muons from new high-mass (multi-TeV) resonances such as heavy neutral gauge bosons, Z', as well as decays from heavy Higgs bosons. At the lowest end of the spectrum, B-physics relies on the reconstruction of muons with momentum down to a few GeV. The resulting requirements are:

- Resolution: the 'golden' decay of the Standard Model Higgs boson into four muons, $H \rightarrow ZZ \rightarrow 4\,\mu$, requires the ability to reconstruct the momentum and thus mass of a narrow two-muon state with a precision at the level of 1%. At the upper end of the spectrum, the goal is to achieve a 10% momentum resolution for 1 TeV muons.
- Wide rapidity coverage: almost two-thirds of the decays of an intermediate-mass Higgs boson to four muons have at least one muon in the region $|\eta| > 1.4$. A hermetic system, which measures muons up to $|\eta| \sim 2.5$, has turned out to be the best compromise.
- Identification inside dense environments, e.g. hadronic jets or regions with high backgrounds.
- Trigger: the ability to measure the momenta of muons online on a stand-alone basis, i.e. without reference to any other detector system, and to select events with muons above 5–10 GeV momentum is of paramount importance.

There are also the requirements which result from the 25 ns spacing in time between successive beam crossings and from the neutron radiation environment of the experimental halls. Good timing resolution and the ability to identify the bunch-crossing in question, as well as redundancy in the measurements, are therefore also demanded of the muon detectors, which represent by far the largest and most difficult system to install in the experiment.

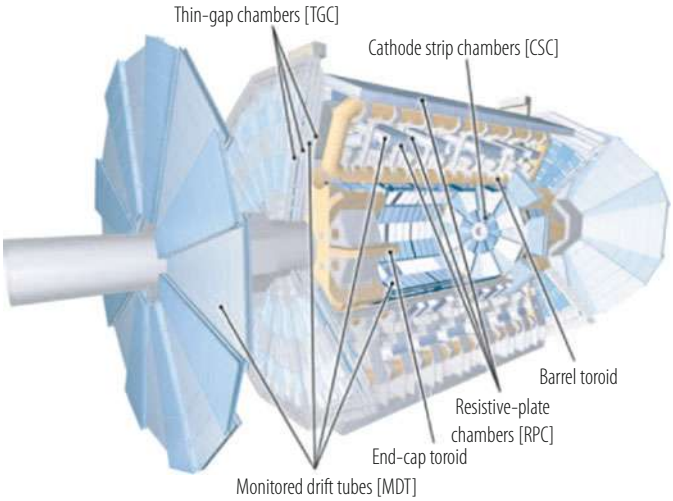

Fig. 16.15 Cut-away view of the ATLAS muon spectrometer system. displaying the regions in which the different muon chamber technologies are used

The conceptual layout of the muon spectrometer is shown in Fig. 16.15 and the main parameters of the muon chambers are listed in Table 16.11. It is based on the magnetic deflection of muon tracks in the large superconducting air-core toroid magnets, instrumented with separate trigger and high-precision tracking chambers. Over the range $|\eta| < 1.4$, magnetic bending is provided by the large barrel toroid. For $1.6 < |\eta| < 2.7$, muon tracks are bent by two smaller end-cap magnets inserted into both ends of the barrel toroid. Over $1.4 < |\eta| < 1.6$, usually referred to as the transition region, magnetic deflection is provided by a combination of barrel and end-cap fields. This magnet configuration provides a field which is mostly orthogonal to the muon trajectories, while minimising the degradation of resolution due to multiple scattering. The anticipated high level of particle flux has had a major impact on the choice and design of the spectrometer instrumentation, affecting performance parameters such as rate capability, granularity, ageing properties, and radiation hardness. In the barrel region, tracks are measured in chambers arranged in three cylindrical layers around the beam axis; in the transition and end-cap regions, the chambers are installed in planes perpendicular to the beam, also in three layers.

16.5.1.1 Muon Chamber Types

Over most of the η-range, a precision measurement of the track coordinates in the principal bending direction of the magnetic field is provided by Monitored Drift Tubes (MDT's). The mechanical isolation in the drift tubes of each sense wire from its neighbours guarantees a robust and reliable operation. At large pseudorapidities, Cathode Strip Chambers (CSC's, which are multiwire proportional chambers with

Table 16.11 Main parameters of the ATLAS muon spectrometer

Monitored drift tubes	MDT
Coverage	$\lvert\eta\rvert < 2.7$ (innermost layer: $\lvert\eta\rvert < 2.0$)
Number of chambers	1088 (1150)
Number of channels	339,000 (354,000)
Function	Precision tracking
Cathode strip chambers	CSC
Coverage	$2.0 < \lvert\eta\rvert < 2.7$
Number of chambers	32
Number of channels	31,000
Function	Precision tracking
Resistive plate chambers	RPC
Coverage	$\lvert\eta\rvert < 1.05$
Number of chambers	544 (606)
Number of channels	359,000 (373,000)
Function	Triggering, second coordinate
Thin gap chambers	TGC
Coverage	$1.05 < \lvert\eta\rvert < 2.7$ (2.4 for triggering)
Number of chambers	3588
Number of channels	318,000
Function	Triggering, second coordinate

Numbers in brackets for the MDT's and the RPC's refer to the final configuration of the detector in 2009

cathodes segmented into strips) with higher granularity are used in the innermost plane over $2 < |\eta| < 2.7$, to withstand the demanding rate and background conditions. The stringent requirements on the relative alignment of the muon chamber layers are met by the combination of precision mechanical-assembly techniques and optical alignment systems both within and between muon chambers.

The trigger system covers the pseudorapidity range $|\eta| < 2.4$. Resistive Plate Chambers (RPC's) are used in the barrel and Thin Gap Chambers (TGC's) in the end-cap regions. The trigger chambers for the muon spectrometer serve a threefold purpose: provide bunch-crossing identification, provide well-defined p_T thresholds, and measure the muon coordinate in the direction orthogonal to that determined by the precision-tracking chambers.

16.5.1.2 Muon Chamber Alignment and B-Field Reconstruction

The overall performance over the large areas involved, particularly at the highest momenta, depends on the alignment of the muon chambers with respect to each other and with respect to the overall detector.

The accuracy of the stand-alone muon momentum measurement necessitates a precision of 30 μm on the relative alignment of chambers both within each projective tower and between consecutive layers in immediately adjacent towers. The internal deformations and relative positions of the MDT chambers are monitored by approximately 12,000 precision-mounted alignment sensors, all based on the optical monitoring of deviations from straight lines. Because of geometrical constraints, the reconstruction and/or monitoring of the chamber positions rely on somewhat different strategies and sensor types in the end-cap and barrel regions, respectively.

The accuracy required for the relative positioning of non-adjacent towers to obtain adequate mass resolution for multi-muon final states, lies in the few millimetre range. This initial positioning accuracy is approximately established during the installation of the chambers. Ultimately, the relative alignment of the barrel and forward regions of the muon spectrometer, of the calorimeters and of the tracker will rely on high-momentum muon trajectories.

For magnetic field reconstruction, the goal is to determine the bending power along the muon trajectory to a few parts in a thousand. The field is continuously monitored by a total of approximately 1800 Hall sensors distributed throughout the spectrometer volume. Their readings are compared with magnetic-field simulations and used for reconstructing the position of the toroid coils in space, as well as to account for magnetic perturbations induced by the tile calorimeter and other nearby metallic structures.

The muon system consists of three large superconducting air-core toroid magnets, which are instrumented with different types of chambers to provide the two needed functions, namely high-precision tracking and triggering. The central (or barrel) region, $|\eta| < 1.0$, is covered by a large barrel magnet consisting of eight coils which surround the hadron calorimeter. In this region, tracks are measured in chambers arranged in three cylindrical layers (stations) around the beam axis. In the end-cap region, $1.4 < |\eta| < 2.7$, muon tracks are bent in two smaller end-cap magnets inserted into both ends of the barrel toroid. The intermediate (transition) region, $1.0 < |\eta| < 1.4$, is less straightforward, since here the barrel and end-cap fields overlap, thus partially reducing the bending power. To keep a uniform resolution in this region, tracking chambers are place in strategic places to improve the quality and accuracy of the measurement. Due to financial constraints, one out of three sets of chambers in this region has been staged, thus leading to an inferior performance in the transition region for the first years of data-taking.

The layout of the ATLAS muon spectrometer system is shown in Fig. 16.15. A total of four types of detectors are used, the choice of technology being driven by the very large surface to be covered, by trigger and precision measurement requirements, and by the different radiation environments. Resistive Plate Chambers (RPC) in the barrel region ($|\eta| < 1.05$) and Thin Gap Chambers (TGC) in the end-cap regions ($1.05 < |\eta| < 2.4$) are used for triggering purposes. These chambers provide a fast response with good time resolution but rather coarse position resolution. The precision measurements are performed by Monitored Drift Tubes (MDT) over most of the coverage. In the regions at large $|\eta|$, where background

conditions are harsher and the rate of muon hits is therefore larger, Cathode Strip Chambers (CSC) are used.

The basic principle of the muon measurement in the ATLAS muon spectrometer is to obtain three segments (or super-points) along the muon trajectory. For momenta up to 300 GeV, the resolution is limited to a few percent by multiple scattering and fluctuations in the energy loss in the calorimeters, and can therefore be improved by combining the momentum measurement with that obtained in the Inner Detector. The momentum resolution goals quoted above at higher momenta imply a very high precision of 80 μm on the individual hits, given the three-point measurement and the available bending power. The required precision on the muon momentum measurement also implies excellent knowledge of the magnetic field. The air-core toroid design leads to a magnetic field, which is modest in average magnitude (0.5 T), but is also inhomogeneous, and must therefore be measured and monitored with high precision (at the level of 20 G). The inhomogeneity of the field and its rapid variations cannot be approximated by simple analytical descriptions and have to be accounted for carefully, thereby enhancing the importance of the use of the inner detector information to reconstruct low-momentum muon tracks with low fake rates.

16.5.1.3 Alignment

Alignment of the muon chambers with respect to each other and with respect to the overall detector is a critical ingredient, key to obtaining the desired performance over the large areas involved, particularly at the highest momenta. The high accuracy of the ATLAS stand-alone measurement necessitates a very high precision of 30 μm on the alignment.

The chambers have however been installed with an accuracy of a few mm, and obviously, no attempt at repositioning the chambers once their installation is completed can realistically be made. Instead, intricate hardware systems have been designed to measure the relative positions between chambers contributing to the measurement of the same tracks, but also to monitor any displacements during the detector operation. These systems are designed to provide continuous monitoring of the positions of the chambers with or without collisions in the accelerator. The very strict requirement of a 30 μm alignment has necessitated the design of a complex system, in which optical sensors are mounted with very high mechanical mounting precision (better than 20 μm in the precise coordinate). The system uses $\sim$5000 alignment sensors, which are either installed on the chambers or in the so-called alignment bars (long instrumented Aluminium cylinders with deformations monitored to within 10 μm, which constitute the alignment reference system in the end-caps). In addition, 1789 magnetic field sensors (3D Hall probes) are also being installed on the chambers to determine with high accuracy the position and shape of the conductors of each coil. From these accurate measurements, the field will be determined throughout the whole volume to an accuracy of about 20 G, provided all magnetic materials are also mapped and described accurately.

The final alignment values will clearly be obtained with the large statistics of muon tracks traversing the muon chambers (rates of about 10 kHz are expected at a luminosity of $10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ for muons with $p_T > 6\,\mathrm{GeV}$).

16.5.2 Construction Experience and Measured Performance in Laboratory and Test Beam

The muon chambers are based on technologies, which were used in previous experiments: drift tubes and CSCs have been used widely in the past; RPCs were used in the L3 and Babar experiments, while TGCs were used in OPAL. Nevertheless, large R&D efforts have been necessary to address the special requirements of the LHC environment.

The high particle fluxes (mainly photons and neutrons) have necessitated searches for the right type of materials and gases, which prevent wire deposits in the case of drift tubes, while new operational modes were developed for the RPCs (proportional regime instead of the streamer regime used in previous experiments) and the TGCs (quasi-proportional mode instead of saturated mode), with the corresponding required changes in the front-end electronics.

In the case of the ATLAS muon spectrometer, the requirement of a precise stand-alone measurement limits the amount of material in order to minimise multiple scattering. This has led to the development of thin but precise Aluminium tubes, which are mounted on very light structures. The deformations of these structures can be monitored by a sophisticated alignment system, as well as the extensive use of paper honeycomb in the trigger chambers to limit the contribution of the detectors in the material description.

Beyond this, the greatest challenge came mostly from the very large, unprecedented areas that the muon chambers had to cover and the correspondingly large numbers of electronic channels. The ATLAS muon system contains approximately $25{,}000\,\mathrm{m}^2$ of active detection planes, and roughly one million electronic channels. The main parameters of the muon chambers are listed in Table 16.11.

The requirement of achieving all this within 'reasonable cost' was actually one of the biggest issues encountered. In terms of lessons learned from the construction process; beyond the general observations made in Sect. 16.2.3, three issues emerge as the most important ones:

- Putting in place, right from the beginning, very tight procedures for quality assurance/quality control (QA/QC). Given the enormous number of elements (wires, strips, tubes, supports) involved, the presence of well-defined and complete QA/QC systems was of the utmost importance. Any and all issues which went unnoticed sooner or later resulted in time and energy-consuming corrective procedures being taken.
- Planning for services. Despite all initial designs and tolerances and safety factors, the cabling procedures always turn out to be more complicated, more time-

consuming and eventually more space-consuming than planned. Whereas the first two issues can, at least in principle, be solved with additional manpower and increased costs, the space issue is a major one, which needs adequate planning right from the start. The space issue has been compounded by the fact that the muon system is traversed by the services of the other detectors, leading to issues of ownership of space and to problems in collecting all the necessary information for proper planning. This major complexity of the actual installation of the services has been one of the major challenges of the Technical Coordination team.

- Uniformity of technologies, power supplies and electronics. As already explained in the introduction, the size of the muon project has necessitated the distribution of the design and construction across different institutes and funding agencies. This necessarily leads to a multitude of different choices for numerous components, from the choice of high-voltage power supplies to basic choices of electronics (ASICs or FPGAs). A strong electronics coordination team is needed to alleviate many of these pressures and lead to an overall system, which will be much easier to maintain.

As for the other detector systems, the ATLAS collaboration has invested a major effort into the validation of the muon spectrometer concept using high-energy test-beam muons. The ATLAS muon test-beam setup had both trigger and tracking chambers placed in the appropriate geometrical positions and equipped with alignment sensors. The most prominent goal (in 2004) was to test the ability to monitor chamber movements and long-term deformations over time-scales of several weeks with the required accuracy, a crucial ingredient for the ultimate accuracy of muon measurements in the TeV range. The test-beam setup included the calculation of deviations from the nominal chamber positions and the storage of the results in a database. These constants were also directly determined by the reconstruction program. The variation of the sagitta as reconstructed in the muon beam, along with that measured from the optical alignment system, was studied over a period covering the thermal fluctuations of a day–night cycle. The spread of the difference between the two distributions was measured to be below 10 μm, i.e. well within the specification of 30 μm. Finally, the correct performance of the trigger was tested with the final trigger electronics prototypes and with all muon systems taking data simultaneously at 40 MHz.

16.6 Trigger and Data Acquisition System

This section briefly describes the main design features and architecture of the ATLAS trigger and data acquisition systems. A few examples of the overall trigger performance expected in the actual configuration of the experiment are presented in Sect. 16.8.4, where it is also compared to the expected performance of the CMS trigger system.

The trigger and data acquisition (DAQ) system of an experiment at a hadron collider plays an essential role because both the collision and the overall data rates are much higher than the rate at which one can write data to mass storage. As mentioned previously, at the LHC, with the beam crossing frequency of 40 MHz, at the design luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, each crossing results in an average of $\sim$23 inelastic p-p collisions with each event producing approximately 1–2 MB of zero-suppressed data. These figures are many orders of magnitude larger than the archival storage as well as the offline processing capability, which correspond to data rates of 200–300 MB/s, or of 100–200 Hz.

The required event rejection power of the real-time system at design luminosity is thus of O(10^7), which is too large to be achieved in a single processing step, if a high efficiency is to be maintained for the physics phenomena of interest. For this reason, the selection task is split into a first, very fast selection step, followed by two steps in which the selection is refined.

The first step (L1 trigger) makes an initial selection based on information of reduced granularity and resolution from only a subset of detectors. This L1 trigger is designed to reduce the rate of events accepted for further processing to less than 100 kHz, i.e. it provides a rejection of a factor $\sim 10^4$ with respect to the collision rate. The figure of 100 kHz is an 'asymptotic' one, to be fully used at the highest luminosities when the beam and experiment conditions demand it, and financial resources allow it. It is expected that at startup, and also during the first years of LHC operation, the L1 trigger will operate at lower rates.

The second step (high-level trigger or HLT) is designed to reduce the L1 accept rate to the final output rate of $\sim 10^2$ Hz. Filtering in the HLT is provided by software algorithms running in large farms of commercial processors, connected to the detector readout system via commercial networks. The physical implementation of the HLT selection is implemented in a two-step process, with independent farms for each of the two steps.

Some key requirements on the overall system are:

- To provide enough bandwidth and computing resources, within financial constraints, to minimise the dead-time at any luminosity, while maintaining the maximum possible efficiency for the discovery signals. The current goal is to have a total dead-time of less than a few (1–2)%. Most of this dead-time is currently planned to occur in the L1 trigger.
- To be robust, i.e. provide an operational efficiency which does not depend significantly on the noise and other conditions in the detector or on changes with time of the calibration and detector alignment constants.
- To provide the possibility of validating and of computing the overall selection efficiencies using only the data themselves, with as little reference to simulation as possible. This implies usage of multiple trigger requirements with overlapping thresholds.
- To uniquely identify the bunch crossing that gave rise to the trigger.
- To allow for the readout, processing and storage of events that will be needed for calibration purposes.

16.6.1 General Considerations

The most important architectural decision in the Trigger/DAQ system is the number of physical entities, or trigger levels, which will be used to provide the rate reduction of $O(10^3)$ from the rate of 100 kHz accepted by the L1 trigger to the final rate to storage of $O(10^2)$ Hz. Current practice for large general-purpose experiments operating at CERN, DESY, Fermilab, KEK and SLAC is to use at least two more entities, colloquially referred to as the L2 and L3 triggers. Some experiments even have a L4 trigger. The higher the level, the more general-purpose the implementation, with the L3 and L4 trigger systems always relying on farms of standard commercial processors.

The implementation of the L2 trigger system varies significantly across experiments, from customised in-house solutions to independent processor farms. The issue encountered by all experiments, which have opted for multiple trigger levels, is the definition of the functionality that the L2 system should provide. Of all the trigger levels after L1, the L2 trigger is the most challenging one, since it has to operate at the highest event rates, often without the benefit of full-granularity and full-resolution data, though with data from more detectors and of higher quality than that used by the L1 Trigger. Decisions that have to be made are the rejection factor that the L2 trigger must provide, the quality of the information it will be provided with, the interconnects between the detector readout, the L1 trigger and the L2 trigger, and finally, the actual implementation of the processing units which will execute the selection algorithms.

Ideally, the High-Level Trigger (HLT) should have no built-in architectural nor design limitations other than the total bandwidth and CPU, which can be purchased based on the experiments resources. Indeed, from very early on, the desire to provide the maximum possible flexibility to the HLT led to the first design principle adopted by ATLAS: the HLT selection should be provided by algorithms executed on standard commercial processors, avoiding all questions and uncertainties related to home-grown hardware processors.

The architecture is depicted schematically in Fig. 16.16. The implementation of the L2 trigger has the advantage that much less data are required to flow into the event filter farm, which in turn has more time to process incoming events. The L2 farm, on the other hand, has to provide a decision on all the events accepted by the L1 trigger. To reduce the data flow into the L2 farm, only a fraction of the detector information is actually transferred from the readout buffers to the L2 processors. This is the concept of the "Region of Interest" (ROI). In brief, the result of the L1 trigger drives the L2 processing, by indicating the regions of the detector which are involved in scrutinising the physics object (electron, muon, jet,...) identified by the L1 trigger. These regions are small, with a total data size of only a few percent of the total event size, so that the full set of data from these regions can be transferred to the L2 farm. The L2 algorithms employ sequential selection and usually not all the data from the ROI in question have to be read in. This farm has tens of *ms* to provide the L2 decision. The events accepted by L2 are sent to the event filter farm, which

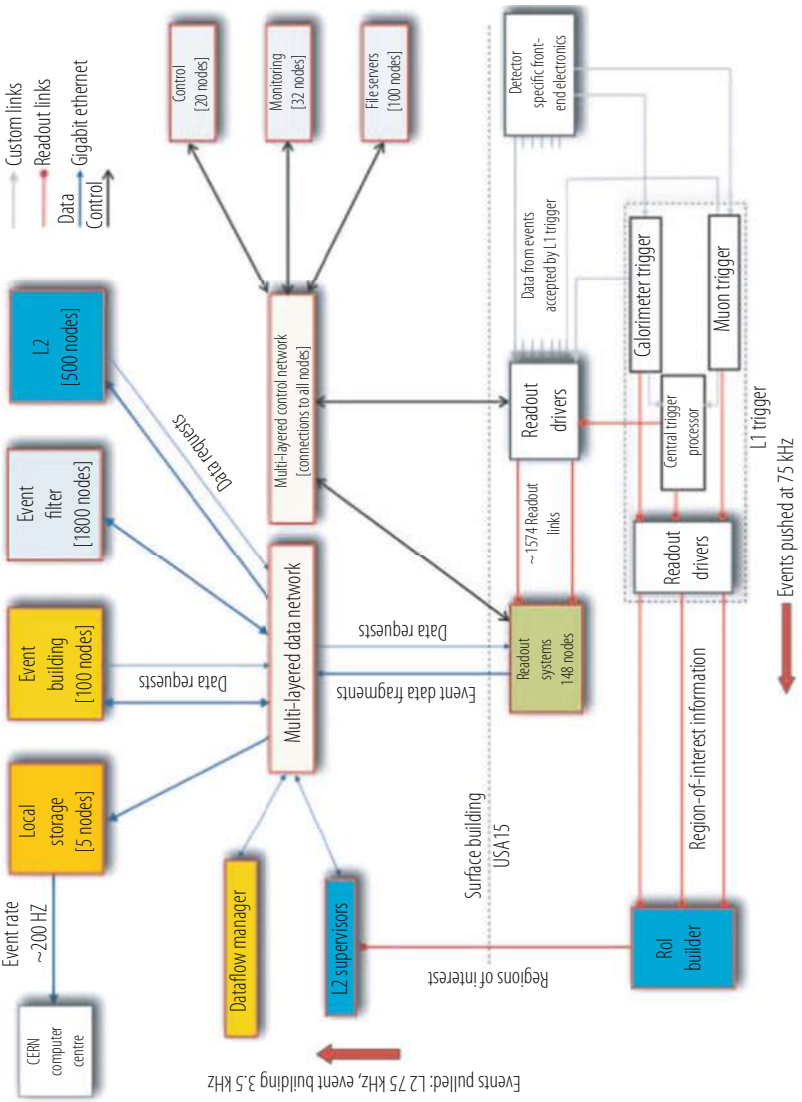

Fig. 16.16 Block diagram of the ATLAS trigger and data acquisition system. Also shown are the different components of the dataflow

now has access to the full event data. This farm runs the final, essentially offline-like selection, "seeding" the reconstruction from the objects previously identified by the L2 trigger in order to reduce the total processing time. The rate input into the event filter farm is a few kHz, so the selection at this level has to provide typically a factor of 10 in rate reduction.

The system relies on commercially available networks for the interconnection between the readout buffers and the HLT farm. The advent of very inexpensive Gbit Ethernet switching fabrics and processor interfaces, along with the rapidly deployable 10 Gbit Ethernet standard, have rendered all early thoughts (back in the mid-1990's) of potential home-grown solutions obsolete.

16.6.2 L1 Trigger System

The L1 trigger has to process information from the detector at the full beam crossing rate of 40 MHz. The very short time between two successive beam crossings (25 ns), along with the wide geographical distribution of the electronic signals from the detector, excludes real-time processing of the full detector data by general-purpose, fully programmable processing elements.

The data are, instead, stored in pipelines awaiting the decision of the L1 trigger within up to 3 μs. The maximum time available for processing in the L1 trigger system is determined by the limited memory resources available in the front-end (FE) electronics which store the detector data during the L1 decision-making process. Technology and financial considerations at the time of the design resulted in a limit of at most 128 bunch crossings, i.e. the equivalent of approximately 3 μs of data, which can be stored in the FE memories. This total latency of 3 μs therefore includes the unavoidable latency components associated with the transfer of the detector information to the processing elements of the L1 trigger and with the latency of the propagation of the L1 decision signals back to the FE electronics. The resulting time available for the actual processing of the data is no more than $\sim 1-1.5\,\mu$s.

In order to avoid dead-time, the trigger electronics must also be pipelined since every process in the trigger must be repeated every 25 ns. The high operational speed and pipelined architecture also imply that only specific data can be brought to the corresponding processing elements in the trigger system. In addition, the data must flow synchronously across the trigger logic in a deterministic manner.

This architecture results in the presence of data from multiple crossings being processed sequentially through the various stages of the trigger logic. To achieve this, most trigger operations are either simple arithmetic operations or functions, which use memory look-up tables, where an address is used to produce rapidly a previously calculated (and stored) result. Moreover, the short time available significantly restricts the data, which can be used in forming the L1 trigger decision, in two ways: on the timing front, the only usable data can come from detectors with very fast response or from slower detectors, which have both good time resolution

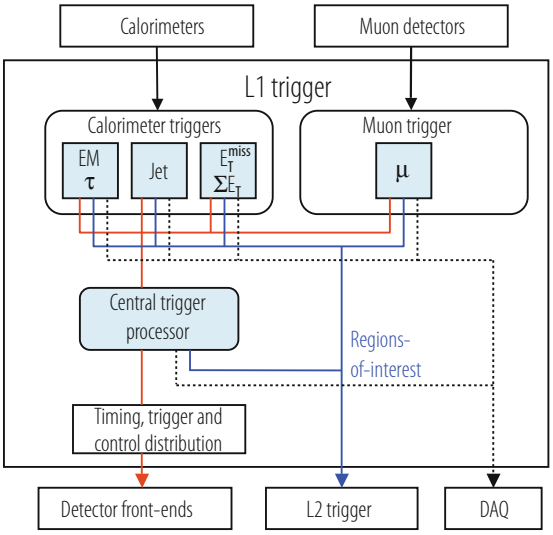

Fig. 16.17 Block diagram of the ATLAS L1 trigger. The overall L1 accept decision is made by the central trigger processor, taking input from calorimeter and muon trigger results. The paths to the detector front-ends, L2 trigger, and data acquisition are shown from left to right in red, blue and black, respectively

and low occupancy; on the volume front, only reduced, coarse information from the calorimeter and muon chambers, corresponding to a smaller fraction of the total volume, and thereby requiring less processing power than e.g. tracker data, can be used.

The block diagram of the ATLAS L1 trigger is shown in Fig. 16.17. It contains a calorimeter trigger, a muon trigger and an overall central trigger processor. The system relies on a Timing, Trigger and Control (TTC) system derived from a precision 40 MHz clock distributed by the LHC accelerator. The different subsystems are essentially independent of each other and the interactions among them are limited to the explicit communication lines in the diagram.

16.6.2.1 Muon Trigger

The L1 muon trigger provides the trigger processor with information on the number, quality and transverse momentum of muon tracks in each event. It consists of a barrel section, two end-cap sections and a part which combines the information from the full system and prepares the input to the central trigger processor. The chambers used in the L1 trigger are used mainly for this purpose, i.e. in the end-cap the L1 muon trigger system uses Thin Gap Chambers (TGC) to cover the region of small angles with respect to the beam axis, whereas, in the barrel, it uses Resistive Plate Chambers (RPC). In both cases, the chambers were selected on their ability to provide signals fast enough for the L1 trigger. Each of the two L1 muon trigger systems has its own trigger logic with different pattern-recognition algorithms.

At the end of processing by the local trigger processors, the muon trigger information from the various sources is collected, and the trigger decision is

prepared before presenting it to the central trigger processor. This intermediate stage carries some significant functionality: the muon trigger to central trigger processor Interface resolves overlaps between chamber sectors in the barrel and between barrel and end-cap chambers and forms the muon candidate multiplicities for the trigger decision.

The final decision on the event is obtained by the central trigger processor itself, using either information from only the muon trigger or in association with other objects in the event (e.g. the presence of a high-p_T electron).

16.6.2.2 Calorimeter Trigger

The L1 calorimeter trigger provides essentially all the L1 trigger streams for the experiment (electrons, photons, QCD jets, $\tau-$jets, missing E_T) except for the muons. The architecture of this trigger contains three elements, namely the generation of the trigger primitives, a local calorimeter trigger which processes information from limited parts of the detector, and a global calorimeter trigger which combines all the information from the local processors, prior to sending the summaries to the central trigger processor. Data from the calorimeters are combined to form trigger towers of approximate size 0.1×0.1 in $\eta - \phi$ space. Analogue sums are formed on the detector and sent through analogue transmission to the counting room.

The information is then digitised and processed to determine the transverse energy E_T in each trigger tower. As discussed previously, most of the ATLAS calorimeters have pulse shapes which extend well beyond a single crossing, so the signals are processed to assign each energy deposition to the correct bunch crossing. Once the transverse energies and the bunch crossing are determined, the algorithms in the local calorimeter trigger take over. The basic features can be summarised as follows:

- Electrons and photons are searched for as peaks in the E_T deposited in a limited $\eta - \phi$ region (neighboring towers) of the EM calorimeter. The corresponding hadronic energy is required to be small, relatively to the EM calorimeter energy. Additional isolation requirements, e.g. by demanding that neighbouring towers do not have energy larger than a certain threshold, may be imposed.
- Jets are formed by adding the energy in a large $\eta - \phi$ region consisting of an array of 4×4 trigger towers/elements. The algorithm provides flexibility in the measurement of the jet energy through the use of a sliding window, but therefore requires an additional processing step to settle jet overlaps and eliminate double-counting.
- τ-jets are formed by demanding very narrow energy depositions in the electromagnetic and hadronic calorimeters. Isolation requirements may also be applied.
- The missing transverse energy (as well as the total transverse energy in the event) is estimated from the sum of the transverse energies of all the calorimeter cells.

The sum of the transverse energies of all jets found in the event is also provided; this will be more stable with increasing luminosity than the sum over all cells.

The results of this local processing, i.e. the electron/photon, τ-jet, and jet candidates are passed on to the central trigger processor. The physics objects are sorted in E_T and finally used in the global decision, possibly in association with other L1 objects in the event.

16.6.3 High-Level Trigger and Data Acquisition Systems

Experience with the data acquisition (DAQ) systems of previous experiments at high-energy lepton and hadron colliders has resulted in the establishment of several fundamental design principles which have been embedded in the architecture from the very beginning.

The technological advances witnessed over the last 20 years have progressed at an extraordinary rate, which until now has remained constant with time. It was decided to invest in these advances of technology and especially in the two main fronts that drive them, processing power and network speed. An additional consideration has been the expected evolution of the experiment and its data acquisition system, rendering a fully programmable HLT system highly desirable to avoid major design changes. The added flexibility provided by the fully programmable environment of a standard CPU also implies that algorithmic changes necessary for the resolution of unforeseen backgrounds or other adverse experimental conditions can be easily introduced. A final consideration was the desire to minimise the amount of non-standard, in-house solutions.

As a result of the above considerations, the data acquisition system relies on industrial standards to the greatest possible extent, and employs commercially available components, if not full-fledged systems, wherever these could meet the requirements. This applies to both hardware and software systems. The benefits of this decision are numerous, with the most important ones being the resulting economies in both the development and production costs, the prompt availability of the relevant components from multiple competing sources, and a maintenance and support mechanism which does not employ significant in-house resources.

Another general design principle, adopted at the very earliest stages of development, is that of maximal scaling. This addresses the fact that the accelerator conditions, the experimental conditions, and finally the physics programme itself are all expected to evolve significantly with time. An easily scalable system is one in which the functions, and thus the challenges as well, are factorised into sub-systems with a performance independent of the rest of the system.

The long difference in time between the design of the systems and their final implementation and deployment implied a development cycle different from that of the other detector projects. In the case of the DAQ systems, the understanding of the required functionality of the various elements of the system was, in many

cases, separated from their performance. The numerous and challenging subsystem components were thus developed along two independent paths. The first development path concentrated on the identification and implementation of the full functionality needed for operation in the final DAQ. The second path concentrated on the issues that arise when the functions identified in the first path are executed at the performance levels required by the final DAQ system.

Following these principles, ATLAS has pursued an R&D programme, which has resulted in a system that could be implemented for the early luminosities of the LHC, and could be scaled to the expected needs at the full design luminosity, since the system architecture is such that in a number of incremental steps, the performance of the system can be increased proportionally.

16.6.3.1 Data Acquisition

The main elements of the ATLAS DAQ system are described in more detail below:

- Detector readout system: this consists of modules which read the data corresponding to a single bunch crossing out of the front-end electronics upon the reception of a L1 trigger accept signal. There are approximately 1600 such modules in the ATLAS readout.
- Event builder: this is the collection of networks, which provide the interconnections between the detector readout and the HLT. It provides (and monitors) the data flow and employs a large switching fabric. ATLAS has two such networks, one for the L2 trigger and one for the event filter.
- HLT systems: these are the processors, which deal with the events provided by the detector readout. They execute the HLT algorithms to select the events to be kept for storage and offline processing.
- Controls and monitors: these consist of all the elements needed to control the flow of data (events) through the DAQ system, as well as the elements needed to configure and operate the DAQ. This includes all the provisions for special runs, e.g. for calibrations, that involve special setups for both the detectors, the trigger and the readout. The other major functionality is the monitoring of the various detector elements, of the operation of the L1 and HLT and of the state of the DAQ system and its elements.

The factorisation of the DAQ function into tasks, which can be made almost independent of each other, facilitates the design of a modular system which can be developed, tested and installed in parallel. To ensure this factorisation, the different operational environments of the four functional stages must be decoupled. This is achieved via the introduction of buffering of adequate depth in between each of these stages. The primary purpose of these buffers is to match the very different operating rates of the elements at each stage. As an example, at a rate of 100,000 events per second, the readout system delivers an event every $10\,\mu s$. On the other hand, the event building process requires, even assuming a 100% efficiency of 2 Gb/s links,

a time of ∼ms to completely read in the event. This is therefore the rate at which the elements of the farm system can operate on events. The two time-scales are very different, and this is where the deep buffers present in the readout system serve to minimise the coupling between the stages.

The design of the DAQ system is very modular, thereby allowing for a staged installation. The event builder has been conceived with the possibility of a phased installation from the very beginning. The operation of the ATLAS experiment has begun with a DAQ system serving only a reduced bandwidth of approximately 20–40 kHz. The deferrals were necessary because of funding pressures, whereas a staged installation of the DAQ was viewed as less damaging to the physics programme, since the initial instantaneous luminosity of the LHC is far below the design value.

16.6.3.2 High-Level Trigger

As mentioned previously, the HLT is a software filtering process executed on standard commercially available processors. The software is drawn from the offline reconstruction software of the experiment. Both levels of the HLT are executed within the offline framework, but in contrast to the event filter which uses the same algorithms as the offline, the L2 trigger processors run more dedicated code (in particular with faster data-preparation algorithms). The trigger software is steered differently from the offline and initiates the reconstruction from the physics candidate objects identified by the previous levels (L1 or L2 trigger). The overall rejection factor is achieved by applying, in software, a number of successive reconstruction and selection steps.

As an example, the HLT electron trigger is typically driven by a L1 electron/photon candidate, which is identified as a high-energy isolated electromagnetic (EM) energy deposition in the calorimeters. At the output of the L1 trigger, the rate is dominated by QCD jets. The first task in reconstructing the electron in the HLT is to rerun the clustering algorithm with access to the full granularity and resolution of the EM calorimeter and to obtain a new, more accurate, measurement of the transverse energy (E_T) of the EM cluster. Given the rapidly falling cross section, this already provides a rejection factor of ≈2 with respect to the input event rate. Further shower-shape and isolation cuts are also applied at this point. The events surviving the EM calorimeter requirements are subsequently subjected to a search for a charged-particle track in the tracking detectors. The matching between track and cluster is a powerful requirement, which yields at least a factor of 10 rejection against jets while maintaining a very high efficiency.

Events selected by the HLT are forwarded to mass storage and from there to the offline system for reconstruction and physics analysis. Given the unprecedented rate of online rejection, another very important task of the HLT is to provide detailed information on the events which have been rejected at each stage of the filtering process.

16.7 Computing and Software

The ATLAS computing and software infrastructure is clearly of paramount importance. The functionality and flexibility of both will determine, to a very large extent, the rate and quality of the physics output of the experiment. As expected, there are numerous challenges to be addressed also in these two areas.

On the computing side, the LHC experiments represent a new frontier in high-energy physics. What is genuinely new at the LHC is that the required level of computing resources can only be provided by a number of computing centres working in unison with the CERN on-site computing facilities. Off-site facilities will thus be vital to ATLAS operation to an extent that is completely different from previous large experiments. Usage of these off-site facilities necessitates the substantial use of Grid computing concepts and technologies [33]. The latter allow for the sharing of the responsibility for processing and storing the data, but also for providing the same level of data access, and making available the same amount of computing resources to all members of the collaboration.

A second challenge for computing is the development and operation of a data storage and management infrastructure which is able to meet the demands of a yearly data volume of O(10) Petabytes and is used by both organised data processing and individual analysis activities, which are geographically dispersed around the world.

The architecture which is now in place is geographically distributed and relies on four levels or tiers, as illustrated in Fig. 16.18. Primary event processing occurs at CERN in the so-called Tier-0 facility. Raw data are archived at CERN and sent (along with the reconstructed data) to the Tier-1 centres around the world. These centres share among themselves the archiving of a second copy of the raw data, while they also provide the reprocessing capacity and access to the various versions of the reconstructed data, and allow scheduled analysis of the latter by physics analysis groups. A more numerous set of Tier-2 centres, which are smaller but still have substantial CPU and disk storage resources, provide capacity for analysis, calibration activities and Monte Carlo simulation. Datasets, which are produced at the Tier-1 centres by physics groups, are copied to the Tier-2 facilities for further analysis. Tier-2 centres rely upon the Tier-1 centres for access to large datasets and secure storage of the new data they produce. A final level in the hierarchy is provided by individual group clusters used for analysis: these are the Tier-3 centres.

The ATLAS collaboration also relies on the CERN Analysis Facility (CAF) for algorithmic development work and a number of short-latency data-intensive calibration and alignment tasks. This facility is also expected to provide additional analysis capacity with, as an example, re-processing of the express-stream data and short turn-around analysis jobs.

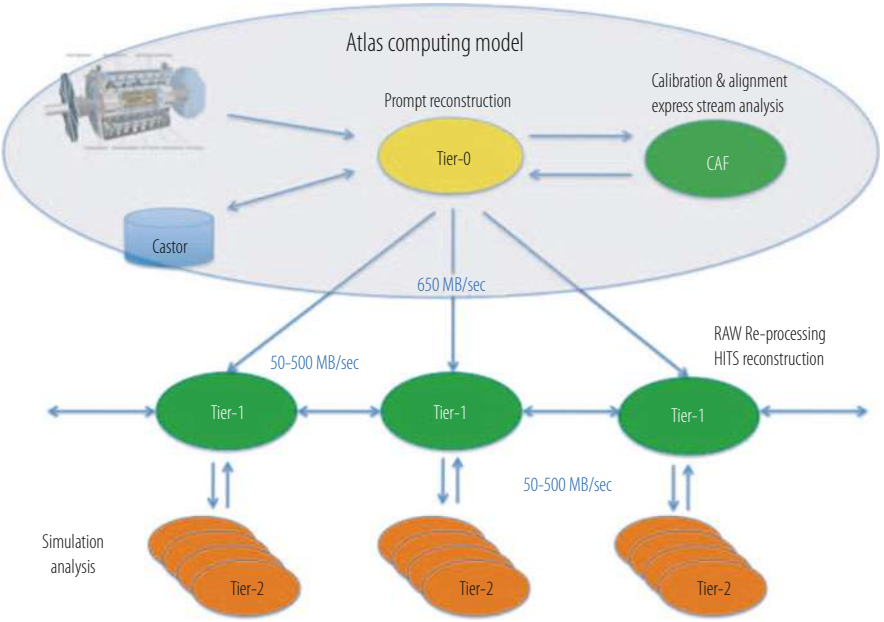

Fig. 16.18 Schematic flow of event data in the ATLAS computing model, illustrating the Tier-0, Tier-1 and Tier-2 connections. Tier-3 centres (typically smaller analysis clusters) are not included

16.7.1 Computing Model

The tasks of archiving, processing and distributing the ATLAS data across a worldwide computing organisation are of an unprecedented magnitude and complexity. The ever-present financial limitations, along with the unpredictability of the accelerator and detector operational details at the start-up, have implied the creation of a very flexible yet cost-effective plan to manage all the computing resources and activities. This plan, referred to as the computing model, was difficult to set up initially since the resources for computing had not been included in the initial funding plan for the LHC experiments. Over the past 5 years, however, a detailed computing model has been put in place and tested thoroughly with large-scale samples of simulated data and various technical computing challenges. This computing model describes as accurately as feasible the flow of data from the data acquisition system of the experiment to the individual physicist desktop [30]. Over the past few years, it has adapted to the evolution of the major parameters which govern it, such as the respective sizes of the various data types, the reality of the resources available at the various Tiers, and the more and more precise understanding of the requirements of the actual analysis in the various physics domains.

The main requirement on the computing model is to provide prompt access to all the data needed to carry out physics analyses. This typically translates to providing all members of the collaboration with access to reconstructed data and appropriate, more limited, access to raw data for organised monitoring, calibration and alignment activities. As already mentioned, the key issue is the decentralisation and wide geographic distribution of the computing resources. Sharing of these resources is possible through the Grid and its middleware, and therefore the interplay with the Grid is built into the models from the very beginning.

The most important elements of the computing model are the event data model and the flow of the various data types to the analysis processes.

16.7.1.1 Event Data Model

The physics event store contains a number of different representations, or levels of detail, of the physics events from the raw (or simulated) data all the way to reconstructed and summary data suitable for massive fast analysis. The different types of data are:

- Raw data: this is the byte-stream output of the High-Level Trigger (HLT) and is the primary input to the reconstruction process. The ATLAS experiment expects ≈1.5 MB of data arriving at a rate of ≈200–300 Hz. Events are transferred from the HLT farm to the Tier-0 in 2 GB files containing events from a data-taking period with the same trigger selections from a single LHC fill. The events will generally not appear in a consecutive order, since they will have undergone parallel processing in the HLT farm beforehand.
- Reconstructed data (referred to as Event Summary Data or ESD): this is the output of the reconstruction process. Most detector and physics studies, with the exception of calibration and alignment procedures, will only have to rely on this format. The data are stored using an object-oriented (OO) representation in so-called POOL-format files [31, 32]. The target size for the ESD files has increased from 500 to 800 kB per event over the past few years.
- Analysis Object Data or AOD: this is derived from the ESD format and is a reduced event representation, intended to be sufficient for most physics analyses. The target size is roughly a factor five smaller than that of the ESD (i.e. 100–200 kB per event) and the contents are physics objects and other high-level analysis elements.

If experience from the Tevatron and initial experience from the experiment commissioning and early data-taking phase are used as a guide, it is expected that in the early stages of the machine and experiment commissioning the ESD format will be in heavy use. The AOD format is expected to become the dominant tool for studies only when both machine and experiments are in steady-state data-taking. Nevertheless, it is planned to commission the AOD format with real collision data as early as possible, since one of the biggest constraints on the computing model will be the access bandwidth to the data. The AOD, in addition to being the format

with the smallest size, has, by construction, the most compact and complete physics information of the event, and is thus going to be indispensable in carrying out high-statistics analyses.

In preparation for the hopefully soon-to-come high-statistics analysis era, ATLAS has defined two further formats, namely a condensed data format for tagging events with certain properties, called TAG, and a Derived Physics Data format(or DPD), which are intended for use in end-user analyses. TAG data are event-level metadata, i.e. thumbnail information about each event to enable rapid and efficient selection for individual analyses. The TAG data are also stored in a relational database to enable various searches via database queries. The average size is a few kB per event. The DPD format corresponds to the highest-level of data representation, with "ntuple"-like content, for direct analysis and display by analysis programs.

These official data formats have been deployed as the vehicle for running physics analyses. As an example, the AOD format and its contents have been the subject of several generations of very extensive sets of tests with different data, conditions, and subsequent uses. Of course, since the AOD format contains only a subset of the information in the event, there will always be analyses that need to refer back to the ESD format. The most critical part of the optimisation of these various formats over the past few years has therefore been to select appropriately the objects to be included in the AOD. There is usually a trade-off between storage cost and CPU to derive the additional objects to be studied, and the details depend very strongly on the sample size required and the number of times the sample is used.

16.7.1.2 Data Flow and Processing

To maximise the physics reach of the experiment, the HLT farms will write events at the maximum possible data rate, which can be supported by the computing resources. Currently, this is expected to be in the range of 200–300 Hz, essentially independent of the instantaneous luminosity of the accelerator. Trigger thresholds will be adjusted up or down to match the maximum data rate, in order to maintain consistency with the data storage and processing capabilities of the offline systems. Extensive test campaigns have shown that the online-offline link and the Tier-0 centre are able to keep up in real-time with the HLT output rate.

The HLT output is streamed according to trigger type for the subsequent reconstruction and physics analysis. In addition, specialised calibration streams allow for independent processing from the bulk of the physics data. These streams are required to produce calibration and alignment constants of sufficient quality to allow a useful first-pass processing of the physics streams with minimum latency. ATLAS also makes use of an express stream, which is a set of physics triggers corresponding to about 5% of the full data rate. These events are selected to tune the physics and detector algorithms and also to provide rapid updates to the calibration and alignment constants required for the first-pass processing.

Streams can be used for a variety of purposes. The primary use, as mentioned previously, is to allow the prioritisation of the processing of the data. As an example, having the di-muon dataset as a independent stream obviously results in a much faster turnaround on any analysis that relies on these data. Streams can also be useful in the commissioning phase, to debug both the software and the overall online and offline computing systems. As an example, a special "debug" stream is dedicated to problematic events, e.g. failing in the HLT step, to facilitate the understanding of errors in the system. Obviously, such streams will be created as the need arises, will be rate-limited, and may even be withdrawn once the primary motivation for them is no longer present.

The first step before full-fledged prompt reconstruction is the actual processing of the calibration data in the shortest possible time. The plan calls for a short 1 to 2-day latency in completing this task. Once the calibration and alignment constants are in place, a first-pass (or prompt) reconstruction is run on the primary event streams, and the resulting reconstructed data (ESD and AOD formats) are archived into the CERN mass storage system.

Upon completion of this step, the data are distributed to the Tier-1 centres. Each Tier-1 site assumes responsibility for a fraction of the reconstructed data. Most of the ESD format data are, however, not available on disk for individual user access. A major role for the Tier-1 centres is the reprocessing of the data, once more mature calibrations and software are available, typically once or twice every year. By shifting the burden of reprocessing to the Tier-1 centres, the experiment can reprocess its data asynchronously and concurrently with data-taking and the associated prompt processing. The Tier-2 centres can obtain partial or full copies of the AOD/DPD/TAG format data, which will be the primary tool for physics analysis. The Tier-2 centres will also be responsible for large-scale simulation tasks, once the Tier-1 sites will be very busy with data reprocessing.

16.7.2 Software

On the software front, there have been two major issues encountered by the LHC experiments, which are either new or simply appear to a much greater extent than in the past: the distributed nature of the development and the maintainability of the code over long time-scales:

- Software development has had to continue down the path established at LEP and at the Tevatron: the code is developed in a distributed manner with responsibilities that span multiple individuals, institutions, countries and time zones. While for the large-scale hardware projects, a factorisation of the overall construction into substantial units has been possible, software, with its much wider contributor base within the collaborations, has a larger degree of fragmentation. This has necessitated the formation of intricate project structures to monitor and steer

the code development. The usual issues which result from relying on multiple institutions and funding agencies have risen here as well (see Sect. 16.2).
- Another major issue has been the maintainability of the systems. Given the expected long lifetime of the LHC programme, it was deemed necessary, from the very beginning, that the software systems be built using object-oriented methodologies. The C++ programming language has been chosen as the major development tool.

At the heart of the software system of the experiment is the software framework, which provides support for all the data-processing tasks. All such tasks, including the simulation, reconstruction, analysis, visualisation, and, very importantly, the high-level trigger operate within this framework. It provides the basic software environment in which code is developed and run, as well as all the basic services (e.g. access to calibration and conditions data, input/output facilities, persistency, to name but a few examples).

All the applications, which are built on top of the framework use a component model, i.e. they have building blocks, which appear to the framework as standard plug-ins. The main advantage of the component model is the factorisation of any one solution into a number of independent software codes, but also a significant flexibility to adapt to changes in the future. The final major architectural and design principle has been the separation of algorithms from the data and the acceptance of different data representations in memory (transient) and file storage (persistent).

16.7.3 Analysis Model

As has been already mentioned, the ESD and AOD/DPD formats are the primary tools for carrying out physics studies. Both formats are stored in POOL files and are processed using the respective software framework of each experiment. The decreasing event size in the event model allows the users to process a much larger number of AOD/DPD events than ESD events. In addition, the AOD/DPD formats will be more accessible, with a full copy at each Tier-1 site and large samples at Tier-2 sites. It is therefore expected that most analyses will be carried out on AOD/DPD data.

To illustrate the ATLAS analysis model with a concrete example, a specific analysis task may begin with a query against the TAG data to select a subset of events for processing using a suitable DPD format. This query might be for events with two leptons, missing transverse energy and at least two jets, all above certain thresholds. The result of this query is then used to define a dataset (or set of files) containing the information for these events. The analysis would then proceed to make further event selection by refining various physics quantities, e.g. the muon isolation or the missing transverse energy calculation. The fine-grained details of how much processing and event selection will be carried out by individuals versus organised physics groups (e.g. the Higgs group) is not frozen yet. It is widely expected that

both modes of operation will occur, i.e. that there will be data samples, which are selected and perhaps processed further in an organised manner by large groups of the collaboration, but also samples created by individuals. The relative fraction of each will be driven to a large extent by the resources that will be available at any given time.

The last element of the analysis model is a distributed analysis system which allows for the remote submission of jobs from any location. This system splits, in an automated way, an analysis job into a number of smaller jobs that run on subsets of the input data. The results of the job may be merged to form an output dataset. Partial results from these jobs are made available to the user before the full set of jobs runs to completion. Finally, the distributed analysis system will ensure that all jobs and resulting datasets are properly catalogued for future reference.

16.8 Expected Performance of Installed Detectors

16.8.1 Tracker Performance

Table 16.12 shows a comparison of the main performance parameters of the ATLAS and CMS trackers, as obtained from extensive simulation studies performed over the years and bench-marked using detailed test-beam measurements of production modules wherever possible. The unprecedentedly large amount of material present

Table 16.12 Main performance characteristics of the ATLAS and CMS trackers

	ATLAS	CMS
Reconstruction efficiency for muons with $p_T = 1\,\text{GeV}$	96.8%	97.0%
Reconstruction efficiency for pions with $p_T = 1\,\text{GeV}$	84.0%	80.0%
Reconstruction efficiency for electrons with $p_T = 5\,\text{GeV}$	90.0%	85.0%
Momentum resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 0$	1.3%	0.7%
Momentum resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 2.5$	2.0%	2.0%
Momentum resolution at $p_T = 100\,\text{GeV}$ and $\eta \approx 0$	3.8%	1.5%
Momentum resolution at $p_T = 100\,\text{GeV}$ and $\eta \approx 2.5$	11%	7%
Transverse i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 0$ [μm]	75	90
Transverse i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	200	220
Transverse i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 0$ [μm]	11	9
Transverse i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	11	11
Longitudinal i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 0$ [μm]	150	125
Longitudinal i.p. resolution at $p_T = 1\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	900	1060
Longitudinal i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 0$ [μm]	90	22–42
Longitudinal i.p. resolution at $p_T = 1000\,\text{GeV}$ and $\eta \approx 2.5$ [μm]	190	70

Examples of typical reconstruction efficiencies, momentum resolutions and transverse and longitudinal impact parameter (i.p.) resolutions are given for various particle types, transverse momenta and pseudorapidities

in the trackers is reflected in the overall reconstruction efficiency for charged pions of low transverse momentum, which is only slightly above 80%, to be compared to 97% obtained for muons of the same transverse momentum. The electron track reconstruction efficiency is even more affected by the tracker material and the numbers shown in Table 16.12 for electrons of 5 GeV transverse momentum are only indicative, since the efficiency obtained depends strongly on the criteria used to define a reasonably well measured electron track. The somewhat lower efficiencies obtained in the case of CMS are probably due to the higher magnetic field, which enhances effects due to interactions in the detector material. The combined performance of the tracker and electromagnetic calorimeter is discussed in Sect. 16.8.2.

The higher and more uniform magnetic field and the better measurement accuracy at large radius of the CMS tracker result in a momentum resolution on single tracks, which is better than that of ATLAS by a factor of almost 3 over the full kinematic range of the fiducial acceptance of the trackers. The impact parameter resolution in the transverse plane is expected to be similar at high momenta for both trackers, because the smaller pixel size in ATLAS is counter-balanced by the charge-sharing between adjacent pixels and the analogue readout in the CMS pixel system. In contrast, the smaller pixel size of the CMS tracker in the longitudinal dimension leads to a significantly better impact parameter resolution in this direction at high momenta.

In summary, the ATLAS and CMS trackers are expected to deliver the performances expected at the time of their design, despite the very harsh environment in which they will operate for many years and the difficulty of the many technical challenges encountered along the way. In contrast to most of the other systems, however, they will not survive nor deliver the required performance if the LHC luminosity is upgraded to $10^{35}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The ATLAS and CMS trackers will therefore have to be replaced by detectors with finer granularity to meet the challenges of the higher luminosity and with an order of magnitude higher resistance to radiation. This will be the major upgrade challenge for both experiments and a lively programme of research and development work has already been launched to this end.

16.8.2 Calorimeter Performance

The performance to be expected in situ for the very large-scale calorimeter systems of ATLAS and CMS is difficult to directly extrapolate from test-beam data. The calibration of these complex electromagnetic and hadronic calorimeter systems can indeed be to some extent ported with high precision from the test-beam measurements to the actual experiment and, more importantly, performed in situ using a set of benchmark physics processes such as $Z \to ee$ decays and $W \to jet - jet$ decays. This situation is somewhat new because of the following reasons:

- for the first time, there will be the possibility to control the absolute scale of hadronic jet energy measurements by using sufficiently abundant statistics from $W \rightarrow jet - jet$ topologies occurring in top-quark decays.
- extensive test-beam measurements in configurations close to that of the real experiment will have been performed at the time of first data-taking.
- it should be possible to constrain the absolute scale of the overall hadronic calorimetry using the measured response to charged pions of energies between 1 and 300 GeV and controlling this scale in situ, using a variety of samples, from single isolated tracks at the lower end of the range to e.g. clean samples of $\tau \rightarrow \pi^{\pm}\nu$ decays.

During the past 15 years, a large-scale and steady software effort has been maintained in the collaborations to simulate in detail calorimeters of this type well before they begin their operation. The complex geometries and high granularities described above and the high energies of the products of the collisions have naturally augmented considerably the computing effort required to produce large-statistics samples of fully simulated events. A few examples are shown below for photon, electron, jet and missing transverse energy measurements.

16.8.2.1 Electromagnetic Calorimetry

Figure 16.19 shows an example of the expected precision with which photon energy measurements will be performed in ATLAS (left) and CMS (right) over the energy range of interest for $H \rightarrow \gamma\gamma$ decays. In the case of ATLAS, the results are shown for all photons (unconverted and converted) and for three values of pseudorapidity. In the case of CMS, the results are shown for dominantly unconverted photons in the barrel crystal calorimeter. The selected photons are required in this latter case to have deposited more than 94.3% of their energy in a 3 by 3 crystal matrix normalised to the 5 by 5 crystal matrix used to compute the total energy. This basically selects unconverted photons and some late conversions with a 70% overall efficiency. For a photon energy of 100 GeV, the ATLAS energy resolution varies between 1.0 and 1.4%, depending on η. These numbers increase respectively to 1.2 and 1.6% if one includes the global constant term of 0.7%. The overall expected CMS energy resolution in the barrel crystal calorimeter is 0.75% for the well-measured photons at that energy (Fig. 16.19 includes the global constant term of 0.5%). This example shows that the intrinsic resolution of the CMS crystal calorimeter is harder to obtain with the large amount of tracker material in front of the EM calorimeter and in the 4T magnetic field: between 20 and 60% of photons in the barrel calorimeter acceptance convert before reaching the front face of the crystals.

Similarly, Fig. 16.20 shows an example of the expected precision with which electron energy measurements will be performed in ATLAS (left) and CMS (right). In the case of ATLAS, the results are shown for electrons at $\eta = 0.3$ and 1.1 in the energy range from 10 to 1700 GeV. The energy of the electrons is always collected in a 3 by 7 cell matrix, which, as for the photons, is wider in the bending direction

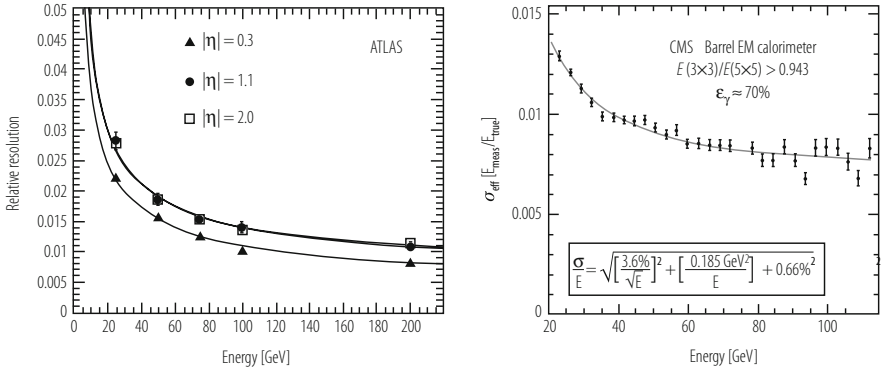

Fig. 16.19 For ATLAS (left) and CMS (right), expected relative precision on the measurement of the energy of photons reconstructed in different pseudorapidity regions as a function of their energy (see text). Also shown are fits to the stochastic, noise and local constant terms of the calorimeter resolution

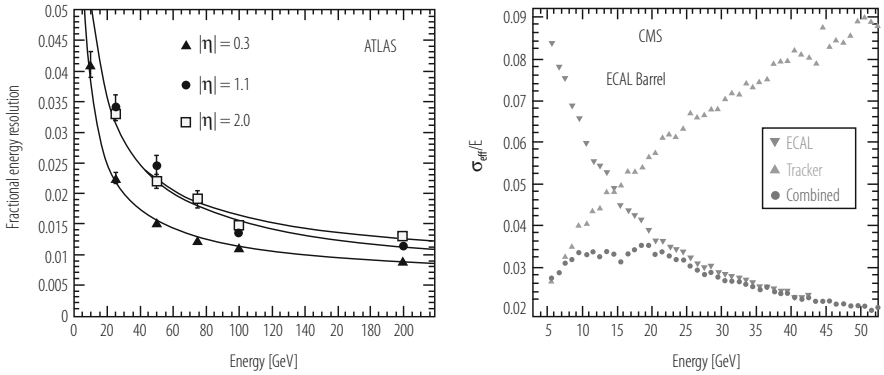

Fig. 16.20 For ATLAS (left) and CMS (right), expected relative precision on the measurement of the energy of electrons as a function of their energy over the energy range of interest for $H \rightarrow ZZ^{(*)} \rightarrow eeee$ decays. In the case of ATLAS, the resolution is shown for three values of pseudorapidity (only the electron energy measurement is used, with the energy collected in a 3 by 7 cell matrix in $\eta \times \phi$ space), together with fits to the stochastic and local constant terms of the calorimeter resolution. In the case of CMS, the combined (tracker and EM calorimeter) effective resolution at low energy, taken as the r.m.s. spread of the reconstructed energy, collected in a 5 by 5 cell matrix and normalised to the true energy, is shown over the acceptance of the barrel crystal calorimeter, together with the individual contributions from the tracker and the EM calorimeter

to collect as efficiently as possible the bremsstrahlung photons while preserving the linearity and low sensitivity to pile-up and noise. In the case of CMS, the effective resolution (r.m.s. spread) is shown for the barrel crystal calorimeter and in the most difficult low-energy range from 5 to 50 GeV. Refined algorithms are used, in both the tracker and the calorimeter, to recover as much as possible the bremsstrahlung tails and thereby to restore most of the excellent intrinsic resolution of the crystal

calorimeter. Nevertheless, for electrons of 50 GeV in the barrel region, the ATLAS energy resolution varies between 1.3% (at $\eta = 0.3$) and 1.8% (at $\eta = 1.1$) without any specific requirements on the performance of the tracker at the moment. In contrast, the CMS effective resolution is estimated to be 2%, demonstrating that it is harder to reconstruct electrons, with a performance in terms of efficiency and energy resolution similar to that obtained in test beam, than photons.

Further performance figures of critical importance to the electromagnetic calorimeters are those related to electron and photon identification in the context of overwhelming backgrounds from QCD jets and of pile-up at the LHC design luminosity, of γ/π^0 separation, of efficient reconstruction of photon conversions and of measurements of the photon direction using the calorimeter alone wherever the longitudinal segmentation provides a sufficiently accurate measurement. All these aspects rely heavily on the details of the longitudinal and lateral segmentation of the EM calorimetry and the reader is referred to the ATLAS and CMS detector performance reports [13, 27] for more information.

Another important issue, especially for the EM calorimeters is the calibration in situ, which will eventually provide the final calibration constants required e.g. for searches for narrow states, such as $H \rightarrow \gamma\gamma$ decays. These can be divided into an overall constant defining the absolute scale and a set of inter-calibration constants between modules or cells:

- the ATLAS EM calorimeter has been shown to be uniform by construction to about 0.4% in areas of 0.2×0.4 or larger in $\Delta\eta \times \Delta\phi$ space. One will therefore have to calibrate in situ only about 440 sectors of this size. The use of the Z mass constraint alone without reference to the tracking should be sufficient to achieve an inter-calibration to better than 0.3% over a few days at low luminosity. If additional problems arise because of the material in the tracker, the use of electrons from W decay to measure E/p will provide additional constraints.
- the CMS crystals could not be pre-calibrated in the laboratory with radioactive sources to better than 4.5%. This inter-calibration spread has been brought down to significantly smaller values using cosmic rays. Without an individual calibration of the crystals in the test beam, one has to rely on in situ calibration for further improvements. Using initially large samples of minimum bias events (including explicit reconstruction of π^0 and η decays) and low E_T jets at fixed η, the inter-calibration could be improved to 1.5% within ϕ-rings of 360 crystals. At a later stage, high statistics samples of W-boson decays to electrons will be needed to reach the target constant term of 0.5%.
- a key issue for both ATLAS and CMS will be to keep the constant term below the respective target values of 0.7 and 0.5% in the presence of the unprecedented amount of material in the trackers. For ATLAS, other major potential contributions to the constant term (each one of the order of 0.2 to 0.3%) are mostly short-range (detector geometry, such as ϕ-modulations, variations of the sampling fraction in the end-caps, absorber and gap thickness fluctuations, fluctuations in the calibration chain, differences between calibration and physics signal), but the more potentially worrisome one is long-range and is related

to the signal dependence on temperature. The LAr signal has a temperature dependence of −2% per degree: the temperature monitoring system in the barrel sensitive volume should therefore track temperature changes above ±0.15°, which is the expected dispersion from the heat influx of 2.5 kW per cryostat. In CMS, the temperature control requirements are even more demanding, since the temperature dependence of a crystal and its readout is about −4.3% per degree for a heat load of 2 W per channel or 160 kW total. The very sophisticated cooling scheme implemented in the super-modules has demonstrated the ability to maintain the temperature to better than ±0.05° and thereby to meet these stringent requirements. Time-dependent effects related to radiation damage of the CMS crystals will have to be monitored continuously with a stable and precise laser system.

16.8.2.2 Hadronic Calorimetry

The expected performance for reconstructing hadronic jets is shown in Fig. 16.21. In the case of ATLAS, the jet energy resolution is depicted for two different pseudorapidity bins over an energy range from 15 to 1000 GeV for two different sizes of the cone algorithm used. The jet energies are computed using a global weighting technique inspired by the work done in the H1 collaboration [28]. In the case of CMS, the jet energy resolution is shown as a function of the jet transverse energy, for a cone size $\Delta R = 0.5$ and for $|\eta| < 1.4$, over a transverse energy range from 15 to 800 GeV. For hadronic jets of typically 100 GeV transverse energy, characteristic for example of jets from W-boson decays produced through top-quark decay, the ATLAS energy resolution varies between 7 and 8%, whereas

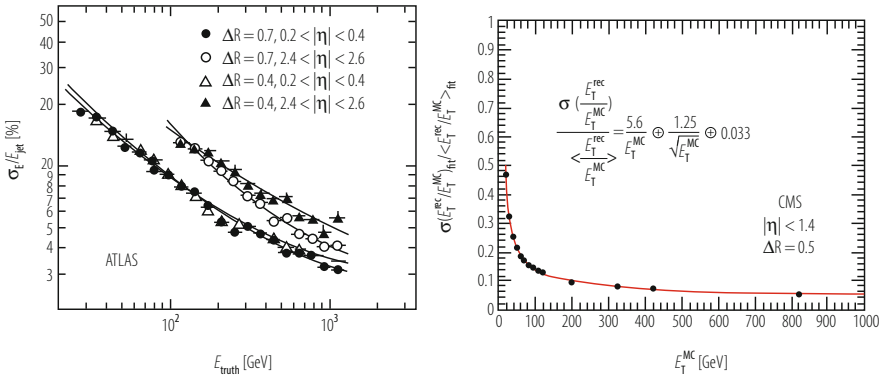

Fig. 16.21 For ATLAS (left) and CMS (right), expected relative precision on the measurement of the energy of QCD jets reconstructed in different pseudorapidity regions as a function of E_{truth}, where E_{truth} is the true jet energy, for ATLAS, and of E_T^{MC}, where E_T^{MC} is the true jet transverse energy, for CMS (see text)

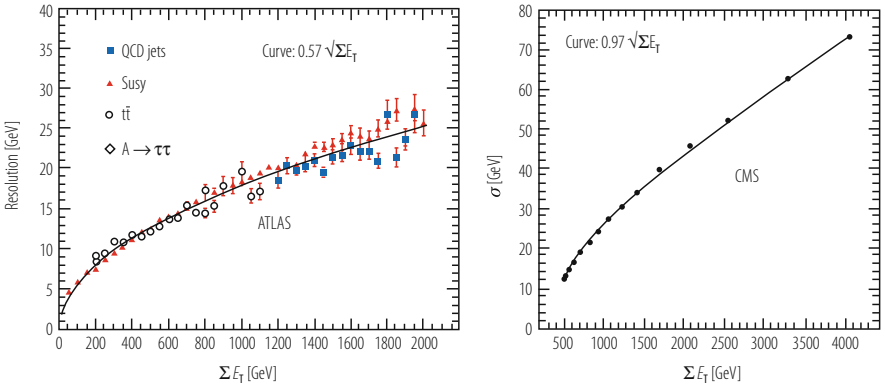

Fig. 16.22 For ATLAS (left) and CMS (right), expected precision on the measurement of the missing transverse energy as a function of the total transverse energy, ΣE_T, measured in the event (see text)

the CMS energy resolution is approximately 14%. The intrinsic performance of the CMS hadron calorimeter can be improved using charge particle momentum measurements, a technique often referred to as particle flow, which was developed at LEP [23]. Initial studies indicate that the jet energy resolution can be significantly improved at low energies, typically from 17 to 12% for $E_T = 50\,\text{GeV}$ and $|\eta| < 0.3$, but such large improvements are not expected for jet transverse energies above 100 GeV or so.

Finally, Fig. 16.22 illustrates a very important aspect of the overall calorimeter performance, namely the expected precision with which the missing transverse energy in the event can be measured in each experiment as a function of the total transverse energy deposited in the calorimeter. The results for ATLAS are expressed as the σ from Gaussian fits to the (x,y) components of the E_T^{miss} vector for events from high-p_T jet production and also from other possible sources containing several high-p_T jets. In the case of CMS, where the distributions are non-Gaussian, the results are expressed as the r.m.s. of the same distributions for events from high-p_T jet production. For transverse momenta of the hard-scattering process ranging from 70 to 700 GeV, the reconstructed ΣE_T ranges from about 500 GeV to about 2 TeV. The difference in performance between ATLAS and CMS is a direct consequence of the difference in performance expected for the jet energy resolution.

16.8.3 Muon Performance

The expected performance of the muon systems has been a subject of very intense study in both experiments. Simulations which take into account a huge amount of

detail from the real geometries of all the chambers and support structures have been refined repeatedly over the years.

In ATLAS, the quality of the stand-alone muon measurement relies on detailed knowledge of the material distribution in the muon spectrometer, especially for intermediate-momentum muons. Reconstruction of these with high accuracy and without introducing a high rate for fake tracks, has to take into account multiple scattering of the muons and thus the details of the material distribution in the spectrometer. This necessitates a very detailed mapping of the detector and the storage of this map for use by the offline simulation and reconstruction programs. The corresponding effect in CMS is much smaller, since the amount of iron in between the muon stations dominates by far and the details of the material are necessary only in the boundaries between the iron blocks.

Figures 16.23 and 16.24 show the expected resolution on the muon momentum measurement. The expected near-independence of the resolution from the pseudorapidity in ATLAS, along with the degradation of the resolution at higher η in CMS are clearly visible. The resolution of the combined measurement in the barrel region is slightly better in CMS due to the higher resolution of the measurement in the tracking system, whereas the reverse is true in the end-cap region due to the better coverage of the ATLAS toroidal system at large rapidities. A summary of the performance of the two muon measurements can be found in Table 16.13 for muon momenta between 10 and 1000 GeV.

The expected performance matches that expected from the original designs. An interesting demonstration of the robustness of the muon systems comes from the reconstruction of muons in heavy-ion collisions. Whereas neither experiment was specifically designed for very high reconstruction efficiency in the very special conditions of heavy-ion collisions, it turns out that they can yield significant physics signals for a few key signatures such as J/ψ and Υ, Υ' production [27].

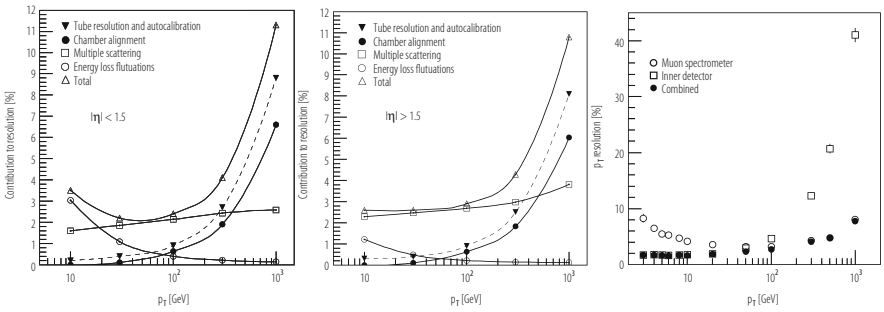

Fig. 16.23 Expected performance of the ATLAS muon measurement. Left: contributions to the momentum resolution in the muon spectrometer, averaged over $|\eta| < 1.5$. Centre: same as left for $1.5 < |\eta| < 2.7$. Right: muon momentum resolution expected from muon spectrometer, inner detector and their combination together as a function of muon transverse momentum

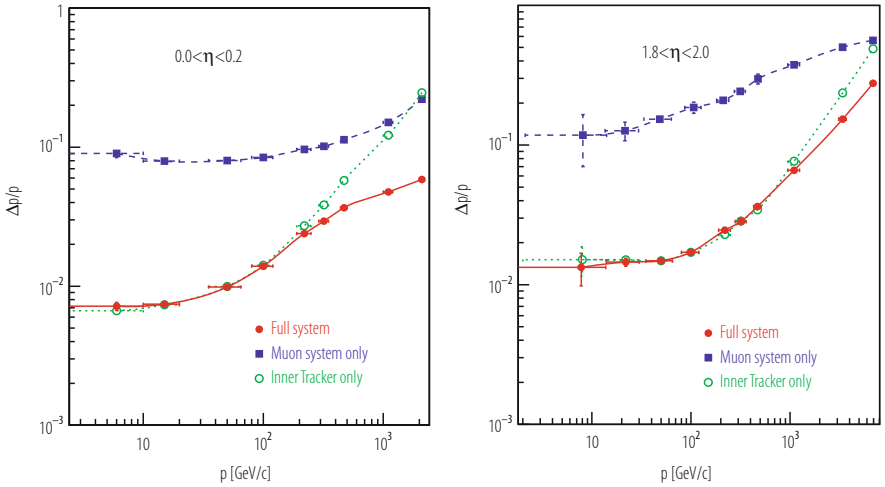

Fig. 16.24 Expected performance of the CMS muon measurement. The muon momentum resolution is plotted versus momentum using the muon system only, the inner tracker only, or their combination (full system) for the barrel, with $|\eta| < 0.2$ (left), and for the end-caps, with $1.8 < |\eta| < 2.0$ (right)

Table 16.13 Main parameters of the ATLAS and CMS muon measurement systems as well as a summary of the expected combined and stand-alone performance at two typical pseudorapidity values (averaged over azimuth)

Parameter	ATLAS	CMS
Pseudorapidity coverage		
Muon measurement	$\|\eta\| < 2.7$	$\|\eta\| < 2.4$
Triggering	$\|\eta\| < 2.4$	$\|\eta\| < 2.1$
Dimensions [m]		
Innermost (outermost) radius	5.0 (10.0)	3.9 (7.0)
Innermost (outermost) disk (z-point)	7.0 (21–23)	6.0–7.0 (9–10)
Segments/super-points per track for barrel (end-caps)	3 (4)	4 (3–4)
Magnetic field B [T]	0.5	2
Bending power (BL [Tm]) at $\|\eta\| \approx 0$	3	16
Bending power (BL [Tm]) at $\|\eta\| \approx 2.5$	8	6
Combined (stand-alone) Momentum resolution at		
$p = 10$ GeV/c and $\eta \approx 0$	1.4% (3.9%)	0.8% (8%)
$p = 10$ GeV/c and $\eta \approx 2$	2.4% (6.4%)	2.0% (11%)
$p = 100$ GeV/c and $\eta \approx 0$	2.6% (3.1%)	1.2% (9%)
$p = 100$ GeV/c and $\eta \approx 2$	2.1% (3.1%)	1.7% (18%)
$p = 1000$ GeV/c and $\eta \approx 0$	10.4% (10.5%)	4.5% (13%)
$p = 1000$ GeV/c and $\eta \approx 2$	4.4% (4.6%)	7.0% (35%)

16.8.4 Trigger Performance

The trigger involves, by design, the selection of only a small fraction of the p−p collisions at the LHC. As a result, a number of compromises on the extent of the physics programme have had to be made. This is an important difference with respect to the experience in e^+e^- machines.

Efficient use of DAQ bandwidth requires that two conditions be fulfilled. First, each level of the trigger attempts to identify physics objects (leptons, photons and jets) as efficiently as possible, while keeping the output bandwidth within requirements. The selected event sample should include all events which would be found by the full offline reconstruction. Hence, the cuts in the trigger must be consistent with those of the offline analysis. Second, since the bandwidth to permanent storage media is limited, events must be selected with care at the final trigger level.

A crucial ingredient of physics analysis is the determination of the trigger efficiency. Three tools allowing the measurement of the requirements imposed by the L1 trigger have been included in the designs. One tool is the presence of overlapping programmable triggers, which allows triggers with different thresholds and cuts to run simultaneously, producing multiple results in parallel. A second tool is prescaled triggers with either lower thresholds or looser requirements (or both) to run in parallel with the main algorithm. A third tool is prescaling of a particular trigger with one of its cuts removed.

Beyond these three tools, another method for measuring the trigger efficiency, which is used extensively, is the use of processes with two physics objects where the trigger selects one of the two. As an example, $Z \rightarrow ee$ decays, selected via the single-electron trigger, can be used to measure the electron trigger efficiency by examining the second, unbiased, electron leg.

A key task is the creation of the trigger tables, i.e. the requirements demanded online, by both the L1 and HLT systems, on the events selected. Table 16.14 lists two examples from ATLAS and CMS, for the L1 trigger. There are, naturally, very significant uncertainties in these rate estimates. At one extreme, CMS allocates only one-third of the assumed DAQ bandwidth to specific triggers. In the ATLAS case, the plan is to absorb any differences in rate via changes in thresholds. Both experiments plan to allocate bandwidth to B physics as well, within the limitations of the total resources available, at the initially low luminosities of the LHC.

The real-time nature of the selections imposes very stringent requirements on the monitoring of the L1 and HLT performance. Initially, many triggers will be run in forced-accept mode, thereby providing the possibility to analyze in detail their performance offline. The trigger monitoring itself will employ a number of tools, including the storage of a small fraction of the events rejected, the comparison of the actual online decisions (as obtained from intermediate hardware calculations that will be stored along with the detector data) and a number of unbiased events, or “minimum-bias” events, which are selected at random, i.e. without any specific requirements on the bunch crossing in question.

Table 16.14 Examples of L1 trigger tables from ATLAS and CMS

	ATLAS		CMS	
Trigger type	Threshold [GeV]	Rate [kHz]	Threshold [GeV]	Rate [kHz]
Inclusive isolated electron/photon	25	12.0	29	3.3
Di-electrons/di-photons	15	4.0	17	1.3
Inclusive isolated muon	20	0.8	14	2.7
Di-muons	6	0.2	3	0.9
Single tau-jet trigger	–	–	86	2.2
Two $\tau-$jets	–	–	59	1.0
Tau-jet * E_T^{miss}	25 * 30	2.0	–	–
1-jet, 3-jets, 4-jets	200, 90, 65	0.6	177, 86, 70	3.0
Jet * E_T^{miss}	60 * 60	0.4	88 * 46	2.3
Electron * Jet	–	–	21 * 45	0.8
Electron * Muon	15 *10	0.1	–	–
Minimum-bias (calibration)			None	0.9
Others (monitor, calibration, ...)		5.0	–	–
Total		25		16

The table corresponds to an instantaneous luminosity of $2 \cdot 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and an assumed total DAQ bandwidth of 25 and 50 kHz respectively. In the case of CMS, only one third of the DAQ bandwidth is allocated, as a safety factor, to account for all the uncertainties in the estimations of the rates. In both cases the threshold corresponds to the point where the efficiency is 95% of the asymptotic efficiency

The trigger systems of the two experiments are also expected to be flexible enough to adapt to changing run and/or coast conditions. As an example, the instantaneous luminosity is expected to drop in the course of a fill, and therefore an optimal allocation of resources might be to change trigger conditions, for instance by lowering trigger thresholds or decreasing pre-scale factors for selected channels. All such changes, along with any other changes in the running conditions, will be logged and the overall online monitoring must record the operational performance as a function of the changes made in real time.

A measure of the performance is given by the efficiency to trigger on single physics objects, namely electrons and photons, muons, jets and tau-jets. The presumed efficiency depends, of course, on the production process and for this reason, Standard-Model processes are used. Table 16.15 lists the efficiencies at L1 and HLT for electrons and muons. For jets, the relevant parameter is not the efficiency which can always reach 100%, but rather the effective threshold needed in order to obtain a fixed efficiency, e.g. 95%, for jets with a certain threshold at the generator level. The situation with τ-jets is more complicated, since the two experiments have studied them in the context of specific physics signatures, which are not directly comparable.

The performance of the L1 trigger and HLT systems has been checked against all the benchmark "major discovery channels" in extensive studies by the two

Table 16.15 Efficiency for triggering on a key physics objects in ATLAS and CMS

Object	ATLAS	CMS
Electrons	$E_T > 25\,\text{GeV}$	$E_T > 29\,\text{GeV}$
L1 efficiency	95%	95%
HLT efficiency	80%	77%
Muons	$P_T > 20\,\text{GeV}$	$P_T > 19\,\text{GeV}$
L1 efficiency	95%	90%
HLT efficiency	80%	77%

The calculations have been performed at different thresholds, which are indicated in the table

experiments. These include all the expected decays of the Standard Model Higgs boson as well as those of the multiple Higgs bosons in the case of supersymmetry. In most cases, the decays involve multiple leptons and can therefore be triggered with very high efficiency. The efficiency to other signatures, such as those expected from supersymmetry is also very high. Overall, current expectations are that the two experiments can address the full physics program that will be made available by the LHC.

16.9 Ten Years of Operation and Physics Analysis in a Nutshell

This section, written 10 years after the previous ones, attempts the impossible, namely to summarise briefly what has been learned at the LHC over the past years. This attempt is limited to the $p-p$ collision data-taking of the ATLAS and CMS experiments, leaving out by necessity entire areas of exciting results obtained in heavy-flavour physics by the LHCb experiment and in heavy-ion physics by ALICE (and also ATLAS and CMS). Most of the examples shown below are taken from ATLAS public results obtained at various stages of the data-taking and physics analysis.

Table 16.16 summarises the different phases of the commissioning and data-taking periods of the ATLAS experiment, as extracted from its already long history of more than 25 years (celebrated in October 2017 in the Bratislava ATLAS week). The first data taken and analysed with the embryonic software under development for the experiment took place in the combined test-beams at the CERN SPS where almost complete slices of the ATLAS detector were exposed to various particle beams over a wide range of energies in the years 2002 to 2006. The next step towards commissioning the experiment took place in the ATLAS cavern itself with combined cosmic runs which illuminated the whole detector, from pixels to outermost muon chambers, and provided a first realistic test-bed for the offline alignment of all sub-systems using the precise measurements of charged-particle tracks in the complex magnetic field of the experiment (silicon sensors, straw tubes, and monitored drift tubes).

Table 16.16 Successive steps in preparation, commissioning, and operation of the ATLAS detector at the LHC

2002 to 2006	Combined test-beams at the CERN SPS
2008 onwards	Combined cosmics
2009	0.9 TeV pp collisions
2010 to 2012	Run-1
2010	7 TeV pp collisions, 36 pb^{-1}
2011	7 TeV pp collisions, 5 fb^{-1}
2012	8 TeV pp collisions, 20 fb^{-1}
2015 to 2018	Run-2
2015	13 TeV pp collisions, 3.2 fb^{-1}
2016	13 TeV pp collisions, 32.8 fb^{-1}
2017	13 TeV pp collisions, 44 fb^{-1}
2018	13 TeV pp collisions, 59 fb^{-1}

The successive years of operation with proton–proton collisions are shown together with the integrated luminosity accumulated each year

16.9.1 Accelerated History: Rediscovering the Standard Model

The first beams at LHC injection energy in 2008 provided huge excitement with only a handful of events called beam splashes produced by single beams interacting in the collimator material just before reaching the experiments. With these events alone, an accurate timing (to $\sim$1 ns) of most of the detector readout channels was achieved, a major step towards commissioning the whole experiment for data-taking with beams. The incident which occurred in the LHC at that point was perceived as a major setback at the time, resulting in a 1 year delay for the LHC to deliver first stable beams with collisions in all experiments. This finally happened in a growing atmosphere of excitement at the end of 2009 at the modest centre-of-mass energy of 0.9 TeV, which corresponds to the injection energy of the proton beams from the CERN SPS into the LHC.

These first few days of data-taking led to the first public results from the LHC experiments and even to a few papers with the first measurements of charged particle multiplicities and differential spectra [34]. The data turned out to be also a wonderful test-bed for rediscovering a large fraction of the very diverse zoo of particles produced in pp interactions. One example is shown in Fig. 16.25 with distinctive peaks at the masses of the π^0 and η mesons in the diphoton spectrum, visible above the combinatorial background from random combinations of pairs of photons reconstructed in the electromagnetic calorimeters.

Another later example of this zoo of particles is shown in Fig. 16.26 based on the first run-2 dataset at 13 TeV from CMS, where one distinguishes clearly among other resonances the narrow J/ψ, Υ, and Z mass peaks used for precise calibration and efficiency measurements of the reconstructed muons across a wide range of energy and pseudorapidities.

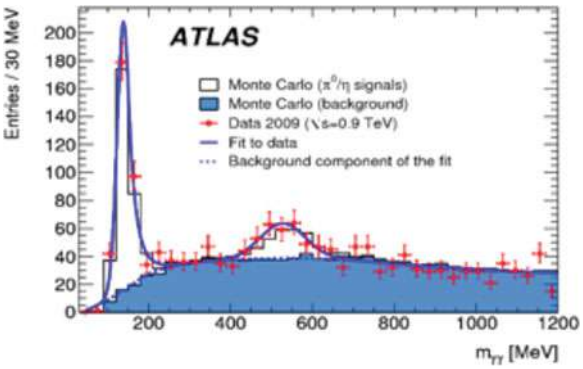

Fig. 16.25 Invariant mass distribution of low-mass diphoton events, as measured in ATLAS with early data at $\sqrt{s} = 0.9\,\text{TeV}$

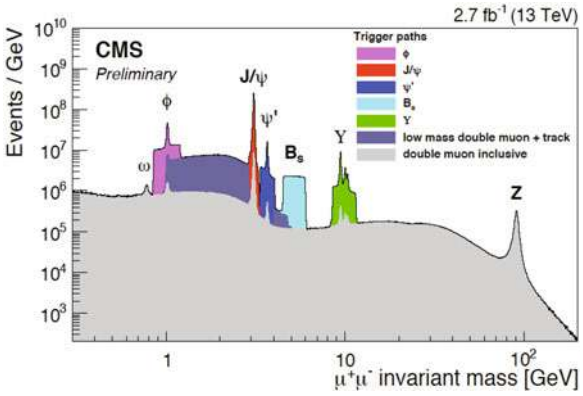

Fig. 16.26 Invariant mass distribution of dimuon events, as measured in CMS with early data at $\sqrt{s} = 13\,\text{TeV}$

In 2010, the very modest accumulated integrated luminosity of $36\,\text{pb}^{-1}$, more than one thousand times smaller than that accumulated in 2017, was nevertheless amply sufficient to observe and measure W/Z-boson production and the production of pairs of top quarks, as shown, respectively, in Figs. 16.27 [35] and 16.28 [36]. Placing LHC measurements on top of the precise predictions from QCD for these production cross-sections as a function of centre-of-mass energy, way beyond previous hadron colliders where these particles were discovered, was the first step in paving the way towards precise tests of the theory with high-statistics measurements based on the very large samples expected in the later years. As of 2019, ATLAS and CMS have accumulated samples of more than 500 million $W \to l\nu$ decays, 50 million $Z \to ll$ decays, and respectively, five million pairs of top quarks with one semi-leptonic top decay and 0.3 million high-purity pairs of top quarks with one electron, one muon, and two b-tagged jets in the final state.

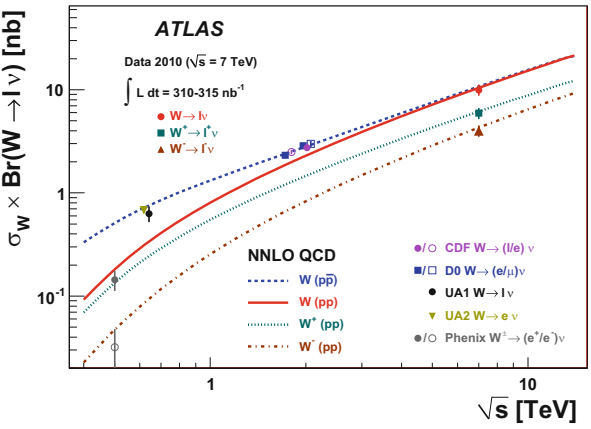

Fig. 16.27 W-boson production cross-section times branching fraction to an electron or muon plus a neutrino, as measured at hadron colliders by PHENIX at RHIC, by UA1/UA2 at the S$p\bar{p}$S, by CDF/D0 at the Tevatron, and by ATLAS at the LHC. The theoretical predictions are shown for both proton-proton and proton-antiproton collisions as a function of the centre-of-mass energy. The ATLAS data correspond to an integrated luminosity of $0.32\,\text{pb}^{-1}$ obtained in 2010 at $\sqrt{s} = 7\,\text{TeV}$

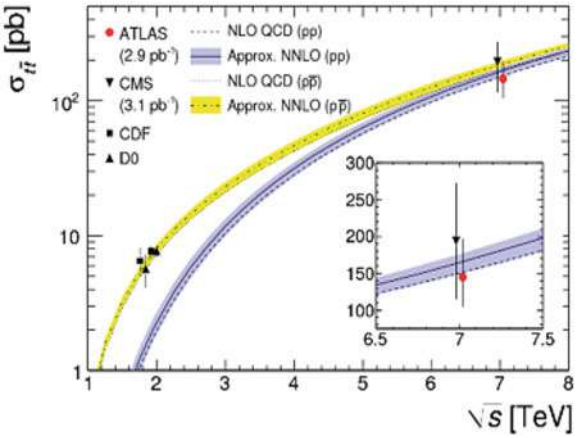

Fig. 16.28 Top quark pair-production cross-section, as measured at hadron colliders by CDF/D0 at the Tevatron and by ATLAS/CMS at the LHC. The theoretical predictions for proton-proton and proton-antiproton collisions assume a top-quark mass of 172.5 GeV and are shown as a function of the centre-of-mass energy. The ATLAS and CMS data correspond to an integrated luminosity of approximately $3\,\text{pb}^{-1}$ obtained in 2010 at $\sqrt{s} = 7\,\text{TeV}$

16.9.2 Precision Measurements

The heavy fundamental particles discussed above are thus an abundant source of prompt isolated electrons and muons, and also, in the case of the Z boson, of hadronically decaying τ-leptons, and have been used extensively in each period

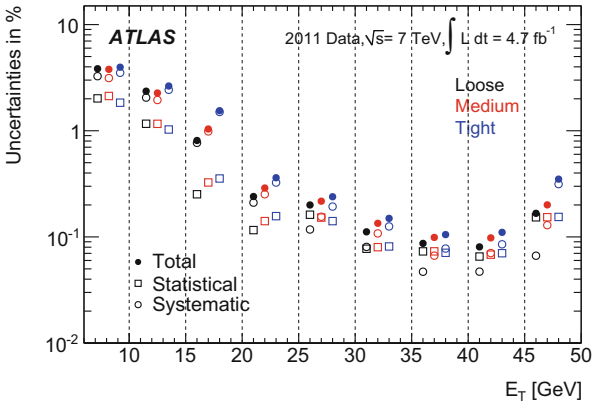

Fig. 16.29 Breakdown of the total uncertainty in the electron combined reconstruction and identification efficiencies, as a function of transverse energy, for the various identification criteria in ATLAS

of data-taking to assess the performance of the detector to reconstruct, identify, and measure their decay products, as well as to provide the most abundant source of triggers for the search for the Higgs boson and for new physics beyond the Standard Model (SM).

Figure 16.29 [37] shows that the efficiencies for reconstructing and identifying prompt isolated electrons could be measured in ATLAS with an overall accuracy ranging from the permil level near the Jacobian peaks from W/Z-boson decays to a few percent in the range 7–10 GeV turned out to be of critical importance for the search for the Higgs boson decaying to four leptons and for still ongoing searches for supersymmetric particles in the electroweak sector.

Figure 16.30 [38] illustrates the calibration accuracy achieved for prompt isolated muons, displayed as a function of the leading muon pseudorapidity for the already very large samples obtained with ATLAS in the run-1 8 TeV data. Tens of millions of J/ψ and Z-boson decays were used to calibrate the data and correct the simulation to reach an overall accuracy at the permil level, leading later on to very precise measurements of the Higgs-boson and W-boson masses. The dimuon events from the intermediate-mass Υ resonance were not used for the calibration itself and served as an independent validation sample to verify the closure of the procedure in terms of its uncertainties.

With sufficiently large samples of prompt isolated electrons, muons and photons, the jets produced in association with these precisely measured objects could be calibrated in situ to a precision far exceeding the initial expectations. Figure 16.31 [39] illustrates this in terms of the overall jet energy scale uncertainty in ATLAS from first run-2 data as a function of jet transverse momentum. The in situ absolute calibration achieves an overall uncertainty at the percent level or even below over a large kinematic range. Uncertainties due to the expected response differences for

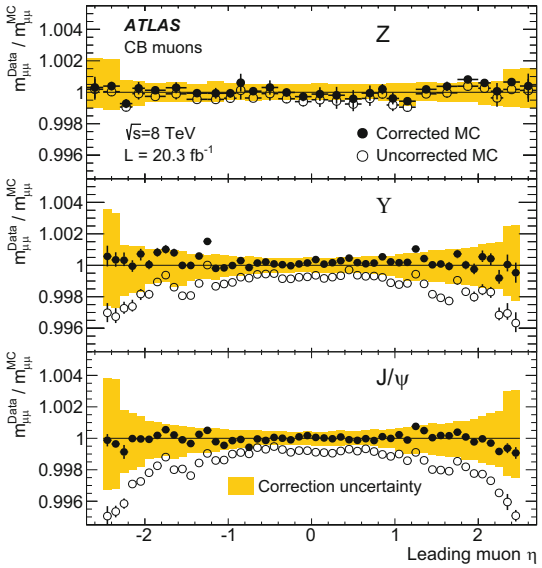

Fig. 16.30 Ratio of the fitted mean mass, $< m_{\mu\mu} >$, for data over simulation (MC), from Z (top), Υ (middle), and J/ψ (bottom) decays to dimuon pairs, as a function of the pseudorapidity of the highest-p_T muon in ATLAS. The ratio is shown for corrected MC (filled symbols) and uncorrected MC (empty symbols). The error bars represent the overall statistical and systematic uncertainty obtained from the mass fits. The bands show the uncertainties in the MC corrections calculated separately for the three samples

quark versus gluon jets and to pile-up at low transverse momenta dominate however the overall uncertainty on the jet energy scale over most of the range.

Precisely measured objects in simple final states lead to precisely measured fiducial differential and integrated cross-sections, which can then be compared to state-of-the-art theoretical predictions and used for example to improve the uncertainties in the parton distribution functions in the proton. Two examples of such ATLAS measurements, among the most precise to-date at the LHC, are shown as an illustration in Figs. 16.32 [40] and 16.33 [41], for inclusive jets as a function of jet transverse momentum in different rapidity ranges and for the integrated $W^{\pm}$ versus Z/γ^* cross-sections, respectively.

These precision measurements together with a wealth of others are not only used to improve the knowledge of the parton distribution in the proton, but also to improve the theoretical modelling of the relevant production processes, thereby reducing theoretical uncertainties which today are dominant when considering the measurement of fundamental Standard Model parameters such as the W-boson mass and the weak mixing angle.

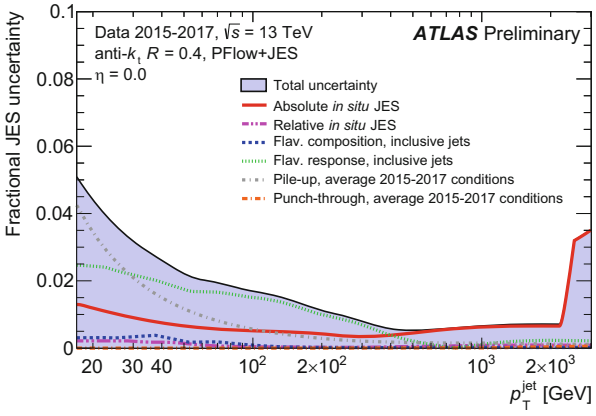

Fig. 16.31 Fractional jet energy scale (JES) systematic uncertainty components as a function of jet transverse momentum, p_T for jets reconstructed at central pseudorapidity from particle flow objects in ATLAS. The total uncertainty (all components summed in quadrature) is shown as a filled region topped by a solid black line. Topology-dependent components are shown under the assumption of a dijet flavour composition. At values of p_T, the uncertainty from the pile-up of $p-p$ interactions in the same or neighbouring bunch-crossings dominates the overall jet energy scale uncertainty. The data shown represent an average over the run-2 period from 2015 to 2017, corresponding to an average number of 30 interactions per bunch crossing

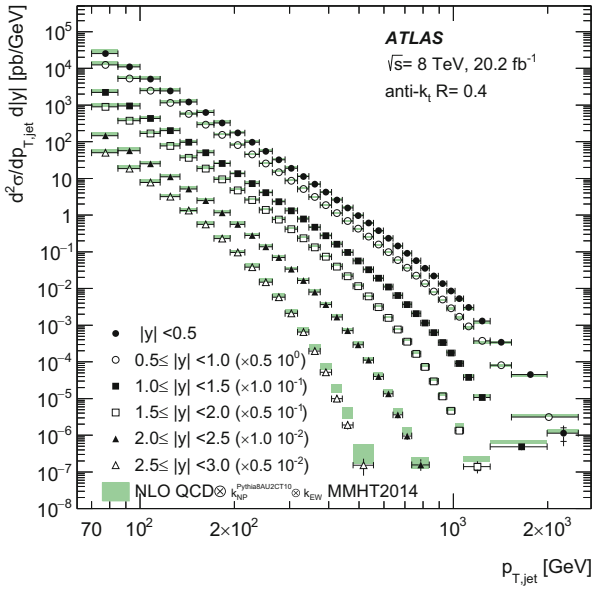

Fig. 16.32 Inclusive jet cross-section as a function of jet transverse momentum, p_T, in bins of jet rapidity. The results are shown for standard jets as measured with ATLAS 8 TeV data. The data are compared to the next-to-leading order QCD predictions with the MMHT2014 parton distribution function set, corrected for non-perturbative and electroweak effects

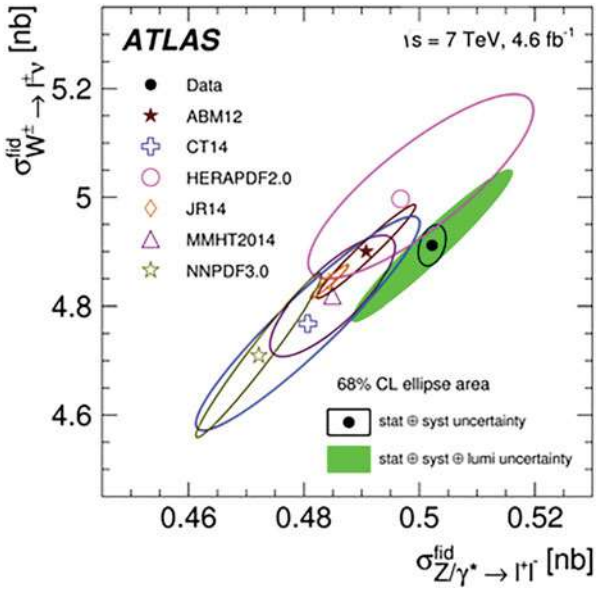

Fig. 16.33 Integrated fiducial cross sections times leptonic branching fractions, σ_W^{fid} versus σ_Z^{fid}, as measured with ATLAS 7 TeV data. The data ellipses display the 68% confidence level coverage for the total uncertainties (full green) and total excluding the luminosity uncertainty (open black). Theoretical predictions based on various parton distribution function (PDF) sets are shown with open symbols of different colours. The uncertainties of the theoretical calculations correspond to the PDF uncertainties only

16.9.3 Discovery and Measurements of the Higgs Boson

The search for the Higgs boson, over a wide mass range, was a major goal and challenge for the LHC physics programme, and the expected signatures from Higgs-boson decays therefore served as benchmarks to optimise the detector design from the very beginning in the late 1980's. These signatures span the full range of physics objects which can be reconstructed, identified and measured precisely in the experiments. The four-lepton $H \to ZZ^* \to 4l$ and the dillepton plus missing transverse energy $H \to WW^* \to l\nu l\nu$ channels were expected to be the most sensitive ones for Higgs-boson masses above 120–130 GeV. For lower values of the Higgs-boson mass, as favoured by the combined precision electroweak fits to the data available before LHC turn-on, the diphoton channel $H \to \gamma\gamma$ channel was expected to be the most sensitive channel.

The expectations for Higgs-boson discovery in the 1990's required integrated luminosities of approximately 30 to 100 fb^{-1} at the nominal LHC centre-of-mass energy of 14 TeV for Higgs-boson discovery in a single decay channel. These were updated before LHC operation with more precise theoretical calculations, resulting in particular in a significant increase of the dominant Higgs-boson

production cross-section through gluon-gluon fusion, to simple combinations of the most sensitive channels, and finally to the reduced 7 TeV centre-of-mass energy of the initial run-1 data. These updated expectations, leading to potential discovery with as little as 5–10 fb^{-1} of integrated luminosity, resulted in a period of great excitement within the ATLAS and CMS experiments, but also in the community at large, from summer 2011 (with 1 fb^{-1} collected by the experiments) to summer 2012 when the Higgs boson was officially announced as having been discovered by each of the two experiments. The evolution of the Higgs-boson signal significance over this period is illustrated in Fig. 16.34. In summer 2011, as shown in Fig. 16.34a, there were no indications of any signal yet and the fluctuations observed as a function of mass were compatible with background fluctuations. At the end of 2011, however, both experiments had excluded a Standard Model Higgs-boson signal over a mass range extending from the LEP limit of 114 to 600 GeV, except for a narrow mass range around 125 GeV in which the largest deviation from background expectations was observed around 125 GeV and corresponded to approximately three standard deviations in each experiment, as shown in Fig. 16.34b. Finally, Fig. 16.34c,d shows the observed significance in summer 2012 when the discovery was claimed and subsequently published by both experiments [42, 43] for 10 fb^{-1} of data at 7 and 8 TeV.

The four-lepton and diphoton channels have always been rightly considered as the two best channels for Higgs-boson discovery, since they both provide a clear and narrow peak for the Higgs-boson signal in the invariant mass distribution of the final state particles on top of a continuous background. In addition the four-lepton channel can be observed above a much smaller continuum background, consisting predominantly of continuum $ZZ* \rightarrow 4l$ final states. These features can be seen in Figs. 16.35 and 16.36 taken from the ATLAS discovery publication [42]. In contrast, the third channel which contributed to the discovery, namely the $H \rightarrow WW* \rightarrow l\nu l\nu$ channel, has a poor mass resolution because of the presence of neutrinos in the final state, as shown in Fig. 16.37.

After the discovery, measurements of the properties of the Higgs boson were performed in successive stages, first focusing on its spin, then on its couplings to bosons and fermions and on possible non-SM contributions to its width. At the end of run-1, ATLAS and CMS produced a combined paper on the Higgs-boson couplings [44], leading to the conclusion that in all production modes and decay channels which had been measured at the time, the Higgs-boson properties were compatible with what one would expect from the SM. More recently, each experiment has produced updated results based also on a large fraction of the run-2 data. This is illustrated in Fig. 16.38, which is based on the most recent run-2 ATLAS Higgs combination results [45] and shows that the strength of the measured Higgs-boson couplings to fermions and bosons follows the expectations from the SM, in which for example the Yukawa fermion coupling is expected to be proportional to the fermion mass. Finally, based on the most recent results from the combined run-1 and run-2 datasets from ATLAS and CMS [46], Table 16.17 shows

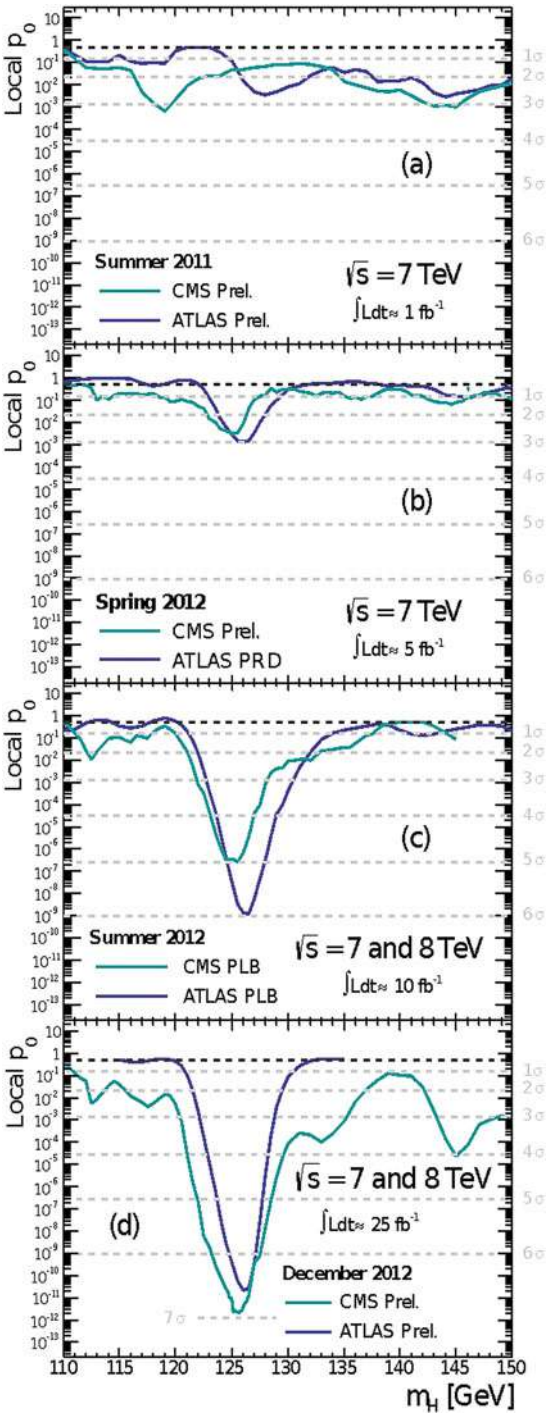

Fig. 16.34 Evolution of the combined significance of the Higgs-boson signal in the ATLAS and CMS experiments from exclusion limits in summer 2011 to discovery in summer 2012

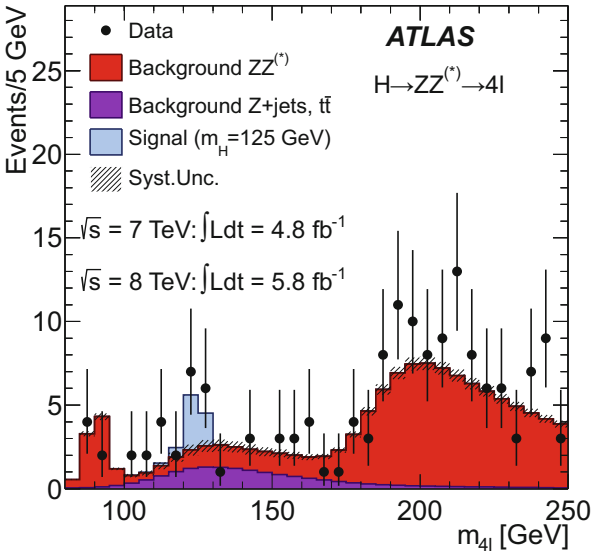

Fig. 16.35 Distribution of the four-lepton invariant mass for the selected candidates in the $H \rightarrow ZZ^* \rightarrow 4l$ channel, as observed by ATLAS at the time of discovery in summer 2012. The expected signal for $m_H = 125\,\text{GeV}$ is shown stacked on top of the overall background prediction

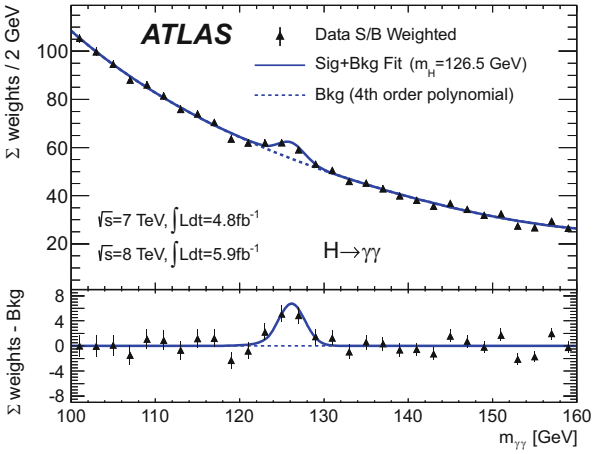

Fig. 16.36 Distribution of the invariant mass of diphoton candidates in the $Htoγγ$ channel, as observed by ATLAS at the time of discovery in summer 2012. The expected signal for $m_H = 125\,\text{GeV}$ is shown stacked on top of the overall background prediction. The residuals of the weighted data with respect to the fitted background is displayed in the bottom panel

that the Higgs couplings to charged third-generation fermions are now all clearly observed unambiguously and measured to be compatible with SM expectations. In contrast to the channels used for the discovery, the vast majority of the signals

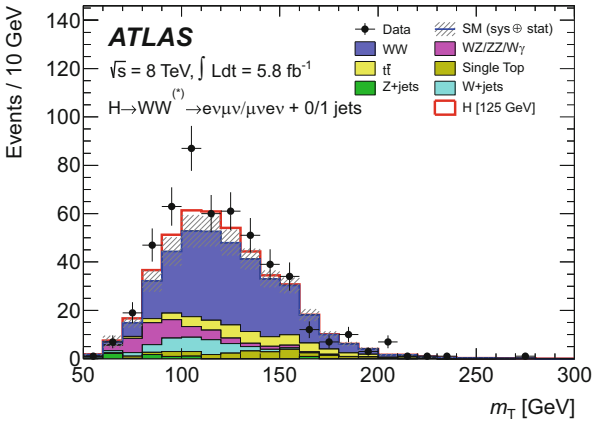

Fig. 16.37 Distribution of the transverse mass of the Higgs boson candidates in the $H \to WW$ decay channel, as observed by ATLAS at the time of discovery in summer 2012. The expected signal for $m_H = 125\,\text{GeV}$ is shown stacked on top of the overall background prediction

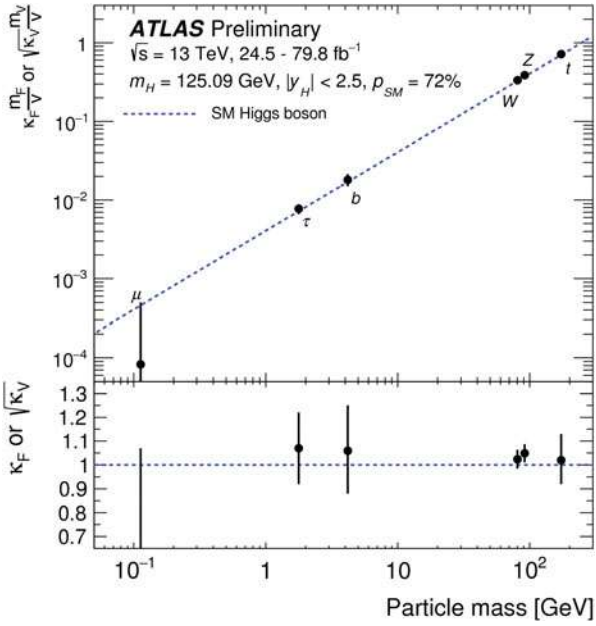

Fig. 16.38 Reduced coupling strength modifiers $\kappa_F m_F/v$ for fermions ($F = t, b, \tau, \mu$) and $\sqrt{\kappa_V} m_V/v$ for weak gauge bosons ($V = W, Z$) as a function of their masses m_F and m_V, respectively, where the vacuum expectation value of the Higgs field $v = 246\,\text{GeV}$. The results are obtained from ATLAS 13 TeV data and the SM prediction is also shown (dotted line). The coupling modifiers κ_F and κ_V are measured assuming that there are no beyond-SM contributions to the Higgs-boson decays or production processes. The lower inset shows the ratios of the measured values to their SM predictions

Table 16.17 Summary of direct measurement of all Yukawa couplings of the Higgs boson to third-generation charged fermions (τ lepton, bottom quark, and top quark) shown for the ATLAS and CMS experiments

		τ lepton	Bottom quark	Top quark
ATLAS	Observed significance	6.4σ	5.4σ	6.3σ
	Expected significance	5.4σ	5.5σ	5.1σ
	Measured to predicted yield ratio	1.09 ± 0.35	1.01 ± 0.20	1.34 ± 0.21
CMS	Observed significance	5.9σ	5.5σ	5.2σ
	Expected significance	5.9σ	5.6σ	4.2σ
	Measured to predicted yield ratio	1.09 ± 0.29	1.04 ± 0.20	1.26 ± 0.28

The expected and observed signal significances are listed, together with the ratios of the observed yields to those predicted by the SM

explored in these cases are among the most difficult Higgs-boson measurements due to the diverse and potentially large backgrounds and to the fact that the signal does not yield a narrow peak above the background.

16.9.4 Search for New Physics: Dashed and Renewed Hopes

The search for signatures from new physics beyond the SM has been ongoing in many directions from the very beginning of LHC data-taking, as has always been the case when an accelerator at the energy frontier begins operation and almost immediately delivers data to the experiments which allow them to supersede the limits from previous searches very quickly in certain cases, such as those obtained at the Tevatron. In the early years of data-taking, the experimental analyses were very much geared towards discovery because each year of data-taking brought either a large increase in integrated luminosity or a significant boost in centre-of-mass energy which is the key to searches at the edge of the available phase space. Examples of such searches are shown in Figs. 16.39 and 16.40, based on very recent results from ATLAS.

Figure 16.39 presents the evolution of the limits set by successive ATLAS searches for one of the simplest signatures of new physics, namely that for a new neutral vector boson, Z', decaying into electron or muon pairs. The limit of $\sim$1 TeV on the mass of the Z' boson in the case of a simple sequential extension of the SM was already competitive in 2010 with the legacy search limits from the CDF/D0 experiments at the Tevatron. With the full run-2 dataset, the limit is now set at 5 TeV [47] and will not extend much further without any further increase of the beam energy. Figure 16.40 shows a similar evolution of the limits set on possible excited quarks decaying into a pair of high transverse momentum jets [48].

Since 2017, however, these golden years for the excitement of searches at the edge of the available phase space are gone, and the focus of the analyses has been more on the more difficult and exotic signatures of new physics. In particular, despite its theoretical beauty before symmetry breaking, supersymmetry, if realised

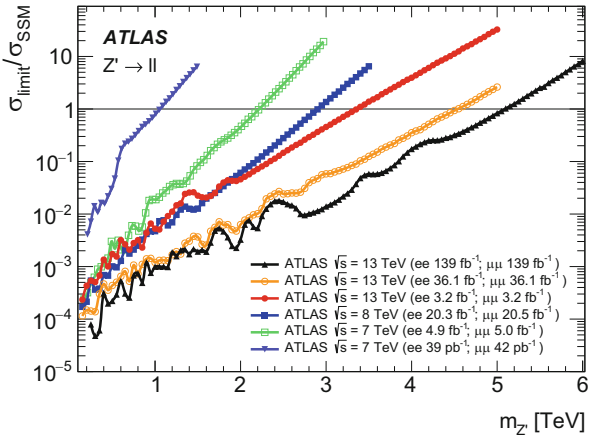

Fig. 16.39 Ratio of the observed cross-section limit to the expected Z' cross-section in the Sequential Standard Model for the combination of the dielectron and dimuon channels. The ratio is shown as a function of the Z' mass for a number of ATLAS searches performed at various LHC centre-of-mass energies from 2010 to 2018

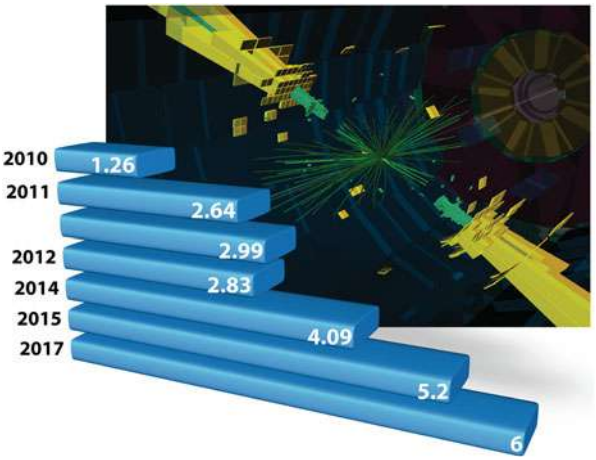

Fig. 16.40 Evolution of exclusion limits in TeV set by ATLAS on dijet resonance searches, interpreted as arising from the decay of an excited quark, from 2010 to 2017. The background image shows a display of one of the highest-mass ATLAS dijet events

in nature, has remained elusive and beyond the reach of the experimental searches in even the most exotic scenarios envisaged for its possible manifestation at the scales at which it is probed. In most models, the third generation supersymmetric partners of the quarks, the so-called stop quarks, are expected to have the smallest mass and therefore to be the most accessible at the LHC. Since their decay signatures involve predominantly top and bottom quarks, the search for these particles has had

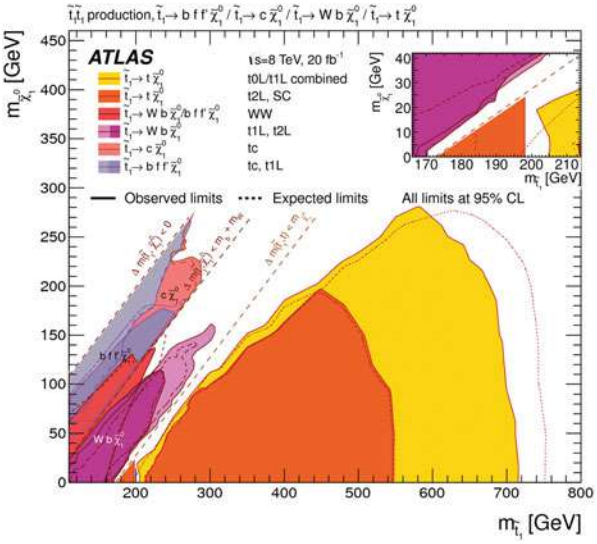

Fig. 16.41 First summary plot based on ATLAS run-1 data at $\sqrt{s} = 7$ and 8 TeV on searches for top squarks, showing the top squark versus lightest supersymmetric particle mass plane

to branch into many complex signatures, leading at first to only a partial coverage of the accessible parameter space in terms of the masses of the lightest stop quark and of the lightest neutralino, assumed to be stable. This is illustrated in Fig. 16.41, based on ATLAS run-1 data [49]. The sensitivity at the time reached at best a mass of 700 GeV and the searches were not yet very sensitive to stop quark masses close to the top-quark mass itself. Eight years later, after several generations of ever more complex and diverse searches for the stop quark, Fig. 16.42 shows that the sensitivity has extended to masses close to 1000 GeV [49], and that most of the plane of possible masses is now excluded for a lightest neutralino mass below 300 GeV.

Perhaps the most striking example of the huge efforts put by ATLAS and CMS into hunting supersymmetry has been the search for the weakly interacting supersymmetric particles, with names such as chargino, neutralino, slepton or Higgsino. It has taken the LHC experiments much longer to supersede the limits from the experiments at the LEP electron-positron collider for some of these hypothetical supersymmetric particles because of the small cross-sections involved and of the rather low energies of the decay products, leading therefore to potentially large backgrounds from SM processes with similar signatures and much larger cross-sections. This is illustrated in Fig. 16.43 which presents the most recent limits on the heavier chargino and neutralino masses as a function of the lightest neutralino mass for cases where the lightest neutralino is assumed to be stable [49].

The few results shown here, together with, for example, the very active ongoing searches for dark matter or long-lived particles, demonstrate that there are many areas still to be covered in the search for new physics at the LHC. The accelerator

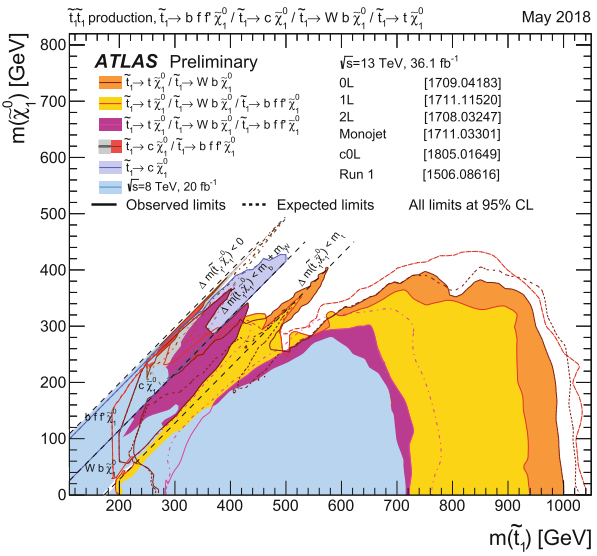

Fig. 16.42 Summary plot based on ATLAS 2015-2016 data at $\sqrt{s} = 13$ TeV on searches for top squarks, showing the top squark versus lightest supersymmetric particle mass plane

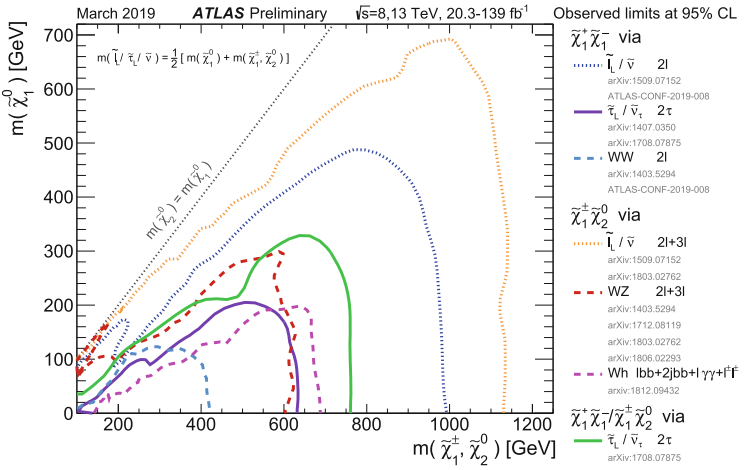

Fig. 16.43 For a variety of ATLAS datasets and search channels, 95% confidence-level exclusion limits on supersymmetric neutralino and chargino production as a function of their mass versus that of the lightest supersymmetric particle (assumed to be stable). Each individual exclusion contour represents one or more analysis in simple merged curves

and all its experiments will remain for many many years to come a wonderful provider of new data in this quest for physics beyond the Standard Model, however elusive it may be.

16.10 Conclusion

The formidable challenge related to the design, construction, installation, and commissioning of the ATLAS and CMS experiments reached a successful conclusion at the end of 2009 with the beginning of data-taking. At the time, the next challenge was as daunting and even more exciting for all the physicists participating in the exploitation phase: understand the performance of these unprecedented detectors as precisely as possible and extract the rich harvest of physics, which would undoubtedly show up once the LHC machine achieved its design goals at high energy and high luminosity.

Ten years later, after taking large amounts of data at centre-of-mass energies of 7, 8 and 13 TeV and operating successfully at luminosities exceeding even the design goals of the machine and the experiments, one can look back with tremendous pride and respect at what has been achieved by the thousands of people involved in the accelerator and the experiments. But we have also been very lucky and should feel huge gratitude towards nature which has offered the ATLAS and CMS experiments the possibility to first observe and later measure the Higgs boson in the somehow miraculous variety of production processes and decay channels with which it manifests itself at the LHC. The searches for new physics at this new frontier have, however, unfortunately not yielded yet any sign of where the solutions of some of the remaining mysteries of nature might lie. Nevertheless, the physics harvest already available from this wonderful tool for fundamental research is already rich beyond belief and the ongoing analyses in the experiments continue to probe the Standard Model predictions to the utmost of our current capabilities. Might new physics still emerge from the expected thirty times larger datasets to be collected over the coming ten to 15 years from the upgraded machine and experiments? The hopes remain high, yet only nature knows.

Acknowledgements The author wishes to thank deeply P. Sphicas, with whom the review article of ref. [4] was written. Much of the contents of this chapter are drawn from that review. May all the ATLAS and CMS colleagues who helped in a significant way to prepare this review find here also the expression of the author's most sincere thanks. It would have been impossible to collect the information in this chapter without the help of many experts in a variety of fields across all the aspects of the design, construction and installation of these very large state-of-the-art experiments in particle physics.

References

1. ATLAS Collaboration, *ATLAS Technical Proposal*. CERN/LHCC/94-43 (1994); CMS Collaboration, *CMS Technical Proposal*. CERN/LHCC/94-38 (1994).
2. ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*. JINST 3 S08003 (2008).
3. CMS Collaboration, *The CMS Experiment at the CERN LHC*. JINST 3 S08004 (2008).
4. Froidevaux, D., Sphicas, P., Annu. Rev. Nucl. Part. Sci. **56** (2006) 375–440.

5. ALICE Collaboration, *The ALICE Experiment at the CERN LHC*. JINST 3 S08002 (2008).
6. LHCb Collaboration, *The LHCb Detector at the LHC*. JINST 3 S08005 (2008).
7. Grunewald, M.W., Invited Talk at EPS Int. Europhysics Conf. on High Energy Physics (HEP-EPS-2005), Lisbon, Portugal (2005). arXiv:hep-ex/0511018.
8. Chanowitz, M.S., Phys. Rev. D **66** (2002) 073002. arXiv:hep-ph/0207123.
9. Barbieri, R., Strumia, A., Phys. Lett. B **462** (1999) 144. arXiv:hep-ph/9905281.
10. Birkedal, A., Matchev, K., Perelstein M., arXiv:hep-ph/0412278.
11. Gianotti, F., Mangano, M., CERN Preprint, CERN-PH-TH/2005-072 (2005). arXiv:hep-ph/0504221.
12. Blaising, B., et al., Contribution to the Zeuthen Briefing Book. http://council-strategygroup.web.cern.ch/council-strategygroup/
13. ATLAS Collaboration, *Detector and Physics Performance Technical Design Report*. CERN/LHCC/99-15 (1999);
Asai, S., et al., Eur. Phys. J. C **32** (2004) 19.
14. ATLAS Collaboration, *Magnet System Technical Design Report*. CERN/LHCC/97-18 (1997);
ATLAS Collaboration, *Barrel Toroid Technical Design Report*. CERN/LHCC/97-19 (1997);
ATLAS Collaboration, *End-cap Toroid Technical Design Report*. CERN/LHCC/97-20 (1997);
ATLAS Collaboration, *Central Solenoid Technical Design Report*. CERN/LHCC/97-21 (1997);
CMS Collaboration, *The Magnet Project - Technical Design Report*. CERN/LHCC/97-10 (1997).
15. ATLAS Collaboration, *Technical Co-ordination Technical Design Report*. CERN/LHCC/99-01 (1999).
16. Akesson, T., et al., *Report of the High-Luminosity Study Group to the CERN Long-Range Planning Committee*, ed. Mulvey, J., CERN Yellow Book, 88-02 (1988).
17. ATLAS Collaboration, *Inner Detector Technical Design Report*, Vol. II. CERN/LHCC/97-16. ISBN 92-9083-103-0 (1997);
ATLAS Collaboration, *Pixel DetectorTechnical Design Report*. CERN/LHCC/98-13, (1998).
18. Lindstrom, G., *Radiation damage in silicon detectors*, Nucl. Instrum. Meth. A **512** (2003) 30–43.
19. Coe, P.A., Howell, D.F., Nickerson, R.B., *Frequency scanning interferometry in ATLAS: remote, multiple, simultaneous and precise distance measurements in a hostile environment*. Meas. Sci. Technol. (2004) 2175–2187.
20. Buttar, C.M., et al., Nucl. Instrum. Meth. A **447** (2000) 126;
Dierlamm, A., Nucl. Instrum. Meth. A **514** (2003) 167.
21. Akesson, T., et al., Nucl. Instrum. Meth. A **522** (2004) 25;
Capeans, M., et al., IEEE Trans. Nucl. Sci. **51** (2004) 960;
Akesson, T., et al., Nucl. Instrum. Meth. A **515** (2003) 166;
Romaniouk A., ATLAS Internal Note, ATL-INDET-98-211 (1998).
22. Gorelov, I., et al., Nucl. Instrum. Meth. A **481** (2002)204;
Alimonti, G., et al., ATLAS Internal Note, ATL-INDET-INT-2005-006 (2005);
Alimonti, G., et al., ATLAS Internal Note, ATL-INDET-INT-2005-007 (2005).
23. ALEPH Collaboration, Nucl. Instrum. Meth. A **360** (1995) 481;
OPAL Collaboration, OPAL Technical Note, OPAL-TN-306 (1995).
24. Drage, L., Parker, M.A., ATLAS Internal Note, ATL-PHYS-2000-007 (2000).
25. ATLAS Collaboration, *Liquid Argon Calorimeter Technical Design Report*. CERN/LHCC/96-41 (1996).
26. Colas, J., et al., Nucl. Instrum. Meth. A **550** (2005) 96.
27. CMS Collaboration, *CMS Physics Technical Design Report*, Vol. I: *Detector Performance and Software*. CERN/LHCC/2006-01 (2006).

28. Braunschweig, H., et al., Nucl. Instrum. Meth. A **265** (1988) 246;
Andrieu, B., et al., DESY Preprint, DESY 93-04 (1993).
29. ATLAS Collaboration, *Muon Spectrometer Technical Design Report*.
CERN/LHCC/97-22 (1997).
30. ATLAS Collaboration, *Computing Technical Design Report*. CERN/LHCC/2005-022 (2005).
31. Chytracek, R., et al., Nucl. Science Symp. Conf. Record, IEEE **4** (2004) 2077;
see also http://lcgapp.cern.ch/project/persist/
32. Brun, R., Rademakers, F., Nucl. Instrum. Meth. A **389** (1997) 81;
see also http://root.cern.ch/
33. The LHC Computing Grid. *Technical Design Report*. CERN/LHCC/2005-024 (2005).
34. ALICE Collaboration, Eur. Phys. J **C65** (2010) 111;
ALICE Collaboration, Phys. Lett. **B693** (2010) 53;
CMS Collaboration, JHEP **02** (2010) 041;
ATLAS Collaboration, Phys. Lett. **B688** (2010) 21.
35. ATLAS Collaboration, JHEP **12** (2010) 060.
36. ATLAS Collaboration, Eur. Phys. J. **C71** (2011) 1577.
37. ATLAS Collaboration, Eur. Phys. J. **C74** (2014) 2941.
38. ATLAS Collaboration, Eur. Phys. J. **C74** (2014) 3130.
39. ATLAS Collaboration, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/JETM-2018-006.
40. ATLAS Collaboration, JHEP **09** (2017) 020.
41. ATLAS Collaboration, Eur. Phys. J. **C77** (2017) 367.
42. ATLAS Collaboration, Science **338** (2012) 1576.
43. CMS Collaboration, Science **338** (2012) 1569.
44. ATLAS and CMS Collaborations, JHEP **08** (2016) 045.
45. ATLAS Collaboration, ATLAS-CONF-2019-005 (2019), https://cds.cern.ch/record/2668375.
46. M. Kado, talk given at the Aspen 2019 Winter Conference in Physics, https://indico.cern.ch/event/748043/contributions/3313769/attachments/1817533/2971936/AspenHiggs.pdf.
47. ATLAS Collaboration, CERN-EP-2019-030 (2019), http://arxiv.org/abs/arXiv:1903.06248, submitted to Phys. Lett. B.
48. ATLAS Collaboration, ATLAS-CONF-2019-007 (2019), http://cdsweb.cern.ch/record/2668385.
49. ATLAS Collaboration, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CombinedSummaryPlots/SUSY.

Chapter 17
Neutrino Detectors Under Water and Ice

Christian Spiering

17.1 Introduction

Underwater/ice neutrino telescopes are multi-purpose detectors covering astrophysical, particle physics and environmental aspects [1–3]. Among them, the detection of the feeble fluxes of astrophysical neutrinos which should accompany the production of high energy cosmic rays is the clear primary goal [4, 5]. Since these neutrinos can escape much denser celestial bodies than light, they can trace processes hidden to traditional astronomy. Different to gamma rays, neutrinos provide incontrovertible evidence for hadronic acceleration. On the other hand, their extremely low interaction cross section makes their detection extraordinarily difficult.

Figure 17.1 shows a compilation of the spectra of dominant natural and artificial neutrino fluxes. Solar neutrinos, burst neutrinos from SN-1987A, reactor neutrinos, terrestrial neutrinos from radioactive decay processes in the Earth and neutrinos generated in cosmic ray interactions in the Earth atmosphere ("atmospheric neutrinos") have been already detected. Two guaranteed—although not yet detected—fluxes are the diffuse flux of neutrinos from past supernovae (marked "background from old supernovae") and the flux of neutrinos generated in collisions of ultra-energetic protons with the 3 K cosmic microwave background [6] (marked GZK after Greisen, Zatsepin and Kuzmin [7] who first considered such collisions). These neutrinos will hopefully be detected in the next decade. Neutrinos in the TeV-PeV range emerging from acceleration sites of cosmic rays (marked AGN after "Active Galactic Nuclei") have been detected in 2013 with IceCube [8]. No practicable idea exists how to detect 1.9 K cosmological neutrinos.

C. Spiering (✉)
DESY, Zeuthen, Germany
e-mail: christian.spiering@desy.de

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_17

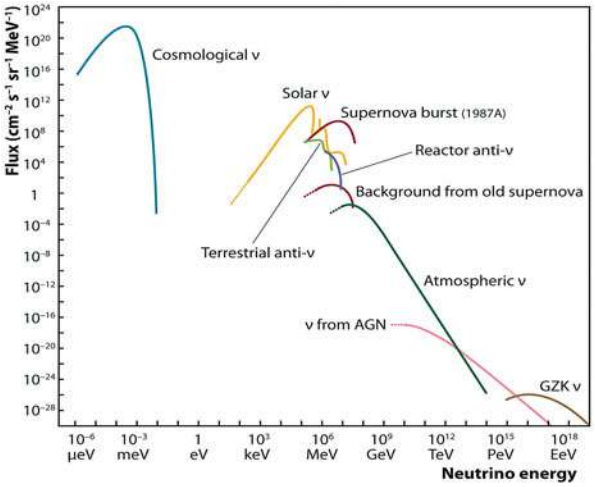

Fig. 17.1 Spectra of natural and reactor neutrinos

The energy range below 5 GeV is the clear domain of underground detectors, notably water Cherenkov, liquid scintillator and radio-chemical detectors (see chapter C4 and [9]) which led to the discovery of solar and atmospheric neutrinos, of neutrino oscillations and of neutrinos from Supernova SN1987A. These detectors, presently with maximal geometrical cross sections of about 1000 m^2, have turned out to be too small to detect the feeble fluxes of astrophysical neutrinos from cosmic acceleration sites. The high energy frontier of TeV and PeV energies is being tackled by much larger, expandable detectors installed in open water or ice, a principle first proposed by M. Markov in 1959 [10]. They consist of arrays of photomultipliers recording the Cherenkov light from charged particles produced in neutrino interactions. Towards even higher energies, novel detectors aim at detecting the coherent Cherenkov radio signals (ice, salt) or acoustic signals (water, ice, salt) from neutrino-induced particle showers. Air shower detectors search for showers with a "neutrino signature". The very highest energies are covered by balloon-borne detectors recording radio emission in terrestrial ice masses, by ground-based radio antennas sensitive to radio emission in the moon crust, or by satellite detectors searching for fluorescence light or radio signals from neutrino-induced air showers. This article focuses on optical detectors in water and ice. The methods for higher energies are sketched in Sect. 17.8. Table 17.1 gives an overview over past, present and future optical detectors in water and ice.

Underwater/ice detectors—apart from searching for neutrinos from cosmic ray sources—also address a variety of particle physics questions (see for reviews [12, 13]). With their huge event statistics, large neutrino telescopes have opened a new perspective for oscillation physics with atmospheric neutrinos, and actually can compete with accelerator experiments [14]. Another example for a particle physics task is the search for muons produced by neutrinos from dark matter annihilation

Table 17.1 Past, present (2018) and future neutrino telescope projects and their main parameters

Experiment	Location	Size (km^3)	Milestones	Remarks
DUMAND	Hawaii		1978/–/–/1996	Terminated due to techn./funding problems
NT200	Lake Baikal	10^{-4}	1980/1993/1998/2015	First proof of principle
NESTOR	Med. Sea off Peloponnes		1991/–/–/–	Data taking with prototype
NEMO	Med. Sea off Sicily		1998/–/–/–	R&D project prototype tests
AMANDA	South Pole	0.015	1990/1996/2000/2009	First deep-ice ν telescope
ANTARES	Med. Sea off Toulon	0.010	1997/2006/2008/2018	First deep-sea ν telescope
IceCube	South Pole	1.0	2011/2005/2010/–	First km^3-sized detector
GVD-1	Lake Baikal	0.4	2012/2015/–/–	High-energy ν astronomy
KM3NeT/ARCA	Med. Sea off Sicily	1–1.5	2013/2015/–/–	High-energy ν astronomy
KM3NeT/ORCA	Med. Sea off Toulon	0.003	2014/2017/–/–	Low-energy configuration for oscillation physics
GVD-2	Lake Baikal	1.5	2012/–/–/–	Extension of GVD-1
KM3NeT Phase 3	Med. Sea	3–5	2013/–/–/–	Planned extension of of KM3NeT
IceCube-Gen2	South Pole	5–10	2014/–/–/–	Planned IceCube extension covering low/high energies, a surface array and radio detection

The milestone years give times of project start, of first data taking with partial configurations, of detector completion, and of project termination. Projects with first data expected past 2025 are in italics (modified after [11])

in the Sun or in the center of the Earth. These searches are sensitive to super-symmetric WIMPs (Weak Interacting Massive Particles) as dark matter candidates. Underwater/ice detectors can also search for relativistic magnetic monopoles, with a light emission 8300 times stronger than that of a bare muon and therefore providing a very clear signature. Other tasks include the search for super-heavy particles like GUT monopoles, super-symmetric Q-balls or nuclearites which would propagate with less than a thousandth of the speed of light and emit light by heating up the medium or by catalyzing proton decays.

The classical operation of neutrino telescopes underground, underwater and in deep ice is recording upward travelling muons generated in a charged current neutrino interaction. The upward signature guarantees the neutrino origin of the muon since no other particle can cross the Earth. A neutrino telescope should be

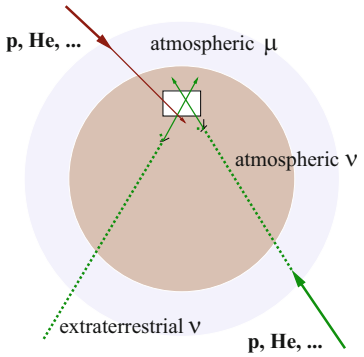

Fig. 17.2 Sources of muons in deep underwater/ice detectors. Cosmic nuclei—protons(p), α-particles(He), etc.—interact in the Earth atmosphere (light-colored). Sufficiently energetic muons produced in these interactions ("atmospheric muons") can reach the detector (white box) from above. Muons from the lower hemisphere must have been produced in neutrino interactions

arranged at >1 km depth in order to suppress the background from misreconstructed downward moving muons which may mimic upward moving ones (Fig. 17.2).

The identification of extraterrestrial neutrino events faces three sources of backgrounds:

- down-going punch-through muons from cosmic-ray interactions in the atmosphere ("atmospheric muons"). This background can be reduced by going deeper.
- random backgrounds due to photomultiplier (PMT) dark counts, ^{40}K decays (mainly in sea water) or bioluminescence (only water), which impact adversely on event recognition and reconstruction. This background can be mitigated by local coincidences of PMTs.
- neutrinos from cosmic-ray interactions in the atmosphere ("atmospheric neutrinos"). Extraterrestrial neutrinos can be separated from atmospheric neutrinos on a statistical basis (due to their harder energy spectrum). For down-going neutrinos interacting within the detector, atmospheric neutrinos can be largely rejected by vetoing accompanying atmospheric muons from the same shower as the atmospheric neutrino.

Atmospheric neutrinos, of course, have an own scientific value: at medium and high energies they are a well-understood "standard candle" to calibrate the detector, at low energies they allow for investigating neutrino oscillations.

17.2 Neutrino Interactions

The behaviour of the neutrino cross section can be approximated by a linear dependence for $E_\nu < 5$ TeV, for energies larger than 5 TeV by an $E_\nu^{0.4}$ dependence [1]. The absolute value of the cross section at 1 TeV is about 10^{-35} cm^2.

The final state lepton follows the initial neutrino direction with a mean mismatch angle θ decreasing with the square root of the neutrino energy [4]:

$$<\theta> \approx \frac{1.5^\circ}{\sqrt{E_\nu[\text{TeV}]}} \tag{17.1}$$

This on the one hand principally enables source tracing with charged current muon neutrinos, but on the other hand sets a kinematical limit to the ultimate angular resolution. It is worse than for high energy gamma astronomy and particularly worse than for conventional astronomy.

The probability $P_{\nu\to\mu}(E_\nu, E_\mu^{min})$ to produce, in a charged current interaction of a muon neutrino with energy E_ν, a muon reaching the detector with a minimum detectable energy E_μ^{min} depends on the cross section $d\sigma_{\nu N}^{CC}(E_\nu, E_\mu)/dE_\mu$ and the effective muon range R_{eff}, which is defined as the range after which the muon energy has decreased to E_μ^{min} [4]:

$$P_{\nu\to\mu}(E_\nu, E_\mu^{min}) = N_A \int_{E_{\mu,min}}^{E_\nu} dE_\mu \frac{d\sigma_{\nu N}^{CC}(E_\nu, E_\mu)}{dE_\mu} \cdot R_{eff}(E_\mu^{min}, E_\mu) \tag{17.2}$$

with N_A being the Avogadro constant. For water and $E_\mu^{min} \approx 1$ GeV one can approximate [4]

$$P_{\nu\to\mu} = 1.3 \cdot 10^{-6} \cdot E_\nu^{2.2} \text{ for } \mathrm{E}_\nu < 1 \text{ TeV} \tag{17.3}$$

$$= 1.3 \cdot 10^{-6} \cdot E_\nu^{0.8} \text{ for } \mathrm{E}_\nu > 1 \text{ TeV} \tag{17.4}$$

(with E_ν given in TeV). This means, that a telescope can detect a muon neutrino with 1 TeV energy with a probability of about 10^{-6}, if the telescope is on the neutrino's path.

The number of events from a flux Φ_ν recorded by a detector with area A within a time T under a zenith angle ϑ is then given by

$$\frac{N_\mu(E_{\mu,min}, \vartheta)}{AT} = \int_{E_{\mu,min}}^{E_\nu} dE_\nu \Phi_\nu(E_\nu, \vartheta) \cdot P_{\nu\mu}(E_\nu, E_{\mu,min}) \cdot e^{-\sigma_{tot}(E_\nu) N_A Z(\vartheta)} \; . \tag{17.5}$$

Here $Z(\delta)$ is the matter column in the Earth crossed by the neutrino. For sub-TeV energies, absorption in the Earth is negligible and the exponential term $\sim$1 (see Fig. 17.5).

17.3 Principle of Underwater/Ice Neutrino Telescopes

Underwater/ice neutrino telescopes consist of a lattice of photomultipliers (PMs) housed in transparent pressure spheres which are spread over a large volume in oceans, lakes or glacial ice. The PMs record arrival time and amplitude, sometimes even the full waveform, of Cherenkov light emitted by muons or particle cascades.

In most designs the spheres are attached to strings which—in the case of water detectors—are moored at the ground and held vertically by buoys. The typical spacing along a string is 10–25 m, and between strings 60–200 m. The spacing is incomparably large compared to Super-Kamiokande (see chapter C4). This allows covering large volumes but makes the detector practically blind with respect to phenomena below 10 GeV. An exception are planned high-density detectors under water and ice which are tailored to oscillation physics and to the determination of the mass hierarchy of neutrinos [15, 16].

17.3.1 Cherenkov Light

Charged particles moving faster than the speed of light in a medium with index of refraction n, $v \geq c/n$, emit Cherenkov light. The index of refraction depends on the frequency ν of the emitted photons, $n = n(\nu)$. The total amount of released energy is given by

$$-\left(\frac{dE}{dx}\right)_c = \frac{2\pi \cdot \alpha}{c} \cdot \int_{\beta \cdot n(\nu) \geq 1} \left(1 - \frac{1}{\beta^2 \cdot n^2(\nu)}\right) d\nu \quad , \tag{17.6}$$

with α being the fine structure constant and $\beta = v/c$. In the transparency window of water, i.e. for wavelength 400 nm $\leq \lambda \leq$ 700 nm, the index of refraction for water is $n \approx 1.33$, yielding about 400 eV/cm, or ≈200 Cherenkov photons per cm. The spectral distribution of Cherenkov photons is given by

$$\frac{dN}{dx d\lambda} = \frac{2\pi \cdot \alpha}{\lambda^2} \cdot \left(1 - \frac{1}{\beta^2 \cdot n^2}\right) \quad . \tag{17.7}$$

The photons are emitted under an angle Θ_C given by

$$\cos \Theta_C = \frac{1}{\beta \cdot n} \quad . \tag{17.8}$$

For water, $\Theta_C = 41.2°$.

17.3.2 Light Propagation

The propagation of light in water is governed by absorption and scattering. In the first case the photon is lost, in the second case it changes its direction. Multiple scattering effectively delays the propagation of photons. The parameters generally chosen as a measure for these phenomena are [17, 18]:

(a) The absorption length $L_a(\lambda)$—or the absorption coefficient $a(\lambda) = 1/L_a$—with λ being the wavelength. It describes the exponential decrease of the number N of non-absorbed photons as a function of distance r, $N = N_0 \cdot \exp(-r/L_a)$.
(b) The scattering length $L_b(\lambda)$ and scattering coefficient $b(\lambda)$, defined in analogy to $L_a(\lambda)$ and $a(\lambda)$.
(c) The scattering function $\chi(\theta, \lambda)$, i.e. the distribution in scattering angle θ.
(d) Often instead of the "geometrical" scattering length $L_b(\lambda)$, the effective scattering length L_{eff} is used: $L_{eff} = L_b/(1 - \langle\cos\theta\rangle)$ with $\langle\cos\theta\rangle$ being the mean cosine of the scattering angle. L_{eff} "normalizes" scattering lengths for different distributions $\chi(\theta, \lambda)$ of the scattering angle to one with $\langle\cos\theta\rangle = 0$, i.e. L_{eff} is a kind of isotropization length. For $\langle\cos\theta\rangle \sim 0.8-0.95$, as for all media considered here, photon delay effects in media with the same L_{eff} are approximately the same.

Table 17.2 summarizes typical values for Lake Baikal [19, 20], oceans [21, 22] and Antarctic ice [23, 24], each are given for the wavelength of their maximum.

Scattering and absorption in water and ice are determined with artificial light sources. The scattering coefficient in water changes only weakly with wavelength. The dependence on depth over the vertical dimensions of a neutrino telescope in water is small, but parameters may change in time, due to transient water inflows loaded with bio-matter or dust, or due to seasonal changes in water parameters. They must therefore be permanently monitored. In glacial ice at the South Pole, the situation is different. The parameters are constant in time but strongly change with depth (see Sect. 17.6.3).

Strong absorption leads to reduced photon collection, strong scattering deteriorates the time information which is essential for the reconstruction of tracks and showers (see Sects. 17.6 and 17.7).

Table 17.2 Absorption length and effective scattering length for different sites

Site	L_a (m)	L_{eff} (m)
Lake Baikal, 1 km depth	18–22	150–250 (seasonal variations)
Ocean, > 1.5 km depth	40–70 (depends on site and season)	200–300 (depends on site and season)
Polar ice, 1.5–2.0 km depth	∼95 (average)	∼20 (average)
Polar ice, 2.2–2.5 km depth	>100	30–40

17.3.3 Detection of Muon Tracks and Cascades

Neutrinos can interact with target nucleons N through charged current, CC ($\nu_l + N \rightarrow l + X$, with l denoting the charged parter lepton of the neutrino) or neutral current, NC ($\nu_l + N \rightarrow \nu_l + X$) processes. A CC reaction of a ν_μ produces a muon track and a hadronic particle cascade, whereas all NC reactions and CC reactions of ν_e produce particle cascades only. CC interactions of ν_τ can have either signature, depending on the τ decay mode.

In most astrophysical models, neutrinos are expected to be produced through the $\pi/K \rightarrow \mu \rightarrow e$ decay chain, i.e. with a flavour ratio $\nu_e : \nu_\mu : \nu_\tau \approx 1 : 2 : 0$. For sources outside the solar system, neutrino oscillations turn this ratio to $\nu_e : \nu_\mu : \nu_\tau \approx 1 : 1 : 1$ upon arrival on Earth. That means that about 2/3 of the charged current interactions appear as cascades.

Figure 17.3 sketches the two basic detection modes of underwater/ice neutrino telescopes.

17.3.3.1 Muon Tracks

In the muon-track mode, high energy neutrinos are inferred from the Cherenkov cone accompanying muons which enter the detector from below or which strat inside the detector. The upward signature guarantees the neutrino origin of the muon since no other particle can cross the Earth. The effective volume considerably exceeds the actual detector volume due to the large range of muons (about 1 km at 300 GeV and 24 km at 1 PeV [4]).

The muon looses energy via ionization, pair production, bremsstrahlung and photonuclear reactions. The energy loss can be parameterized by [4, 25]

$$-\frac{dE_\mu}{dx} = a + b \cdot E_\mu \quad . \tag{17.9}$$

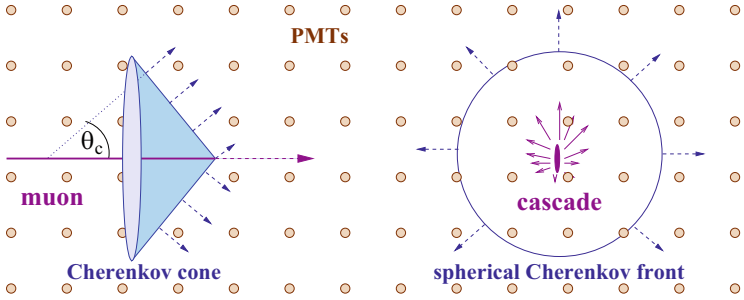

Fig. 17.3 Detection of muon tracks (left) and cascades (right) in underwater detectors

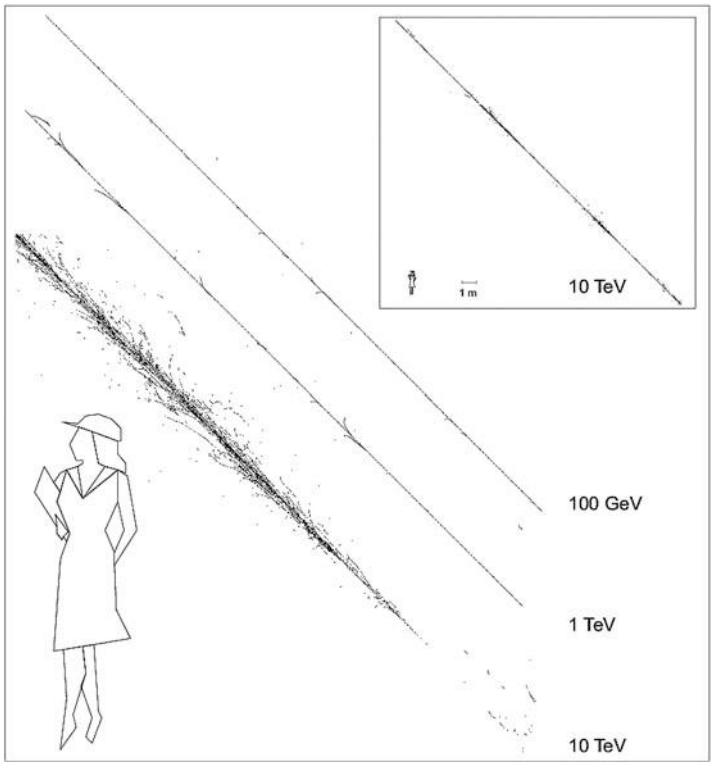

Fig. 17.4 Typical muon tracks and charged secondaries in water above the Cherenkov threshold, for different muon energies (a) 10 TeV (box), and zoomed: 10 TeV, 1 TeV and 100 GeV [26]

For water, the ionization loss is given by $a = 2\,\text{MeV/cm}$, the energy loss from pair production, bremsstrahlung and photonuclear reactions is described by $b = (1.7 + 1.3 + 0.4) \cdot 10^{-6}\,\text{cm}^{-1} = 3.4 \cdot 10^{-6}\,\text{cm}^{-1}$ and rises linearly with energy [25]. Figure 17.4 shows muons tracks with the corresponding secondaries from the last three processes [26]. A detailed description of the muon propagation through matter has to take into account the stochastic character of the individual energy loss processes, which leads to separated cascades of secondaries along the muon track.

Underwater/ice telescopes are optimized for the detection of muon tracks and for energies of a TeV or above, by the following reasons:

(a) The flux of neutrinos from cosmic accelerators is expected to be harder than that of atmospheric neutrinos, yielding a better signal-to-background ratio at higher energies.
(b) Neutrino cross section and muon range increase with energy. The larger the muon range, the larger the effective detector volume.
(c) The mean angle between muon an neutrino decreases with energy like $E^{-0.5}$, resulting in better source tracing and signal-to-background ratio at high energy.

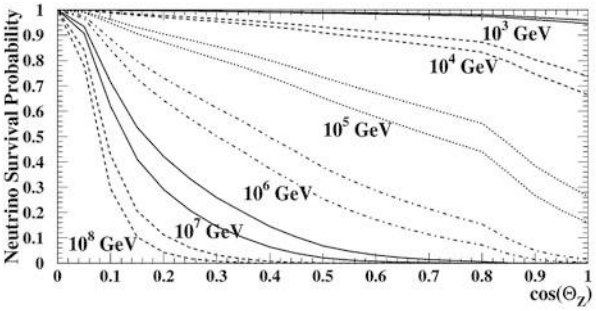

Fig. 17.5 Transmission of the Earth for neutrinos of different energy, as a function of zenith angle [27]

(d) For energies above a TeV, the increasing light emission allows estimating the muon energy with an accuracy of $\sigma(\log E_\mu) \sim 0.3$. By unfolding procedures, a muon energy spectrum can be translated into a neutrino energy spectrum.

Muons which have been generated in the Earth's atmosphere above the detector and punch through the water or ice down to the detector outnumber neutrino-induced upward moving muons by several orders of magnitude (about 10^6 at 1 km depth and 10^4 at 4 km depth) and have to be removed by careful up/down assignment.

At energies above a few hundred TeV, where the Earth is going to become opaque even to neutrinos, neutrino-generated muons arrive preferentially from directions close to the horizon, at EeV energies essentially only from the upper hemisphere (Fig. 17.5). The high energy deposition of muons from PeV-EeV extraterrestrial neutrinos provides a handle to distinguish them—on a statistical basis—from downward going atmospheric muons (those with a spectrum decreasing rather steeply with energy). A different case are down-going muon tracks or cascades starting within the detector. They must be due to neutrino interactions. If the neutrino has been generated in the atmosphere, it will be accompanied in most cases by muons from the same air shower, the higher the energy, the more frequently. Therefore one can apply a veto against accompanying down-going muons and thereby remove most atmospheric neutrinos. This method has been first applied in [8].

17.3.3.2 Cascades

Neutral current interactions and charged current interactions of electron and (most) tau neutrinos do not lead to high energy muons but to electromagnetic or hadronic cascades. Their length increases only like the logarithm of the cascade energy (Fig. 17.6). Cascade events are therefore typically "contained" events.

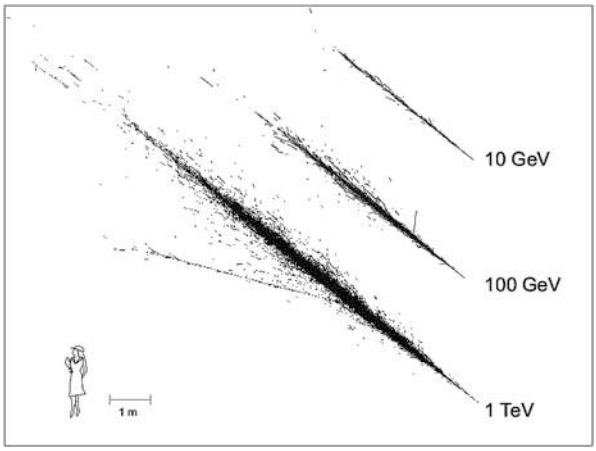

Fig. 17.6 Typical electromagnetic cascades in fresh water for 10 GeV, 100 GeV and 1 TeV [26]

With 5–20 m length in water, and a diameter of the order of 10–20 cm, cascades may be considered as quasi point-like compared to the spacing of the strings along which the PMs are arranged (again with the exception of high-density arrays tailored to oscillation physics). The effective volume for the clear identification of isolated cascades from neutrino interactions is close to the geometrical volume of the detector. For first-generation neutrino telescopes it is therefore much smaller than that for muon detection. However, for kilometer-scale detectors and not too large energies it can reach the same order of magnitude as the latter. The total amount of light is proportional to the energy of the cascade. Since the cascades are "contained", they do not only provide a dE/dx measurement (like muons) but an E-measurement. Therefore, in charged current ν_e and ν_τ interactions, the neutrino energy can be determined with an accuracy of 10–30% (depending on energy and PM spacing). While this is much better than for muons, the directional accuracy is worse since the lever arm for fitting the direction is negligibly small. The background from atmospheric electron neutrinos is much smaller than in the case of extraterrestrial muon neutrinos and atmospheric muon neutrinos. All this taken together, makes the cascade channel particularly interesting for searches for diffuse high-energy excesses of extraterrestrial neutrinos over atmospheric neutrinos.

17.4 Effective Area and Sensitivity

The detection efficiency of a neutrino telescope is quantified by its effective area, e.g., the fictitious area for which the full incoming neutrino flux would be recorded. Fig. 17.7 shows the effective area of the IceCube detector for the detection mode of through-going muons. The increase with E_ν is due to the rise of neutrino cross

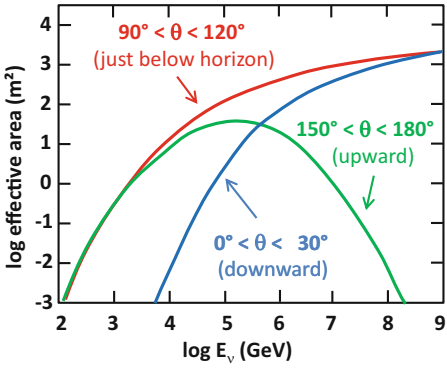

Fig. 17.7 Effective area of the IceCube detector for neutrinos, $A_{eff}(\nu)$, assuming the detection mode of through-going muons. The zenith angle θ is counted 0°/180° for vertically downward/upward moving muons. The effective area is strongly increasing with energy due to increasing neutrino cross section and muon range. The decrease at high energy and large zenith angles is due to the opacity of the Earth to neutrinos with energies above ≈100 TeV. Identification of downward-going neutrinos requires strong cuts against atmospheric muons, hence the cut-off towards low E_ν for $\theta < 30°$

section and muon range, while neutrino absorption in the Earth causes the decrease at large zenith angle θ. Identification of downward-going neutrinos requires strong cuts against atmospheric muons, hence the cut-off towards low E_ν.

Due to the small cross section, the effective area is many orders of magnitude smaller than the geometrical dimension of the detector; a muon neutrino with 1 TeV, e.g., can be detected with a probability of the order 10^{-6} if the telescope is on its path. Note that the detection efficiency for cascades or muons starting within the detector are much smaller since these detection modes do not profit from the potentially large range of muons coming from outside.

Even cubic kilometer neutrino telescopes reach only effective areas between a few square meters and a few hundred square meters, depending on energy. This has to be compared to several ten thousand square meters typical for air Cherenkov telescopes which detect gamma ray-initiated air showers. A ratio 1:1000 (10 m^2:10 000 m^2) may appear desperately small. However, one has to take into account that Cherenkov gamma telescopes can only observe one source at a time, and that their observations are restricted to moon-less, cloud-less nights. Neutrino telescopes observe a full hemisphere, 24 h per day. Therefore, cubic kilometer detectors reach a flux sensitivity similar to that which first-generation Cherenkov gamma telescopes like Whipple and HEGRA [28, 29] had reached for TeV gamma rays, namely $\Phi(>1\,\text{TeV}) \approx 10^{-12}\,\text{cm}^{-2}\,\text{s}^{-1}$.

17.5 Reconstruction

In this section, some relevant aspects of event reconstruction are demonstrated for the case of muons tracks [30, 31]. For cascades, see [32, 33]. The reconstruction procedure for a muon track consists of several consecutive steps:

1. Rejection of noise hits
2. Simple pre-fit procedures providing a first-guess estimate for the following iterative maximum-likelihood reconstructions
3. Maximum-likelihood reconstruction
4. Quality cuts in order to reduce background contaminations and to enrich the sample with signal events. This step is strongly dependent of the actual analysis—diffuse fluxes at high energies, searches for steady point sources, searches for transient sources etc.

An infinitely long muon track can be described by an arbitrary point $\vec{r}_0$ on the track which is passed by the muon at time t_0, with a direction $\vec{p}$ and energy E_0. Photons propagating under the Cherenkov angle θ_c and on a straight path ("direct photons") are expected to arrive at PM i located at $\vec{r}_i$ at a time

$$t_{geo} = t_0 + \frac{\vec{p} \cdot (\vec{r}_i - \vec{r}_0) + d \cdot \tan\theta_c}{c}, \tag{17.10}$$

where d is the closest distance between PM i and the track, and c the vacuum speed of light. The time residual t_{res} is given by the difference between the measured hit time t_{hit} and the hit time expected for a direct photon t_{geo}:

$$t_{res} = t_{hit} - t_{geo}. \tag{17.11}$$

Schematic distributions for time residuals are given in Fig. 17.8. An unavoidable symmetric contribution in the range of a nanosecond comes from the PM/electronics time jitter, σ_t. An admixture of noise hits to the true hits from a muon track adds a flat pedestal contribution like shown in top right of Fig. 17.8. Electromagnetic cascades along the track lead to a tail towards larger (and only larger) time residuals (bottom left). Scattering of photons which propagate in water loaded with bio-matter and dust or in ice can lead to an even stronger delay of the arrival time (bottom right). These residuals must be properly implemented in the probability density function for the arrival times.

The simplest likelihood function is based exclusively on the measured arrival times. It is the product of all N_{hit} probability density functions p_i to observe, for a given value of track parameters $\{a\}$, photons at times t_i at the location of the hit PMs:

$$L_{time} = \prod_{i=1}^{Nhit} p(t_{res,i}|\{a\}) \tag{17.12}$$

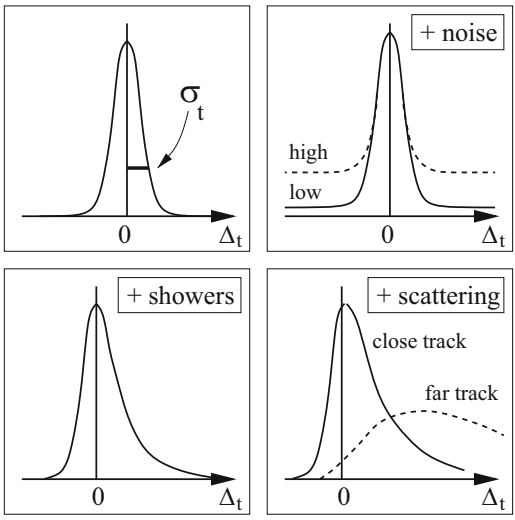

Fig. 17.8 Schematic distributions of arrival times for different cases (see text)

More complicated likelihoods include the probability of hit PMs to be hit and of non-hit PMs to be not hit, or the amplitudes of hit PMs. Instead for referring only to the arrival time of the first photon for a given track hypothesis, and the amplitude for a given energy hypothesis, one may also refer to the full waveform from multiple photons hitting the PM. For efficient background suppression, the likelihood may also incorporate information about the zenith angular dependence of background and signal (Bayesian probability). The reconstruction procedure finds the best track hypothesis by maximizing the likelihood.

17.6 First Generation Neutrino Telescopes

The development of the field was pioneered by the project DUMAND (Deep Underwater Muon And Neutrino Detection Array) close to Hawaii [34]. First activities started in 1975. With the final goal of a cubic kilometre array, the envisaged first step was a configuration with 216 optical modules at 9 strings, 30 km offshore the Big Island of Hawaii, at a depth of 4.8 km. A test string with 7 optical modules (OMs) was deployed in 1987 from a ship, took data at different depths for several hours and measured the depth dependence of the muon flux [35]. A shore cable for a stationary array was laid in 1993 and a first string with 24 OMs deployed. It failed due to water leakages. Financial and technical difficulties led to the official termination of the project in 1996. Therefore the eventual breakthrough and proof of principle came from the other pioneering experiment located in Lake Baikal. See for the history of neutrino telescopes [36].

17.6.1 The Baikal Neutrino Telescope NT200

The Baikal Neutrino Telescope NT200 was installed in the Southern part of Lake Baikal [37]. The distance to shore is 3.6 km, the depth of the lake at this location is 1366 m, the depth of the detector about 1.1 km.

The BAIKAL collaboration was not only the first to deploy three strings (as necessary for full spatial reconstruction), but also reported the first atmospheric neutrinos detected underwater [38, 39] (see also Fig. 17.9, right).

NT200 was an array of 192 optical modules (OMs), completed in April 1998. It is sketched in Fig. 17.9, left. The OMs were attached to eight strings carried by an umbrella-like frame consisting of 7 arms each 21.5 m in length. The strings were anchored by weights at the bottom and held in a vertical position by buoys at various depths. The geometrical dimensions of the configuration were 72 m (height) and 43 m (diameter). Detectors in Lake Baikal are deployed (or hauled up for repairs) within 6–7 weeks in February/April, when the lake is covered with a thick ice layer providing an ideal, stable working platform. They are connected to shore by several cables which allow operation over the full year.

The time calibration of NT200 was done with several nitrogen lasers, one sending short light pulses via optical fibres of equal length to each individual OM pair (top

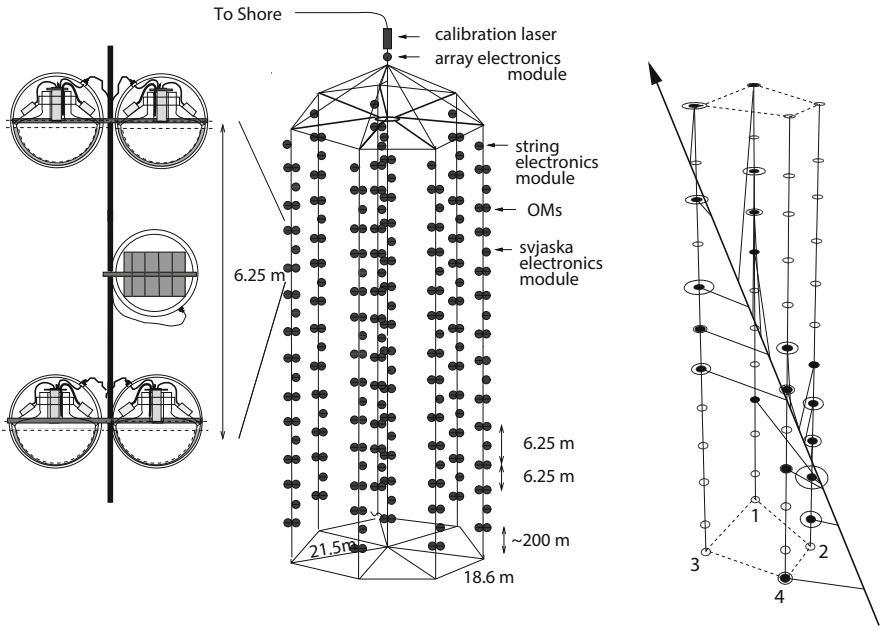

Fig. 17.9 Left: The Baikal Neutrino Telescope NT200. Right: One of the first upward moving muons from a neutrino interaction recorded with the 4-string stage of the Lake Baikal detector in 1996 [40]. The muon fires 19 channels

of Fig. 17.9), the other, the light pulses of the other laser below the array (not shown in the figure) propagate to the OMs through the water.

The OMs consisted of a pressure glass housing equipped with a QUASAR-370 phototube and were grouped pair-wise along a string. In order to suppress accidental hits from dark noise (∼30 kHz) and bio-luminescence (typically 50 kHz but seasonally raising up to hundreds of kHz), the two PMs of a pair were switched in coincidence, defining a *channel*, with only ∼0.25 kHz noise rate. The basic cell of NT200 consisted of a *svjaska* (Russian for "bundle"), comprising two OM pairs and an electronics module which was responsible for time and amplitude conversion and slow control functions (Fig. 17.9, left). A majority trigger was formed if $\geq m$ channels were fired within a time window of 500 ns (this is about twice the time a relativistic particle needed to cross the NT200 array), with m typically set to 4. Trigger and inter-string synchronization electronics were housed in an array electronics module at the top of the umbrella frame. This is less than 100 m away from the OMs, allowing for easy nanosecond synchronization over copper cable.

Figure 17.10 shows the phototube and the full OM [41]. The QUASAR-370 consisted of an electro-optical preamplifier followed by a conventional PM (type UGON). In this hybrid scheme, photoelectrons from a large hemispherical cathode (K_2CsSb) with $>2\pi$ viewing angle are accelerated by 25 kV to a fast, high gain scintillator which is placed near the centre of the glass bulb. The light from the scintillator is read out by the small conventional PM. One photoelectron emerging from the hemispherical photocathode yields typically 20 photoelectrons in the conventional PM. This high multiplication factor results in an excellent single electron resolution of 70%, a small time jitter (2 ns) and a small sensitivity to the Earth's magnetic field. The OM contains the QUASAR, the HV supply for the small PM (2 kV) and the large tube (25 kV) and a LED. The signal from the last dynode and the anode is read out via two penetrators, the two other penetrators pass the signal driving the calibration LED and the low voltages for the HV system and the preamplifiers. The optical contact between QUASAR bulb and glass housing is

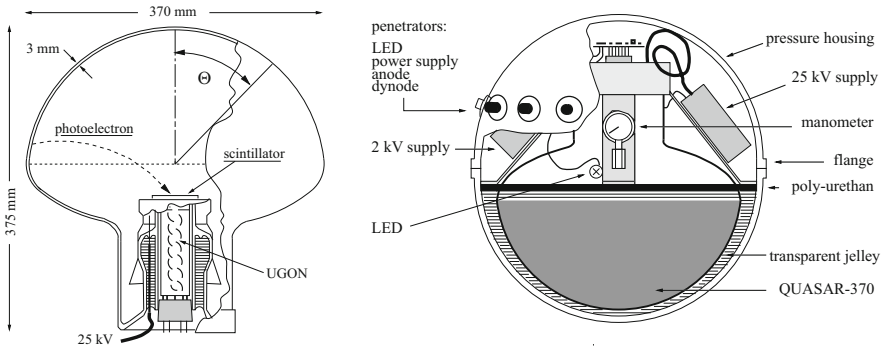

Fig. 17.10 Left: The QUASAR phototube. Right: a full Baikal optical module [41]

made by liquid glycerine sealed with a layer of polyurethane, in later versions with a silicon gel.

Due to the small lever arm, the angular resolution of NT200 for muon tracks was only 3–4°. On the other hand, the small spacing of modules led to a comparably low energy threshold for muon detection of ∼15 GeV. The total number of upward muon events collected over 5 years was only about 400, due not only to the small dimensions of the array, but also to its unstable operation. Still, NT200 could compete for some time with the much larger AMANDA, by searching for high energy cascades *below* NT200, surveying a volume about ten times as large as NT200 itself [42].

17.6.2 AMANDA

Rather than water, AMANDA (Antarctic Muon And Neutrino Detection Array) was using the 3 km thick ice layer at the South Pole as target and detection medium [43, 44]. AMANDA was (actually still *is*, although switched off) located some hundred meters away from the Amundsen–Scott station which provides the necessary infrastructure. Holes of 60 cm diameter were drilled with pressurized hot water, and strings with OMs were deployed in the column of molten water and frozen into the ice. South Pole installation operations are performed in the Antarctic summer, November to February, when temperatures rise to up to −25 °C. For the rest of the time, two operators (of a winter-over crew of 30–40 persons in total) maintain the detector, connected to the outside world via satellite communication.

Figure 17.11, left, shows the configuration of AMANDA. A first shallow test array with 80 OMs at 4 strings (not shown in the figure) was deployed in the Antarctic season 1993/1994, at depths between 800 and 1000 m [45]. It turned out that the effective scattering length L_{eff} was desperately small, 40 cm at 830 m depth, but increased with depth (80 cm at 970 m depth). The scattering was due to remnant bubbles and made track reconstruction impossible. The tendency of scattering decreasing with depth, as well as results from ice core analyses at other places in Antarctica, suggested that bubbles should disappear below 1300 m. This expectation was confirmed with a second 4-string array which was deployed in 1995/1996. The effect of bubbles disappeared, with the remaining scattering being mostly due to dust (see Fig. 17.11, right). The scattering length averaged over 1500–2000 m depth is $L_{eff} \approx 20$ m, still considerably worse than for water but sufficient for track reconstruction [30, 46]. The array was upgraded stepwise, completed in January 2000 and eventually comprised 19 strings with a total of 677 OM, most of them at depth between 1500 and 2000 m.

Figure 17.11, right, gives absorption and scattering coefficient as a function of depth and wavelength for glacial ice at the South Pole. The variations with depth are due to (a) bubble remnants at shallow depth leading to very strong scattering, (b) dust and other scattering and absorbing material transported in varying climate epochs to Antarctica. The depth dependence complicates the evaluation of the

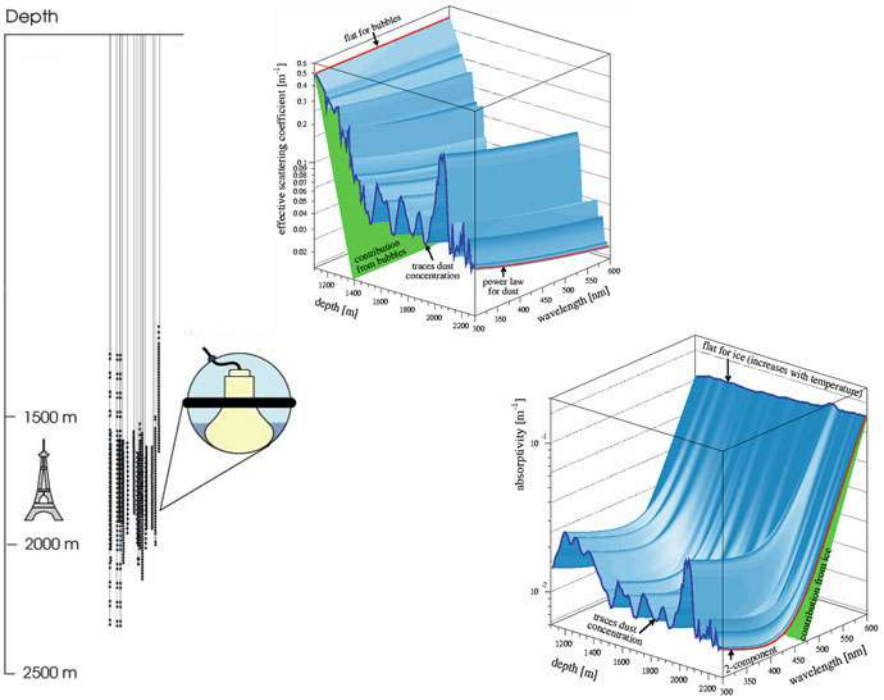

Fig. 17.11 Left: The AMANDA configuration. Three of the 19 strings have been sparsely equipped towards larger and smaller depth in order to explore ice properties, one string got stuck during deployment at too shallow depth and was not used in analyses. Right: scattering coefficient (top) and absorption coefficient (bottom) as a function of depth and wavelength

experimental data. Furthermore, the strong scattering leads to strong delays in photon propagation, resulting in worse angular resolution of deep ice detectors compared to water. On the other hand, the large absorption length, with a cut-off below 300 nm instead at 350–400 nm (water), results in better photon collection than in water. The quality of the ice improves substantially below a major dust layer at 2000–2100 m, with a value for the scattering length about twice as large as for the shallower region above 2000 m.

The short distance between OM and surface electronics allowed for a unique technical solution: the analogue PM anode signals were not digitized in the depth, but driven over 2 km cable to surface. This requires a large output signal of the PM, a specification met by the 8-inch R5912-2 from Hamamatsu with 14 dynodes and an internal amplification of 10^9. The first ten strings used coaxial (string 1–4) and twisted pair (string 6–10) cables for both HV supply and signal transmission, for the last 9 strings the anode signal was fed to an LED, and the light signal transmitted via optical fibre to surface. Naturally, the electrical signal transmission suffered from strong dispersion, widening the anode signal to several 100 ns. However, applying an amplitude correction to time flags from a constant fraction discriminator, a time

jitter of 5–7 ns was achieved. Given the strong smearing of photon arrival times due to light scattering in ice, this jitter appeared to be acceptable. For optical signals, dispersion was negligible. An event was defined by a majority trigger formed in the surface counting house, requesting $\geq$8 hits within a sliding window of 2 µs.

Time calibration of the AMANDA array was performed with a YAG laser at surface (wavelengths >450 nm), sending short pulses via optical fibres of well defined length to each OM. This laser system was also used to measure the delay of optical pulses propagating between strings and to determine the ice properties as well as the inter-string distances. A nitrogen laser (337 nm) at 1850 m depth, halogen lamps (350 and 380 nm) and LED beacons (450 nm) extended the information about ice properties across a large range of wavelengths (see Fig. 17.11, right). The measured time delays were fitted and the resulting parameterizations implemented in the probability density functions for the residual times t_{res}.

One big advantage compared to underwater detectors is the small PM noise rate, about 1 kHz in an 8-inch PM, compared to 20–40 kHz due to K^{40} decays and bioluminescence in lakes and oceans. The contribution of noise hits to the true hits from a particle interaction is therefore small and makes hit cleaning procedures much easier than in water.

The angular resolution of AMANDA for muon tracks was 2–2.5°, with an energy threshold of ≈50 GeV. Although better than for Lake Baikal (3–4°), it was much worse than for ANTARES (<0.5°, see below). This is the result of the strong light scattering which deteriorates the original information contained in the Cherenkov cone. The effect is even worse for cascades, where the angular resolution achieved with algorithms of that time was only ≈25 deg (compared to 5–8° in ANTARES).

In 2008, AMANDA had established a series of record upper limits, e.g. for diffuse extraterrestrial neutrino fluxes using muon as well as cascade searches, for the flux of relativistic magnetic monopoles or for neutrinos from point sources (see for a review [47]). The final AMANDA point source analysis was based on 6595 neutrinos collected in the years 2000–2006 [48]. AMANDA was switched off in 2009.

17.6.3 Mediterranean Projects: ANTARES

Mediterranean efforts to build an underwater neutrino telescope are related to three locations:

(a) a site close to Pylos at the Peloponnesus, with available depths ranging from 3.5 to 5 km for distances to shore of 30–50 km,
(b) a site close to Capo Passero, Sicily, at a depth of 3.5 km and 70 km, distance to shore,
(c) a site close to Toulon, at a depth of 2.5 km and 40 km distance to shore.

All of these sites are considered locations for a future distributed infrastructure of a total volume of few cubic kilometres. All sites have physics and infrastructural

pros and cons. For instance, large depth is a challenge for long-term ocean technology and bears corresponding risks, but has convincing physics advantages: less background from punch-through muons from above, less bio-luminescence, less sedimentation.

Historically, the first Mediterranean project was NESTOR [49] off the Greek coast. It was conceived as a tower-like structure with 12 floors, 300 m in height and 32 m in diameter. A prototype hexagonal floor with 14 PMs (15-inch Hamamatsu) was deployed in 2004 and took data for a few weeks. The project is terminated meanwhile. NEMO, close to Sicily [50], focused on technology development and feasibility studies for a cubic kilometer array. The basic unit of NEMO have been towers composed by a sequence of floors. The floors consist of rigid horizontal structures, 15 m long, each equipped with four 10-inch PMs. The floors are tilted against each other and form a three-dimensional structure.

In the following, ANTARES [51, 52] is described, being the one of the three projects which made it to a full telescope of AMANDA-size.

Figure 17.12 shows a schematic view of the detector. It consists of 12 strings, each anchored at the seabed and kept vertical by a top buoy. The minimum distance

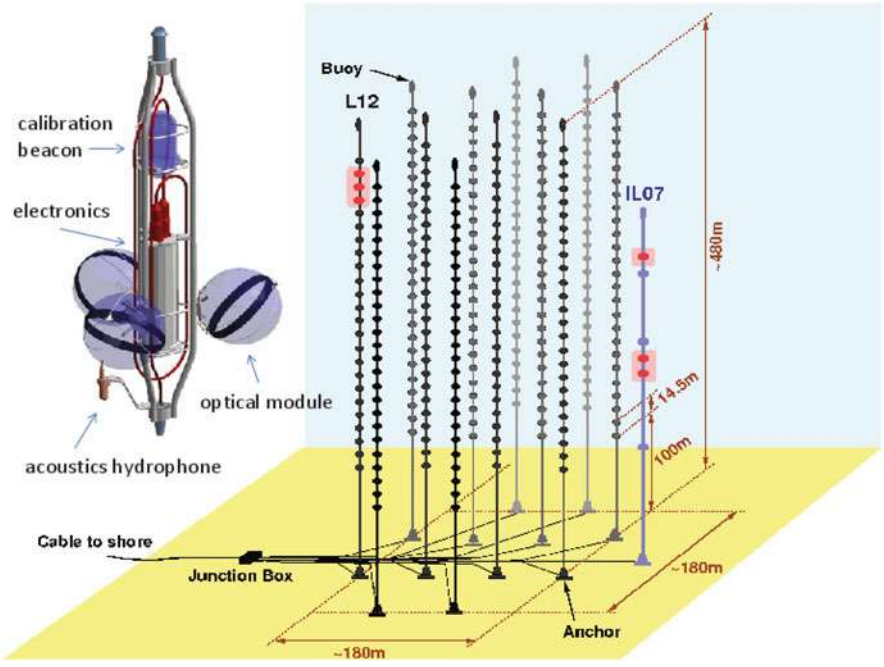

Fig. 17.12 Right: Schematic view of the ANTARES detector, with 12 detector lines L1–L12, and an extra-line with environmental equipment (IL07). L12 and ILO7 carry test equipment for basic tests toward acoustic neutrino detection. Left: A storey with three optical modules and the metallic cylinder housing the Local Control Module (LCM). Every fifth storey carries a LED beacon (above the LCM) and a hydrophone (bottom left) for acoustic triangulation[52]

between the strings is 60 m. Each string is composed of 25 storeys. A storey is equipped with three 10-inch PMs Hamamatsu R7091-20 housed in 13-inch glass spheres. The PMs are oriented at 45° with respect to the vertical. A mu-metal grid reduces the influence of the Earth magnetic field. The storeys are spaced by 14.5 m, the lowest being located about 100 m above seabed. The storeys are connected by an electro-optical cable, including 21 optical fibres for digital communications [53, 54].

From a functional point of view, each string is divided into five sectors, each containing five storeys. A storey is controlled by a Local Control Module (LCM) which maintains the data communication between its sector and the shore. A String Controller Module (SCM), located at the basis of each string, interfaces the string to the rest of the detector. The string cables are led to a junction box to which the shore cable is connected.

The signals from the PMTs are digitized by an Analogue Ring Sampler (ARS). The ARS produces "hits" by time-stamping the PMT signal and by integrating the PMT anode current over a programmable time interval (25–80 ns). The time stamp is provided by the local clock of the LCM. The master clock signal of 20 MHz is generated at shore and distributed through optical fibres to the LCM clocks. Sub-nanosecond precision is achieved by a time-to-voltage converter which allows interpolation between two subsequent clock pulses. The output voltage is digitized with an 8-bit ADC. The maximally achievable time resolution is therefore $1/(20\,\text{MHz} \times 256) \sim 0.2\,\text{ns}$.

The timing calibration is performed with calibration pulses between shore clock and LMC clocks, and with LED beacons which fire both the ARS (electrically) and the PMT (optically) and correct for the varying PMT transit time. The position calibration is particularly import since the string positions change due to water current. It is performed with compasses and tiltmeters along the strings, and with a acoustic triangulation system based on transmitters at the bottom of the strings and hydrophones along the strings. The relative positions of the OMs can be determined with an accuracy of a few centimetres.

The Monte Carlo angular resolution for muons is 0.2° at 10 TeV. At low energies the neutrino tracing is limited by the angle between muon and neutrino, 0.7° at 1 TeV and 1.8° at 100 GeV (median mismatch angle for those muons triggering the detector [55]).

Naturally, the angular resolution for cascades is worse than for muons. Simple reconstruction algorithms give 10° median mismatch angle above 5 TeV, however, with proper quality cuts, values below 4° can be achieved, with 20–40% passing rates for signals [33].

ANTARES is operated in its full configuration since 2008 and is planned to continue data taking until the follow-up project KM3NeT has surpassed ANTARES w.r.t. to its sensitivity, i.e. at least through 2018.

17.7 Second Generation Neutrino Telescopes

17.7.1 IceCube

IceCube [56] is the successor of AMANDA. It consists of 5160 digital optical modules (DOMs) installed on 86 strings at depths between 1450 and 2450 m in the Antarctic ice [57], and 320 DOMs installed in IceTop [58], detectors in pairs on the ice surface directly above the strings (see Fig. 17.13). AMANDA was integrated into IceCube as a low-energy sub-detector, but later was replaced by DeepCore, a high density, six-string sub-array at large depths (i.e. in best ice) at the centre of IceCube. The energy threshold is about 100 GeV for the full IceCube array and about 10 GeV for DeepCore.

The thermal power of hot-water drill factory is increased to 5 MW, compared to 2 MW for AMANDA, reducing the average time to drill a 60 cm hole to 2450 m depth down to ≈35 h. The subsequent installation of a string with 60 DOMs requires

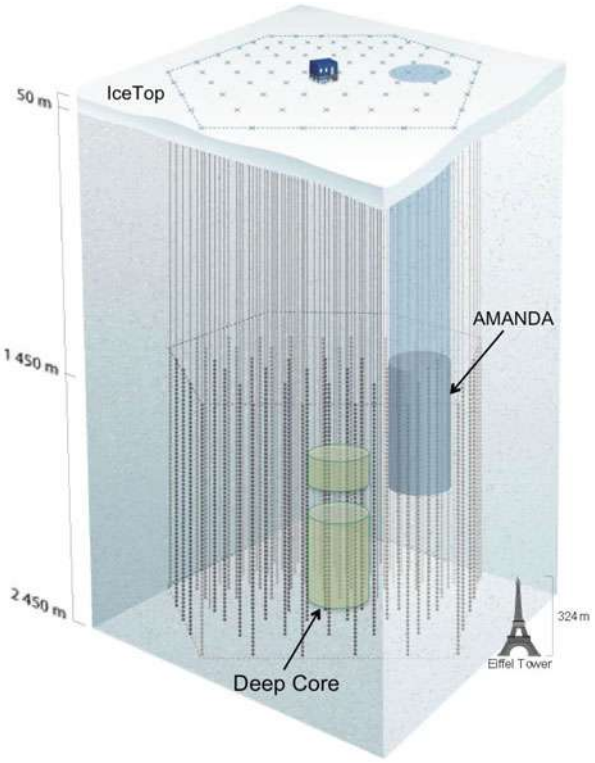

Fig. 17.13 Schematic view of the IceCube Neutrino Observatory. Since 2009, AMANDA is replaced by DeepCore, a nested low-threshold array. At the surface are the air shower detector IceTop and the IceCube counting house

typically 12 h. A record number of 20 strings was deployed in the season 2009/2010. The detector was completed in December 2010.

The components are not accessible after refreezing of the holes. Therefore—as for AMANDA—the architecture has to avoid single-point failures in the ice. A string carries 60 DOMs, with 30 twisted copper pair cables providing power and communication. Two sensors are operated on the same wire pair. Neighbouring DOMs are connected to enable fast local coincidence triggering in the ice [57].

A schematic view of a DOM is shown in Fig. 17.14. A 10-inch PMT Hamamatsu R7081-02 is embedded in a 13-inch glass pressure sphere [59]. A mu-metal grid reduces the influence of the Earth's magnetic field. The programmable high voltage is generated inside the DOM. The average PMT gain is set to 10^7. Signals are digitized by a fast analogue transient waveform recorder (ATWD, 3.3 ns sampling) and by a FADC (25 ns sampling). The PM signal is amplified by 3 different gains to extend the dynamic range of the ATWD to 16 bits, resulting in a linear dynamic range of 400 photoelectrons in 15 ns; the dynamic range integrated over 2 μs is about 5000 photoelectrons.

The digital electronic on the main board are based on a field-programmable gate array (FPGA). It communicate with the surface electronics, new programs can be downloaded. The LEDs on the flasher board emit calibration pulses at 405 nm which can be adjusted over a wide range up to $\sim 10^{11}$ photons.

All digitized PM pulses are sent to the surface. The full waveform, however, is only sent for pulses from local (neighbour or next to neighbour) coincidences in order to apply data compression for isolated hits which are mostly noise pulses. All DOMs have precise quartz oscillators providing local clock signals, which are synchronized every few seconds to a central GPS clock. The time resolution is about

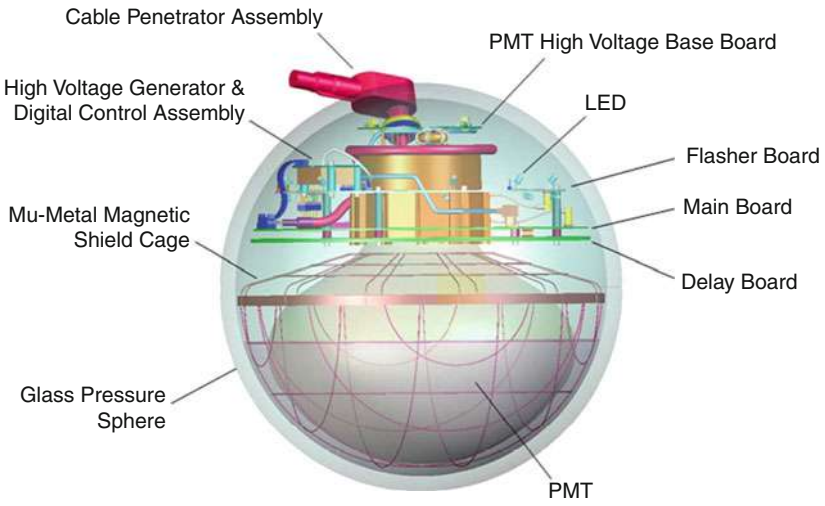

Fig. 17.14 Schematic view of an IceCube digital optical module

2 ns. The noise rate for DOMs in the deep ice is ∼540 Hz, if a deadtime of 250 μs is applied only ∼280 Hz. The very low noise rates are critical for the detection of the low-energy neutrino emission associated with a supernova collapse (see below).

At the surface, 8 custom PCI cards per string provide power, communication and time calibration. Subsequent processors sort and buffer hits until the array trigger and event builder process is completed [61]. The architecture allows deadtime free operation. The design raw data rate of the full array is of the order of 100 GB/day which are written to tapes. Online processing in a computer farm allows extraction of interesting event classes, like all upgoing muon candidates, high-energy events, IceTop/IceCube coincidences, cascade events, events from the direction of the muon or events in coincidence with Gamma Ray Bursts (GRB). The filtered data stream (∼20 GB/day) is then transmitted via satellite to the Northern hemisphere.

The muon angular resolution is about 1° for 1 TeV tracks and below <0.5° for energies of 10 TeV and higher. The very good ice below 2100 m has a particular potential for improved resolution. This will be even more important for the angular reconstruction of cascades. The presently achieved angular resolution for cascades is only 25°, much worse than for water, with the inferiority being mainly due to light scattering in ice.

IceCube is the only detector which can be permanently operated together with a surface air shower array, IceTop [58]. It consists of tanks filled with ice, each instrumented with 2 DOMs. The comparison of air shower directions measured with IceTop and directions of muons from these showers in IceCube allows an angular calibration of IceCube (absolute pointing and angular resolution). IceTop can measure the spectrum air showers up to primary particle energies of $\sim 10^{18}$ eV. Combination of IceTop information (reflecting dominantly the electron component of the air shower) and IceCube information (muons from the hadronic component) allows estimating the mass range of the primary particle.

Last but not least, IceCube allows for another mode of operation which is essentially only possible in ice: the detection of burst neutrinos from a supernova [60]. The low dark counting rate of PMs (∼280 Hz, see above) allows detecting of the feeble increase of the summed count rates of all PMs during several seconds, which would be produced by millions of MeV neutrino interactions from a supernova burst. IceCube records the counting rate of all PMs in millisecond steps. A supernova in the centre of the Galaxy would be detected with extremely high confidence and the onset of the pulse could be measured in unprecedented detail. Even a 1987A-type supernova in the Large Magellanic Cloud would result in a 5σ effect and be sufficient to provide a trigger to the SuperNova Early Warning System, SNEWS [62].

The following figures show displays of some events recorded with IceCube. Figure 17.15, left, is a typical muon track crossing the detector from below. The event on the right side is a cascade event, actually the fully contained cascade event with the highest energy recorded, about 2 PeV. The analysis employed containment conditions and an atmospheric muon veto for suppression of down-going atmospheric neutrinos ("High-Energy Starting Event" analysis, HESE). The HESE events cannot be explained by atmospheric neutrinos and misidentified

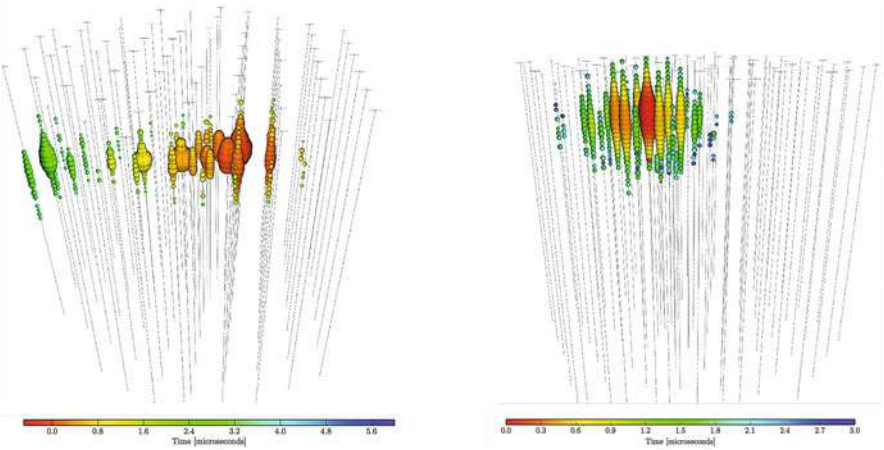

Fig. 17.15 Left: A through-going upward muon track. Right: The highest-energy cascade event detected (status 2018) with IceCube, with ≈2 PeV energy released in the detector [68]. The size of the symbols reflect the recorded amount of light, the color indicates the signal timing (red: early; green: late), see the scale at the bottom

atmospheric muons alone: with 6 years of data, the excess has a significance of $> 7\sigma$, i.e. a flux of extraterrestrial neutrinos could be safely confirmed.

Also, events with through-going muons show a corresponding excess of cosmic origin [69]—see the display of the highest-energy track-like neutrino event in Fig. 17.16.

In its final configuration, IceCube takes data since spring 2011, with a duty cycle of more than 99%. It collects almost 10^5 clean neutrino events per year, with nearly 99.9% of them being of atmospheric origin. The failure rate of DOMs is only about one per year, out of more than 5000.

17.7.2 KM3NeT

KM3NeT has two main, independent objectives: (a) the discovery and subsequent observation of high-energy cosmic neutrino sources and (b) precise oscillation measurements and the determination of the mass hierarchy of neutrinos [63, 64]. For these purposes the KM3NeT Collaboration plans to build an infrastructure distributed over three sites: off-shore Toulon (France), Capo Passero (Sicily, Italy) and Pylos (Peloponnese, Greece). In a configuration to be realized until 2021/2022, KM3NeT will consist of three so-called building blocks ("KM3NeT Phase-2").

A building block comprises 115 strings, each string with 18 optical modules.Two building blocks will be sparsely configured to fully explore the IceCube signal with a comparable instrumented volume, different methodology, improved resolution and complementary field of view, including the Galactic plane. These two blocks will

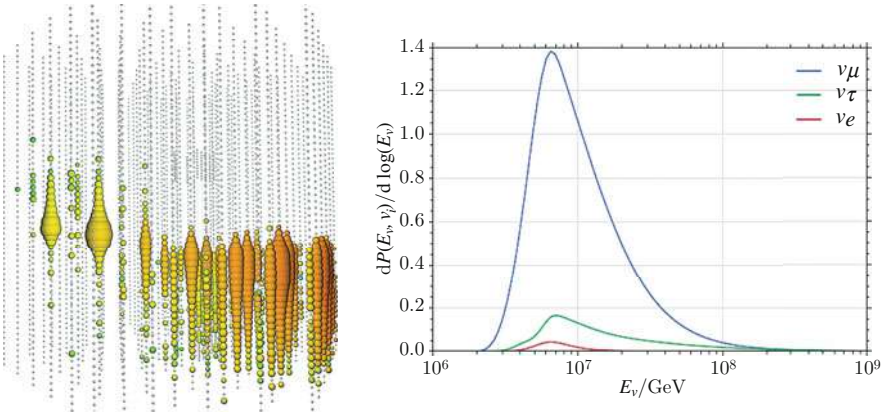

Fig. 17.16 Left: Event view of the PeV track-like event recorded by IceCube on June 11, 2014. Color code like in the previous figure. Note that the scaling is non-linear and a doubling in sphere size corresponds to one hundred times the measured charge. This event deposited an energy of 2.6 ± 0.3 PeV in the detector volume. Right: Probability distribution of primary neutrino energies that could result in the observed multi-PeV track-like event, assuming an E_ν^{-2} spectrum. The total probabilities for the different flavors are 87.7, 10.9 and 1.4% for ν_μ, ν_e and ν_τ, respectively. The most probable energy of the primary neutrino is between 8 and 9 PeV

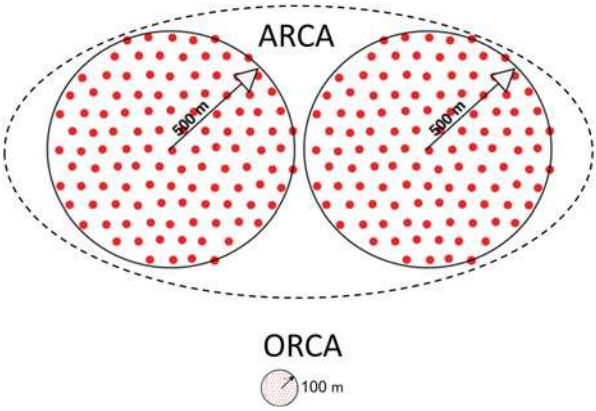

Fig. 17.17 The two incarnations of KM3NeT. The two ARCA blocks (top) have diameters of 1 km and a height of about 600 m and focus to high-energy neutrino astronomy. ORCA (bottom) is a shrinked version of ARCA with only 200 m diameter and 100 m height. Both ARCA and ORCA have 115 strings with 18 optical modules (OMs) per string

be deployed at the Capo Passero site and are referred to as ARCA: Astroparticle Research with Cosmics in the Abyss. The third building block will be densely configured to precisely measure atmospheric neutrino oscillations. This block, being deployed at the Toulon site, is referred to as ORCA: Oscillation Research with Cosmics in the Abyss (see Fig. 17.17).

Fig. 17.18 View and a cross-sectional drawing of a KM3NeT-DOM with its 31 small PMs inside [64]

A novel concept has been chosen for the KM3NeT optical module: The 43 cm glass spheres of the DOMs will be equipped with 31 PMs of 7.5 cm diameter, with the following advantages: (a) The overall photocathode area exceeds that of a 25 cm PM by more than a factor three; (b) The individual readout of the PMs results in a very good separation between one- and two-photoelectron signals which is essential for online data filtering; (c) some directional information is provided. This technical design has been validated with in situ prototypes. A view and a cross-sectional drawing of the DOM are shown at the top of Fig. 17.18.

Rather than digitizing the full waveform (like for the one large PM per DOM in IceCube), for each of the analogue pulses from 31 small PMs which pass a preset threshold, the time of the leading edge and the time over threshold are digitized (referred to as a *hit*). Each hit corresponds to 6 Bytes of data (1 B for PM address, 4 B for time and 1 B for time over threshold, with the least significant bit of the time information corresponding to 1 ns). All hits are sent to shore (all-data-to-shore concept). The total rate for a single building block with its 64,170 PMTs amounts to about 25 Gb/s which are sent via optical fibers to shore. To limit the number of fibres, wavelength multiplexing is used.

At shore, the physics events are filtered from the background. To maintain all available information for the offline analysis, each event contains a snapshot of all the data during that event. The filtered data (with a rate reduced by a factor of about 10^5 with respect to the data arriving at shore), are stored at disks.

KM3NeT-ARCA is conceived as the European counterpart to IceCube and will preferentially observe the Southern instead of the Northern hemisphere, including the Galactic Centre [63]. With a fully equipped ARCA, IceCube's cosmic neutrino flux could be detected with high-significance within 1 year of operation. In practise the detector will be deployed in stages allowing to reach the 1 year sensitivity of two clusters much before the second cluster is fully installed.

ORCA will continue along the venue opened by IceCube-DeepCore and perform precision measurements of neutrino oscillations. In particular, it could determine the neutrino mass hierarchy with at least 3σ significance after 3 years of operation.

17.7.3 GVD

Based on the long-term experience with the NT200 detector and on extended prototype tests, the Baikal Collaboration has started the stepwise installation of a kilometer-scale array in Lake Baikal, the Giant Volume Detector, GVD [65, 66].

The optical modules of Baikal-GVD are equipped with 10-inch PMs of the type Hamamatsu R7081-100, with a quantum efficiency of ≈35%. The OMs are mounted on vertical strings, fixed to the bottom with anchors. Each eight strings form a cluster, with 36 OMs per string, i.e. 288 OMs per cluster. Each cluster is a full functional detector which is capable of detecting a physical event both in standalone mode and as part of the full-scale array. The first phase of GVD (GVD-1) is planned to be be completed by 2021, with eight clusters carrying $\approx 2.3 \times 10^3$ OMs in total and a volume of 0.3–0.4 km^3. Figure 17.19 shows a schematic view of GVD-1. In a second phase, Baikal-GVD is conceived to be extended to an array of about 10^4 OMs with an instrumented volume of 1–2 km^3.

The OMs are vertically spaced by 15 m, with the lowest OM at a depth of 1275 m (about 100 m above the bottom of the lake) and the top OM at 750 m below the lake surface. The seven strings of a cluster are arranged at a radius of 60 m around a central string. The distances between the centers of the clusters are 300 m.

A string is composed of three *sections*, each comprising 12 OMs with analog outputs. A *Central section Module (CM)* converts the analog signals into a digital code, using a 12-bit ADC with a sampling frequency of 200 MHz. Coincidences of signals from any pairs of neighbouring OMs are used as a local trigger of the section (signal request), with average frequencies of the section request signals in the range

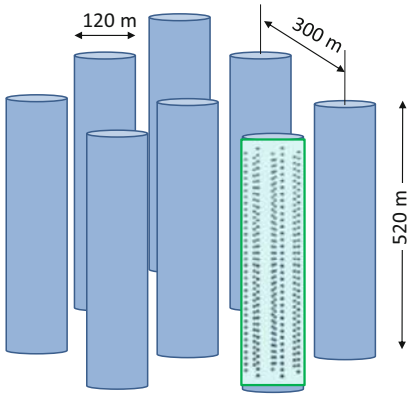

Fig. 17.19 Schematic view of phase-1 of Baikal-GVD, consisting of 8 clusters, each with 120 m diameter and 520 m height. A cluster consists of eight strings with 36 optical modules along each string

of 2–10 Hz, dependent on signal thresholds and level of water luminescence. The request signals from three sections are combined in the *Control Module* of the string (CoM) and transferred to the cluster Cluster DAQ center where a global trigger is formed. The Cluster DAQ center is arranged close to the water surface at a depth of 25 m and connected to the Shore DAQ Center by hybrid electro-optical cable.

Calibration is performed by LEDs and lasers. LEDs installed in each OM provide amplitude and time calibration of the OMs, separate underwater modules equipped with LEDs are used for time calibration between sections. A high-power laser is used arranged between clusters ensures calibration of the cluster as a whole and calibration between neighboured clusters. The coordinates of the optical modules are determined using an acoustic positioning. Each cluster has its own acoustic positioning system, with four acoustic modems per string, the lowest at the bottom of the string, the highest 538 m higher. The transit time between acoustic sources at the lake bed and the acoustic modems gives the coordinates of the acoustic modems with an accuracy of $\approx$2 cm.

17.7.4 IceCube-Gen2

The progress from IceCube will be limited by the modest numbers of cosmic neutrinos measured, even in a cubic kilometer array. In [67] a vision for the next-generation IceCube neutrino observatory is presented. At its heart is an expanded array of optical modules with a volume of 7–10 km^3. This high-energy array will mainly address the 100 TeV to 100 PeV scale. For point sources, it will have five times better sensitivity than IceCube, and the rate for events at energies above a few hundred TeV will be ten times higher than for IceCube. It has the potential to deliver first GZK neutrinos, of anti-electron neutrinos produced via the Glashow resonance, and of PeV tau neutrinos, where both particle showers associated with the production and decay of the tau are observed ("double bang events").

Another possible component of IceCube-Gen2 is the PINGU sub-array. It targets—similar to ORCA—precision measurements of the atmospheric oscillation parameters and the determination of the neutrino mass hierarchy. The facility's reach would further be enhanced by exploiting the air-shower measurement and vetoing capabilities of an extended surface array. Moreover, a radio array ("ARA", for Askarian Radio Array, see below) will achieve improved sensitivity to neutrinos in the $10^{16}-10^{20}$ eV energy range, including GZK neutrinos.

17.8 Physics Results: A 2018 Snapshot

The 2018 status of the field is dominantly defined by the IceCube results. ANTARES significantly contributes to searches for neutrinos from the Southern hemisphere and the central parts of the Galaxy. These are the main results obtained over the last 5 years:

- Both IceCube and ANTARES have measured the flux of "conventional" atmospheric neutrinos from π and K decay up to a few hundred TeV and found it in agreement with predictions [70, 71]. Tight upper limits have been set for the flux of "prompt" atmospheric neutrinos from charm and bottom decays.
- At energies below 50 GeV, the oscillation of atmospheric neutrinos passing through the Earth has been observed both by IceCube and ANTARES. The IceCube constraints on the neutrino mixing parameters are meanwhile as tight as those derived from accelerator experiments [72].
- In 2013, IceCube has detected a diffuse flux of astrophysical neutrinos with a very high confidence (meanwhile larger than 7σ). This observation can be considered a real breakthrough, 53 years after the first ideas on underwater neutrino detectors have been proposed [73].
- ANTARES and IceCube have jointly analysed their data to identify a neutrino excess from the Galactic Plane and can exclude that more than 8.5% of the observed diffuse astrophysical flux comes from the Galactic plane [74].
- No steady neutrino point sources could be identified, neither using 8 years of IceCube data, with 497,000 upward muons from neutrino interactions, nor with ANTARES data. The derived limits on point source fluxes are a fantastic factor 3000 below those obtained in 2000 with AMANDA data [75].
- Also, various analyses where many sources belonging to a certain source class are "stacked" did not yield significant excesses. For instance, latest IceCube results exclude that more 6% of the observed diffuse astrophysical muon neutrino flux could come from blazars (active galaxies with their jet pointing to the Earth) [76]. Blazars have been considered since long as top-candidate neutrino sources. The same applies to neutrinos from Gamma Ray Bursts (GRB). IceCube could exclude at more than 90% confidence those models which assume that GRBs are the dominant source of the measured cosmic-ray flux at highest energies [77].
- An alert issued by IceCube on September 22, 2017, led to the first coincident observation of a high-energy energy neutrino with X-ray, gamma-ray and optical information. These electromagnetic follow-up observations identified a blazar named TXS 0506+056 in its active state as the likely source of the neutrino. IceCube examined its archival data in the direction of TXS 0606+056 and found an additional 3.5σ evidence for a flare of 13 neutrinos starting at the end of 2014 and lasting about 4 months. This is considered the first compelling evidence for flaring source of neutrinos [78, 79].
- No neutrinos from cosmic-ray interactions with the 3K-microwave background radiation could yet be identified. Their observation will need multi-km^3 detectors like IceCube-Gen2 or even radio detectors as discussed in the next Sect. [80].

- Record limits have been derived for neutrino fluxes from dark matter annihilations in the Earth, the Sun or the Galactic halo and for the flux of magnetic monopoles (which, if at relativistic velocity, could be identified via their high light emission) and to the coupling of hypothetical sterile neutrinos to normal neutrino states (see [81] for a review of results on particle physics with IceCube).

17.9 Technologies for Extremely High Energies

The technologies described in this section are tailored to signals which propagate with km-scale attenuation. Consequently, they allow for the observation of much larger volumes than those typical for optical neutrino telescopes. 100 km^3 scale detectors are necessary, for instance, to record more than just a few GZK neutrinos, with a typical energy range of 100 PeV to 10 EeV.

17.9.1 Detection via Air Showers

At energies above 10^{17} eV, large extensive air shower arrays like the Pierre Auger detector in Argentina [82] or the Telescope Array in Utah/USA [83] are seeking for horizontal air showers due to neutrino interactions deep in the atmosphere (showers induced by charged cosmic rays start on top of the atmosphere). Figure 17.20 explains the principle. AUGER consists of an array of water tanks spanning an area

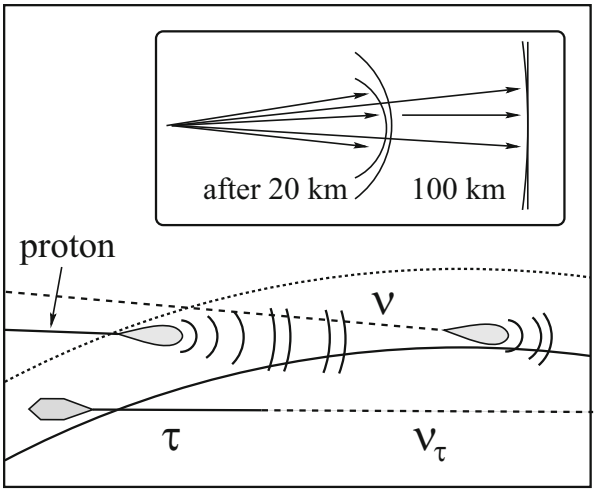

Fig. 17.20 Detection of particles or fluorescence light emitted by horizontal or upward directed air showers from neutrino interactions

of 3000 km^2 and recording the Cherenkov light of air-shower particles crossing the tanks. It is combined with telescopes looking for the atmospheric fluorescence light from air showers (see chapter on cosmic ray detectors). The optimum sensitivity window for this method is at 1–100 EeV, the effective detector mass is up to 20 Gigatons. An even better sensitivity might be obtained for tau neutrinos, ν_τ, scratching the Earth and interacting close to the array [84, 85]. The charged τ lepton produced in the interaction can escape the rock around the array, in contrast to electrons, and in contrast to muons it decays after a short path into hadrons. If this decay happens above the array or in the field of view of the fluorescence telescopes, the decay cascade can be recorded. Provided the experimental pattern allows clear identification, the acceptance for this kind of signals can be large. For the optimal energy scale of EeV, the present differential single-flavor limit (2017) is about $2 \times 10^{-8} E_\nu^{-2}$ GeV^{-1} cm^{-2} s^{-1} sr^{-1} [86].

A variation of this idea is to search for tau lepton cascades which are produced by horizontal PeV neutrinos hitting a mountain and then decay in a valley between target mountain and an "observer" mountain [87].

17.9.2 Radio Detection

Electromagnetic cascades generated by high energy neutrino interactions in ice or salt emit coherent Cherenkov radiation at radio frequencies. The effect was predicted in 1962 [88] and confirmed by measurements at accelerators [89, 90]. Electrons are swept into the developing shower, which acquires an electric net charge from the added shell electrons. This charge propagates like a relativistic pancake of 1 cm thickness and 10 cm diameter. Each particle emits Cherenkov radiation, with the total signal being the convolution of the overlapping Cherenkov cones. For wavelengths larger than the cascade diameter, coherence is observed and the signal rises proportional to E_ν^2, making the method attractive for high energy cascades. The bipolar radio pulse has a width of 1–2 ns. In ice, attenuation lengths of up to a kilometer are observed, depending on the frequency band and the ice temperature. Thus, for energies above a few ten PeV, radio detection becomes competitive or superior to optical detection (with its attenuation length of $\sim$100 m) [91].

A prototype Cherenkov radio detector called RICE was operated at the South Pole, with 20 receivers and emitters buried at depths between 120 and 300 m. From the non-observation of very large pulses, limits on the diffuse flux of neutrinos with E$>$100 PeV and on the flux of relativistic magnetic monopoles have been derived [92].

Three groups are working towards detectors with 100–300 km^3 active volume: The Askarian Radio Array (ARA [93]) at the South Pole, the Antarctic Ross Iceshelf Antenna Neutrino Array (ARIANNA [94]) on the Antarctic Ross ice shelf—both in the phase of tests with engineering arrays—and the Greenland Neutrino Observatory

(GNO [96]) which is conceived to be deployed near the USA Summit Station in Greenland.

The current ARA proposal [93] envisages and array of 37 stations, each consisting of 16 antennas, buried up to 200 m depth below the firn ice. The stations are spaced by 2 km. As of 2018, five of them are deployed. ARIANNA [94] will observe the 570 m thick ice covering the Ross Sea. The smooth ice-seawater interface reflects radio waves; therefore ARIANNA might have a better sensitivity for downward moving and horizontal neutrinos. However, the ice is warmer than at the South Pole, reducing the attenuation length for GHz radio waves from 800–900 m (South Pole) to about 400 m (ice shelf). ARIANNA antennas face downward and are arranged just below the ice surface, with about thousand antennas for the ultimate array, spread over an area of $\approx$1000 km^2. One can reasonably expect that only one of these two projects can be funded in its full size.

ANITA (Antarctic Impulsive Transient Array [95] is an array of radio antennas which has been flown at a balloon on an Antarctic circumpolar path in 2006 and 2008/2009 (see Fig. 17.21).

From 35 km altitude it searches for radio pulses from neutrino interactions in the thick ice cover and monitored, with a threshold in the EeV range and a volume of the order of 10^6 Gigatons. This corresponds to a much larger volume than that of ARA and ARIANNA and can be achieved only for the price of an energy threshold about two orders of magnitude above that of ARA and ARIANNA. With its dual-polarization horn antennas it scanned the ice out to 650 km away. Neutrino signals would be vertically polarized, while background signals from down-going cosmic-ray induced air showers are preferentially horizontally polarized. Signals pointing to known or suspected areas of human activity are rejected. The ANITA 90% C.L. integral flux limit on a pure E^{-2} spectrum, integrating over $10^{18}-10^{23.5}$ eV, is $E^2 \times 1.3 \cdot 10^{-7}$ GeV cm^{-2} s^{-1} sr^{-1}, presently (2018) the most stringent limit on the GZK neutrino flux.

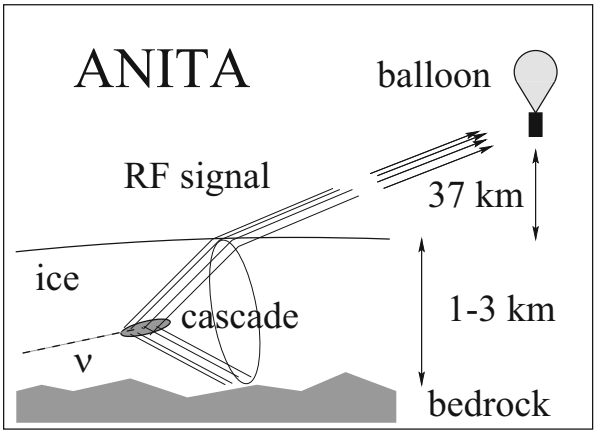

Fig. 17.21 Principle of the ANITA balloon experiment

Even higher energies are are addressed when searching for radio emission from particle cascades induced by neutrinos or cosmic rays skimming the moon surface. An example is the GLUE project (Goldstone Ultra-high Energy Neutrino Experiment [97]) which used two NASA antennas and reached maximum sensitivity at several ZeV (1 ZeV = 1000 EeV). With the same method, the NUMOON experiment at the Westerbork Radio Telescope searched for extremely energetic neutrinos [98], and the LUNASKA experiment which uses the Parkes and ATCA radio telescopes [99]. LUNASKA stands for "Lunar Ultra-high Neutrino Astrophysics with the SKA", indicating the final purpose: to use the Square Kilometer Array SKA to perfrom a lunar neutrino search.

17.9.3 Acoustic Detection

Production of pressure waves by fast particles passing through liquids was predicted in 1957 [100] and experimentally proven with high intensity proton beams two decades later [101]. A high energy cascade deposits energy into the medium via ionization losses which is immediately converted into heat. The effect is a fast expansion, generating a bipolar acoustic pulse with a width of a few 10 µs in water or ice (Fig. 17.22). Transversely to the pencil-like cascade, the radiation propagates within a disk of about 10 m thickness (the length of the cascade) into the medium. The signal power peaks at 20 kHz where the attenuation length of sea water is a few kilometres, compared to 50 m for light. The threshold of this method is however very high, in the several-EeV range. Acoustic detection was also considered an option for ice, where the signal itself is higher and ambient noise is lower than in water. A test array, SPATS (South Pole Acoustic Test Setup), has been deployed at the South Pole in order to determine attenuation length and ambient noise [102]. Another test configuration has been deployed together with the ANTARES detector (see Fig. 17.22). Tests are also performed close to Sicily, close to Scotland and in Lake Baikal. Another project has been using a very large hydrophone array of the US Navy, close to the Bahamas [109]. The existing array of hydrophones spans an area of 250 km and has good sensitivity at 1–500 kHz and can trigger on events above 100 EeV with a tolerable false rate.

17.9.4 Hybrid Arrays

Best signal identification would be obtained by combining signatures from two of the three methods, optical, radio and acoustic [110]. Naturally, radio detection does not work in water. The threshold for acoustic detection is so high that coincidences from a 100 km^3 acoustic array and a 1 km^3 optical array would be rare and a true hybrid approach not promising. The hybrid principle may be applicable at the South Pole, since the overlap between optical and radio methods (threshold for radio ∼10–

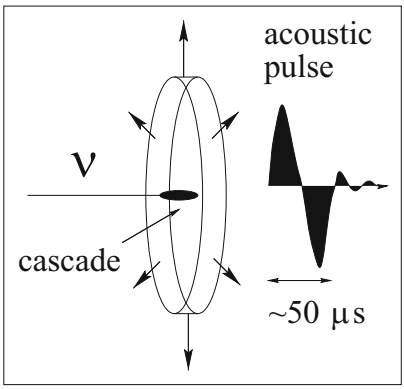

Fig. 17.22 Acoustic emission of a particle cascade

100 PeV) is significant. A nested hybrid array with optical-radio coincidences is therefore be conceivable and is actually part of the IceCube-Gen2 proposal (see previous section).

Overviews on acoustic and radio detection can be found in the proceedings of the workshops on "Acoustic and Radio EeV Neutrino Detection Activities" (ARENA) [103–108].

References

1. T.K. Gaisser, F. Halzen and T. Stanev, Phys. Rep. 258 (1995) 173.
2. J.G. Learned and K. Mannheim, Ann. Rev. Nucl. Part. Sci. 50 (2000) 679.
3. U.F. Katz and C. Spiering, Prog. in Part. Nucl. Phys. 67 (2012) 651 and arXiv:1111.0507.
4. R. Engel, T. Gaisser and E. Resconi, *Cosmic Rays and Particle Physics*, Cambridge University Press 2016.
5. T. Gaisser and A. Karle (eds.),*Neutrino Astronomy*, World Scientific 2017.
6. V.S. Berezinsky, G.T. Zatsepin, Phys. Lett. B28(1969) 423.
7. K. Greisen, Phys. Rev. Lett. 16 (1966) 748; G.T. Zatsepin and A.A. Kuzmin, J. Exp. Theor. Phys. Lett. 4 (1966) 78.
8. M. Aartsen et al. (IceCube Coll.), Science 342 (2013) 2342856.
9. A.B. McDonald et al., Rev. Sci. Instrum. 75 (2004) 293, and arXiv:0311343.
10. M.A. Markov, Proc. ICHEP, Rochester (1960) 578.
11. U. Katz and C. Spiering in C. Patrignani et al. (Particle Data Group), Chin. Phys. C. 40 (2016) 10001.
12. L. Anchordoqui, F. Halzen, Annals Phys. 321 (2006) 2660.
13. M. Ahlert, C. de los Heros and K. Helbing, review to appear in Europ. Phys. Journ. C.
14. M. G. Aartsen et al. (IceCube Coll.), arXiv:1707.07081.
15. P. Coyle for the KM3NeT Coll., J. Phys. Conf. Ser. 888 (2017) no.1, 012024 and arXiv:1701.01382.
16. M.G. Aartsen et al. (IceCube Coll.), J. Phys. G44 (2017) 054006 and arXiv:1707.02671.
17. N.G. Jerlov, Marine Optics, Elsevier Oceanography Series 5 (1976).
18. H. Bradner and G. Blackington, Appl. Opt. 23 (1984) 1009.
19. V. Balkanov et al., Appl. Opt. 33 (1999) 6818.
20. V. Balkanov et al., Nucl. Instrum. Meth. A298 (2003) 231.

21. J.A. Aguliar et al., Astropart. Phys. 23 (2005) 131, and arXiv:astro-ph/0412126.
22. G. Riccobene et al., Astropart. Phys. 27 (2007) 1, and arXiv:astro-ph/0603701 (see also for further references therein).
23. M. Ackermann et al., J. Geophys. Res. 111 (2006) D13203.
24. J. Lundberg et al., Nucl. Instrum. Meth. A581 (2007) 619.
25. W. Lohmann et al. CERN Yellow Report 85–03.
26. C.H.W. Wiebusch, PhD thesis, preprint PITHA 95/37.
27. I. Albuquerque, J. Lamoureux and G.F. Smoot, Astrophys. J. Suppt. 114 (2002) 195 and arXiv:hep-ph/0109177.
28. T.C. Weekes et al., Astrophys. J.343 (1989) 379.
29. D. Petry et al., J. Astron. Astrophys. 311 (1996) L13.
30. J. Ahrens et al., Nucl. Instrum. Meth. A524 (2004) 169.
31. Y. Becherini for the Antares coll., Proc. 30th Int. Cosmic Ray Conf. Merida 2007, and arXiv:0710.5355 (see also references therein).
32. M. Ackermann et al., Astropart. Phys. 22 (2004) 127, and arXiv:astro-ph/0405218.
33. B. Hartmann, PhD thesis, Erlangen 2008, see arXiv:astro-ph/06060697.
34. A. Roberts, Rev. Mod. Phys. 64 (1992) 259.
35. E. Babson et al.,Phys. Rev. D42 (1990) 3613.
36. C. Spiering Eur. Phys. J. H37 (2012) 515 and arXiv:1207.4952.
37. I.A. Belolaptikov et al., Astropart. Phys. 7 (1997) 263.
38. CERN Courier, Sept. 1996, p.24.
39. C. Spiering for the Baikal Coll., Prog. Part. Nucl. Phys. 40 (1998) 391.
40. R.V. Balkanov et al., Astropart. Phys. 12 (1999) 75, and arXiv:astro-ph/9705244.
41. R. Bagduev et al., Nucl. Instrum. Meth. A420 (1999) 138.
42. V. Aynutdinov et al., Astropart. Phys. 25 (2006) 140, and arXiv:astro-ph/0508675.
43. E. Andres et al., Astropart. Phys. 13 (2000) 1.
44. E. Andres et al., Nature 410 (2001) 441.
45. P. Askebjer et al., Science 267 (1995) 1147.
46. M. Ackermann et al., J. Geophys. Res. 111 (2006) D13203.
47. T. DeYoung, Journ. of Physics Conf. Series 136 (2008) 042058.
48. R. Abbasi et al. Phys. Rev. D79 (2009) 062001.
49. G. Aggouras et al., Nucl. Instr. Meth. A552 (2005) 420.
50. E. Migneco et al., Nucl. Instr. Meth. A588 (2008) 111.
51. ANTARES homepage, http://antares.in2p3.fr
52. M. Ageron et al. (ANTARES Coll.) Nucl. Instr. Meth. A656 (2011) 11 and arXiv:1104.1607.
53. J. Aguilar, Astropart. Phys. 26 (2006) 314.
54. M. Ageron et al., Astropart. Phys. 31 (2009) 277. and arXiv:0812.2095.
55. T. Montaruli, J. of Modern Physics A, arXiv:0810.3933.
56. IceCube homepage, http://icecube.wisc.edu.
57. M.G. Aartsen et al. (IceCube Coll.) JINST 12 (2017) P03012 and arXiv:1612.05093.
58. R. Abbasi et al. (IceCube Coll.), Nucl. Instr. Meth. A 700 (2013) 188 and arXiv:1207.6326.
59. R. Abbasi et al. (IceCube Coll.), Nucl. Instr. Meth. A 618 (2010) 139 and arXiv:1002.2442.
60. R. Abbasi et al. (IceCube Coll.) Astron. Astrophys. 535 (2011) A109 and arXiv:1108.0171.
61. R. Abbasi et al. (IceCube Coll.), Nucl. Instr. Meth. A 601 (2009) 294 and arXiv:0810.4930.
62. SNEWS: P. Antonioli et al., New Journ. Phys. 6 (2004) 114 and arXiv:astro-ph/0406214.
63. KM3NeT homepage: http://www.km3net.org.
64. S. Adrian Martinez et al. (KM3NeT Coll.) J.Phys. G43 (2016) no.8, 084001 and arXiv:1601.07459.
65. http://baikalweb.jinr.ru, including a full english project description.
66. V. Avronin et al. (Baikal Coll.), Nucl. Instr. Meth. A742 (2014) 82. Nucl. Instr. Meth. A602 (2009) 227, and arXiv:0811.1110.
67. M.G. Aartsen et al. (IceCube Coll.), arXiv:1412.5106.
68. M.G. Aartsen et al. (IceCube Coll.), Phys. Rev. Lett. 113 (2014) 101101.
69. M.G. Aartsen et al. (IceCube Coll.) Astrophys. J. 833 (2016) no.1, 3 and arXiv:1607.08006.

70. R. Abbasi et al. (IceCube Coll.) Phys. Rev. D83 (2011) 012001 and arXiv:1010.3980.
71. S. Adrian-Martinez et al. (Antares Coll.) Eur. Phys. J. C73 (2013) 2606 and arXiv:1306.1599.
72. M.G. Aartsen et al. (IceCube Coll.) Phys. Rev. Lett. 120/no.7 (2018) and arXiv:1707.07081.
73. M.G. Aartsen et al. (IceCube Coll.) Science 342 (2013) 1242856 and arXiv:1311.5238.
74. A. Albert et al. (Antares and IceCube Coll.) Astrophys. J. 868 (2018) L20 and arXiv:1808.03531.
75. M.G. Aartsen et al. (IceCube Coll.) arXiv:1811.07979.
76. M.G. Aartsen et al. (IceCube Coll.) Astrophys. J. 835/no.1 (2017) 45 and arXiv:1611.0374.
77. M.G. Aartsen et al. (IceCube Coll.) Astrophys. J. 843/no.2 (2017) 112. and arXiv:1702.06862.
78. M.G. Aartsen et al. (IceCube, Fermi-LAT, MAGIC, AGILE, ASAS-SN, HAWC, H.E.S.S., INTEGRAL, Kanata, Kiso, Kapteyn, Liverpool Telescope, Subaru, Swift NuSTAR, VERITAS and VLA/17B-403 Collaborations) Science 361/no.6389 (2018) eaat1378 and arXiv:1807.08816.
79. M.G. Aartsen et al. (IceCube Coll.) Science 361/no.6389 (2018) 147 and arXiv:1807.08794.
80. M.G. Aartsen et al. (IceCube Coll.) Phys. Rev. D98 (2018) 062003 and arXiv:1807.01820
81. M. Ahlers, K. Helbig and C. Perez de los Heros, Eur. Phys. J.C 78 (2018) 924 and arXiv:1806.05695.
82. A. Aab et al., Nucl. Instr. Meth. A798 (2015) 172.
83. H. Tokuno et al., Nucl. Instr. Meth. A676 (2012) 54.
84. Letessier Selvon, Proc. AIP Conf. 566 (2001) 157.
85. D. Fargion, Astrophys. Journ. 570 (2002) 909.
86. A. Aab et al., Phys.Rev. D91 (2015) no.9, 092008.
87. G.W.S. Hou and M.A. Huang, astro-ph/0204145.
88. G.A. Askaryan, Sov.Phys. JETP 14 (1962) 441.
89. D. Saltzberg et al., Phys. Rev. Lett. 86 (2001) 2802.
90. P. Gorham et al., Phys. Rev. Lett. 99 (2007) 171101.
91. B. Price, Astropart. Phys. 5 (1996) 43.
92. I. Kravtchenko et al., Phys. Rev. D73 (2006) 082002.
93. P. Allison et al. (ARA Coll.), Phys.Rev. D93 (2016) no. 8, 082003 and arXiv:1507.08991.
94. S. Barwick et al. (ARIANNA Coll.) Astropart. Phys. 90 (2017) 50 and arXiv:1612.04473.
95. P. Gorham et al. (ANITA Coll.) Phys. Rev. Lett. 103 (2009) 051103.
96. S.A. Wissel et al., PoS (ICRC 2015) 1150. Astroparticle Physics 90 (2017) 50 and arXiv:1612.04473.
97. P. Gorham et al., Phys. Rev. Lett. 93 (2004) 0041101.
98. O. Scholten et al., in [Proc. 2008 ARENA Workshop, Rome 2008, Nucl. Instr. Meth. 2009, ed. A. Capone] and arXiv:0810.3426.
99. J. D. Bray, Phys. Rev. D91 (2015) no.6, 063002 and arXiv:1502.03313.
100. G.A. Askaryan, Sov. Journ. Atom. Energy 3 (1957) 921.
101. see e.g. J.G. Learned, Phys. Rev. D1, 19 (1979) 3293.
102. R. Abbasi et al. (IceCube Coll.) Astropart.Phys. 34 (2011) 382 and arXiv:1004.1694; R. Abbasi et al. (IceCube Coll.), Astropart.Phys. 35 (2012) 312 and arXiv:1103.1216.
103. Proc. Int. Workshop on Acoustic and Radio EeV Neutrino detection Activities (ARENA) DESY, Zeuthen 2005, World Scientific 2006, eds. R. Nahnhauer and S. Boeser.
104. Proc. 2006 ARENA Workshop, Univ. Northumbria 2006, Journ. of Phys., Conf. Series 81 (2007), eds. L. Thomson and S. Danaher.
105. L. Thompson, Nucl. Instr. Meth. 558 (2008) 155.
106. Proc. 2008 ARENA Workshop, Rome 2008, Nucl. Instr. Meth. 2009, ed. A. Capone.
107. Proc. 2012 ARENA Workshop, eds. R. Lahmann, Th. Eberl, K. Graf, C. James, T. Huege, T. Karg and R. Nahnhauer, AIP Conference Collection Volume 1535, Erlangen 2012.
108. Proc. 2017 ARENA Workshop, eds. S. Buitnik, J.R. H"orandel, S. de Jong, R. Lahmann, R. Nahnhauer, O. Scholten, EPJ Web Conf. 135 (2017).
109. N. Lehtinen et al., Astropart. Phys. 17 (2002) 279 and astro-ph/010433.
110. D. Besson et al., in [Proc. 2008 ARENA Workshop, Rome 2008, Nucl. Instr. Meth. 2009, ed. A. Capone] and arXiv:0811.2100.

Chapter 18
Spaceborne Experiments

Roberto Battiston

18.1 Introduction: Particle Physics from Ground to Space

The Universe is the ultimate laboratory to understand the laws of nature. Under the action of the fundamental forces, lasting infinitesimal times or billion of years, matter and energy reach most extreme conditions. Using sophisticated instruments capable to select the signals reaching us from the depths of space and of time, we are able to extract information otherwise not obtainable with the most sophisticated ground based experiments. The results of these observations deeply influence the way we today look at the Universe and try to understand it.

During hundreds of thousands of years we have observed the sky only using our eyes, accessing in this way only the very small part of the electromagnetic radiation which is able to traverse the atmosphere, the visible light. The first telescope observations by Galileo in 1609, which dramatically changed our understanding of the solar system, yet were based only on the visible part of the electromagnetic spectrum. Only during the second half of the twentieth century we started to access wider parts of the spectrum. After the end of the war, using the new radar related technology, the scientists developed the radio telescopes to record the first radio images of the galaxy. But only in the 60's, with the advent of the first man made satellites, we began to access the much wider e.m. spectrum, including infrared, UV, X-ray and γ-ray radiation.

A similar situation happened with the charged cosmic radiation. Cosmic Rays, discovered by Hess in 1912 [1] using electrometers operated on atmospheric balloons, for about 40 years were the subject of very intense studies. The discovery of a realm of new particles using CR experiments, gave birth to particle physics

R. Battiston (✉)
Dipartimento di Fisica, Università di Trento, Povo, Italy
e-mail: roberto.battiston@pg.infn.it; roberto.battiston@unitn.it

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_18

and to high energy physics, successfully performed, since the 50's, at particle accelerators. However, the study of the cosmic radiation performed within the atmosphere, deals only with secondary particles. The primary radiation can only be studied with stratospheric balloons or using satellites. In 1958 Van Allen [2] and collaborators studied for the first time the charged cosmic radiation trapped around the Earth, and, since, the measurement of Cosmic Rays from space has become an important tool for the study of the Universe.

A third, more recent example is the discovery of ravitational waves (GW) [3], one hundred years after the prediction of Einstein [4]. Direct observation of gravitational waves opened a new era in astrophysics, adding to the spectrum of electromagnetic radiation the new messenger represented by GW. Since the pioneering attempts of Weber [5] in the 60's, using resonating bars, the GW community has developed in the 90's a network of $O(1)$ km arms, ground based interferometers to search for GWs in the frequency range 10 Hz to 100s of Hz [6],[7]. The detected signals confirm the prediction of General Relativity but also validate the sensitivity of the interferometer technologies. GW are expected much more abundantly in the frequency range $O(0.001)$ Hz to a $O(1)$ Hz. This range can be studied with a $5 \cdot 10^6$ km arm, space based interferometer, as the proposed ESA/NASA LISA mission. The successful LISA technology demonstrator, the ESA lead LISA-Pathfinder (LISA-PF) [8] flown in 2016, opened the way for the LISA[9] adoption, to be developed and implemented during the 20's to start operating at the end of the 20's or a the beginning of the 30's.

During the last century, particle detectors developed on ground have been adapted or designed to be used on stratospheric balloons and on space born experiments. Space, however, is a hostile environment and launching a payload is a very expensive endeavour. For these reasons, the design and the testing of a spaceborn detector requires particular care. In this chapter we deal with this topic.

We begin discussing the properties of the space environment from the upper atmosphere, the transition from the atmosphere to the magnetosphere and from the magnetosphere to the deep interplanetary space.

We then address the requirements for hardness and survivability of space born instrumentation.

We subsequently turn to the issue of manufacturing of hardware to be operated in space, with particular care to the issue of the space qualification tests.

We will also discuss modern spaceborne high energy radiation detectors, mainly from the point of view of the design characteristics related to the operation in space. We will make no attempt to cover the historical development or to cover low energy radiation instrumentation, in particular X-ray space borne detectors.

18.2 The Space Environment

18.2.1 The Neutral Component

Although there are some notable exceptions, a good fraction of scientific satellites which observe the different kinds of radiations emitted by the universe operate on LEO (Low Earth Orbit), namely between 200 and 2000 km from the Earth surface. Below 200 km the atmospheric drag dramatically reduces the lifetime of satellites, above 700 km the radiation environment, due to the Van Allen belts, becomes more and more hostile.

When operating close to the lower limit of LEO orbits, the external surfaces of the payloads are affected by the heat produced by the upper atmosphere drag and by the corrosion due to the presence of highly reactive elements such as atomic oxygen. Above ∼600 km drag is sufficiently weak not to influence anymore the lifetime of most satellites.

Altitudes below ∼600 km are within the Earth's *thermosphere*, the region of the atmosphere where the absorption of the solar UV radiation induces a fast rate of temperature increase with the altitude. At ∼200–250 km the temperature of the tenuous residual atmosphere reaches a limiting value, the *exospheric temperature* ranging from ∼600–1200 K over a typical solar cycle. The thermosphere temperature can also quickly change during the geomagnetic activity.

Atomic oxygen is the main atmospheric constituent from ∼ 200–600 km, since it is lighter than molecular nitrogen and oxygen. Figure 18.1 shows the altitude profiles of atomic oxygen for different solar activities. Atomic oxygen plays an important role in defining the properties of LEO space environment. Since this form of oxygen is highly reactive, surfaces covered with thin organic films, advanced composites or thin metallized layers can be damaged [10]. Kapton, for example, erodes at a rate of approximately 2.8 μm for every 10^{24} atoms/m^2 of atomic oxygen fluence [11], with the fluence during a time interval t being defined as $F_0 = \rho_N v t$, ρ_N being the number density of atomic oxygen and v the satellite velocity. Chemical reaction involving atomic oxygen can in turn produce excited atomic states emitting significant amount of e.m. radiation, creating effects such as the *shuttle glow* which are interfering with optical instrumentation.

18.2.2 The Thermal Environment

From a thermal point of view a spacecraft orbiting around the Earth is exposed to various heat sources; direct sunlight, sunlight reflected off the Earth or other planets (*albedo*) and infrared radiation emitted by the planet atmosphere or surface. The spacecraft in turn loses energy by radiation to deep space, which acts as a sink at 2.7 K.

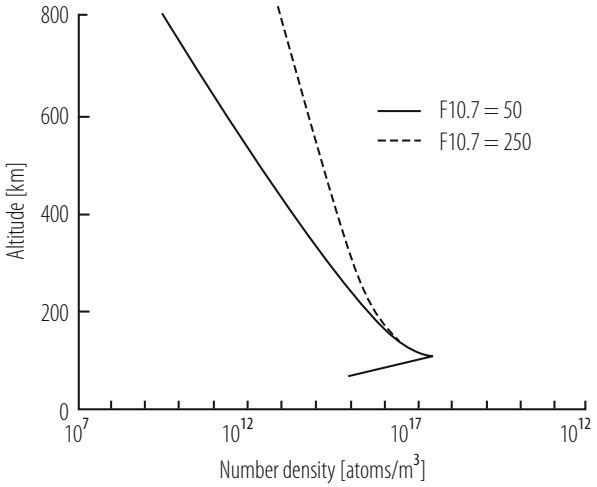

Fig. 18.1 Altitude profiles of number density of atomic oxygen at solar minimum (solid line) and solar maximum (dashed line) [12]

18.2.2.1 Direct Sunlight

A main source of thermal energy is of course the Sun, which acts as a black body at a temperature of 5777 K. The Sun is a very stable source of energy: at the Earth the energy flux varies from 1414 W/m^2 during winter time to 1322 W/m^2 during summer time. The mean intensity at 1 AU is called *solar constant* and is equal to 1367 W/m^2. The spectral energy distribution is approximatively 7% UV, 46% visible and 47% near-IR.

18.2.2.2 Albedo

Albedo refers to the sunlight reflected by a planet. It is highly variable with the conditions of the surface. For spacecraft orbiting close to the Earth, the *albedo* can reach a significant fraction, up to 57%, of the Earth emitted radiation, which in turn is 200–270 W/m^2, depending on the latitude and of the orbit inclination. The Earth itself is a blackbody radiating at around 255 K. This energy cannot be reflected away form the spacecraft which is approximatively at the same temperature. This energy can only be rejected through the spacecraft thermal control system. It is a non negligible amount of radiation: for example, when the Shuttle bay area looks at the illuminated surface of the Earth, its temperature reaches values close to 250 K even if the back of the spacecraft sees the 2.7 K of deep space.

18.2.3 The Charged Component

18.2.3.1 The Low Energy Plasma

At typical Shuttle altitudes, ~300 km, about 1% of the atmosphere is ionized. This fraction increases to 100% at geosynchronous altitudes. This plasma environment can easily charge up satellite components, both on the surface and on the interior of the spacecraft. If the charging exceeds the electric breakdown and discharges are produced they can damage the satellite electronics. The charged component of the radiation is heavily influenced by the existence of the Earth magnetic field. The Earth magnetic field is roughly dipolar:

$$B(R, \theta) = (1 + \sin^2 \theta_M)^{1/2} B_0 / R^3 \tag{18.1}$$

where B is the local magnetic field intensity, θ_M is the magnetic latitude, R is the radial distance measured in Earth radii (R_E) and B_0 is the magnetic field at the equator and at $R = 1$, $B = 0.30$ G. The interaction between the solar wind and the Earth's magnetic field results in a magnetic field structure much more elongated on the night side than it results on the day side, known as *magnetotail*. The resulting magnetic structure is called *magnetosphere* (Fig. 18.2).

The electrical potential of a spacecraft or payload is measured with respect to the nearby plasma when the net charge flow is zero. This current is the sum of the various exchanges of charge between the plasma and the spacecraft including photo-extraction and secondary emission from the spacecraft surfaces. The single

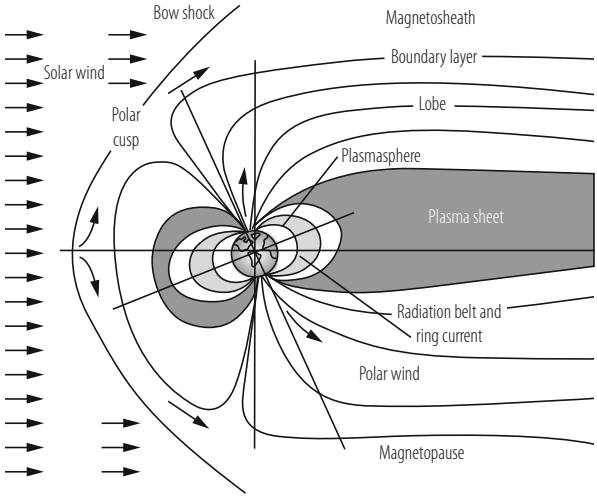

Fig. 18.2 Cross section of the Earth's magnetosphere, showing the key plasma and energetic particle populations [12]

component voltage to the spacecraft ground depends on the element capacitance to the nearby materials. Space charging is particular detrimental in orbits where electron energies in the 10 to 20 keV range dominate the current from the plasma to the spacecraft. At low altitudes this happens only at high latitudes where there are energetic auroral electrons [13]. At other low altitudes locations, low energy electrons are sufficiently abundant to keep the electric fields below the breakdown levels.

The situation is different in higher orbits, such as geosynchronous, where surface charging occurs during magnetospheric substorms between the longitudes corresponding to midnight and dawn [14]. The design of spacecrafts capable to keep a small differential potential with respect to the plasma or to tolerate electrostatic discharges is necessary for these orbits. Design rules and material selection criteria have been developed to help reducing the effect of surface charging on spacecrafts and payloads [15, 16].

It should be noted that, although in the equatorial regions of LEO differential charging is small, the potential of the spacecraft with respect to the surrounding plasma can reach a level close to 90% of the solar array voltage. This should be taken into account when designing experiments aimed to study the plasma properties or when dealing with high voltage power supplies.

18.2.3.2 The Trapped Radiation

Well inside the magnetosphere lie the radiation belts, regions where energetic ions and electrons experience long-term magnetic trapping [17]. Since this trapping requires stable magnetic fields, near the magnetopause the magnetic field fluctuations induced by solar wind prevent long term trapping. On the low altitude side the atmosphere limits the radiation belts to the region above 200 km. The magnetic geometry limits the trapping volume to magnetic latitudes of about 65°. A $magnetic\ L-shell$ is the surface generated by rotating a magnetic field line around the Earth dipole axis and L is measured in units of Earth radius. Trapped particles spiral along paths centered on a given shell. The shell surface can be approximately described as: $R = L\cos^2\theta_M$ [18]. Electrons preferentially populate the toroidal region centered on $L \sim 1.3$ (inner zone) while protons populates the region around $L \sim 5$ (outer zone). The energy of these trapped particles is greater than 30 keV and can reach hundreds of MeV. The intensity of the trapped radiation flux can reach the maximum intensity of $10^8 - 10^9\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ at a distance of $\sim 2\ R_E$ for $E_k > 0.5$ MeV electrons and of $\sim 3\ R_E$ for $E_k > 0.1$ MeV protons. Satellite components, in particular electronics, can be damaged by this penetrating charged form of radiation. A dramatic example of this occurred in 1962 when several satellites ceased to operate after their solar cells were damaged by the increase of radiation belts intensity from high altitude nuclear explosions. Since the basic principles of the trapping are well understood, radiation belts can be modeled quite accurately: a standard model of the Van Allen Belts is available by the National Space Science Data Center [20]. It should be noted, however, that due

to the structure of the Earth magnetic field, which has a dipolar structure not aligned with the Earth angular momentum, the radiation belts are only approximatively of toroidal shape: in the vicinity of the South Atlantic, the structure of the belts is strongly affected and the bouncing altitude of the trapped particles decreases very significantly (South Atlantic Anomaly, SAA). This leads to a region which, although located at LEO altitudes, is characterized by a very intense particle flux, since it is basically within the belts.

Energetic particles, such as electrons from 200 to 1.5 MeV, can implant in the dielectrics and produce discharges within the components themselves (*bulk charging*). At even higher energies, above few MeV, charged particles are highly penetrating and release their energy in the form of ionization deep inside materials. The damages induced by this penetrating radiation can be divided into:

- total dose effects which can degrade the material properties of microelectronics devices, optical elements (lenses, mirrors), solar arrays, sensors, ...
- Single Event Effects or Phenomena (SEE or SEP), effects induced by single particles creating short circuits which can temporarily or permanently damage microelectronics components. They are further subdivided into
  - *Single Event Upset (SEU)* or *bitflip* which although do not damage the electronics may influence the operation of onboard software.
  - *Single Event Latch-up (SEL)*, causing sudden low resistance paths and subsequent drift on the power lines of electronics components which start to operate abnormally until the correct voltage is restored. Depending on the power supply performances SEL can be recovered or could cause permanent damages.
  - *Single Event Burnout* (SEB), causing permanent failures of electronic devices.

18.2.3.3 Solar Particle Events

The Solar Particle Events (SPE) occur in association with solar flares. They consist in an increase of the flux of energetic particles, mostly protons, ($\sim$1 MeV to $\sim$1 GeV) over time scales of minutes, lasting from few hours to several days. Although SPEs occur at a rate of few per year, they are very dangerous for payloads and astronauts, due to the intense radiation dose they deliver, several orders of magnitude higher than in normal conditions (see Fig. 18.3). The global time structure of a SPE is somewhat characteristic (see Fig. 18.4), although the detailed structure depends on the evolution of the original solar flare. X-rays reach the Earth within minutes together with the most relativistic part of the proton spectra; lower energy particles diffuse over time scales of several hours. The fast component of a SPE can be used as early warning to protect the most delicate parts of a payload by switching them off, by using radiation shields or changing the satellite attitude or operational mode.

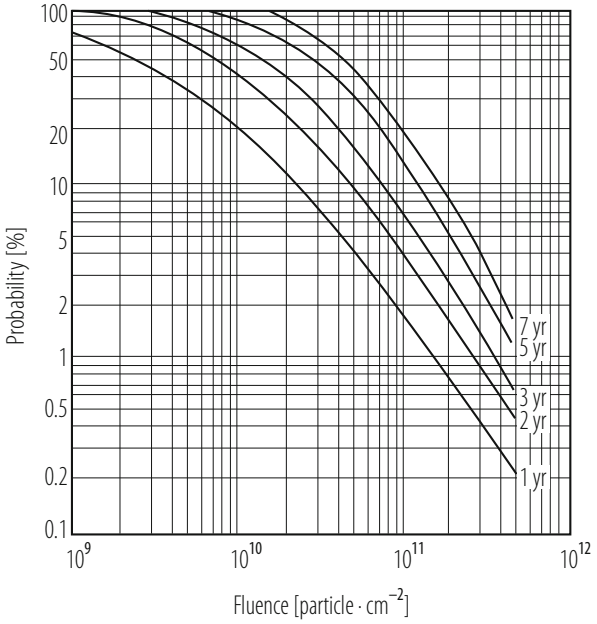

Fig. 18.3 Probability of exceeding a given fluency level as a function of mission duration [12]

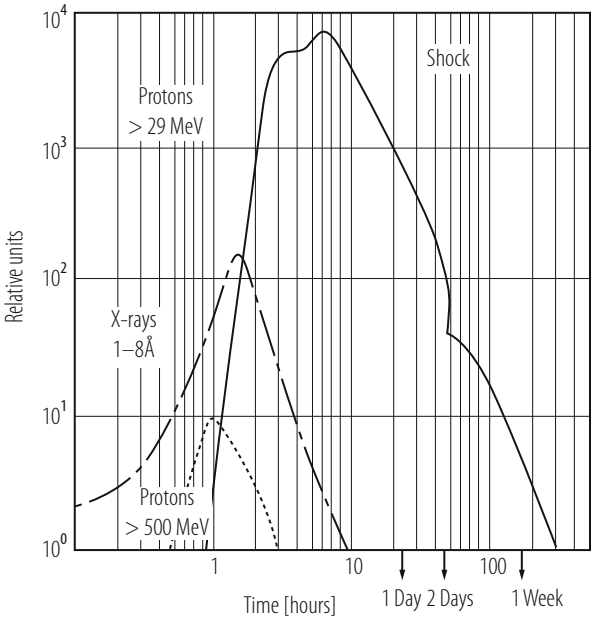

Fig. 18.4 Typical time evolution of a Solar Particle Event (SPE) observed from the Earth [12]

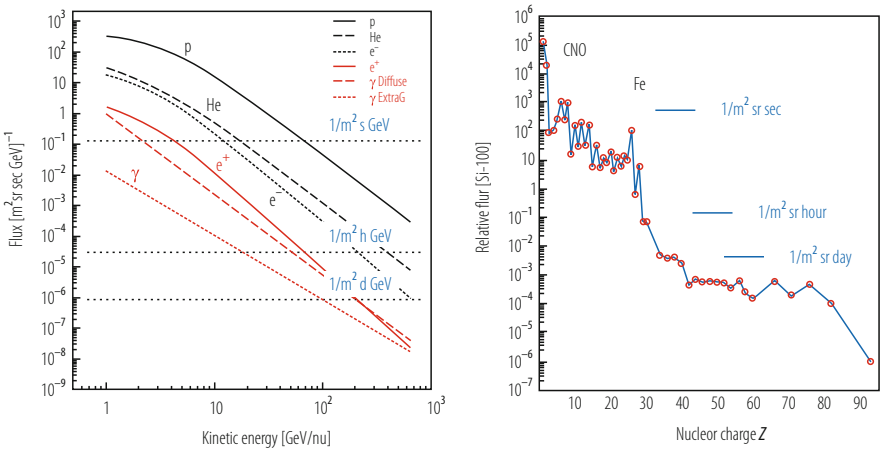

Fig. 18.5 Galactic Cosmic Rays Composition. Left, differential flux of H and He nuclei compared with e^-, e^+ and γ rays. Right, total flux of the nuclear component of galactic Cosmic Rays as a function of the electric charge Z

18.2.3.4 Galactic Cosmic Rays

Galactic Cosmic Rays (GCR) are high energy charged particles reaching the Earth from outside the solar system. The GCR composition is similar to the composition of the energetic particles within the solar system but extend to much higher energy (see Fig. 18.5 and Sect. 18.2.3.4). Their energy ranges from $O(100\,\text{MeV})$ to $10^6\,\text{GeV}$ or more, with an energy spectrum falling as $\sim E^{-2.7}$ for $E > 1\,\text{GeV}$. These particles are very penetrating, loosing their energy only by ionization. Nuclear interaction phenomena are indeed negligible in space for what concerns radiation damages. The ionization losses can create $SEEs$ as discussed above. $GCRs$ have a significant content of high Z particles, fully ionized nuclei with charge extending up to Iron ($Z = 25$). Since ionization losses are proportional to Z^2, high Z GCR can be very effective in causing $SEEs$.

18.2.4 Space Debris

Orbiting spacecraft are subjected to hypervelocity (several km/s) impacts with micron size or larger pieces of dust or debris, both of natural ($micrometeorites$) and artificial ($orbital\ debris$) origin. These impacts can have dramatic effects on a space mission. The probability of a catastrophic impact can be assessed for a given mission and payload. Some measures can be implemented to reduce the effect of the space debris protecting the most important parts with screens made of multilayered materials which can absorb and dissipate the energy of the incoming fragments.

18.3 Types of Orbits

The choice of the orbit heavily influences satellites and payload design.

Many scientific applications are operating on Earth-Referenced Spacecraft orbits. Depending on their typical altitude we talk of *Low Earth Orbits (LEO)*, which are mostly below the Van Allen Belts (typically below 1000 km of height), and of *Geosynchronous Orbits (GEO)* which are well above the Van Allen Belts. Payloads spending substantial time within the Van Allen belts are exposed to high doses of radiation and requires particular care designing and protecting the electronics from SEE and total doses effect.

Table 18.1 shows the types of specialized Earth-Referenced orbits.

Higher orbits are typical of interplanetary missions; for these missions the typical doses received by the satellite payloads are significantly higher than for LEO but lower than within the Van Allen Belts. Far away from the Earth satellites are not anymore shielded from $SPEs$ by the Earth shadow nor by the screening effect of its magnetic field. SEE due to heavy ions and low energy protons should be carefully taken into account when designing the payload electronics.

The space radiation environment remains one of the primary challenges and concerns for space exploration, in particular for deep space missions of long duration, i.e., when the the combined shield due to Earth magnetosphere and atmosphere vanishes. In the inner heliosphere, major sources of radiation are Galactic Cosmic Rays, Solar Particles and Jovian Electrons. Furthermore, in the space nearby Earth particles (mainly electrons and protons) are trapped within the Van Allen radiation belts. Particles populating such a space environment induce single event and cumulative dose in spacecraft materials and, eventually, create electronics hazards.

Table 18.1 Specialized Earth-Referenced orbits

Orbit	Characteristics	Application
Geosynchronous (GEO)	Maintains nearly fixed position over equator	Communication, weather
Sun-synchronous	Orbit rotates so as to maintain approximately constant orientation with respect to Sun	Earth resources, weather
Molniya	Apogee/perigee do not rotate	High latitude communications
Frozen Orbit	Minimizes changes in orbit parameters	Any orbit requiring stable conditions
Repeated Ground Track	Sub orbits repeats	Any orbit where constant viewing angles are desirable

18.4 Space Mission Design

18.4.1 The Qualification Program

Since repairing in space is extremely expensive, if at all possible, designing and building spacecraft and payloads which maximal reliability is a must in the field of space engineering. It follows that quality control is an essential part during the various phases of the program. The $Qualification\ Program$ adds to the cost of the space hardware construction, sometime very significantly, but it makes sure that the program is not headed for failure.

Qualification tests must be designed and implemented to check that the spacecraft/payload can withstand the challenges of launch, deployment and operation in space. Subsystems and components environmental tests include vibration, shock and thermal vacuum, electromagnetic compatibility and radiation hardness.

Although the goal is the same, testing strategies are not unique. There are indeed various testing methods:

- $dedicated\ qualification\ hardware$ (QM): a set of qualification components is built and tested at qualification levels. A set of flight components (FM) is then built and launched after passing a qualification test a lower levels;
- $proto - flight$ approach: a set of flight components is tested at qualification level then assembled into a subsystem or payload which is tested at qualification levels and then launched;
- $similarity$ approach: demonstrate that the components and the environment are identical to previously qualified hardware.

A typical test sequence includes a series of functional tests preceeding/following each environmental test, for example:

- functional test;
- vibration test (levels depending on the mission);
- functional test;
- shock test (levels depending on the mission);
- functional test;
- thermal-vacuum tests, including some functional tests during exposure;
- Electro Magnetic Compatibility (EMC) tests (if required);
- flash X rays with functional tests during exposure (if required);

18.4.2 Vibration and Shock Test

A payload must withstand vibrations caused by the launch vehicle and transmitted through its structural mount. During launch, payload components may experience shocks due to the explosives used for the separation of the various stages. In case reentry is foreseen, they do experience shocks when entering the atmosphere as

well as during the landing phase. In order to understand the dynamical behavior of the payload and of its mounting under these circumstances, Finite Element Analysis (FEA) dynamic and numerical analysis together with Computer Aided Design (CAD) simulation should be performed. In this way it is possible to search for resonances of the mechanical structures, identifying conditions where the material could be stressed or damaged. Following an iterative process the mechanical design of the payload and of its mountings can be improved until all negative margins are eliminated. Dynamic and vibration tests are then performed on a qualification model, using for example an electro-dynamical shaker operating at frequencies between 5 and 3000 Hz, with a spectrum which depends on the mission characteristics. Table 18.2 shows a typical acceleration spectrum expected for payload launched using the Shuttle transportation system. Qualification levels are typically higher by factor 2 to 4. Shock tests are performed using a similar strategy. For example Fig. 18.6 shows shock levels used to simulate the launch of an Alpha-Centaur rocket.

Table 18.2 Maximum expected flight levels for a shuttle mission

	Frequency range	Frequency dependence
X axis	20–58 Hz	0.0025 g^2/Hz
	58–125 Hz	+9 dB/Octave
	125–300 Hz	0.025 g^2/Hz
	300–900 Hz	−9 dB/Octave
	900–2000 Hz	0.001 g^2/Hz
Overall = 3.1 Grms		
Y axis	20–90 Hz	0.008 g^2/Hz
	90–100 Hz	+9 dB/Octave
	100–300 Hz	0.01 g^2/Hz
	300–650 Hz	−9 dB/Octave
	850–2000 Hz	0.001 g^2/Hz
Overall = 3.1 Grms		
Z axis	20–45 Hz	0.009 g^2/Hz
	45–125 Hz	+3 dB/Octave
	125–300 Hz	0.025 g^2/Hz
	300–900 Hz	−9 dB/Octave
	900–2000 Hz	0.001 g^2/Hz
Overall = 3.1 Grms		

18.4.3 Environmental Tests

The environmental qualification campaign of a space component can be divided into three main steps. The first step consists in the development of requirements and constraints related to the payload and to the mission. The second step is

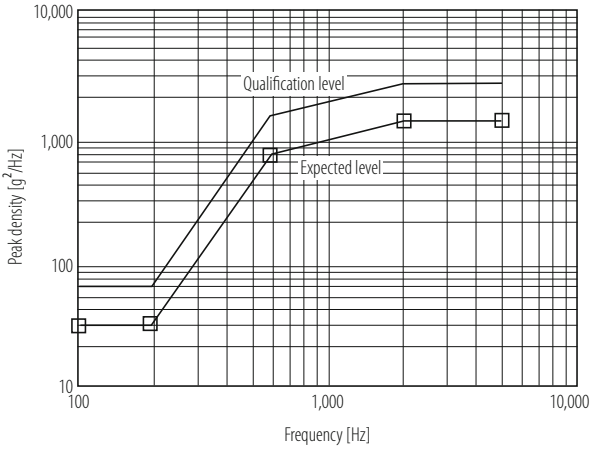

Fig. 18.6 Shock levels simulating the launch environment of an Alpha-Centaur rocket: qualification levels are designed to be greater than the expected design values [12]

to determine and define the space environment (in terms of temperatures, heat transfer ways, worst hot and cold case, etc.) that will characterise thermal conditions throughout the entire life of the component. An important part of the process of qualification is then the thermal analysis which can be conducted using FEA techniques. Once an acceptable thermal model has been developed, test predictions can be calculated to correlate thermal verification tests with the test results. If this correlation is found acceptable, the thermal model is then used to perform flight predictions. If, instead, the correlation is poor the thermal analysis and the hardware configuration need to be carefully checked to understand whether the actual configuration (hardware) requires modifications or the thermal model needs to be updated. Payload temperature requirements derive from the spacecrafts thermal design and the orbital environment and attitude. The purpose of these tests is to demonstrate that the subsystems comply with the specification and perform satisfactorily in the intended thermal environment with sufficient margins. The test environment should be based either on previous flight data, often scaled for differences in mission parameters, or, if more reliable, on analytical prediction or by a combination of analysis and flight data. A margin can include an increase in level or range, an increase in duration or cycles of exposure, as well as any other appropriate increase in severity of the test. Humidity and thermal qualification tests in climatic rooms are performed to test the behaviour of the electronic components and mechanical structures under thermal and humidity changes. The tests are conducted using climatic chambers, with temperature ranges depending on the mission parameters: for a LEO mission typical range lays within $-80\,°\mathrm{C}$ and $+120\,°\mathrm{C}$ for a planetary mission wider intervals are required. Components should be switched on and work both at temperature extremes or during transition, following the mission specifications (see Fig. 18.7).

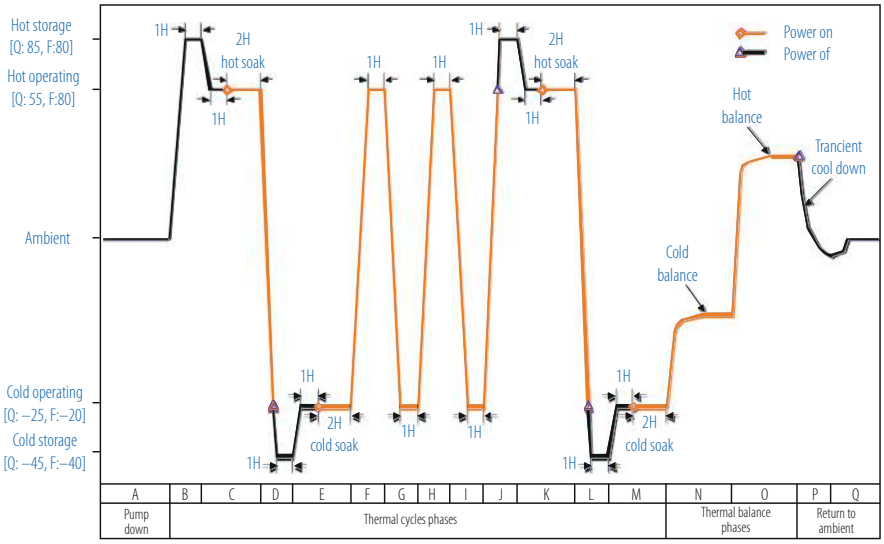

Fig. 18.7 Typical Thermal Vacuum Cycling test profile for a payload to be operated on a LEO orbit on the ISS [54]

18.4.4 EMC Tests

Another area where space payloads are submitted to extensive testing is the compatibility to Electro Magnetic fields, either radiated or received. During the test of radiated EM the device under test is powered and operated in standard operating condition in an EM anechoic chamber. Through suitable antennas and filters read by receivers, the intensity of the emitted radiation is measured as a function of the frequency. The results are compared with the limits requested by specific standards or design rules. If the limits are exceeded, then the electrical grounding or design of the device should be modified. During the received EM test, the device is operated within an EM anechoic chamber while EM radiation, monochromatic or with a specific spectral structure, is generated at a predetermined intensity using special antennas located nearby. The purpose of the test is to check that the item under test does not exhibits anomalies when illuminated by beams of EM radiation, typically emitted by a communication antenna or a nearby electronic device. Figure 18.8 show a typical the result for an EMC radiated test on a payload to be operated on the ISS.

18.4.5 Radiation Hardness Tests

As discussed in the previous paragraphs, the space environment is particularly harsh for operating microelectronics devices, due to the presence of single, heavily

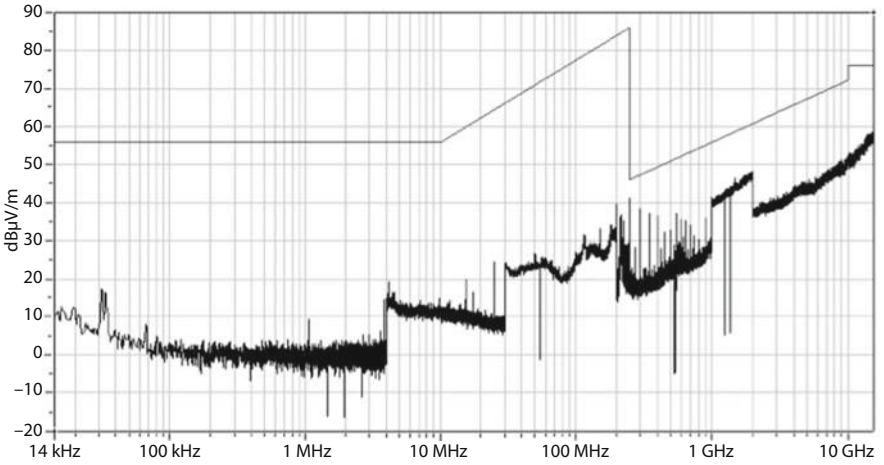

Fig. 18.8 Typical radiated EMC tests result for a payload to be operated on a LEO orbit on the ISS [19]. The thin line presents the e.m. field limits which should not be exceeded as described in Table 18.3

Table 18.3 Radiated EMC limits for the tests described in Fig. 18.8

Frequency range	Emission [dBμV/m]	Antenna
14 kHz–10 MHz	56	rod—vertical
10 MHz–259 MHz	56–86 (16 dB/decade)	biconical—horiz/vert
259 MHz–10 GHz	46–72 (16 dB/decade)	double ridge—horiz/vert
10 GHz–20 GHz	76	horn—horiz/vert

ionizing particles which can deposit large amount of charge in the bulk, inducing short circuits or spurious currents in the solid state circuits. The total dose collected during a space flight is relatively small, mostly in the $krad$ range, so the radiation damages are mostly due to Single Event Effects (SEE).

Depending on the type of circuits and on their construction technology, the sensitivity to ionizing radiation can be very different. In order to select families of commercial circuits which are more insensitive than others it is necessary to run testing campaigns, comparing the behavior of several different chips when exposed to low energy ion beams. The type circuits which shows *latch up* sensitivity or abnormal behavior only at high Linear Energy Transfer (LET) are the one which are radiation hard and can be used in space. In order to ensure statistical significance of the radiation hardness measurement, several chips of each type should be tested (typically > 5).

Often, it is possible to protect the circuit by limiting the current which can flow through the power lines, by the use of an active switch which temporarily cut off the voltage to stop the *latch up* effect. In order to develop and implement a protection scheme it is then important to understand which area of the chip is sensitive. Nowadays it is also possible to perform in laboratory part of these test

by mean of IR laser beams which are absorbed by the silicon and can deposit a controlled amount of energy into the bulk simulating the charge released by a low energy ion [21].

All microelectronics components used in a space experiment must be radiation hard. In addition, the design of the on board electronics should include multiple redundancy since the radiation damage due to SEE is a stochastic process.

Space qualification of radiation resistant devices for a mission not only requires the understanding of damage mechanisms [22], but also the knowledge of local particles (and species) intensities [23, 24] and, in addition, of dose amounts, deposited via ionization and non-ionization energy loss (NIEL) processes. The latter mechanism is that one responsible for displacement damages particularly relevant for semiconductor devices. Only recently, the SR (screened relativistic) NIEL treatment framework has allowed a comprehensive calculation of NIEL doses imparted by electrons, protons, ions and neutrons in any material and compound [25]. SR-NIEL treatment is currently embedded in ESA transport codes, like GRAS [27] and MULASSIS [28] as well as in GEANT4 and it is available at the SR-NIEL and SPENVIS websites.

18.5 Design of a Space Particle Detector

Space born radiation detectors for a space application are, in most cases, adaptation to the space environment of detection techniques used at accelerators or nuclear laboratories.

The environmental conditions discussed in the previous sections obviously influence the detectors design. Particularly important examples are the design of a controlled temperature environment and, for gas detectors, the establishment of controlled pressure conditions.

However, a space born particle experiment has specific limits of different nature which are basically not existing in the case of a laboratory experiment. They are:

- $Weight$. Each kg transported in orbit is very expensive in terms of propellent, costing from 10.000 to 50.000€/kg, depending on the size of the satellite and orbit of deployment (larger satellites cost less than smaller satellites per kg, higher orbits cost more than lower orbits per kg). This is a substantial limitation for the size of a payload. In addition today space transportation systems have a maximum capacity of about 10 to 20 t in LEO;
- $Power$. The basic source of power in space is the solar energy transformed into electrical power by solar panels. The power can be accumulated in batteries for the periods of the orbit where the spacecraft is shadowed by the Earth. The amount of power consumed by a payload is thus proportional to the area of the panels. One kW of power in space is a large amount of energy consumption. For instance, the entire International Space Station (ISS) power capability does not exceed 110 kW.

- *Volume*. The largest transportation systems can carry payloads which must fit within a cylindrical volume having a maximum radius of about 3 m and a maximum length of about 10 m. Most particle detectors have much smaller sizes. Once in orbit, the size of the payload can increase very significantly, when solar panels, mirrors or radio antennas are expanded from the launch configuration.
- *Accessibility*. Because of the huge cost involved, most of the payload are not accessible during their lifetime in space. Very rare exceptions are the Hubble Space Telescope and the ISS. It follows that the reliability of the instrumentation is essential.
- *Consumables*. Due to the reasons listed above, the amount of consumables is limited. If consumables are needed, e.g. gas for a wire detector or cryogenics for a low temperature payload, the lifetime of the instrument will be limited. In orbit servicing is being developed nowadays for refurbishing the most expensive satellites, but it is still an emerging technology.

These limitations require the ingenuity of the scientist and the knowledge of the engineers to developed most advanced detectors within the available resources.

The reduction of weight calls for the most advanced techniques of CAD (Computer Aided Design) and FEA (Finite Element Analysis) to design structural elements which minimize the amount of material used while tolerating the mechanical stresses and shocks with margins of safety of 2 or more. The techniques used here are typical of aeronautics. The use of light advanced structural materials is mandatory e.g. aluminum, carbon fiber and in general composite materials. Once the structural properties are well defined, static and dynamic FEA is used to identify which part of the structure contribute to the weight without contributing to the structural properties. These parts are normally machined away during the construction. With the advent of Computer Additive Manufacturing (CAM) the weight optimization of structural elements and the integration of functional&structural elements is developing quickly to the advantage of the reduction of the mass of new payloads.

The reduction of power consumption calls for low power electronics and motors. The low power requirement is typical of consumer portable electronics. For this reason modern space experiments make extensive use of electrical devices (VLSI chips, actuators, motors,. . .) used in commercial applications. *Uprating* these parts to be used in space must be a part of the qualification process, in particular from the point of view of radiation hardness, which is not a requirement for consumer electronics. This approach of using $COTS$ (Component Off The Shelf) can reduce significantly the cost of a payload while producing very performant space instrumentation.

Due to the limited accessibility, reliability is a must in space born instrumentation. Reliability is the result of design, manufacturing, integration techniques which must be implemented since the early phases of the development of a payload. During the design phase, redundancy must be implemented in particular in the most critical areas. Special software allows, for example, to measure the probability of the failure of a given circuit, starting from the failure probability of its different components.

Typically the overall probability for a catastrophic failure must be in the range of 1% or less. Single point failures, namely parts of a circuit which are so critical that their failure would generate unacceptable level of malfunctioning, must be avoided. Redundancy of mission critical elements should at least be three to four fold. Similar techniques are applied to test the on board software, exploring all possible software states so to avoid unexpected software conditions which might degrade the payload performances. During the manufacturing phase and integration phases particular care should be given to Quality Assurance (QA), to ensure that the quality of the workmanship of the flight and qualification units and of the fully integrated payload matches the requirements of space standards and specifications. During the testing and qualification campaigns, all possible conditions to be encountered by the payload are simulated to make sure it will operate correctly under any circumstance. QA requires the operators to follow procedures written in advance, perform special tests and report all results and anomalies through written documents which can be verified and used by all the people involved in the various phases of development, commissioning and operation of the payload.

18.6 Space Borne Particle Detectors

The development of modern particle space borne detectors (both for charged particles and photons) has been preceded/accompanied by decades of development of particle detectors for ground based nuclear and particle physics detectors, followed by extended use on stratospheric balloons [29–38].

Small particles detectors have been routinely used on satellites mission to explore the Earth magnetosphere and heliosphere [39, 40].

Modern particle experiments in space can be grouped in three broad categories: (1) experiments measuring the composition, rates and energy spectra of the charged component, (2) experiments detecting single energetic photons and (3) interferometers designed to measure Gravitational Waves in space.

In the first category we find various types of magnetic spectrometers, in the second experiments are based on high granularity tracking calorimeters while the third category cover multiple arms laser interferometers. In the following paragraphs we will briefly discuss some of the most significant space particle detectors developed during the last 10 years, namely AMS-01/02 [41, 42] and PAMELA [43] for the charged component Agile [44] and Fermi [45, 46], for the electromagnetic component and LISA-PF[8] for measuring GW. We will underline the main differences with their ground based counterparts currently used at accelerators experiments. Details of the detection principles, readout electronics or on board software will not be given since they have been addressed in other chapters of this book.

18.6.1 Magnetic Spectrometers

The purpose of a space borne particle detector is to identify the basic properties of the charged cosmic radiation, namely its composition, the energy spectra of the various components and the corresponding fluxes. Thus, the components of a space born magnetic spectrometer are very similar to modern ground based spectrometers, namely:

- a magnet, permanent or superconducting, to measure the sign of the charge by bending the particles path;
- a precise tracking device to measure the particle signed rigidity ($R = B\rho = pc/Ze$), where B is the magnetic field and ρ is the radius of curvature;
- a scintillator based system to trigger the experiment and measure the Time of Flight;
- particle identification (ID) detectors like:
  - Transition Radiation Detectors (TRD) to separate e^+ and e^- from hadrons;
  - Cherenkov Ring Imaging detectors to measure the absolute value of the charge, Z, and the velocity;
  - Electromagnetic Calorimeters to identify the electromagnetic component within the cosmic radiation and measure its energy;
  - Neutron Counters to improve the calorimetric rejection the hadronic CR component.

The first magnetic spectrometers were flown on stratospheric balloons in the 80's. The magnets were based on superconducting coils. The magnets where switched on ground and operated cryogenically for a period of order of 1 day [29, 31–34]. Recently balloon cryostats were able to operate for order of few weeks making possible Long Duration Balloon flights (LDB) around the South Pole [35–38]. Pressurized stratospheric balloons are also beginning to operate Ultra Long Duration Flights (ULDB) which could eventually reach several months duration [90].

The first space borne large magnetic spectrometer, AMS-01 [41] was built only in the mid 90's, due to difficulty of developing a large magnet to be used in space. AMS-01 was the precursor flight of the AMS-02 spectrometer [42], approved by NASA to be flown to and operated on the international space station. (ISS): the engineering model, AMS-01, was operated during the 12 days Shuttle STS91 mission in June 1998 [47]. AMS-02, initially was based on a superconducting magnet to be installed on the ISS in the early 2000's, to be operated for about 3 years, namely for the estimated duration of the superfluid Helium consumable, with the possibility to be reflown after Helium refilling on Earth. The 2003 Challenger disaster forced the earlier retirement of the Shuttle fleet and a modification of the AMS-02 manifest: AMS-02 has been then flown to ISS in 2011 based again on a permanent magnet configuration to benefit of the longest possible exposure ensured by the ISS lifetime. In 2006 a smaller spectrometer, PAMELA [43] also based on a permanent magnet, was launched on a Resurs DK1 Russian satellite to operate in LEO.

One important difference between ground based or balloons magnetic spectrometers and the space borne version is related to the issue of the coupling between the Earth magnetic field and the magnet dipole moment. Since the payload attitude is not a relevant parameter for balloon spectrometers, superconducting magnets exhibiting significant dipole moment can be operated without problems. In space the situation is completely different: the magnetic coupling would affect the attitude of the entire satellite or platform, requiring continuous steering to keep a stable, outward looking attitude. It is then mandatory to design magnets having special geometries (see the following paragraph) and exhibiting minimal magnetic dipole moments.

18.7 Space Spectrometers Based on a Permanent Magnet

All space borne magnetic spectrometers which have been operated in space, AMS-01 (1998), Pamela (2006) and AMS-02 (2011) were based on permanent magnets.

18.7.1 The Alpha Magnetic Spectrometer on Its Precursor Flight (AMS-01)

AMS is an international project involving 16 countries an 56 institutes [42], operated under a NASA-DOE agreement, to install on the ISS a large magnetic spectrometer for the search of nuclear antimatter and to study the origin of dark matter. The first version of the spectrometer was built around a cylindrically shaped, permanent magnet having 800 mm of height and an inner diameter of 1115 mm, resulting in a geometrical acceptance of $0.82\,\mathrm{m}^2$ sr. Figure 18.9 shows the dimensions of the AMS-01 flight magnet. The magnet was made from 64 sectors. Each sector was composed of 100 $5 \times 5 \times 2.5\,\mathrm{cm}^3$ high grade NdFeB blocks. Figure 18.10 shows the arrangement of the field directions of the 64 sectors (left) and the resulting magnetic field map on the middle plane (right). This magnetic configuration is called *magic ring*, and ensures, theoretically, a small magnetic dipole field. To build this magnet the highest grade NdFeB available at the time was used, with an energy level of $(BH)_{max} = 50 \cdot 10^6$ GOe. This configuration resulted in an internal dipole field of 0.15 T and a negligible dipole moment. The total weight of the magnet including the support structure was 2.2 t. The magnetic field, directed orthogonally to the cylinder axis, provided an analyzing power of $BL^2 = 0.15\,\mathrm{Tm}^2$. Outside the magnet the field becomes less then 3–4 G anywhere at a distance larger than 2 m from the magnet center.

Before the construction of full scale magnets, many smaller magnets were built to confirm and measure the field inside the bore, the dipole moment and the flux leakage [41]. Three full scale magnets were built:

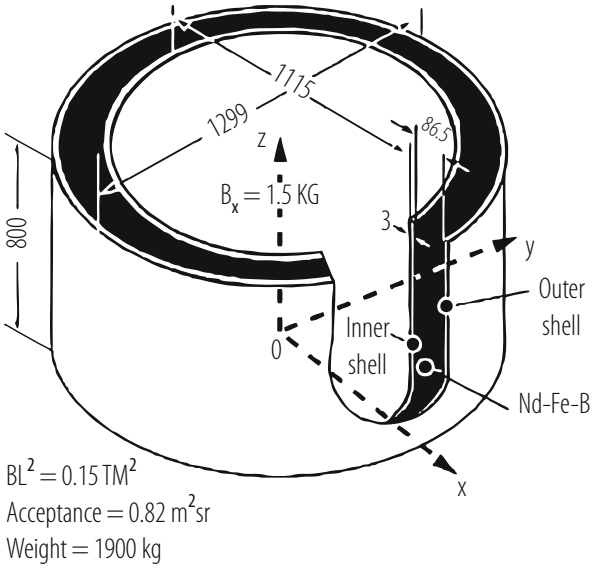

Fig. 18.9 Properties of the AMS-01 flight magnet (dimensions in mm)

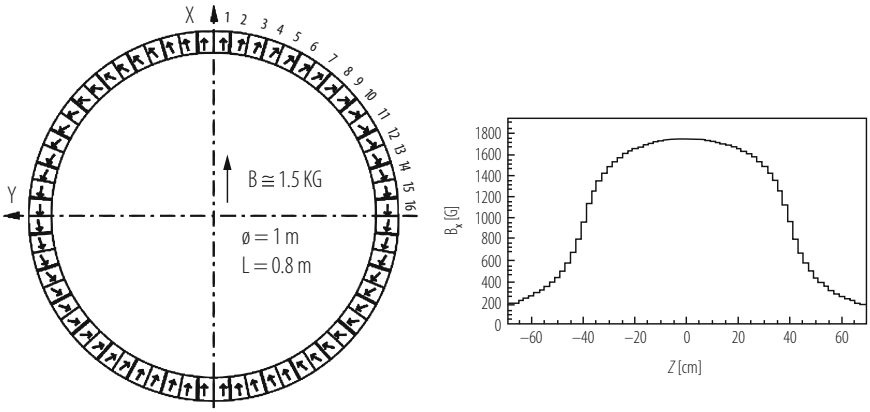

Fig. 18.10 Magnetic field orientation of the AMS-01 magnet sectors (left); B_x field map along the vertical axis ($x = 0$, $y = 0$) (right)

(a) The first magnet was used in acceleration and vibration tests for space qualification.
(b) The second magnet was the flight magnet.
(c) The third magnet was built without glue for NASA safety tests.

The magnet, the supporting structure and space qualification testing were completed by the Institute of Electrical Engineering [48] and the Chinese Academy of Launch Vehicle Technology ($CALT$) [49]. Figure 18.11 shows the first magnet

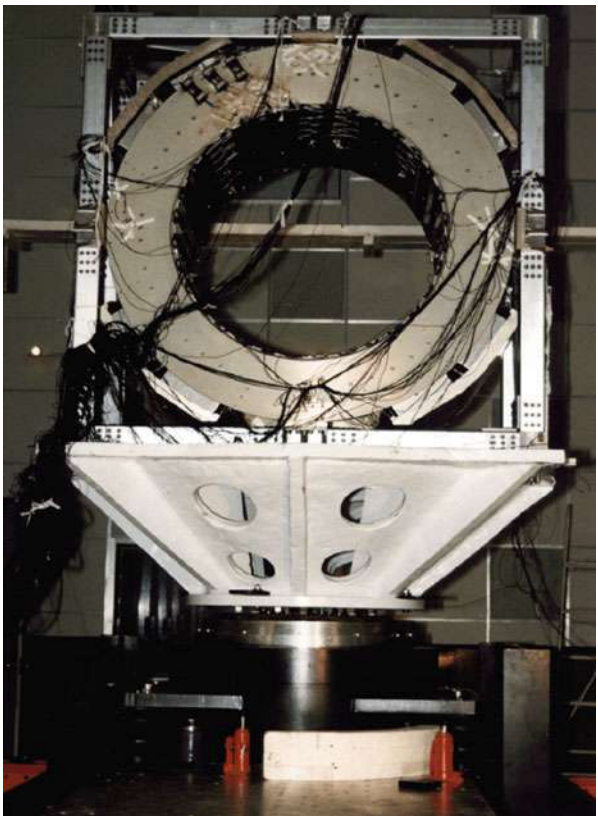

Fig. 18.11 AMS-01 magnet during vibration tests at the Beijing Institute of Spacecraft Environment and Engineering in Beijing, China

Fig. 18.12 AMS-01 magnet undergoing centrifuge (static load) testing at the Laboratory for Centrifugal Modeling in Beijing, China. The picture is blurred since it has been taking through a thick glass window

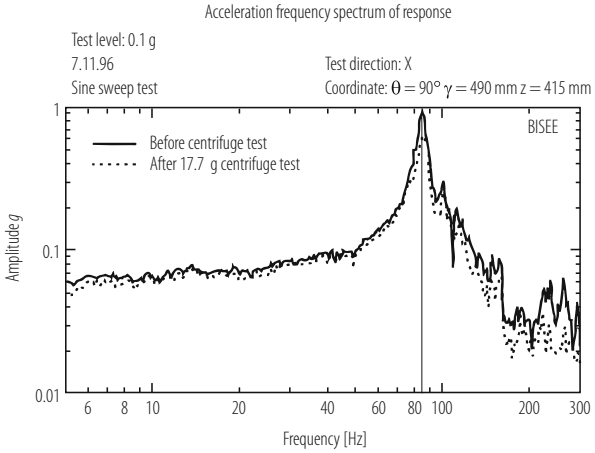

Fig. 18.13 Sine sweep test frequency spectrum response of AMS-01 magnet before and after 17.7 *g* centrifuge test

undergoing vibration testing. Figure 18.12 shows it undergoing centrifuge testing up to 17.7 *g*. Figure 18.13 shows the comparison of the sine sweep test results before and after the 17.7 *g* centrifuge test. The test results indicate that there is no deformation in the detector before and after this test and that the eigenfrequency for the magnet is above the ∼50 Hz region, where the spectral power of the random vibrations produced by shuttle is the highest, as imposed by the NASA safety requirements. The third full scale magnet was built because of the lack of knowledge of the glue performance over an extended period in the space environment. This magnet without any glue was to be tested to destruction to ensure that AMS could be returned on the Shuttle to Earth even if the glue completely failed. The result of the test shows that even with stresses 310 times higher than expected according to analysis the magnet would not break.

During spring of 2006 a smaller but sophisticated magnetic spectrometer, Pamela was launched from Baikonur on a Resource DK Rocket and inserted on a LEO for a 3 years mission. The Pamela experiment is built by an INFN-led international collaboration, and it was launched and operated under an Italian-Russian agreement. The magnet consists of 5 modules of permanent magnets, made of a sintered NdBFe alloy, interleaved by 6 silicon detector planes. The available cavity is 445 mm tall with a section of $1.31 \cdot 10^5\,\mathrm{mm}^2$, giving a geometrical factor of $20.5\,\mathrm{cm}^2$ sr. The mean magnetic field inside the cavity is 0.4 T, providing an analyzing power $BL^2 = 0.1\,\mathrm{Tm}^2$ resulting in a Maximum Detectable Rigidity of 740 GV/c, assuming a spatial resolution of 4 μm along the bending view [43]. The apparatus is 1.3 m high, has a mass of 470 kg and an average power consumption of 355 W. The layout of the magnet and the experiment is shown in Fig. 18.14.

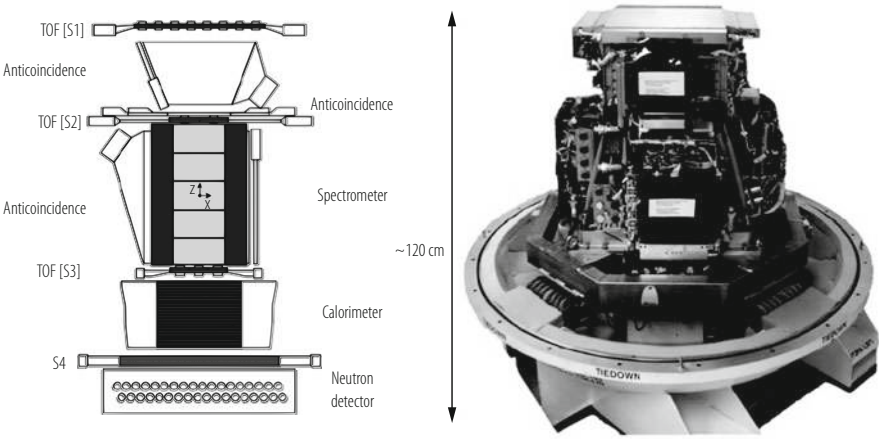

Fig. 18.14 Schematic lateral view of the PAMELA detector (left) and a photograph of it (right) taken before the delivery of the instrument for the integration to the Resours satellite. The geometrical acceptance of the detector is 20.5 cm^2 sr [63]

18.7.1.1 Superconducting Space Spectrometers

The sensitivity to new physics requires spectrometers able to explore higher Cosmic Ray energies while collecting large statistical samples. For this reason ground based modern spectrometers are routinely built using large superconducting magnets which measure particles with momenta in the multi-TeV range [50, 51]. It is of course much more difficult to design a superconducting magnet instead of a permanent magnet to be operated in space. Large facilities like the International Space Station could however provide the necessary infrastructure in terms of power, payload weight and size, data transfer and so on, to install and operate an superconducting spectrometer devoted to high energy particle physics in space. Already in the 80's a proposal was made to install on the Space Station a superconducting spectrometer, ASTROMAG [52]. ASTROMAG was designed around two parallel, large superconducting coils having opposite dipole moments, providing a highly non-uniform magnetic field but an almost zero residual dipole moment. The downsizing of the initial Alpha Station design which took place at the end of the 80's, put the ASTROMAG on indefinite hold status. In 1994 a new proposal was presented through DOE to NASA by the AMS Collaboration, to install and operate a large magnetic spectrometer on the ISS for at least 3 years. This proposal was based on a cylindrical magnetic geometry (*magic ring*), providing much more uniform magnetic field for the particle spectrometer and an almost zero magnetic dipole moment. After the successful flight of the AMS-01 permanent magnet in 1998, the AMS Collaboration proposed to DOE and NASA to upgrade the permanent magnet to a superconducting one having identical geometrical properties but an almost one order of magnitude stronger field (Fig. 18.15).

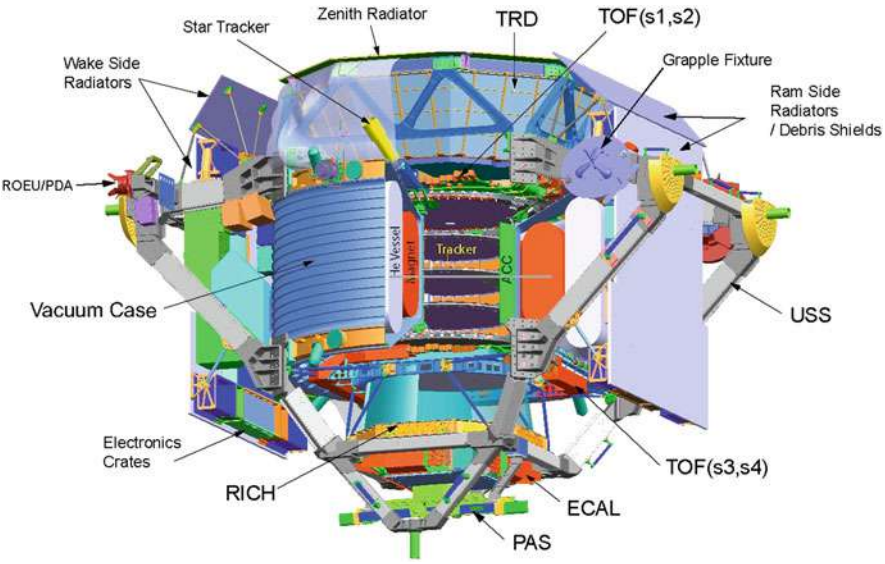

Fig. 18.15 Schematic 3D of the AMS spectrometer in the superconducting version

The project which developed during the years 2000–2010, consisted in the design, construction and extensive testing of the first space qualified superconducting magnet, including thermal-vacuum test in the large ESA-ESTEC space simulator in April 2010. In order to be compatible with the payloads designed for AMS-01, this magnet had identical inner dimensions to the AMS-01 permanent magnet, making the two magnet interchangable with the particle identification detectors. This fact has been instrumental to allow for the switch back to the permanent magnet when it became clear that the early retirement of Shuttle would not have allowed refilling of superfluid ^4He as initially planned. AMS-02 on a permanent magnet configuration has been largely benefitting of the longest possible exposure ensured by the ISS lifetime, which is particularly important in the search of ultrarare events (Fig. 18.16).

The AMS-02 superconducting magnet has been the first designed for operating in space. For this purpose a number of unique challenges had to be solved. Among them:

- endurance: how to maintain the magnet in the superconducting state for the longest possible time, of the order of 3 years, without cryogenic refill;
- safety: how to safely handle the large amount of energy (O(MJ)) stored in the magnet in case of a quench;
- mechanical stability: how to build a structure able to withstand large magnetic forces while being as light as possible.

Two magnets have been built. One is the flight magnet and the other is used for space qualification tests. The magnet system consists of superconducting coils, a

Fig. 18.16 The AMS-02 spectrometer during its integration at CERN in 2009 in the final flight configuration with the permanent magnet

superfluid helium vessel and a cryogenic system, all enclosed in a vacuum tank. The magnet operates at a temperature of 1.8 K, cooled by superfluid helium stored in the vessel. It was designed to be launched at the operating temperature, with the vessel full of 2600 liters of superfluid helium. Four cryocoolers operating between ∼300 and ∼80 K help to minimize the heat losses maximizing the endurance.

The magnet was designed to be launched with no field since it would be charged only after installation on the ISS. Because of parasitic heat loads, the helium will gradually boil away throughout the lifetime of the experiment. After a projected time of 3 years, the helium would be used up and the magnet would warm up and be no longer be operable. Three years of operation in space would indeed correspond to a continuum heat load into the superfluid Helium of about 100 mW, quite a small amount for a magnet which has a volume of about 14 cubic meters.

The coil system consists of a set of 14 superconducting coils arranged, as shown in Fig. 18.17, around the inner cylinder of the vacuum tank. The coil set has been designed to provide the maximum field in the appropriate direction inside the cylindrical bore, while minimizing the stray field outside the magnet. As a result, with the bore geometry identical to the geometry of the AMS-01 magnet, AMS-02 with the superconducting magnet would have had a field almost one order of magnitude larger. A single large pair of coils generates the magnetic dipole field perpendicular to the experiment axis. The twelve smaller flux return coils control the stray field and, with this geometry, they also contribute to the useful dipole field. The magnetic flux density at the geometric centre of the system is 0.73 T. The superconducting wire was developed specifically to meet the requirements of the AMS cryomagnet [53]. The current is carried by tiny (22.4 μm diameter) filaments of niobium titanium (NbTi) which are embedded in a copper matrix, which

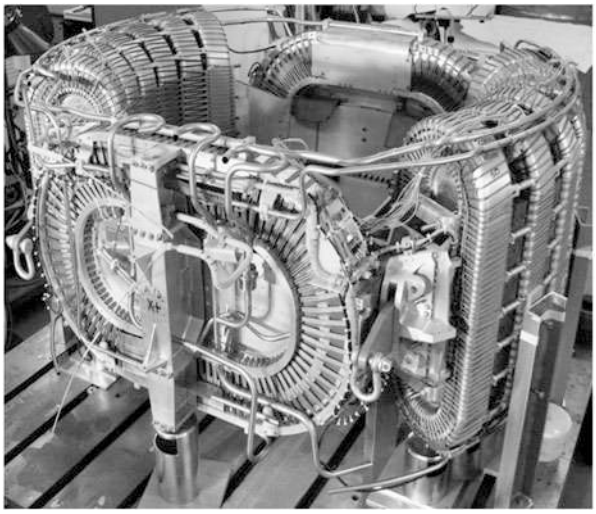

Fig. 18.17 The AMS-02 superconducting magnet: the dipole and the return coils are clearly visible, arranged in the characteristic cylindrical geometry

Table 18.4 AMS-02 superconducting magnet parameters

Parameter	Value
Central magnetic field B_x (at $x = y = z = 0$)	0.750 T
Dipole bending power	0.750 Tm^2
Maximum stray magnetic field at $R = 2.3$ m	13.2 mT
Maximum stray magnetic field at $Y = 2.3$ m	6.62 mT
Maximum stray magnetic field at $R = 3.0$ m	3.4 mT
Peak magnetic field on the dipole coils	5.75 T
Peak magnetic field on the racetrack coils	5.14 T
Maximum torque in geomagnetic field	0.237 Nm
Maximum stray magnetic field at $R = 3.0$ m	3.4 mT
Nominal operating magnet current	400 A
Stored energy	3.72 MJ
Nominal magnet inductance	48 H

is encased in high-purity aluminium. The copper is required for manufacturing reasons, but the aluminium is thermally highly conductive and much less dense, thus providing maximum thermal stability for the same weight. The characteristics of the AMS-02 superconducting magnet are listed in Table 18.4.

The current density in the superconductor is 2300 or 157 $\mathrm{A/mm}^2$ including the aluminium. The 14 coils are connected in series, with a single conductor joint between each pair of adjacent coils. The magnet is designed for a maximum current of 459.5 A, although it is operated at ∼85% of this value. The coils are not coupled thermally. All the coils are constantly monitored by an electronic protection system. If the onset of a quench is detected in any coil, heaters are powered in the other coils

to quench all 14 coils simultaneously. This distributes the stored energy between the coils, preventing any single coil from taking a disproportionate amount of energy which could otherwise result in degradation. The operation of these quench heaters is an important part of the testing and qualification procedure for the magnet coils.

This SC magnet is cooled by superfluid helium, since the thermal conductivity of the superfluid state is almost 6 orders of magnitude higher than in the normal state; in addition, the specific latent heat of the superfluid helium is higher than in normal liquid helium and this can also be used to extend the magnet operation time.

Safety of the AMS magnet had to be assured in ground handling operations, during launch, on orbit and during landing. All cryogenic volumes, as well as the vacuum tank, are protected by burst discs to prevent excessive pressures building up in any fault conditions. Some of the burst discs have to operate at temperatures below 2 K have been the subject of a special development and testing program. In addition, extra protection is provided to mitigate the effect of a catastrophic loss of vacuum. All parts of the AMS magnet system are subject to a battery of tests to ensure their quality, integrity and their suitability for the mission. Every one of the 14 superconducting coils have been tested before assembly into the final magnet configuration. A special test facility has been constructed which allows the coil to be operated under cryogenic conditions as close as possible to the launch. Tests have also been carried out on prototype burst discs. Discs for protecting the vacuum tank have undergone vibration testing followed by controlled bursts. These tests have shown that the discs are not affected by the levels of vibration encountered during a launch. Further tests have been carried out on discs for protecting the helium vessel, which operate at 1.8 K. These discs have been shown to have extremely good leak tightness against superfluid helium.

Mechanical tests of the qualification magnet were done at various facilities: study of the low frequency non-linear behavior were done on a special slip table set up at the SERMS Laboratory [54], in Italy, while static tests were done at IABG [55], in Germany, using a mechanically high fidelity replica of the AMS-02 experiment.

The main characteristics of the AMS01/02 and Pamela magnetic systems are listed in Table 18.5.

Table 18.5 Space borne magnets

Parameter	AMS-01/02	PAMELA	AMS-02[a]
Type of magnet	Permanent	Permanent	Superconducting
MDR [TV]	0.55	0.80	2.6
Magnetic field [T]	0.12	0.48	0.75
Dipole bending power [Tm^2]	0.12	0.085	0.75
Maximum torque in geomagnetic field [Nm]		0.0021	0.24
Maximum geometrical acceptance [cm^2 sr]	5000	20.5	5000

[a]Not deployed in space

Table 18.6 Space borne magnetic spectrometers

Particle ID	AMS-01	PAMELA	AMS-02
Transition radiation detector	No	No	Yes
Time of flight	Yes	Yes	Yes
Silicon tracker	Yes	Yes	Yes
Ring imaging Cherenkov	Yes	No	Yes
Electromagnetic calorimeter	No	Yes	Yes
Neutron counter	No	Yes	No

18.7.2 Particle Identification

High precision study of primary energetic Cosmic Rays requires reliable particle identification. Similar detectors to the one used at the accelerators have been developed and qualified for space usage. With respect to accelerators, however, the task of identification a given particle against its background is significantly different, since, at accelerators, the goal is mostly the identification of short lived particles, while in space short lived particles are irrelevant while the goal is the identification of stable particles and long lived isotopes.

Table 18.6 compare the properties of AMS01/02 and Pamela spectrometers.

18.7.2.1 Tracking Detectors

Silicon detectors, commonly used as tracking devices in ground-based accelerator experiments, offer the best resolution in terms of position measurement. However, a large scale application of these devices in space was never made before AMS-01 [56] in 1998. The AMS-02 silicon tracker [57] (Fig. 18.18) is composed by double-sided micro-strip sensors similar to those used for the L3 [58] micro-vertex detectors at the Large Electron-Positron collider (LEP) at CERN, but the technology and the assembly procedures were qualified for the operation in space. The silicon detectors were produced at Colibrys, SA Switzerland [59] and FBK-irst, Italy [60]. The silicon detectors are assembled together forming ladders up to 60 cm long: particular care was taken to control the readout noise produced by these large silicon assemblies, both from the point of view of the capacitive noise as well as from the point of view of the number of defects, which was requested to be below 10^{-3}. The tracker consists of 8 planes of silicon sensors providing 10 μm (30 μm) position resolution in the bending (non-bending) plane of the 0.15 T field of the magnet. The detectors measure both crossing position and energy loss of charged cosmic ray particles. The readout strips of the silicon sensors are ac-coupled to the low noise, high dynamic range, radiation hard, front-end readout chip, the version Hdr9A of the original Viking design, via 700 pF capacitor chips [61]. Once the charge is known, the momentum is determined by the coordinate measurements in the silicon, which are used to reconstruct the trajectory in the magnet field.

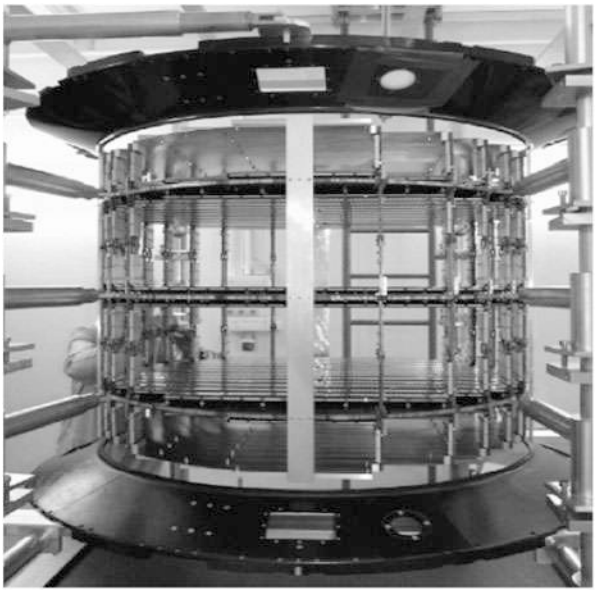

Fig. 18.18 The 8 layers Silicon Tracker of the AMS-02 experiment: the inner planes consists of three double layers of silicon detectors

A similar approach was followed by the PAMELA collaboration. Here the tracking device [62] was based on high accuracy double sided silicon microstrip detectors organized in 12 cm long silicon ladders, produced by Hamamatsu Photonics [64] while low noise, low power, VLSI VA1 chips were used for the front-end section. The use of low-noise front-end electronics is of great importance since the spatial resolution of the detector is strongly related to its signal-to-noise ratio. The applied position finding algorithm gives a spatial resolution of $2.9 \pm 0.1\,\mu\mathrm{m}$ [63]. The junction side shows a larger signal-to-noise ratio ($S/N = 49$) and a better spatial resolution. For this reason this side was used to measure the position along the bending view.

18.7.2.2 Time of Flight Detector

The Time-of-Flight (ToF) measurement is typically associated with the experiment trigger, and, in case of compact magnetic spectrometers, these detectors operates in presence of significant magnetic fields. Figure 18.19 show a schematics of the AMS-02 [65] ToF system, the largest of such systems built to date for space operation. This design follows the experience gained with the AMS-01 detector [66], modified to take into account the different conditions in AMS-02, in particular the stronger stray magnetic field at the photomultiplier tubes (PMTs) which can reach several hundred of G. Each scintillating paddle is instrumented with two PMTs at each

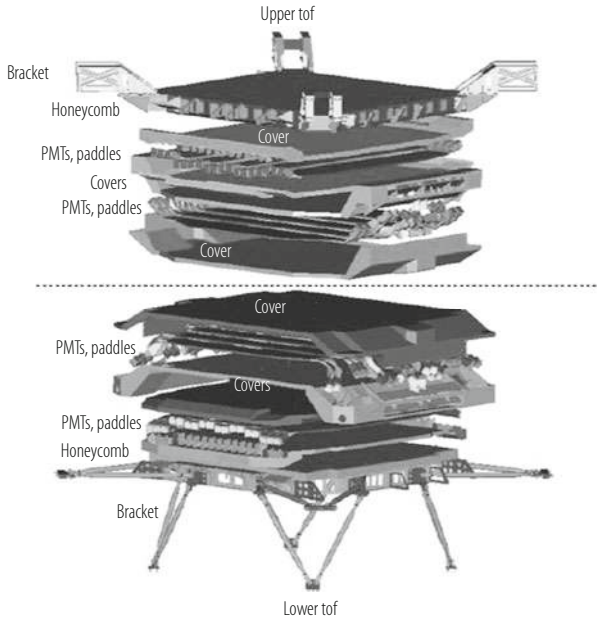

Fig. 18.19 Exploded view of the AMS02 ToF system

end. The time resolution needs to satisfy the physics requirements is 160 ps. The scintillator paddles are 1 cm thick, a compromise between minimum thickness and the light output needed to reach this resolution. Downward going charged particles are distinguished from upward going at the level of 10^9. The system measures the energy loss by a charged particle (to first order proportional to the square of the particle charge) with a resolution sufficient to distinguish nuclei up to charge $Z \sim 20$. Taking into account the attenuation along the counters and the need to have a good measurement of singly charge particles, a dynamic range of more than 10,000 in the measurement of the pulse height is required.

Each paddle is encased in a mechanically robust and light-tight cover and the support structure conforms to the NASA specifications concerning resistance to load and vibrations. The electronics withstands the highly ionizing low Earth orbit environment. Moreover the system guarantees redundancy, with two PMTs on each end of the paddles and double redundant electronics. The system can operate in vacuum over the temperature range -20 to $+50\,^{\circ}$C, it has a weight of less than 280 kg and a power consumption, including all electronics, lower than 170 W. System components have been qualified for use in space and have been extensively tested with particle beams.

18.7.2.3 Transition Radiation Detector

Because of their low mass, Transition Radiation Detectors (TRD) are well suited for utilization in primary Cosmic Ray experiments to separate leptons (electrons) from hadrons (protons) up to hundreds of GeV of energy. The principle of the TRD is very well understood and these detectors are used in large particle physics experiments like ATLAS [67] and ALICE [68] at CERN, and HERA-B at DESY [69]. However, $TRDs$ are gas based detectors and the new challenge is to operate such a large gas detector safely and reliably in space. This has been achieved in the design and construction of the large AMS-02 TRD [70]. The TR photons are detected in straw tubes, filled with a Xe:CO_2 (80%:20%) gas mixture and operated at 1600 V. With a probability of about 50% TR photons are produced in the radiator, 20 mm thick fleece located above each straw layer. Figure 18.20 shows the TRD on top of the magnet vacuum case. The gas tightness of the straw modules is the most critical design issue. The available supplies of gas, 49.5 kg of Xe and 4.5 kg of CO_2, will have to last for 3 years of operation. Using as standard conditions 1 bar and 298 K, this corresponds to 84201 of Xe and 25301 of CO_2. The CO_2 leak rate for one meter of straw-tube was measured to be $0.23 \cdot 10^{-6}$ l mbar/s with the TRD gas Xe:CO_2 80:20 mixture. This leak rate is attributed to diffusion through the straw walls. It corresponds to $1.85 \cdot 10^{-5}$ l mbar/s per module-meter or $9.3 \cdot 10^{-3}$ l mbar/s for the full TRD (500 module meters). A single polycarbonate end piece has a CO_2 leak rate of $0.9 \cdot 10^{-5}$ l mbar/s, for all 328*2 end pieces this totals to $5.9 \cdot 10^{-3}$ l mbar/s. Summing, the total TRD CO_2 leak rate of $1.5 \cdot 10^{-2}$ l mbar/s would correspond to a loss of CO_2 over 3 years of 287 l or a safety factor of 8.8 with respect to the CO_2 supply. This low leak rate has been verified on the completely integrated detector, which could then operate in space for about 26 years. Fabricated TRD modules are accepted if they have a leak rate better than a factor 4 with respect to the overall detector limit. This can only be assured by testing each of the 5248

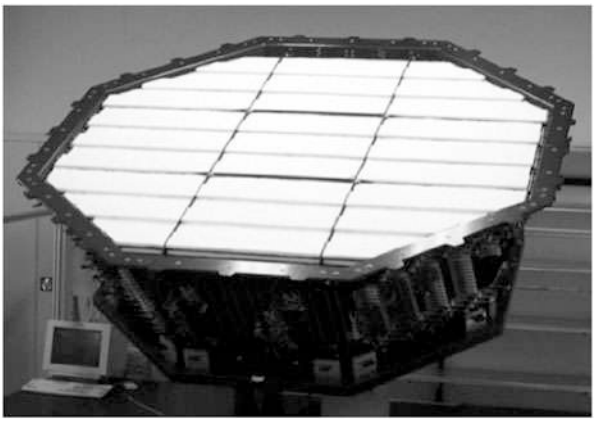

Fig. 18.20 The AMS02 Transition Radiation Detector (TRD) system

straws individually before producing a module [70]. The optimized AMS-02 TRD design with a diameter of 2.2 m and 5248 straw tubes arranged in 20 layers weighs less than 500 kg.

The thermal stability of the TRD is essential for the performance of the detector as temperature variations change the gas density and hence the gas gain. To keep these variations below the 5% level, comparable to other module to module inter-calibration uncertainties, temperature gradients within the TRD should not exceed $\pm 1\,^{\circ}\mathrm{C}$. To keep the spatial and temporal orbit temperature gradient below $1\,^{\circ}\mathrm{C}$ the TRD will be fully covered in multi-layer-insulation (MLI), including the front end electronics. Thermal simulations for orbit parameters which will give the highest TRD temperature swing have been done and prove the effectiveness of this approach. Nonetheless, this has been backed up by a full scale thermal vacuum test in the large volume space simulator at $ESA\ ESTEC$, Holland.

18.7.2.4 Ring Cherenkov Imaging Detector

Cherenkov light is very useful in measuring the velocity and the charge of particles up to tens of GeV of energy, providing a precise measurement to be used together with the momentum determination provided by the spectrometer to identify the different isotopes in the CR flux. The mass of a particle, m, is related to its momentum, p, and velocity, β, through the expression $m = (p/\beta)\sqrt{(1-\beta^2)}$ and its determination is based on the measurement of both quantities. In the AMS spectrometer, the momentum is determined from the information provided by the Silicon Tracker with a relative accuracy of 2% over a wide range of momenta. This entails an error of the same order on the mass of the particle if the velocity is measured with a relative accuracy of about 1 per mil: this is achieved by fitting the shape of the Cherenkov rings measured on the focal plane by high granularity ($4 \times 4\,\mathrm{mm}^2$) pixel photomultipliers located on the focal plane. For this purpose a Ring Imaging Cherenkov Detector (RICH) [71] has been designed with a large geometrical acceptance to operate in the environmental conditions of the outer space. The velocity is determined from the measurement of the opening angle of the Cherenkov cone produced within a radiator layer and the number of detected photons will provide an independent estimation of the charge of the incoming particle.

The measured distribution of charges in the beam is shown in Fig. 18.21 where the structure of individual ion peaks up to $Z = 26$ (Fe) is clearly visible (protons have been suppressed). This spectrum has been fitted to a sum of Gaussian distributions and from their widths we have estimated the charge resolution for each of the ions.

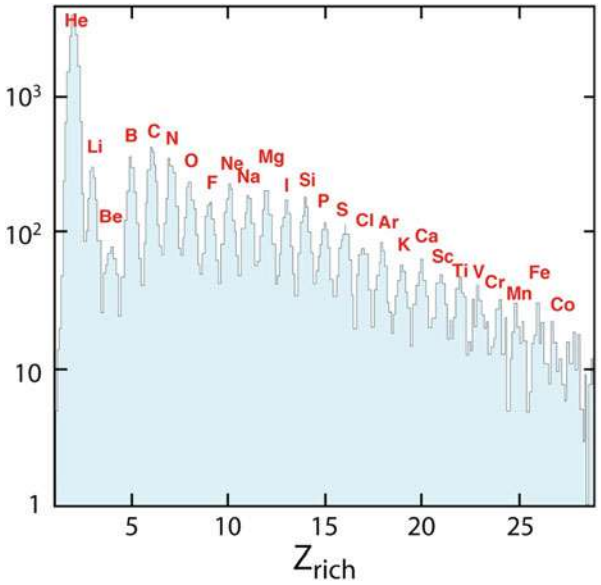

Fig. 18.21 Charge separation of the AMS02 Ring Imaging Cerenkov Detector (RICH) system

18.7.2.5 Electromagnetic Calorimeters

Protons and electrons dominate the positively and negatively charged components of CR, respectively. The main task of the calorimeter is helping the magnetic spectrometer to identify positrons and antiprotons from like-charged backgrounds which are significantly more abundant. Positrons must be identified from a background of protons that increases from about 10^3 times the positron component at 1 GeV/c to $5 \cdot 10^3$ times at 10 GeV/c, and antiprotons from a background of electrons that decreases from $5 \cdot 10^3$ times the antiproton component at 1 GeV/c to less than 10^2 times above 10 GeV/c.

The Electromagnetic Calorimeter ($ECAL$) of the AMS-02 experiment is a fine grained lead-scintillating fiber sampling calorimeter with a thickness corresponding to about 17 radiation lengths [72, 73]. This configuration allows precise, three-dimensional imaging of the longitudinal and lateral shower development, providing at the same time high ($>10^6$) electron/hadron discrimination in combination with the other AMS-02 detectors and good energy resolution, in the range $\sim$1 to $\sim$1000 GeV when the maximum of the e.m. shower is still within the calorimeter. The $ECAL$ also provides a standalone photon trigger capability to AMS. The mechanical assembly has met the challenges of supporting the intrinsically dense calorimeter during launch and landing with minimum weight. The light collection system and electronics are optimized for the calorimeter to measure electromagnetic particles over a wide energy range, from GeV up to TeV.

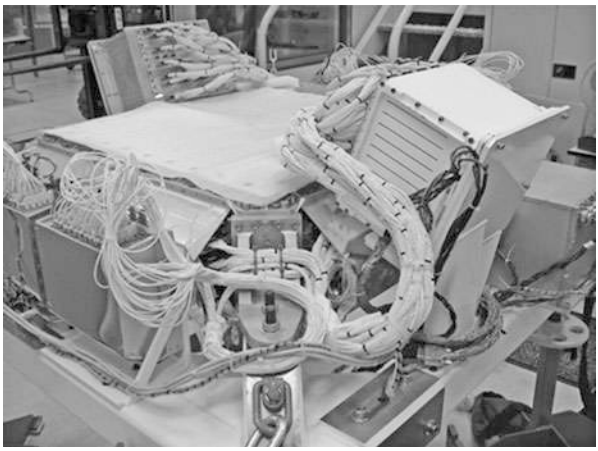

Fig. 18.22 The AMS02 Electromagnetic Calorimeter (ECAL) system

The calorimeter has a total weight of 496 kg. The ECAL mechanical assembly, shown in Fig. 18.22, supports the calorimeter, PMTs and attached electronics. It is designed to minimum weight with a first resonance frequency above 50 Hz, a capability to withstand accelerations up to 14 g in any direction and thermal insulation limiting the gradients (the external temperature ranges from −40 to +50 °C).

The PAMELA ECAL system is a sampling electromagnetic calorimeter comprising 44 single-sided silicon sensor planes (380 μm thick) interleaved with 22 plates of tungsten absorber [74]. Each tungsten layer has a thickness of 0.26 cm, which corresponds to 0.74 X_0 (radiation lengths), giving a total depth of 16.3 X_0 (0.6 nuclear interaction lengths). Each tungsten plate is sandwiched between two printed circuit boards upon which the silicon detectors, front-end electronics and ADCs are mounted. The $(8 \times 8)\,\mathrm{cm}^2$ silicon detectors are segmented into 32 read-out strips with a pitch of 2.4 mm. The silicon detectors are arranged in a 3×3 matrix and each of the 32 strips is bonded to the corresponding strip on the other two detectors in the same row (or column), thereby forming 24 cm long read-out strips. The orientation of the strips of two consecutive layers is orthogonal and therefore provides two-dimensional spatial information ($views$). Figure 18.23 shows the calorimeter prior to integration with the other PAMELA detectors.

More recently other space experiments based on fine grained calorimeters have been developed and are operating in space to study the spectrum of high energy electrons and positrons: CALET [75] on the Japanese segment of the ISS and Dampe [76] on a Chinese satellite.

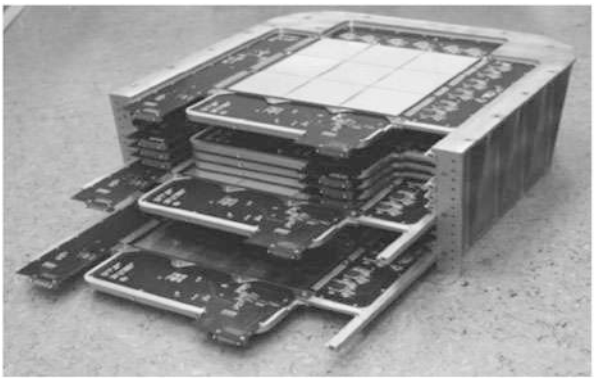

Fig. 18.23 The PAMELA electromagnetic calorimeter. The device is approximately 20 cm tall and the active silicon layer is about $24 \times 24\,\mathrm{cm}^2$. Some of the detecting planes are seen partially, or fully, inserted

18.8 Gamma Rays Detectors

During the last 30 years astrophysicists have discovered the high energy sky, namely sources emitting gamma rays with energy exceeding 1 MeV. The first space borne detector detecting MeV gamma rays were SAS-2 [77] and COS-B [78], followed by the EGRET instrument [79] which extended the energy range to hundreds of MeV with the Compton Gamma-Ray Observatory (CGRO) [80]. More recently Agile [44] and Fermi [45, 46] extended the energy reach to the GeV and hundreds of GeV scale, respectively, closing the gap with the ground based Cherenkov detectors operating from hundreds GeV to tens of TeV.

At these energies the quantized nature of photons is obvious and optical focusing is not anymore possible: high-energy gamma-rays cannot be reflected or refracted and they are detected by their conversion into an e^+e^- pair using techniques developed in nuclear and particle physics. Since both the gamma rays incoming direction and the energy are important informations, the instrument used are a combination of tracking and calorimetric detectors.

EGRET performed the first all-sky survey above 50 MeV and made breakthrough observations of high energy γ-ray blazars, pulsars, delayed emission from Gamma Ray Bursts ($GRBs$), high-energy solar flares, and diffuse radiation from our Galaxy and beyond that have all changed our view of the high-energy Universe. The EGRET instrument (Fig. 18.24), however, was based on detector technologies developed in the 80's: the tracking was provided by a streak chamber while the energy was measured with crystal based NaI calorimeter. In order to eliminate the background due to the charged CRs, about 10^5 times more frequent, the whole instrument was surrounded by a monolithic anti-coincidence counter. This design had two main limitations. First the limited operation time since the tracking device based on a consumable, the gas mixture. Second at increasing photon energy the anti

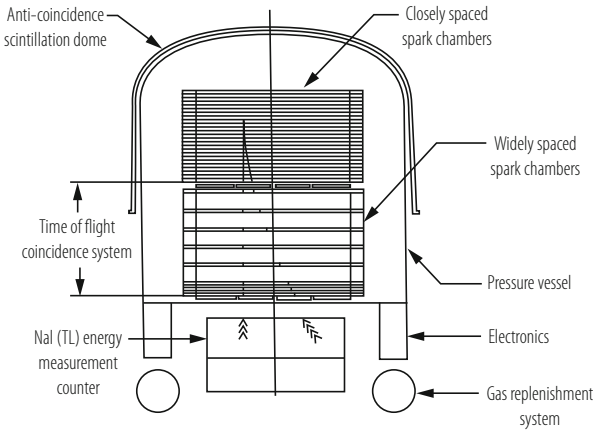

Fig. 18.24 Schematic view of the Energetic Gamma Ray Experiment Telescope (EGRET) on the Compton Gamma Ray Observatory (CGRO)

coincidence system was making the instrument increasingly inefficient due to back scattered particles created in the calorimetric section.

The follow up missions of EGRET, AGILE and Fermi, were based on modern technologies: in these payloads tracking is provided by solid state, imaging calorimeters based on silicon detectors, while the veto system is segmented in several sub elements suitably interconnected within the trigger electronics.

AGILE is a small mission of the Italian Space Agency (ASI), which was launched in April 23rd, 2007. The detector consists on an imaging silicon calorimeter, followed by a thin crystal calorimeter and covered by a coded mask layer to image hard X-rays sources. Its main parameters are listed in Table 18.7.

The Large Area Telescope (LAT) on the Fermi Gamma-ray Space Telescope (Fermi), see Fig. 18.25, formerly the Gamma-ray Large Area Space Telescope (GLAST), was launched by NASA on June 11th, 2008. The LAT is a pair-conversion, high granularity, silicon based imaging telescope made of 16 adjacent towers, followed by an electromagnetic crystal calorimeter. Some of the design choices of Fermi are similar to AGILE, although the detector geometric factor is much larger: each of the 16 Fermi imaging calorimetric towers is equivalent to the whole area of the AGILE detector. In addition the crystal calorimeter section of Fermi is much thicker, providing a much better energy determination. Table 18.8 shows the parameters of the Large Area Telescope instrument.

The self-triggering capability of the LAT tracker is an important new feature of the LAT design made possible by the choice of silicon-strip detectors, which do not require an external trigger, for the active elements [45, 46]. This feature is of essence for the detection of gamma rays in space. In addition, all of the LAT instrument subsystems utilize technologies that do not use consumables such as gas. Upon triggering, the DAQ initiates the read out of these 3 subsystems and utilizes on-board event processing to reduce the rate of events transmitted to ground to a rate

Table 18.7 Agile instrument parameters

Parameter	Value or range
Gamma-ray imaging detector (GRID)	
Energy range	30 MeV–50 GeV
Field of view	~ 2.5 sr
Flux sensitivity ($E > 100$ MeV, 5σ in 10^6 s)	$3 \cdot 10^7$ ph cm^{-2} s^{-1}
Angular resolution	
At 100 MeV (68% cont. radius)	3.5°
At 400 MeV (68% cont. radius)	1.2°
Source location accuracy (high Gal. lat., 90% C.L.)	15 arcmin
Energy resolution (at 400 MeV)	$\Delta E/E \sim 1$
Absolute time resolution	2 μs
Deadtime	~100–200 μs
Mini-calorimeter	
Energy range	0.35–50 MeV
Energy resolution (at 1.3 MeV)	13% FWHM
Absolute time resolution	~3 μs
Deadtime (for each of the 30 CsI bars)	~20 μs

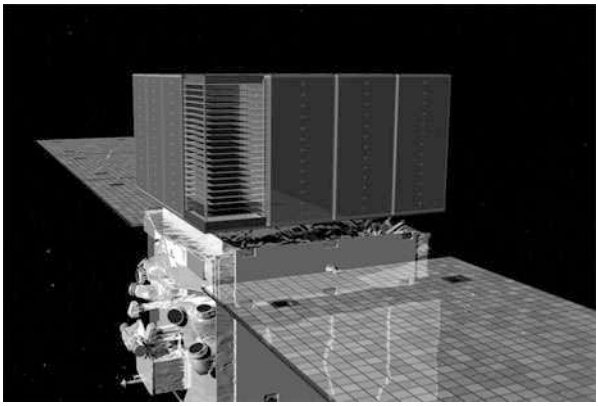

Fig. 18.25 The Fermi Large Area Telescope

compatible with the 1 Mbps average downlink available to the LAT. The on-board processing is optimized for rejecting events triggered by cosmic-ray background particles while maximizing the number of events triggered by gamma-rays, which are transmitted to the ground. Heat produced by the tracker, calorimeter and DAQ electronics is transferred to radiators through heat pipes. The overall aspect ratio of the LAT tracker (height/width) is 0.4, allowing a large field of view and ensuring that nearly all pair conversion events initiated in the tracker will pass into the calorimeter for energy measurement.

Table 18.8 Fermi Large Area Telescope (LAT) parameters [45]

Parameter	Value or range
Energy range	20 MeV–300 GeV
Effective area at normal incidence	9.500 cm^2
Energy resolution (equivalent Gaussian 1σ)	
100 MeV–1 GeV (on axis)	9–15%
1–10 GeV (on axis)	8–9%
10–300 GeV (on-axis)	8.5–18%
>10 GeV (>60° incidence)	≤6%
Single photon angular resolution (space angle)	
On-axis, 68% containment radius	
>10 GeV	≤0.15°
1 GeV	0.6°
100 MeV	3.5°
On-axis, 95% containment radius	$<3 \times \theta_{68\%}$
Off-axis containment radius at 55°	<1.7×(on-axis value)
Field of View (FoV)	2.4 sr
Timing accuracy	<10 μs
Event readout time (dead time)	26.5 μs
GRB location accuracy on-board	$<10'$
GRB notification time to spacecraft	<5 s
Point source location determination	$<0.5'$
Point source sensitivity (>100 MeV)	$3 \cdot 10^{-9}$ ph cm^{-2} s^{-1}

18.9 Gravitational Waves Detectors

Gravitational Waves (GW) are the analogous of the electromagnetic waves for gravitation. They propagate at the speed of light temporarily deforming the texture of space time. Predicted by Albert Einstein [4] on the basis of his theory of General Relativity, gravitational waves transport energy as gravitational radiation, and have been discovered exactly 100 years later by ground based laser interferometers [3]. They are emitted by massive bodies undergoing acceleration. A two body orbiting system, with masses m_1 and m_2, emits a power P:

$$P = \frac{dE}{dt} = -\frac{32}{5}\frac{G^4}{c^5}\frac{(m_1 m_2)^2(m_1 + m_2)}{r^5}. \tag{18.2}$$

Emitted power is really small in most gravitating systems. For example, in the case of the Sun–Earth system, it amounts to about 200 W, about $5 \cdot 10^{-25}$ times less that the electromagnetic power emitted by our star. The GW spectrum extends from frequencies corresponding to the inverse of the age of the universe to few hundreds of Hz (Fig. 18.26).

Their detection has only recently been demonstrated on ground but there are solid reasons to believe that the S/N ratio will be much larger for space borne

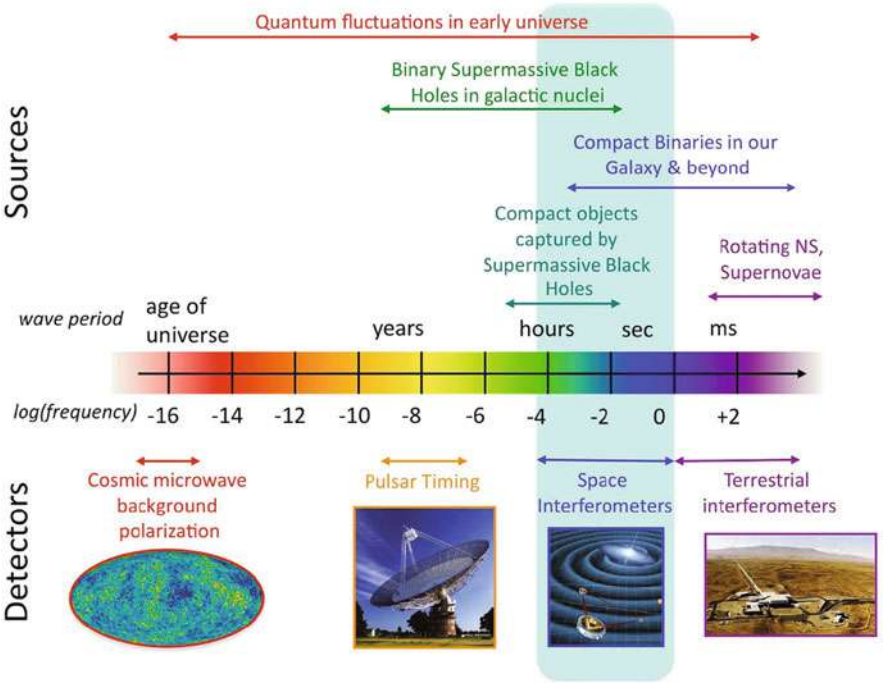

Fig. 18.26 Gravitational wave spectrum and detection techniques

interferometers. LISA is a three-arm space interferometer studied by ESA and NASA up to formulation level for more than 10 years. With the success of the LISA Pathfinder experiment [8], ESA is on track to develop LISA [9] which could be operational towards the beginning of the 30's and detect signals coming from supermassive black-hole mergers, compact objects captured by supermassive black holes and compact binaries (Fig. 18.27).

Once deployed, LISA would measure (a) the orbital period of the binary system, (b) the chirp mass $M = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}$, discriminating between white dwarf, neutron star and black hole binaries and determining the distance for most binary sources with an accuracies better than 1%.

18.9.1 Space-Borne GW Detectors

Measurement of space-time curvature using light beams requires an emitter and a receiver which are perfectly free falling. In flat space-time, the length of proper time between two light-wave crests is the same for the emitter and for the receiver. GW curvature gives oscillating relative *acceleration* to local inertial frames if wave-

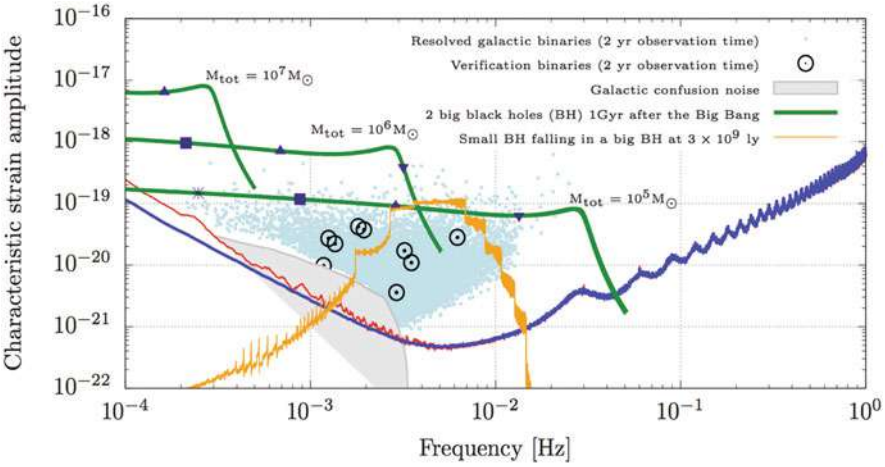

Fig. 18.27 LISA sensitivity to gravitational waves

front is used as a reference: it follows that the receiver sees frequency oscillating. Acceleration of receiver and/or emitter relative to their respective inertial frame produces the same effect of a curvature and should be carefully avoided.

In order to detect gravitational waves via the slowly-oscillating (T up to hours), relative motion they impose onto far apart free bodies, one needs (a) an instrument to detect tiny oscillations, of the size on atom peak-to-peak, ensuring (b) that only gravitational waves can put your test-bodies into oscillation and (c) eliminating all other forces above the weight of a bacteria.

The motion detector (a) is provided by a laser interferometer, as for ground based GW detectors, detecting relative velocities by measuring the Doppler effect through the interferometric pattern variation. Using very stable laser light one can reach the accuracy of 1 atom size in 1 h.

The free falling bodies (b) cannot be touched or supported, at least in an ordinary way. They must be shielded against all other forces (c), in particular, one needs to suppress gravity of the Earth (and of the Sun). The gravity force can be turned off by falling with it, a condition achievable for long periods only on an orbiting satellite. For all other forces, the satellite body would neutralize solar radiation and plasma pressure, actively and precisely following the test mass inertial motion. In order to ensure non contacting (drag-free) behavior, the spacecraft position relative to the test mass is measured by a local interferometer, and it is kept centered on the test mass by acting on micro-Newton thrusters.

The specifications of the LISA GW interferometer design are

- LISA
  - 3 arms, each 5 Mkm
  - $10\,pm/\sqrt{Hz}$ single-link interferometry @1 mHz

– Forces (per unit mass) on test masses $<3\,fm/(s^2\sqrt{Hz})$ @ 0.1 mHz
– 3 non-contacting ("drag-free") satellites

A basic concept of LISA is that the satellites follow independent heliocentric orbits and no formation keeping is needed. In addition the three satellites constellation rotates with respect to the fixed stars providing gravitational waves source location. In the case of the LISA instrument, the implementation of the requirements (a)–(c) is provided by the following main elements:

- the Gravitational Reference Sensor (GRS) with the test mass (also called Inertial Sensor): the GRS is drag-free along sensitive direction, while the other degrees of freedom are controlled via electrostatic forces through a 3–4 mm clearance between test mass and electrodes (Fig. 18.28);
- the Optical Bench with the complete interferometry: it carries all needed interferometry on a monolithic ultra-stable structure obtained by silica hydroxyl bonding (Fig. 18.29);
- a telescope allowing to exchange light with other satellites.

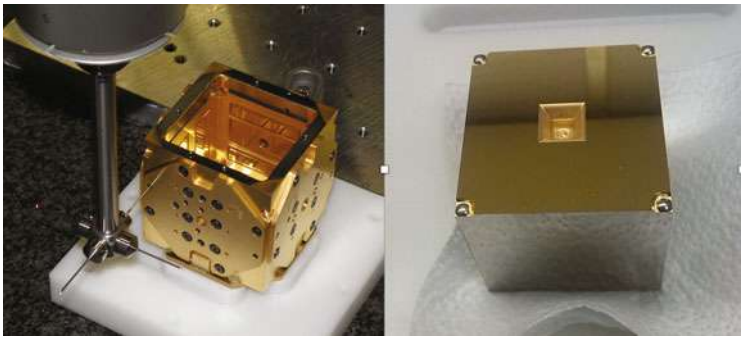

Fig. 18.28 The GRS; left: reference mass housing, right: reference mass

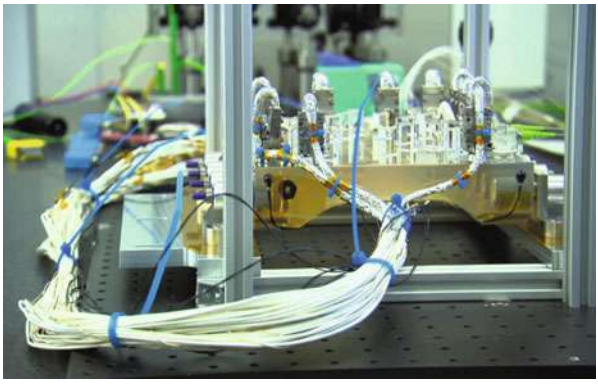

Fig. 18.29 LISA-pathfinder optical bench

18.9.2 LISA Pathfinder

In order to test in space most of the techniques needed for a LISA class space interferometer, the LISA Pathfinder mission has been built, launched in December 3rd 2015 and successfully operated in space during about 8 months, starting from March 1st 2016.

The LISA Pathfinder is based on squeezing of one arm of the final interferometer to within a $O(1)$ m optical bench. This was implemented by removing the long-arm interferometer and replacing the long-arm laser beam reference with a second (quasi-) free test mass. In this miniature implementation of one LISA arm two Au-Pt test masses and two interferometers were placed on the same optical bench. The two masses were not contacting the satellite but the second test mass was forced to follow the first at very low frequency by electrostatic forces (this is different from LISA).

LISA Pathfinder can be seen as a remotely controlled gravitational laboratory operating in space conditions. The GRS consists in two light test masses (2 kg, 46 mm) with a very high density homogeneity ($<<$ 1 μm pores), so that the position of the CoG at geometrical center is known within ± 2 μm. It has a very low magnetic susceptibility $\chi = -(2.3 \pm 0.2) \cdot 10^{-5}$ as well as a negligible magnetic moment $< 4\,\mathrm{nAm}^2$.

Many subtle physical effects apply unwanted forces to test-bodies [81], such as:

- impact with the few molecules that still surround the bodies in high vacuum [82, 83];
- spontaneous electric fields generated by surrounding bodies;
- fluctuating electrical charge from cosmic rays [84];
- changing gravitation generated by thermal deformation of satellite;
- impact with wandering photons;
- fluctuations of the interplanetary magnetic field;

These effects have been studied over the years in the laboratory, pushing forward knowledge in different fields of physics. The results published by the LISA-PF [8] shows that the mission has been very successful, exceeding the predicted accuracy and demonstrating that sub-femto-g differential accelerometry can be achieved, which is an improvement of orders of magnitude with respect to sensors used in the field of experimental gravitation. LISA-PF results confirm the projected LISA sensitivity to the bulk of GW sources present in our galaxy (blue line in Fig. 18.27): a green light for an ESA LISA class mission which could start operating at the beginning of the 30's.

18.10 Future Space Experiments

During the last 20 years an increasing number of modern experiments devoted to particle physics in space have been developed, providing a wealth of new data about CRs composition, high energy astrophysics and gravitational waves. The success of these programs opens the way to the proposal of new, more ambitious projects, designed to measure more accurately the properties of the cosmic radiation.

The universe contains the most powerful particle accelerators, able to accelerate particles to energies inaccessible to ground based laboratories. However these accelerators are quite inefficient and the differential flux of these energetic particles decreases quickly, typically with the third power of the energy. Above a few TeV for the charged component and few hundred GeV for gamma rays, it becomes impractical to develop space instruments having a sufficiently large geometric aperture. For this reason space scientists are considering experiments where the medium where the particle interactions take place is separated from the detector, similarly to what happens for ground based Cherenkov Telescopes, under water or under ice neutrino detectors or Extremely Energetic Cosmic Rays detector arrays, where Cerenkov and fluorescence light produced in the atmosphere, water or ice, respectively, is measured using photon detectors. In the case of these space experiments the medium could be the atmosphere [85, 86], the Moon surface [87] or the magnetosphere [88]: extremely large sensitivities to rare events can be reached by using our whole planet, the Earth, or its satellite, the Moon, as detecting media observable from space borne detectors, collecting emitted light or radio waves by using suitable instrumentation. Discussing these projects is outside the scope of this chapter, however it is interesting to note here a pattern of development which might in the future drive the development of space borne particle experiments devoted to extremely rare events.

18.11 Balloons Experiments

For nearly 40 years, until the mid of the 90's, experiments on stratospheric balloons have been instrumental to study primary CR composition. The advantage of balloons experiments over space experiment is a much lower cost, in the range of 10 M€/mission or less. The main disadvantage is the limited duration of the mission: in the early days it was limited to a day or two, while with the advent of circumpolar flights, the duration has increased to nearly a month/mission. NASA is developing a pressurized balloons technology which would allow for Ultra Long Duration Balloon missions (ULDB) [89, 90] which would reach several months of operation at stratospheric altitudes. In the meantime balloons demonstrated the ability to operate payloads weighting in excess of 1 t, powered by solar panels. It is quite clear that stratospheric balloons missions will be complementary and may become competitive to space missions, in particular when they will last for

several months close to the top of the atmosphere. Most considerations concerning detector developments are quite similar to what has been discussed for space missions: experiments must operate at extreme temperature conditions, withstand shocks, minimize weight and power consumption. Balloons payloads operates in an atmospheric environment, although very rarefied: thermal properties and design should be optimized taking into account also the convective contribution to heat transfer.

References

1. Hess, V., *Über Beobachtungen der durchdringenden Strahlung bei Sieben Freiballonfahrten*, Phys. Z. 13 (1912) 1084.
2. Van Allen, J.A., Ludwig, G.H., Ray, E.C., McIlwain, C.E., *Observations of high intensity radiation by satellites 1958 Alpha and Gamma*, Jet Propulsion 28 (1958) 588–592.
3. Abbott, B. P. et al., *Observation of Gravitational Waves from a Binary Black Hole Merger*, Phys. Rev. Lett. 116 (2016) 061102.
4. Einstein, A.,*Die Feldgleichungen der Gravitation*, Sitzungsberichte der Preussischen Akademie der Wissenschaften zu Berlin, 844–847 (1915).
5. Weber, J., *Gravitational-Wave-Detector Events*, Phys. Rev. Lett. 20 (1968) 1307.
6. Barish, B. C., Weiss, R., *LIGO and the Detection of Gravitational Waves*, Physics Today. 52 (1999) 10.
7. Acernese F. et al., *Advanced Virgo: a second-generation interferometric gravitational wave detector*, Classical and Quantum Gravity, Volume 32, Number 2 (2014).
8. Armano, M. et al., *Sub-Femto-g Free Fall for Space-Based Gravitational Wave Observatories: LISA Pathfinder Results*, P.R.L 116 (2016) 231101.
9. LISA Consortium, *LISA: Laser Interferometer Space Antena*, 20 January 2017.
10. Visentine, J.T. (ed.), *Atomic Oxygen Effects Measurements for Shuttle Missions STS-8 and 41-G*, Vols. I-III, NASA TM-100459 (1988).
11. Leger, L.J., Visentine J.T., Kuminecz, J.F., *Low Earth Orbit Oxygen Effects on Surfaces*, AIAA 22nd Aerospace Sciences Meeting, Reno, NV, Jan. 9–12, 1984.
12. Wertz, J.R., Larson, W.J. (eds.), *Space Mission Analysis and Design*, Microcosm Press and Kluwer Academic Publisher (1999).
13. Gussenhoven, M.S., Hardy, D.A., Rich, F., Burke, W.J., Yeh, H.C., *High Level Spacecraft, Charging in the Low-Altitude Polar Auroral Environment*, J. Geophys. Res. 90 (1985) 11009.
14. Fennel, J.F., Koons, H.C., Leung, M.S., Mizera, P.F., *A Review of SCATHA Satellite Results: Charging and Discharging*, ESA SP-198, Noordwijk, The Netherlands (1983).
15. Purvis, C.K., Garrett, H.B., Witthlesey, A.C., Stevens, N.J., *Design Guidelines for Assessing and Controlling Spacecraft Charging Effects*, NASA Technical Paper 2361 (1984).
16. Vampola, A.L., *The Nature of Bulk Charging and Its Mitigation in Spacecraft Design*, paper presented at WESCON, Anaheim, CA, Oct. 22–24, 1996.
17. Walt, M., *Introduction to Geomagnetically Trapped Radiation*, Cambridge University Press (1994).
18. McIlwain, C.E., *Coordinates for Mapping the Distribution of Magnetically Trapped Particles*, J. Geophys. Res. 66 (1961) 3681–3691.
19. Cervelli, F., et al., *The space qualified read-out electronics for the e.m. calorimeter (ECAL) of the AMS-02 experiment*, IEEE, TNS-00184-2009.
20. National Space Science Data Center web site: http://nssdc.gsfc.nasa.gov/.
21. Alpat, B., et al., *A pulsed nanosecond IR laser diode system to automatically test the Single Event Effects in Laboratory*, Nucl. Instrum. Meth. A 485 (2002) 183–187.

22. see Chapter 3 on *Managing Space Radiation Risk in the New Era of Space Exploration*, Committee on the Evaluation of Radiation Shielding for Space Exploration of the Aeronautics and Space Engineering Board (National Research Council, USA), National Academies Press, Washington DC (2008), ISBN 9780309113830.
23. SPENVIS, *ESA's SPace ENVironment Information System* (2018), available at https://www.spenvis.oma.be/.
24. OMERE software (2018), *Outil de Modèlisation de l'Environnement Radiatif Externe*, the code is developed by TRAD with the support of the CNES and is available at http://www.trad.fr/en/space/omere-sotftware/.
25. SR-NIEL Calculator: *Screened Relativistic (SR) Treatment for Calculating the Displacement Damage and Nuclear Stopping Powers for Electrons, Protons, Light- and Heavy- Ions in Materials* by Boschini, M.J., Rancoita, P.-G. and Tacconi, M., current version 3.9.3 (October 2017) is available at http://www.sr-niel.org/; the treatment can be comprehensively found in Chapters 2, 7 and 11 of [26].
26. Leroy, C. and Rancoita, P.-G., *Principles of Radiation Interaction in Matter and Detection* 4th Edition, World Scientific (Singapore) 2016, ISBN 9789814603188.
27. GRAS *(Geant4 Radiation Analysis for Space)* code is available at ESA website upon registration; the original article is by Santin, G., Ivanchenko, V., Evans, H., Nieminen, P. and Daly, E., IEEE Trans. Nucl. Sci. 52, Issue 6, 2005, pp 2294–2299.
28. MULASSIS - *MUlti-LAyered Shielding SImulation Software, available at ESA website*: the original article is by Lei, F., Truscott, R.R., Dyer, C.S., Quaghebeur, B., Heynderickx, D., Nieminen, P., Evans, H. and Daly, E., IEEE Transactions on Nuclear Science Vol 49 No 6 (2002) P2788–2793.
29. Streitmatter, R.E., *ISOMAX: A Balloon-borne Instrument to Study Berlyllium and Other Light Isotopes in the Cosmic Radiation*, Proc. 23th Int. Cosmic Ray Conf., Calgary 1993.
30. Mitchell, J.W., et al., *(IMAX) Isotope Matter-Antimatter Experiment*, Proc. 23rd Int. Cosmic Ray Conf., Calgary 1993, Vol. 1, p. 519.
31. Carlson, P., Francke, T., Suffert, M., Weber, N., *A RICH counter for antimatter and isotope identification in the cosmic radiation*, Proc. 23th Int. Cosmic Ray Conf., Calgary 1993, Vol. 2, p. 504.
32. Yamamoto, A., et al., *Balloon-Borne Experiment with a Superconducting Solenoidal Magnet Spectrometer*, Adv. Space Res. 14(2) (1994) 75–87.
33. Barwick, S.W., et al., *The High-Energy Antimatter Telescope (HEAT): an instrument for the study of cosmic-ray positrons*, Nucl. Instrum. Meth. A 400 (1997) 34–52.
34. Beatty, J.J., et al., *Cosmic Ray Energetics And Mass (CREAM): A Detector for Cosmic Rays near the Knee*, Proc. 26th Int. Cosmic Ray Conf., Salt Lake City 1999, Vol. 5, pp. 61–64.
35. Isbert, J., et al., *ATIC, a Balloon Borne Calorimeter for Cosmic Ray Measurements*, Proc. 10th Int. Conf. Calorimetry in Particle Physics, Pasadena, CA, March 25, 2002, pp. 89–94.
36. Boyle, P., et al., *Cosmic Ray Energy Spectra of Primary Nuclei from Oxygen to Iron: Results from the TRACER 2003 LDB Flight*, 30th Int. Cosmic Ray Conf., Merida, Mexico (2007).
37. Yoshimura, K., et al., *The First BESS-Polar Flight over Antarctica*, Proc. 25th Int. Symp. Space Technology and Science, Kanazawa, Japan (2006), pp. 1132–1137.
38. Seo, E.S., et al., *CREAM: 70 days of flight from 2 launches in Antarctica*, Advances in Space Research 42 (2008) 1656–1663.
39. Baker, D.N., Mason, G.M., Figueroa, O., Colon, G., Watzin, J.G., Aleman, R.M., *An Overview of the Solar, Anomalous, and Magnetospheric Particle Explorer (SAMPEX) Mission*, IEEE Trans. Geosci. Remote Sens. 31 (1993) 531–541.
40. ESA's Report to the 30th COSPAR Meeting, Hamburg, Germany, July 1994, European Space Agency, Paris, (1992) 47–57.
41. Ahlen, S.P., et al., *An Antimatter spectrometer in space*, Nucl. Instrum. Meth. A 350 (1994) 351–367.
42. AMS Collaboration, Aguilar, M., et al., *The Anti Matter Spectrometer (AMS-02): A particle physics detector in space*, Nucl. Phys. Proc. Suppl. 166 (2007) 19–29.

43. Bonvicini, V., et al., *The PAMELA experiment in space*, Nucl. Instrum. Meth. A 461 (2001) 262–268.
44. Tavani, M., et al., Astron. Astrophys. 502 (2009) 995.
45. Atwood, W.B., et al., *The Large Area Telescope on the Fermi Gamma-ray Space Telescope Mission*, Astrophys. J. 697 (2009) 1071–1102.
46. Meegan, C., et al., *The Fermi Gamma-Ray Burst Monitor*, Astrophys. J. 702 (2009) 791–804.
47. AMS Collaboration, Aguilar, M., et al., *The Alpha Magnetic Spectrometer (AMS) on the International Space Station. Part I: Results from the Testflight on the Space Shuttle*, Physics Reports 366 (2002) 331–405.
48. Institute of Electrical Engineering, IEE, Chinese Academy of Sciences, 100080 Beijing, China.
49. Chinese Academy of Launching Vehicle Technology, CALT, 100076 Beijing, China.
50. CMS Physics, Technical Design Report, Volume I: CERN-LHCC-2006-001, Feb. 2, 2006.
51. ATLAS detector and physics performance, Technical Design Report, Volume I, May 25, 1999.
52. Jones, W. V., *Astromag - Particle astrophysics magnet facility for Space Station Freedom*, IAF, 40th Int. Astronautical Congress, Malaga, Spain, Oct. 7–13, 1989.
53. Blau, B., et al., Grav. Cosmol. Suppl. 5 (2000) 1; IEEE Trans. Appl. Supercond. 12 (2002) 349.
54. SERMS, Via Pentima 4, 05100 Terni, Italy; Bertucci, B., *The S.E.R.M.S. laboratory. A research and test facility for space payloads and instrumentation*, Memorie della Societa Astronomica Italiana 79 (2008) 818.
55. IABG mbH, Einsteinstrasse 20, 85521 Ottobrunn, Germany.
56. Battiston, R., *A silicon tracker for the antimatter spectrometer on the International Space Station ALPHA*, Proc. 1st Arctic Workshop Future Physics and Accelerators, Saariselka, Finland, Aug. 21–26, 1994, (1994) 138–156; Alcaraz, J., et al., *A silicon microstrip tracker in space: Experience with the AMS silicon tracker on STS-91*, Nuovo Cimento A 112 (1999).
57. Alcaraz, J., et al., *The alpha magnetic spectrometer silicon tracker: Performance results with protons and helium nuclei*, Nucl. Instrum. Meth. A 593 (2008) 376–398, Erratum: *ibid.* 597 (2008) 270.
58. Acciarri, M., et al., *The L3 silicon microvertex detector*, Nucl. Instrum. Meth. A 351 (1994) 300–312.
59. Colibrys (Switzerland) Ltd, Maladière 83, 2000 Neuchâtel, Switzerland.
60. FBK-irst, Via Sommarive, 18, 38050 Povo (Trento), Italy.
61. Toker, O., et al., Nucl. Instrum. Meth. A 340 (1994) 572.
62. Picozza, A., et al., Astroparticle Phys. 27 (2007) 296–315.
63. Straulino, S., et al., *Spatial resolution of double-sided silicon microstrip detectors for the PAMELA apparatus*, Nucl. Instrum. Meth. A 556 (2006) 100–114.
64. 5000, Hirakuchi, Hamakita-ku, Hamamatsu City, Shizuoka Pref., 434-8601, Japan.
65. Bindi, V., et al., *The AMS-02 time of flight system. Final design*, Proc. 28th Int. Cosmic Ray Conf., Tsukuba, Japan, July 31 - Aug. 7, 2003.
66. Baldini, L., *The AMS time-of-flight system*, Proc. 27th Int. Cosmic Ray Conf., Hamburg, Germany, Aug. 7–15, 2001.
67. The ATLAS TRT collaboration, Abat, E., et al., J. Instrum. 3 (2008) P02014.
68. ALICE TRD Collaboration, *The ALICE transition radiation detector*, Nucl. Instrum. Meth. A 502 (2003) 127–132.
69. Saveliev, V., *The HERA-B Transition Radiation Detector*, Nucl. Instrum. Meth. A 408 (1998) 289–295.
70. Siedenburg, T., et al., *A transition radiation detector for AMS*, Nucl. Phys. Proc. Suppl. 113 (2002) 154–158.
71. Casaus, J., et al., *The AMS RICH detector*, Nucl. Phys. Proc. Suppl. 113 (2002) 147–153.
72. Adinolfi, M., et al., *The KLOE electromagnetic calorimeter*, Nucl. Instrum. Meth. A 482 (2002) 364–386.
73. Cadoux, F., et al., *The AMS-02 electromagnetic calorimeter*, Nucl. Phys. Proc. Suppl. 113 (2002) 159–165.
74. Bonvicini, V., et al., *A silicon-tungsten imaging calorimeter for PAMELA*, Proc. 26th Int. Cosmic Ray Conf. (ICRC 99), Salt Lake City, Utah, Aug. 17–25, 1999, Vol. 5, pp. 187–190.

75. Torii S. et al., *Calorimetric electron telescope mission. Search for dark matter and nearby sources*, Nucl. Instr. and Meth. A 630 (2011) 55-7; Torii S. et al., Proc. of 33rd ICRC (2013) 245
76. Chang J et al., *Dark Matter Particle Explorer: The First Chinese Cosmic Ray and Hard γ-ray Detector in Space*, Chin. J. Space Sci. 34 (2014); Chang J et al., *The DArk Matter Particle Explorer mission*, Astropart. Phys. 95 (2017) 6
77. Derdeyn, S.M., Ehrmann, L.H., Fichtel, G.J., Kniffen, D.A., Ross, R.W., Nucl. Instrum. Meth. 98 (1972) 557.
78. Bignami, G.F., et al., Space Sci. Instrum. 1 (1975) 245.
79. Thompson, D.J., et al., Astrophys. J. 415 (1993) L13.
80. Thompson, D.J., et al., Astrophys. J. Suppl. Ser. 86 (1993) 629.
81. Carbone, L. et al., *Thermal gradient-induced forces on geodesic reference masses for LISA*, P.R.L. D76 (2007) 102003
82. Carbone L., et al., *Achieving Geodetic Motion for LISA Test Masses: Ground Testing Results* P.R.L. 91 (2003) 151101; Erratum-ibid. P.R.L. 91 (2003) 179903
83. Cavalleri, A., *Increased Brownian Force Noise from Molecular Impacts in a Constrained Volume*, P.R.L. D103 (2009) 140601
84. Antonucci, E. et al., *Interaction between Stray Electrostatic Fields and a Charged Free-Falling Test Mass*, P.R.L. D108 (2012) 181101
85. G. D'Ali Staiti, G., et al., *EUSO: A space mission searching for extreme energy cosmic rays and neutrinos*, Nucl. Phys. Proc. Suppl. 136 (2004) 415–432.
86. Takahashi, Y., *A Giant natural TPC (500 km)^3 to observe extremely high energy cosmic particles - JEM EUSO telescope on International Space Station*, J. Phys. Conf. Ser. 65 (2007) 012022.
87. Battiston, R., Brunetti, M.T., Cervelli, F., Fidani, C., Menichelli, M., *A Moon-borne electromagnetic calorimeter*, Astrophys. Space Sci. 323(4) (2009) 357–366.
88. Gusev, A.A., et al., *Detector for electron spectrum measurements in TeV region on synchrotron radiation in geomagnetic field*, Proc. 21st Int. Cosmic Ray Conf., Adelaide 1990, Vol. 3, pp. 245–248; Anderhub, H., et al., *Preliminary results from the prototype Synchrotron Radiation Detector on Space Shuttle mission STS-108*, Nucl. Phys. Proc. Suppl. 113 (2002) 166–169.
89. *NASA Stratospheric Balloons Pioneers of Space Exploration and Research*, Report of the Scientific Ballooning Planning Team, Oct. 2005; http://sites.wff.nasa.gov/code820/uldb.html.
90. Wiencke, L., *EUSO-Balloon mission to record extensive air showers from near space*, PoS ICRC 2015, (2016) 631.

Chapter 19
Cryogenic Detectors

Klaus Pretzl

19.1 Introduction

Most calorimeters used in high energy physics measure the energy loss of a particle in form of ionization (free charges) or scintillation light. However, a large fraction of the deposited energy in form of heat remains undetected. The energy resolution of these devices is therefore mainly driven by the statistical fluctuations of the number of charge carriers or photoelectrons involved in an event. In contrast, cryogenic calorimeters are able to measure the total deposited energy including the heat in form of phonons or quasi-particles in a superconductor. With the appropriate phonon or quasi-particle detection system much higher energy resolutions can be obtained due to the very large number of low energy quanta (meV) involved in the process. This feature makes cryogenic calorimeters very effective in the detection of very small energy deposits (eV) with resolutions more than an order of magnitude better than for example semiconductor devices.

During the last two decades cryogenic detectors have been developed to explore new frontiers in physics and astrophysics. Among these are the quest for the dark matter in the universe, the neutrinoless double beta decay and the mass of the neutrino. But other fields of research have also benefited from these developments, such as astrophysics, material and life sciences.

The calorimetric measurement of deposited energy in an absorber dates back to 1878, when the American astronomer S.P. Langley invented the bolometer [1]. With this device he was able to measure the energy flow of the sun in the far infrared region of the spectrum and to determine the solar constant. Since then the bolometer has played an important role to measure the energy of electromagnetic radiation

K. Pretzl (✉)
Laboratory for High Energy Physics - Albert Einstein Center for Fundamental Physics, University of Bern, Bern, Switzerland
e-mail: pretzl@lhep.unibe.ch

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_19

of celestial objects. At the turn of the century radioactivity was discovered and P. Curie and A. Laborde made a first attempt in 1903 to measure the energy released in radioactive decays using a calorimetric device [2]. Thereafter micro-calorimeters were developed by C.D. Ellis and A. Wooster in 1927 [3] and independently by W. Orthmann and L. Meitner in 1930 [4] to determine the average energy of the electron in the beta-decay of ^{210}Bi. The differential micro-calorimeter developed by W. Orthmann allowed to measure heat transfers of the order of μW. Using this true calorimetric technique, he and L. Meitner were able to determine the average energy of the continuous beta spectrum in ^{210}Bi to 0.337 MeV with a 6% accuracy. These measurements contributed greatly to the notion of a continuous beta-spectrum leading to W. Pauli's neutrino hypothesis in 1930.

In 1935 F. Simon [5] suggested to measure the energy deposited in radioactive decays with low temperature calorimeters. He claimed that with a calorimeter of 1 cm^3 tungsten in a liquid helium bath at 1.3 K, one could measure a heat transfer of nW, which is about 1000 times more sensitive than the calorimeter of W. Orthmann. The argument is that at low temperatures the heat capacity C of a micro-calorimeter is low and a small energy loss E of a particle in the calorimeter can lead to an appreciable temperature increase $\Delta T = E/C$. Later in 1949, D.H. Andrews, R.D. Fowler and M.C. Williams [6] reported the detection of α-particles from a Po source with a bolometer made of a superconducting strip of NbN mounted on a copper base. The operating temperature was chosen 15.5 K, which corresponded to the center of the transition halfway between the superconducting and the normal state of NbN. However, at this stage of the experiment no energy information of the alpha particles could be extracted from the signals, since the signal to background ratio was not sufficient. Their bolometer was used only as a particle counter. In 1969, G.H. Wood and B.L. White [7] were able to measure the energy of the emitted alpha particles from a polonium source with a superconducting tunnel junction (STJ). The energy was derived from the tunneling current, which is proportional to the excitations of the quasi-particles induced by the energy loss of the α-particles in the junction.

H. Bernas et al. [8] introduced 1967 superheated superconducting granules (SSG) to measure beta radiation. Used as an energy threshold detector the energy loss of an electron in a granule could suffice to drive the granule from a super-conducting into a normal state. This phase transition would induce a signal in a pickup coil due to the Meissner effect, provided the granules are sitting in an external magnetic field. A. Drukier and C. Vallette [9] were able to detect charged particles with a SSG device. Later in 1984, A. Drukier and L. Stodolsky [10] suggested the use of SSG detectors for neutrino and astrophysics experiments.

In early 1970 a new type of bolometer, the so-called composite one, was developed by N. Coron, G. Dambier and J. Leblanc [11]. It consisted of an absorber and a thermally coupled thermometer in form of a semiconductor thermistor. Later in 1974, T. Niinikosky and F. Udo [12] proposed cryogenic calorimeters for the detection of neutrinos. E. Fiorini and T. Niinikoski [13] explored in 1984 the possibility of using low temperature bolometers to improve the limits on neutrinoless double beta decays. At this time D. McCammon, S.H. Moseley, J.C. Mather and R. Mushotzky [14] published first results with a cryogenic calorimeter

for X-ray spectroscopy. In 1985 N. Coron et al. [15] developed a cryogenic composite bolometer as a charged particle spectrometer.

Based on all these interesting ideas and developments, a first workshop on low temperature detectors (LTD1) was held in 1987 at Ringberg-Castle on Lake Tegernsee in southern Bavaria. Due to the success of this workshop and the growing interest in this field, further workshops have been organized in Europe, the USA and Japan. Much of the original work in this field can be found in the proceedings of these LTD workshops [16–31]. There exist also excellent review articles [32–37] as well as a textbook [38] on this subject.

19.2 General Features of Cryogenic Calorimeters

A typical cryogenic calorimeter is shown in Fig. 19.1. It consists of three basic elements: an absorber, which confines the interaction volume, a thermometer, which is thermally well coupled to the absorber and which measures the temperature increase due to the energy loss of a particle in the absorber, and a thermal bath, which has a weak thermal link to the absorber and restores the temperature in the absorber to a defined base value. Particles interacting in the absorber material lose their energy in producing atomic and solid state excitations. These excitations produce electrons, photons (photoelectrons) and phonons. Phonons are quantized lattice vibrations which behave like particles and propagate with the speed of sound. The energy of these particles will degrade in time via electron-phonon and phonon-

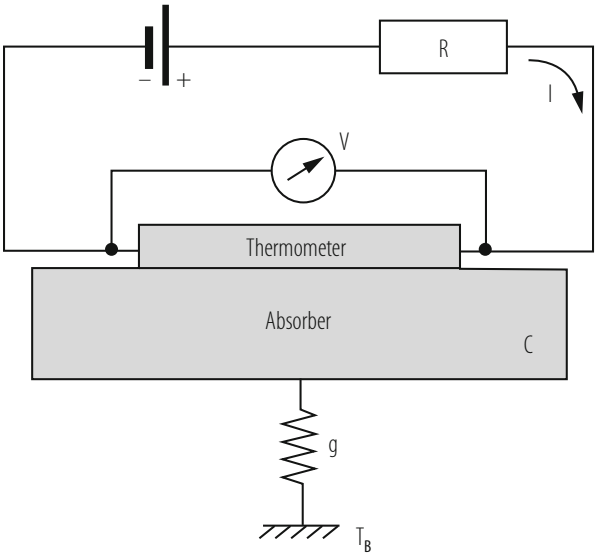

Fig. 19.1 The principle of a cryogenic calorimeter is shown

phonon interactions as well as via interactions with lattice irregularities until the system settles in thermal equilibrium. Calorimeters operating in the equilibrium mode (i.e. being sensitive to thermal phonons) offer in principle the best energy resolution, because the number of thermal phonons, with typical energies of meV, is large and the statistical fluctuations are small. For some applications thermal detectors can also be used in a non-equilibrium mode being sensitive to only high energy, so-called quasi-ballistic, phonons. These devices have the advantage of being intrinsically faster, but with energy resolutions inferior to equilibrium detectors. In calorimeters made from superconducting materials, such as superconducting tunnel junctions, the excitation energy is transformed into phonons as well as quasi-particles. These devices operate in the non-equilibrium mode, since the excitations (quasi-particles) are measured before they settle in thermal equilibrium. As described in more detail in the following paragraphs, most low temperature calorimeters differ in the way they are converting the excitation energy into a measurable signal.

Assuming that the deposited energy E of a particle in the absorber is fully thermalized the temperature rise ΔT is given by:

$$\Delta T = \frac{E}{C_{tot}}\,, \tag{19.1}$$

where $C_{tot} = cV$ is the heat capacity of an absorber with the volume V and the specific heat c. Cryogenic detectors operate at low temperatures because the heat capacity of many absorber materials becomes very small leading to an appreciable temperature rise. In addition the absorber volumes are kept as small as possible, in some cases of mm^3 or cm^3 size. Therefore they are also often called micro-calorimeters. Applying the Debye model to calculate the internal energy of the lattice vibrations (phonons), the specific heat of a dielectric crystal absorber comes out to be:

$$c_{dielectric} = \beta(\frac{T}{\theta_D})^3\,, \tag{19.2}$$

with $\beta = 1944\ \mathrm{J\ mol^{-1}\ K^{-1}}$ and θ_D the Debye temperature of the crystal. The cubic dependence on temperature demonstrates a strong decrease of the phonon specific heat at low temperatures. In a metal absorber there are two components which determine the specific heat: lattice vibrations and thermally excited conduction electrons. The specific heat of a normal conducting material at low temperatures is given by:

$$c_{metal} = \beta(\frac{T}{\theta_D})^3 + \gamma T\,, \tag{19.3}$$

with γ being a material dependent constant (Sommerfeld constant). At temperatures below 1 K the electronic specific heat dominates. Therefore the total specific heat

decreases only linearly with temperature. Another frequently used absorber is a superconductor. In this case the specific heat consists of a term due to lattice vibrations and a second term which reflects the number of thermally excited electrons across the energy gap of a superconductor Δ. The latter diminishes exponentially with temperature due to the decrease of the quasi-particle density:

$$c_{supercond.} = \beta(\frac{T}{\theta_D})^3 + a \exp\left\{(-\frac{b\Delta}{k_B T})\right\}, \tag{19.4}$$

with a and b being material constants and k_B is the Boltzmann constant. Therefore at very low temperatures the specific heat of a superconductor is dominated by lattice vibrations.

The characteristics of an ideal cryogenic calorimeter can be described by the heat capacity C of the absorber and the thermal conductivity g of the link to the heat bath with the temperature T_B. In the event of a particle losing an energy E in the absorber the temperature in the absorber will according to Eq. (19.1) rise by ΔT and then decay back to its starting temperature, which corresponds to the bath temperature T_B. The time constant of this process is given by $\tau = C/g$. The temperature rise in the absorber will change the resistance of the thermometer, which is measured by recording a voltage drop across it when passing a current I through the thermometer (Fig. 19.1). The same device can also be used to measure a continuous power input P in form of electromagnetic radiation for example. In this case the temperature rise is given by $\Delta T = P/g$. Such a device is usually referred to as a bolometer. Bolometers have a long tradition in detecting infrared radiation from astrophysical objects. They have also been used in the measurements of the cosmic microwave background radiation.

Cryogenic calorimeters can be made from many different materials including superconductors, a feature which turns out to be very useful for many applications. They can be used as targets and detectors at the same time. Due to the very small energy quanta involved they reach much higher energy resolutions than conventional ionization or solid state devices. For example, it takes only of the order of 1 meV to break a Cooper pair in a superconductor whereas a few eV are needed to create an electron-hole pair in a solid-state device. Cryogenic calorimeters are able to detect very small energy transfers, which makes them sensitive also to non-ionizing events.

The intrinsic energy resolution of a cryogenic calorimeter is limited by the thermal energy fluctuations due to the phonon exchange between the absorber and the heat sink. The mean square energy fluctuation is given by Chui et al. [43]:

$$< \Delta E^2 >= k_B T^2 C\,. \tag{19.5}$$

It is independent of the absorbed energy E, the thermal conductivity g of the heat link and the time constant τ. The above equation can intuitively be understood when assuming that the effective number of phonon modes in the detector is $N = C/k_B$, the typical mean energy of one phonon is $k_B T$ and the rms fluctuation of one phonon is one. Then the mean square energy fluctuation is $N(k_B T)^2 = k_B T^2 C$.

For a practical cryogenic calorimeter the energy resolution is therefore given to first order by

$$\Delta E_{FWHM} = 2.35\xi\sqrt{k_B T^2 C}\,, \tag{19.6}$$

where ξ is a parameter which depends on the sensitivity and noise characteristics of the thermometer and can have values between 1.2 and 2.0. The best resolution obtained so far with cryogenic calorimeters is ∼2 eV at 6 keV.

The use of superconductors as cryogenic particle detectors was motivated by the small binding energy 2Δ (order of meV) of the Cooper pairs. The breaking of a Cooper pair results in the creation of two excited electronic states the so-called quasi-particles. A particle traversing a superconductor produces quasi-particles and phonons. As long as the energy of the quasi-particles and the phonons is higher than 2Δ, they break up more Cooper pairs and continue to produce quasi-particles until their energy falls below the threshold of 2Δ. Particles which lose the energy E in an absorber produce ideally $N = E/\Delta$ quasi-particles. Thus the intrinsic energy resolution of a superconducting cryogenic calorimeter with $\Delta \approx 1$ meV is given by

$$\Delta E_{FWHM} = 2.35\sqrt{\Delta F E} \sim 2.6\,\text{eV} \tag{19.7}$$

for a 6 keV X-ray assuming a Fano factor $F = 0.2$, which is representative for most superconductors and which takes the deviation from Poisson statistics in the generation of quasi-particles into account.

For comparison, the energy resolution of a semiconductor device is typically:

$$\Delta E_{FWHM} = 2.35\sqrt{w F E} \sim 110\,\text{eV} \tag{19.8}$$

for a 6 keV X-ray, where w is the average energy necessary to produce an electron hole pair. It has a typical value of $w \approx 3$ eV. The Fano factor is $F = 0.12$ for Silicon. Because of the larger number of free charges a super-conducting device has a much better energy resolution.

In Fig. 19.2 X-ray spectra obtained with a state of the art Si(Li) solid-state device (dashed line) and a cryogenic micro-calorimeter (solid line) using a Bi absorber and an Al-Ag bilayer superconducting transition edge thermometer are compared. The micro-calorimeter has been developed at the National Institute of Standards and Technology (NIST) in Boulder (USA) [42].

19.3 Phonon Sensors

Phonons produced by a particle interaction in an absorber are far from thermal equilibrium. They must decay to lower energy phonons and become thermalized before the temperature rise ΔT can be measured. The time required to thermalize

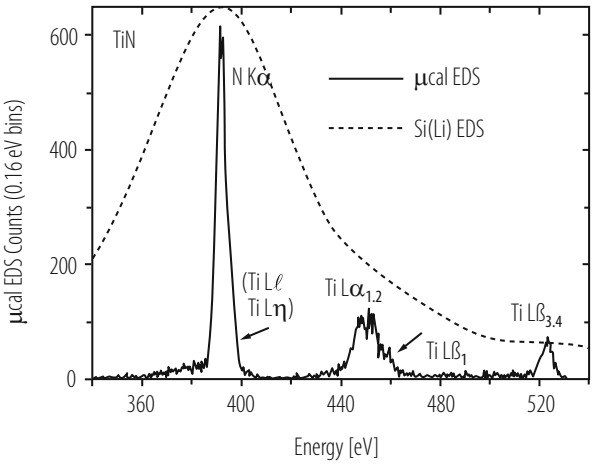

Fig. 19.2 TiN X-ray spectra obtained with a cryogenic micro-calorimeter (solid line) from the NIST group (see text) and with a state of the art Si(Li) solid-state device (dashed line) are compared. EDS stands for energy dispersive spectrometer. TiN is an interconnect and diffusion barrier material used in semiconductor industry

and the long pulse recovery time ($\tau = C/g$) limits the counting rate of thermal calorimeters to a few Hz. The most commonly used phonon sensors are resistive thermometers, like semiconducting thermistors and superconducting transition edge sensors (TES), where the resistance changes as a function of temperature. These thermometers have Johnson noise and they are dissipative, since the resistance requires power to be read out, which in turn heats the calorimeter (Joule heating). However, the very high sensitivities of these calorimeters can outweigh to a large extent these disadvantages. There are also magnetic thermometers under development, which do not have readout power dissipation.

19.3.1 Semiconducting Thermistors

A thermistor is a heavily doped semiconductor slightly below the metal insulator transition. Its conductivity at low temperatures can be described by a phonon assisted electron hopping mechanism between impurity sites. This process is also called "variable range hopping" (VRH) [44]. For temperatures between 10 mK and 4 K the resistance is expected to follow $R(T) = R_0 \exp\left\{(\frac{T_0}{T})^{\frac{1}{2}}\right\}$. This behavior is observed in doped Si and Ge thermistors. However, depending on the doping concentrations of the thermistor and the temperature range of its use, deviations from this behaviour have also been discovered. An important requirement for the fabrication of thermistors is to achieve a good doping homogeneity and reproducibility. Good uniformity of doping concentrations has been achieved either

with ion implantation or with neutron transmutation doping (NTD). In the latter case, thermal neutrons from a reactor are captured by nuclei which transform into isotopes. These can then be the donors or acceptors for the semiconductor. NTD Ge thermistors are frequently used because of their reproducibility and their uniformity in doping density. Furthermore they are easy to handle and commercially available.

It is convenient to define a dimensionless sensitivity of the thermometer:

$$\alpha \equiv \frac{d \log R}{d \log T} = \frac{T}{R} \frac{dR}{dT} . \tag{19.9}$$

The energy resolution of these devices is primarily driven by the heat capacity C of the absorber, the sensitivity of the thermometer α, the Joule heating, the Johnson noise of the load resistor and the amplifier noise. The bias current through the resistor can be optimized in such a way that it is kept high enough to provide a suitable voltage signal and low enough to minimize the Joule heating. If also the Johnson noise and the amplifier noise can be kept sufficiently low, the energy resolution of an ideal calorimeter can be described to first order by Eq. (19.6), where ξ is approximately $5(1/\alpha)^{1/2}$ [39]. For large values of α the energy resolution can be even much better than the magnitude of the thermodynamic fluctuations provided no power is dissipated by the temperature measurement of the sensor. Semiconducting thermistors have typically α values between 6 and 10, while superconducting transition edge sensors (TES) have values which are two orders of magnitude higher. A detailed description of the noise behavior and the energy resolution of cryogenic detectors can be found in [39–41].

19.3.2 Superconducting Transition Edge Sensors (TES)

A frequently used phonon sensor is the so-called transition edge sensor (TES). It consists of a very thin superconducting film or strip which is operated at a temperature in the narrow transition region between the superconducting and the normal phase, where its resistance changes between zero and its normal value R_N, as shown in Fig. 19.3a. TES sensors are usually attached to an absorber, but they can also be used as absorber and sensor at the same time. The very strong dependence of the resistance change on temperature, which can be expressed in the dimensionless parameter α of Eq. (19.9), makes the TES calorimeter sensitive to very small input energies. Superconducting strips with low T_c can have α values as high as 1000. This requires very high temperature stability. The Munich group has developed one of the first TES sensors, which was made from tungsten with a transition temperature of 15 mK [45].

The TES sensor can be operated in two different modes: the current and the voltage biased mode. In a current biased mode of operation a constant current is fed through the readout circuit as shown in Fig. 19.3b. A particle interaction in the absorber causes a temperature rise and a corresponding increase of the

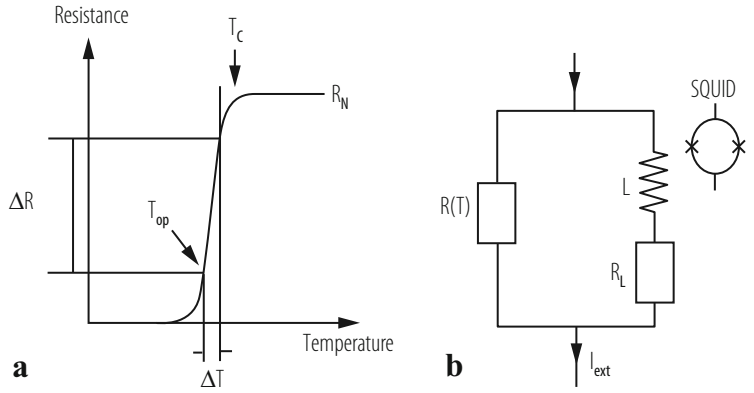

Fig. 19.3 (**a**) The temperature versus resistance diagram of a superconducting strip close to the transition temperature T_c is shown. (**b**) The dc-SQUID readout of a transition edge sensor is shown

resistance $R(T)$ of the attached TES sensor. The change of the resistance forces more current through the parallel branch of the circuit, inducing a magnetic flux change in L which is measured with high sensitivity by a superconducting quantum interference device (SQUID). However, in this mode Joule heating by the current through the sensor and small fluctuations in the bath temperature can prevent to achieve good detector performances. To solve this problem K.D. Irwin [46] has developed a so-called auto-biasing electro-thermal feedback system (ETF), which works like a thermal equivalent to an operation amplifier and keeps the temperature of the superconducting strip at a constant value within its transition region. When operating the transition edge sensor in a voltage biased mode (V_B), a temperature rise in the sensor causes an increase in its resistance and a corresponding decrease in the current through the sensor, which results in a decrease of the Joule heating ($V_B \cdot \Delta I$). The feedback uses the decrease of the Joule heating to bring the temperature of the strip back to the constant operating value. Thus the device is self-calibrating. The deposited energy in the absorber is given by $E = V_B \int \Delta I(t)dt$. It can directly be determined from the bias voltage and the integral of the current change. The use of SQUID current amplifiers allows for an easy impedance matching to the low resistance sensors and opens the possibility to multiplex the read out of large arrays of TES detectors. Another advantage of ETF is that in large pixel arrays the individual channels are self-calibrating and temperature regulated. Most important, ETF shortens the pulse duration time of TES by two orders of magnitude compared to thermistor devices allowing for higher count rates of the order of 500 Hz.

The intrinsic energy resolution for an ideal TES calorimeter is given by Eq. (19.6) with $\xi = 2(1/\alpha)^{1/2}(n/2)^{1/4}$, where n is a parameter which depends on the thermal impedance between the absorber (phonons) and the electrons in the superconducting film [46, 47]. For thin films and at low temperatures the electron-phonon decoupling dominates in the film and n is equal to 5. The best reported energy resolutions with TES devices so far are a little below 2 eV at 6 keV.

The observed transition width of TES ΔT in the presence of a typical bias current is of the order of a few mK. Large bias currents usually lead to transition broadenings due to Joule heating and self-induced magnetic fields. In order to achieve best performance of TES in terms of energy resolution or response time for certain applications, specific superconducting materials have to be selected. Both superconductors of type I and type II qualify in principle. However, the physics of the phase transition influences the noise behavior, the bias current capability and the sensitivity to magnetic field of the TES. Sensors made from high temperature superconductors have a much lower sensitivity than low temperature superconductors due to the larger gap energies Δ and heat capacities C. Thermal sensors made from strips of Al (with $T_c = 1.140\,\mathrm{K}$), Ti (0.39 K), Mo (0.92 K), W (0.012 K) and Ir (0.140 K) have been used. Ti and W sensors have been developed in early dark matter detectors [48–51]. But also other transition edge sensors, made from proximity bi-layers such as Al/Ag, Al/Cu, Ir/Au, Mo/Au, Mo/Cu, Ti/Au, or multi-layers such as Al/Ti/Au [56], have been developed to cover transition temperatures in the range between 15 and 150 mK. Although not all of these combinations are chemically stable, good detector performances have been obtained with Ir/Au bi-layers [52, 53] at transition temperatures near 30 mK. Methods to calculate bi-layer T_c can be found in [54, 55]. Another method to suppress the T_c of a superconducting film is to dope it with magnetic ions, like for example Fe (<100 ppm) [57]. However, there is a concern that the magnetic impurities may drastically increase the heat capacity of the film.

TES is also sensitive to non-thermal phonons with energies well above 2Δ. While losing energy these phonons produce quasi-particles before they thermalize. Since this process is very much faster than thermalization, signals of the order of μs can be achieved, enhancing considerably the counting rate capability of these devices as compared to thermal phonon sensors. Due to its high resolution and timing capabilities as well as versatile applications, TES sensors are currently among the most frequently used devices in calorimetric measurements. A detailed description of the performance of TES and ETF-TES can be found in [46, 47].

19.3.3 Magnetic Sensors

The magnetic properties of many materials are strongly dependent on temperature. This feature has been used to build very sensitive magnetic calorimeters applying thermal sensors made from thin paramagnetic strips, placed in a small magnetic field, which are in strong thermal contact with a suitable particle absorber. The energy deposited in the absorber leads to a temperature rise and a corresponding decrease in magnetization of the sensor. The change of magnetization is given by

$$\Delta M = \frac{dM}{dT}\frac{\Delta E}{C_{tot}} \tag{19.10}$$

with C_{tot} the total heat capacity of the thermometer and the absorber. It can be very accurately measured with a high bandwidth dc-SQUID magnetometer. The use of magnetism as thermal sensor was first developed by Buehler and Umlauf [58] and Umlauf and Buehler [59]. In these first attempts magnetic calorimeters were using the magnetization of 4f ions in dielectric host materials to measure temperature changes. Due to the weak coupling of the magnetic moments to the phonons at low temperatures these devices exhibited a too slow response time (order of seconds) for many applications. This problem was overcome by introducing sensors which use magnetic ions in metallic base material [60]. This type of device is called metallic magnetic calorimeter (MMC). In metals the relaxation times due to interactions between conduction electrons and magnetic moments are orders of magnitude faster than in dielectrics. However, the presence of conduction electrons increases the heat capacity of the sensor and leads to an enhanced interaction amongst magnetic moments. Nevertheless, very promising results were obtained with a metallic magnetic calorimeter [61]. It consisted of two thin Au disc sensors (50 µm in diameter and 25 µm thick) containing 300 ppm enriched ^{166}Er and a gold foil ($150 \times 150 \times 5\,\mu m^3$) as an X-ray absorber. The calorimeter reached an energy resolution of 3.4 eV at 6 keV, which is quite comparable to TES and thermistor calorimeters. An important property of MMC is that its inductive read out, which consists of a primary detector SQUID and a secondary SQUID amplifier, does not dissipate power into the system [61]. This feature makes MMC very attractive for many applications, in particular where large pixel arrays are of interest. The energy resolution of MMC is primarily driven by the thermal conductance between the absorber and the temperature bath and between the absorber and the sensor. For an ideal MMC the energy resolution is then given by Eq. (19.6), where ξ is approximately $\xi = 2\sqrt{2}(\tau_0/\tau)^{1/4}$ with τ_0 (typically order of µs) the relaxation time between the absorber and the sensor and τ (typically order of ms) the relaxation time between the absorber and the bath temperature [62]. It is further assumed that the heat capacities of the absorber and the sensor are approximately equal. In this case the heat capacity C in Eq. (19.6) represents the heat capacity of the absorber. MMC devices have potential applications in X-ray spectroscopy and are under further development for large pixel array cameras.

19.4 Quasiparticle Detection

The physics of superconducting detectors are based on Cooper pair breaking and quasi-particle production. Quasi-particles created by the absorption of X-rays or by the energy loss of a transient particle in a superconducting absorber can be measured with a Superconducting Tunnel Junction (STJ). The STJ device is in principle the same as the more widely known Josephson junction [63]. When biasing the STJ at a suitable voltage the tunneling current through the junction is proportional to the excess number of quasi-particles produced. To be able to measure these excess quasi-particles above the thermal background one has

to go to very low temperatures $T < 0.1\,T_c$. Arrays of STJs are also used to measure high energy, non-thermal (ballistic) phonons produced in either a dielectric or superconducting absorber. A new detector concept, called microwave kinetic inductance detector (MKID), has been developed which allows a frequency-domain approach to multiplexing and results in a dramatic simplification of the array and its associated readout electronics. Another detection scheme is based on small superheated superconducting granules (SSG) embedded in an external magnetic field. They are kept just below the phase-transition border and will change from the superconducting to the normal-conducting phase upon thermal excitation, which leads to the breaking of Cooper pairs and the penetration of the external magnetic field into the granule, causing a magnetic flux change (Ochsenfeld-Meissner effect). The flux change can be measured with an appropriate pickup coil. All these detectors are non-equilibrium devices.

19.4.1 Superconducting Tunnel Junctions (STJ)

The pioneering work of the groups of the Paul Scherrer Institute (at Villigen, Switzerland) [64] and of the Technical University Munich (Germany) [65] and their promising first results have stimulated other institutes to further develop STJs for high resolution X-ray detection. A typical STJ consists of two superconducting films S1 and S2 with a thickness of a few nm separated by a thin, 1–2 nm thick, tunnel barrier, which is usually the oxide of one of the superconductors. Because of its structure the device is frequently referred to as SIS (superconductor/insolator/superconductor) junction, also sometimes called Giaever junction. Typical junction areas are of the order of $100 \times 100\,\mu\text{m}^2$. As a quasi-particle detector, the STJ is operated with a bias voltage which is usually set to be less than Δ/e, where e is the charge of an electron. The principle processes taking place in a STJ are illustrated in Fig. 19.4, which is taken from [67]. In the event of an incident particle

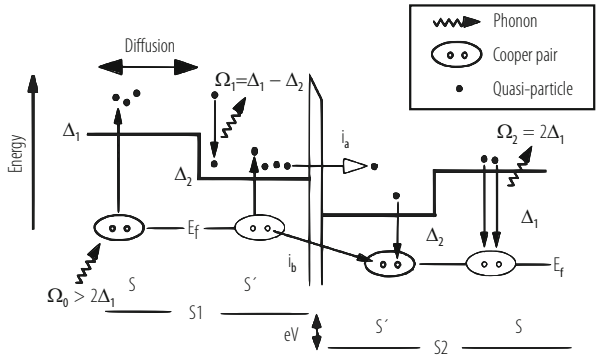

Fig. 19.4 The processes in a superconducting tunnel junction (ST) are illustrated

or X-ray interaction in film S1 the quasi-particle density is increased. This will lead to an increase of a net quasi-particle transfer from S1 to S2 and consequently to an increase of the tunneling current. However, not all quasi-particles will reach and pass through the junction barrier. Depending on the geometry and structure of the junction there will be losses. Quasi-particles can recombine to Cooper pairs radiating phonons as consequence of the relaxation process. If the phonon energy is high enough $\Omega_0 > 2\Delta_1$ to break new Cooper pairs, this process can lead to quasi-particle multiplication enhancing the signal output of the STJ. If, however, the phonon energy is below the energy threshold for breaking a Cooper pair $\Omega_0 < 2\Delta_1$ the quasi-particle will be lost and will not contribute to the signal. Quasi-particles will also be lost when they diffuse out of the overlap region of the junction films into the current leads instead of crossing the junction barrier. N. Booth [68] proposed a scheme which allows to recover some of these losses by quasi-particle trapping and in some cases quasi-particle multiplication. Quasi-particle trapping can be achieved by introducing bi-layers of superconducting materials (S(Δ_1) and S'(Δ_2)) with different gap energies $\Delta_2 < \Delta_1$ [68]. For example, an X-ray absorbed in the superconductor S produces phonons of energy $\Omega_0 > 2\Delta_1$ breaking a number of Cooper pairs. Some of the produced quasi-particles diffuse to the superconducting film S' of the STJ with a smaller gap energy Δ_2. By falling in that trap they relax to lower energies by emitting phonons, which could generate additional quasi-particles in the film S' (quasi-particle multiplication) if their energy is larger than $2\Delta_2$. However, the relaxed quasi-particles cannot diffuse back into the superconductor S because of their lower energy. They are trapped in S' and will eventually tunnel through the STJ, contributing to the signal with the tunneling current i_a. In order for quasi-particle trapping to be effective superconducting absorber materials with long quasi-particle lifetimes have to be selected (for example Al). Back tunneling, the so-called Gray effect, is also enhancing the signal [69]. In this Cooper pair mediated process a quasi-particle in film S2 recombines to form a Cooper pair at the expense of a Cooper pair in film S1. In this case the quasi-particle current i_b is also running in the direction of decreasing potential. Thus both excess quasi-particle currents i_a and i_b have the same sign. This feature allows to record signals from X-rays absorbed in either superconducting films S1 or S2 with the same sign. However, their signal shapes may not necessarily be the same due to different quasi-particle and tunneling losses in the two films. There are two other ways of electrical transport through the tunnel barrier which need to be suppressed when the STJ is used as a particle detector. One is the so-called dc Josephson current of Cooper pairs through the tunnel barrier. This current can be suppressed by applying a magnetic field of the order of a few Gauss parallel to the insulating barrier. The second is the tunnel current generated by thermally excited excess quasi-particles. The number density of these quasi-particles is decreasing with decreasing temperature according to $N_{th} \sim T^{1/2} \exp(-\Delta / k_B T)$. In order to obtain a significant signal to background ratio the operating temperature of a STJ detector should be typically lower than $0.1\, T_c$. The intrinsic energy resolution of the excess quasi-particles in a STJ device is given by Eq. (19.7), where Δ has to be replaced by ϵ, the effective energy needed to create one excited state. It turns out that $\epsilon \sim 1.7\, \Delta$ for Sn and Nb superconductors,

reflecting the fact that only a fraction of the absorbed energy is transfered into quasi-particles [70]. The number of quasi-particles generated in the STJ by an energy loss E of a particle in the superconductor is thus $N = E/\epsilon$. For a Nb superconductor with a Fano factor $F = 0.2$ and $\epsilon = 2.5\,\text{meV}$ one would expect from Eq. (19.7) an energy resolution of 4 eV at 6 keV. However, the best resolution observed so far is 12 eV at 6 keV. In order to estimate a more realistic energy resolution, quasi-particle loss and gain processes have to be taken into account. The two most important parameters driving the energy resolution of the STJ are the tunneling rate $\Gamma_t \equiv \tau_t^{-1}$ and the thermal recombination rate $\Gamma_r \equiv \tau_r^{-1}$. The temperature dependence of the thermal recombination rate is given by

$$\tau_r^{-1}(T) = \tau_0^{-1}\sqrt{\pi}\left(\frac{2\Delta}{k_B T_c}\right)^{5/2}\sqrt{\frac{T}{T_c}}\exp\left\{(-\frac{\Delta}{k_B T})\right\} \tag{19.11}$$

where τ_0 is the characteristic time of a superconductor. It has the values $\tau_0 = 2.3\,\text{ns}$ for Sn, $\tau_0 = 438\,\text{ns}$ for Al and $\tau_0 = 0.15\,\text{ns}$ for Nb [71].

The recombination rate $\Gamma_r \equiv \tau_r^{-1}$ can be minimized when operating the detector at sufficiently low temperatures, typically at $0.1\,T_c$, where the number of thermally excited quasi-particles is very small. The tunneling rate of a symmetric STJ is given by de Korte et al. [72]

$$\tau_t^{-1} = (4\,R_{norm}\cdot e^2\cdot N_0\cdot A\cdot d)^{-1}\frac{\Delta + eV_b}{\sqrt{(\Delta + eV_b)^2 - \Delta^2}} \tag{19.12}$$

where R_{norm} is the normal-conducting resistance of the junction, N_0 is the density of states of one spin at the Fermi energy, A is the junction overlap area, d the thickness of the corresponding film and V_b the bias voltage of the STJ. In practice, the tunneling time has to be shorter than the quasi-particle lifetime. For a $100\times100\,\mu\text{m}^2$ Nb-Al tunnel junctions with $R_{norm} = 15\,\text{m}\Omega$ a tunneling time of $\tau_t = 220\,\text{ns}$ and a recombination time of $\tau_r = 4.2\,\mu\text{s}$ has been measured [66]. The recombination time was determined from the decay time of the current pulse. Thus quasi-particles tunneled on average 19 times. In order to achieve even shorter tunneling times, one would have to try to further reduce R_{norm}. However, there is a fabricational limit avoiding micro-shorts in the insulator between the superconducting films. The STJ counting rate capability is determined by the pulse recovery time, which depends on the quasi-particle recombination time and can have values between several µs and up to ∼50 µs. Typical count rates of STJs are 10^4 Hz. An order of magnitude higher count rates can still be achieved, but not without losses in energy resolution. The total quasi-particle charge collected in the STJ is to first order given by

$$Q = Q_0\frac{\Gamma_t}{\Gamma_d} \tag{19.13}$$

with $Q_0 = Ne$ and $\Gamma_d = 2\Gamma_r + \Gamma_{loss}$ the total quasi-particle loss rate. The factor 2 in the recombination rate takes into account the loss of two excited electronic states and Γ_{loss} stands for all the other quasi-particle losses, mainly due to diffusion. These effects can be parametrized into an effective Fano factor which is added into the equation for the energy resolution

$$\Delta E_{FWHM} = 2.35\sqrt{\epsilon(F+G)E}\,. \tag{19.14}$$

For a symmetric tunnel junction with equal tunneling probabilities on both sides the G factor is given by $G = (1 + 1/\bar{n})$ with $\bar{n} = Q/Q_0 = \Gamma_t/\Gamma_d$ [66, 67]. It emphasizes the importance of a large tunneling rate. Still the energy resolution in Eq. (19.14) is only approximative since it neglects gain factors like quasi-particle multiplication due to relaxation phonons and loss factors due to cancellation currents, which becomes important at low bias voltage.

From Eq. (19.12) it is clear that in order to achieve a high tunnel rate the STJ detector has to be made from very thin films with a small area A. These dimensions also determine the capacitance which should be kept as small as possible in order not to degrade the detector rise time and the signal to noise ratio. For the very thin films the quantum efficiencies at X-ray energies are very low. This can be changed by separating the absorber and detector functions in fabricating devices with a larger size superconducting absorber as substrate to a STJ. Quasi-particle trapping will be achieved when choosing substrate materials with a higher energy gap with respect to the junction.

Quasi-particle trapping was first demonstrated using a Sn absorber ($975 \times 150 \times 0.25\,\mu m^3$) with an energy gap of $\Delta_{Sn} = 0.58\,meV$ and with an Al-Al_2O_3-Al STJ at each end of the absorber [65]. Quasi-particles generated by an event in the Sn absorber diffuse into the aluminum junctions, where they stay trapped because of the smaller energy gap $\Delta_{Al} = 0.18\,meV$ of Al with respect to Sn. The excess quasi-particle tunnel current was then measured with the two Al STJs. In order to prevent diffusion losses out of the Sn absorber the common contact leads to the STJs and to the Sn absorber where made from Pb, which has an even higher energy gap of $\Delta_{Pb} = 1.34\,meV$. It turned out that more than 99.6% of the quasi-particles which tried to diffuse out of the absorber where rejected at the Pb barrier and hence confined to the absorber. Currently the best energy resolution of 12 eV at 6 keV has been achieved with a single Al-Al_2O_3-Al STJ using a superconducting Pb absorber ($90 \times 90 \times 1.3\,\mu m^3$) with an absorption efficiency of ≈50% [73]. It turns out that Al is an ideal material for STJ because it allows to fabricate a very uniform layer of the tunnel barrier and has a very long quasi-particle lifetime. These are features which are essential for a high performance STJ. A very good description of the physics and applications of STJ detectors can be found in [67].

Arrays of STJs have been developed for astronomical observations and other practical applications as discussed in the chapters below. The Naples collaboration [76] produced an array of circular shaped STJs [76, 77]. This device allows the operation of STJs without external magnetic field. Position sensitive devices have been developed for reading out large pixel devices [65, 74, 78].

19.4.2 Microwave Kinetic Inductance Detector

A new detector concept, called microwave kinetic inductance detector (MKID) has been introduced with the aim to develop multi-pixel array cameras for X-ray and single photon detection [79, 80]. MKID is, like STJ and SSG, a non-equilibrium detector which is based on Cooper pair breaking and the production of quasi-particles. The basic element of the device consists of a thin superconducting film, which is part of a transmission line resonator.

The principle of detection is shown in Fig. 19.5, taken from [79]: A photon absorbed in a superconducting film will break up Cooper pairs and produce quasi-particles (a). The increase of quasiparticle density will affect the electrical conductivity and thus change the inductive surface impedance of the superconducting film, which is used as part of a transmission line resonator (b). At resonance, this will change the amplitude (c) and the transmission phase of the resonator (d).

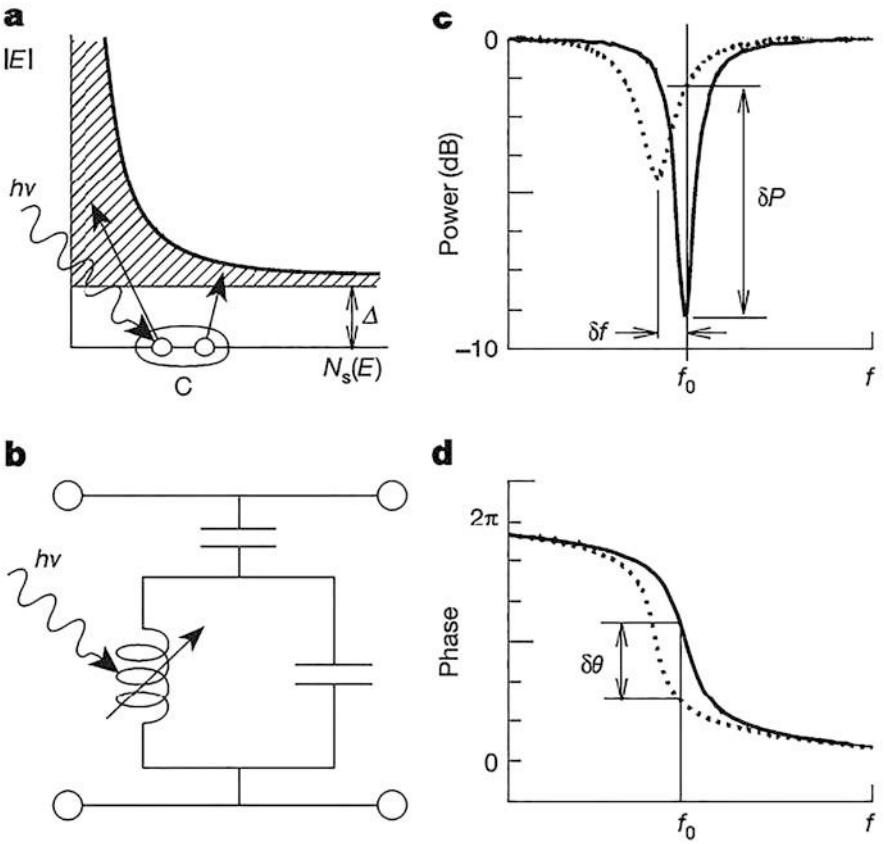

Fig. 19.5 The basic operation of a MKID (Microwave Kinetic Inductance Detector) is shown

The change in the transmission phase is proportional to the produced number of quasi-particles and thus to the photon energy. First measurements with an X-ray source yielded an energy resolution of 11 eV at 6 keV. MKID detectors find many applications where a large number of pixels are demanded. As compared to other devices multiplexing can be realized rather easy by coupling an array of many resonators with slightly different resonance frequencies to a common transmission line. A single amplifier is needed to amplify the signals from a large number of detectors. Due to its interesting features MKID is under development for many applications in ultraviolet, optical and infrared imaging [80].

19.4.3 Superheated Superconducting Granules (SSG)

Superheated superconducting granules (SSG) have been developed for X-ray imaging, transition radiation, dark matter as well as solar and reactor neutrino detection [10, 94]. A SSG detector consists of billions of small grains (typically 30 µm in diameter), diluted in a dielectric material (e.g. Teflon) with a volume filling factor of typically 10%. The detector is operated in an external magnetic field. Metastable type-1 superconductors (e.g. Sn, Zn, Al, Ta) are used, since their phase transitions from the metastable superconducting state to the normal-conducting state are sudden (in the order of 100 ns) allowing for a fast time correlation between SSG signals and those of other detectors. Its energy threshold is adjustable by setting the external magnetic field at a certain value ΔH just below the phase transition border. The phase diagram of a type-1 superconductor is schematically shown in Fig. 19.6, where H_{sh} is the superheating field, H_{sc} is the supercooling field and H_c is the critical thermodynamic field which is approximately given by $H_c(T) = H_c(0)(1 - (\frac{T}{T_c})^2)$. The region below H_{sc} is the superconducting and above H_{sh} the normal-conducting phase, while the region between the two is the so-called meta-stable phase, which is characteristic for superconductors of type-1. In order to keep the heat capacity as low as possible the SSG detector is operated at a temperature much below the critical temperature T_c at typically $T_0 \approx 100$ mK. Particles interacting in a granule produce quasi-particles. While spreading over the volume of the granule the quasi-particles are losing energy via electron-phonon interactions, thereby globally heating the granule up to a point where it may undergo a sudden phase transition (granule flip). The temperature change experienced by the granule is $\Delta T = \frac{3\Delta E}{4\pi c r^3}$, with ΔE the energy loss of the particle in the grain, c the specific heat and r the radius of the grain. The phase transition of a single grain can be detected by a pickup coil which measures the magnetic flux change $\Delta\Phi$ due to the disappearance of the Ochsenfeld-Meissner effect. In case of a single grain located in the center of the pickup coil the flux change is given by

$$\Delta\Phi = 2\pi B n \frac{r^3}{\sqrt{4R^2 + l^2}} \tag{19.15}$$

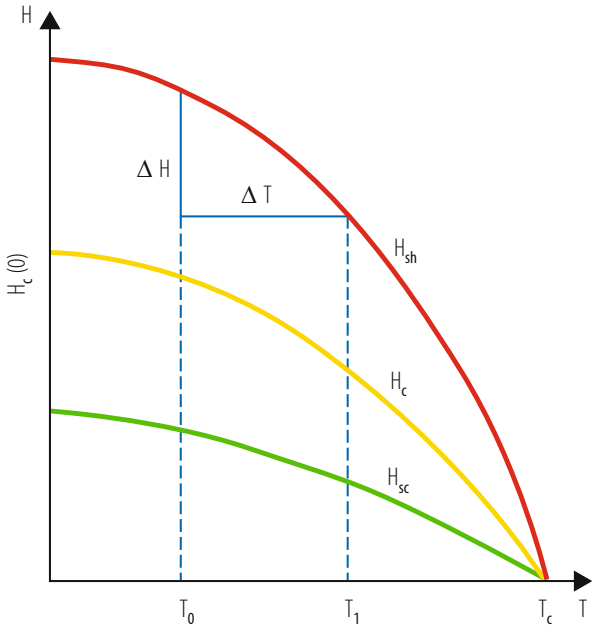

Fig. 19.6 The phase-diagram of a superconductor type1 is shown. H_{sh} is the superheating field, H_{sc} is the supercooling field and H_c is the critical thermodynamic field

with B the applied magnetic field, n the number of windings, R the radius and l the length of the pickup coil. It should be noted that one coil may contain a very large number of grains. If the flipping time τ is small compared to the characteristic time of the readout circuit ($\tau \ll 2\pi\sqrt{LC}$) the flux change induces a voltage pulse in the pick-up coil

$$V(t) = \frac{\Delta\Phi}{\omega LC} e^{-t/2RC} \sin(\omega t)\,, \quad \omega^2 = \frac{1}{LC} - \frac{1}{(2RC)^2} \tag{19.16}$$

with ω, L, R and C being parameters of the pick-up circuit. A detailed description of a readout concept using conventional pick-up coils and electronics including noise estimation is given in [81, 82]. Besides conventional readout coils more sensitive Superconducting Quantum Interference Devices (SQUID) were introduced [83–85]. The SQUID readout allows the detection of single flip signals from smaller size granules and/or the usage of larger size pickup coils. Granules of 20 µm diameter were measured in a large size prototype [85].

Small spherical grains can be produced at low cost by industry using a fine powder gas atomization technique. Since after fabrication the grains are not of a uniform diameter, they have to be sieved to select the desired size. A grain size selection within ±2 µm was achieved.

The Bern Collaboration has built and operated a dark matter SSG detector, named ORPHEUS, which consisted of 0.45 kg of spherical Sn granules with a diameter of ≈30 µm [81]. The detector was read out by 56 conventional pick-up coils, each 6.8 cm long and 1.8 cm in diameter. Each pick-up coil contained ≈80 million granules. The phase transition of each individual grain could be detected with a typical signal to noise ratio of better than 10. The principle to detect small nuclear recoil energies with SSG was successfully tested prior to the construction of the ORPHEUS detector in a neutron beam of the Paul Scherrer Institute (Villigen, Switzerland) [86]. The special cryogenics required for the ORPHEUS detector is described in [87]. The detector is located in the underground facility of the University Bern with an overburden of 70 meter water equivalent (m.w.e.). In its first phase the ORPHEUS dark matter experiment did not reach the sensitivity of other experiments employing cryogenic detectors, as described below. Further improvements on the superconducting behavior of the granules and on the local shielding are necessary.

SSG is a threshold detector. Its resolution depends on the sharpness $\delta H/H$, respectively $\delta T/T$, of the phase transition. It was found that the phase transition smearing depends on the production process of the grains. Industrially produced grains using the atomization technique exhibited a smearing of $\delta H/H \sim 20\%$. By using planar arrays of regularly spaced superheated superconducting microstructures which were produced by various sputtering and evaporation techniques the transition smearing could be reduced to about 2% [88–92]. The improvement of the phase transition smearing is one of the most important developments for future applications of SSG detectors. It looks promising that large quantities of planar arrays can be produced industrially [92].

There is a mechanism by which the energy transfered to a grain can be measured directly. If the grain is held in a temperature bath just below the H_{sc} boundary and the energy (heat) transfer to the grain is large enough to cross the meta-stable region to become normal-conducting, it will after some time cool down again to the bath temperature and become superconducting again. During this process the granule will provide a "flip" signal when crossing the H_{sh} border, and an opposite polarity "flop" signal when crossing the H_{sc} border. The elapsed time between the flip and the flop signal is a measure of the deposited energy in the grain. This effect has been demonstrated with a 11 µm Sn grain bombarded with α particles [93]. It offers the possibility to build an energy resolving and self-recovering SSG detector.

The practical realization of a large SSG detector is still very challenging. Nevertheless, the detector principle offers several unique features:

(a) The large list of suitable type-1 superconductor materials allows to optimize SSG for specific applications.
(b) Very low energy thresholds (eV) can be achieved.
(c) The inductive readout does not dissipate any power into the grains. Therefore the sensitivity of SSG is essentially determined by the grain size and the specific heat of the grain material.

(d) The sudden phase transitions are beneficial for coincident timing with other signals. Generally speaking, SSG detectors are among the most sensitive devices to detect very low energy transfers, i.e. nuclear recoils. A detailed description of SSG can be found in [10, 94, 95].

19.5 Physics with Cryogenic Detectors

19.5.1 Direct Dark Matter Detection

Among the most challenging puzzles in physics and cosmology is the existence of dark matter and dark energy. Dark matter, which was first inferred by Fritz Zwicky in 1933 [96], shows its presence by gravitational interaction with ordinary matter. It holds numerous galaxies together in large clusters and it keeps stars rotating with practically constant velocities around the centers of spiral galaxies. Dark energy, which was discovered by the Supernovae type 1a surveys in 1998 [97, 98], is driven by a repulsive force quite in contrast to the attractive gravitational force and causes the universe to expand with acceleration. The most recent information about the matter/energy content of the universe was gained from the Cosmic Microwave Background radiation (CMB) measurements by the Planck satellite [99]. According to these observations the universe contains 69.4% dark energy, 30.6% matter (including baryonic and dark matter) and 4.8% baryonic matter in form of atoms. The true nature of the dark energy and the dark matter, which fills about 95% of the universe, is still unknown. The direct detection of the dark energy, which is related to Einstein's cosmological constant, seems not to be in reach with present technologies. However, the direct detection of dark matter, if it exists in form of particles, is encouraged by the large expected particle flux which can be deduced under the following assumptions. In an isothermal dark matter halo model the velocity of particles in our galaxy is given by a Maxwell Boltzmann distribution with an average value of $< v >= 230\,\mathrm{km\,s^{-1}}$ and an upper cutoff value of $575\,\mathrm{km\,s^{-1}}$ corresponding to the escape velocity. The dark matter halo density in our solar neighborhood is estimated to be $\rho = 0.3\,\mathrm{GeV\ c^{-2}\,cm^{-3}}$. From that one expects a flux of $\Phi = \rho < v > /m_\chi \sim 7 \cdot 10^6/\mathrm{m}_\chi\,\mathrm{cm^{-2}\,s^{-1}}$ with m_χ the mass of the dark matter particle in GeV $\mathrm{c^{-2}}$. However, since neither the mass nor the interaction cross section of these particles are known one is forced to explore a very large parameter space, which requires very sensitive and efficient detection systems. The most prominent candidates for the dark matter are: massive neutrinos, WIMPs (weakly interacting massive particles) and axions. Neutrinos are among the most abundant particles in the universe, but their masses seem to be too small to contribute significantly to the missing mass. Neutrinos being relativistic at freeze out are free streaming particles, which cluster preferentially at very large scales. Therefore massive neutrinos would enhance large-scale and suppress small-scale structure formations. From hot dark matter and cold dark matter model calculations

fitting the power spectrum obtained from Large Scale Structure (LSS) surveys one obtains a value for the ratio neutrino density to matter density Ω_ν / Ω_m. From this value and Ω_m obtained from CMB an upper limit for the sum of the neutrino masses $\sum m_\nu \leq 0.234\,\mathrm{eV\ c^{-2}}$ can be derived. However, this and the results from direct neutrino mass experiments, as described below, indicate, that neutrinos have a mass to low to qualify for the dark matter. The introduction of axions was not motivated by cosmological considerations, but rather to solve the charge conjugation and Parity violation (CP) problem in Quantum Chromo-Dynamics (QCD) [100]. Nevertheless axions would be produced abundantly during the QCD phase transition in the early universe when hadrons were formed from quarks and gluons. A recent review of axion searches can be found in [101]. The most favored candidate for a WIMP is the neutralino, which is predicted by some Super Symmetric Theories (SUSY) to be the lightest stable SUSY particle. If the neutralino were to be discovered by the Large Hadron Collider (LHC) at CERN, it still would need to be confirmed as a dark matter candidate by direct detection experiments. However, up to now no sign of SUSY-particles has been observed at the LHC [102]. In the following the WIMP searches with some of the most advanced cryogenic detectors are described.

The direct detection of WIMPs is based on the measurement of nuclear recoils in elastic WIMP scattering processes. In the case of neutralinos, spin-independent coherent scatterings as well as spin-dependent scatterings are possible. The expressions for the corresponding cross sections can be found in [103, 104]. In order to obtain good detection efficiencies, devices with high sensitivity to low nuclear recoil energies (eV) are needed. WIMP detectors can be categorized in conventional and cryogenic devices. Most of the conventional WIMP detectors use NaI, Ge crystals, liquid Xenon (LXe) or liquid Argon (LAr). These devices have the advantage that large detector masses (∼ton) can be employed, which makes them sensitive to annual modulations of the WIMP signal owing to the movement of the earth with respect to the dark halo rest frame. Annual modulation, if observed, would provide strong evidence for a WIMP signal, assuming it is not faked by spurious modulated background signals. However, due to quenching of the ionization signals, conventional detectors have lower nuclear recoil detection efficiencies than cryogenic devices.

Cryogenic detectors are able to measure small recoil energies with high efficiency because they measure the total deposited energy in form of ionization and heat. They can be made of many different materials, like Ge, Si, TeO_2, sapphire (Al_2O_3), LiF, $CaWO_4$ and BGO, including superconductors like Sn, Zn, Al, etc. This turns out to be an advantage for the WIMP search, since for a given WIMP mass the resulting recoil spectra are characteristically different for detectors with different materials, a feature which helps to effectively discriminate a WIMP signal against background. If the atomic mass of the detector is matched to the WIMP mass better sensitivity can be obtained due to the larger recoil energies. In comparison to conventional detectors, however, cryogenic detectors are so far rather limited in target mass (∼kg).

Dark matter detectors have to be operated in deep underground laboratories in order to be screened from cosmic-ray background. In addition they need to be shielded locally against radioactivity from surrounding rocks and materials. The shielding as well as the detector itself has to be fabricated from radio-poor materials, which turns out to be rather expensive and limited in its effectiveness. Nevertheless, cryogenic detectors are capable of active background recognition, which allows to discriminate between signals from background minimum ionizing particles, i.e. Compton electrons, and signals from genuine nuclear recoils by a simultaneous but separate measurement of phonons and ionization (or photons) in each event. For the same deposited energy the ionization (or photon) signal from nuclear recoils is highly quenched compared to signals from electrons. The dual phonon-ionization detection method, which was first sugested by Sadoulet [105] and further developed by the CDMS and EDELWEISS collaborations, increases the sensitivity for WIMP detection considerably. A similar idea using scintillating crystals as absorbers and simultaneous phonon-photon detection was introduced by Gonzales-Mestres and Perret-Galix [106] and further developed by the ROSEBUD [107] and CRESST II [108] collaborations. The principle of the method is demonstrated in the scatterplot of Fig. 19.7, taken from [108]. It shows the energy equivalent of the pulse heights measured in the light detector versus those measured in the phonon detector. The scintillating $CaWO_4$ crystal absorber was irradiated with photons and electrons (using Cobalt and Strontium sources respectively) as well as with neutrons (using an Americium-Beryllium source). The photon lines visible in Fig. 19.7 were used

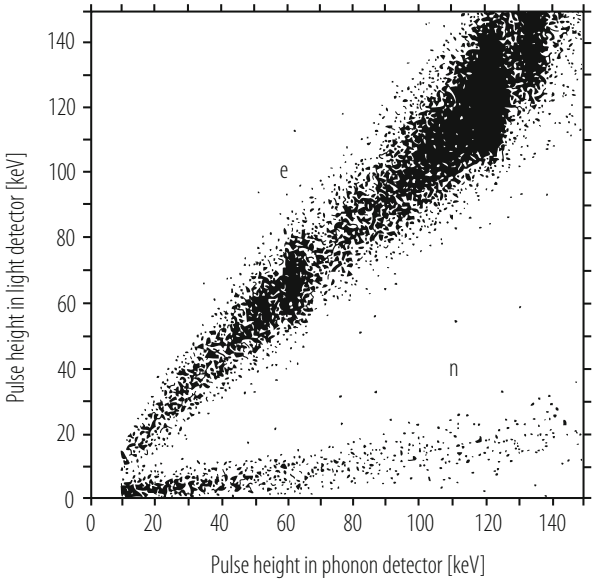

Fig. 19.7 The energy equivalent of the pulse heights measured with the light detector versus those in the phonon detector under electron, photon (e) and neutron (n) irradiation are shown

for the energy calibration in both the light and the phonon detector. The upper band in Fig. 19.7 shows electron recoils (e) and the lower band the nuclear recoils (n). Above an energy of 15 keV 99.7% of the electron recoils can be recognized and clearly distinguished from the nuclear recoils. Active background rejection was also practiced with the ORPHEUS SSG dark matter detector, since minimum ionizing particles cause many granules to flip, while WIMPs cause only one granule to flip (flip meaning a transition from superconducting to normal state) [81]. In the following some of the most sensitive cryogenic WIMP detectors in operation are described.

The CDMS experiment [109] is located at the Soudan Underground Laboratory, USA, with an overburden of 2090 meter water equivalent (m.w.e.). In an early phase of the experiment the cryogenic detectors consisted of 4 towers of 250 g Ge absorbers which where read out by NTD germanium thermistors, so called Berkeley Large Ionization and Phonon (BLIP) detectors, and two towers of 100 g Si absorbers, which where read out by TES sensors, the so-called Z-sensitive Ionization and Phonon based (ZIP) detectors. The ZIP detectors utilize tungsten aluminum Quasi-particle trapping assisted Electrothermal feedback Transition edge sensors (QET). This type of sensor covers a large area of the Si absorber with aluminum phonon collector pads, where phonons are absorbed by breaking Cooper pairs and forming quasi-particles. The quasi-particles are trapped into a meander of tungsten strips which are used as transition edge sensors. The release of the quasi-particle energy in the tungsten strips increases their resistance, which will be observed as a current change in L detected with a SQUID as indicated in Fig. 19.3b. The transition edge device is voltage biased to take advantage of the electrothermal feedback (ETF). The signal pulses of the ZIP detector have rise times of a few µs and fall times of about 50 µs. They are much faster than the signals of the BLIP detector since the ZIP detectors are sensitive to the more energetic non thermal phonons. Their sensitivity to non thermal phonons and the pad structure of the sensors at the surface of the crystal allows for a localization of the event in the x-y plane. A separate circuit collects ionization charges, which are drifted by an electric field of 3 V/cm and collected on two concentric electrodes mounted on opposite sides of the absorber. The ratio of the ionization pulse height to the phonon pulse height versus the pulse height of the phonon detector allows to discriminate nuclear from electron recoils with a rejection factor better than 10^4 and with full nuclear recoil detection efficiency above 10 keV. In order to further improve the sensitivity of the experiment CDMS II is operating at present 19 Ge (250 g each) and 11 Si (100 g each) ZIP type detectors in the Soudan Underground Laboratory at a temperature of about 40 mK. Each detector is 7.62 cm in diameter and 1 cm thick. Limits on the direct detection of WIMPs obtained with the Ge and Si detectors are published in [110] and [111] respectively. The CDMSlite (Cryogenic Dark Matter Search low ionization threshold experiment) uses the Neganov-Luke effect, which leads to an amplification of the phonon signal and allows for lower energy thresholds (56 eV) to be reached [112–114]. In this mode a large detector bias voltage is applied to amplify the phonon signals produced by drifting quasi particle charges. This opens the possibility to extend the WIMP search to masses well below 10 GeV c^{-2}.

Recent results on low mass WIMP searches for spin independent and spin dependent interactions are published in [115].

The EDELWEISS experiment [116], which is located in the Frejus tunnel (4800 m.w.e.), South of France, uses a technique similar to the CDMS BLIP detectors. It consists of 3 towers of 320 g Ge absorbers which are read out by NTD germanium thermistors. For the ionization measurement the detectors are equipped with Al electrodes which are directly sputtered on the Ge absorber crystal. For some data taking runs the EDELWEISS group used towers with amorphous Ge and Si films under the Al electrodes. More data were collected between 2005 and 2011 with the EDELWEISS II detector which contains an array of ten cryogenic Ge detectors with a mass of 400 g each [117]. As an upgrade of EDELWEISS II the collaboration developed EDELWEISS III with 36 FID (Fully Inter-Digital) detectors based on cylindrical Ge crystals with a mass of about 800 g each operating at 18 mK [118].

Early experience with the ionization measurements showed a severe limitation of the background separation capability due to insufficient charge collection of surface events. The effect can be attributed to a plasma screening of the external electric field of the electrodes. As a result surface interactions of electrons can fake nuclear recoil events. One way to solve the problem was developed by the Berkeley group by sputtering films of amorphous Si or Ge on the absorber surface before deposition of the Al electrodes [119]. Due to the modified energy band gap of the amorphous layer the charge collection efficiency was largely improved. Fast phonon detectors like ZIP allow to identify surface events by measuring the relative timing between the phonon and ionization signals as demonstrated by the CDMS experiment [120]. The surface event problem completely disappears when using dual phonon-photon detection. This method was chosen by the CRESST II collaboration.

The CRESST experiment [121] is located in the Gran Sasso Underground Laboratory (3800 m.w.e.) north of Rome, Italy. It uses scintillating $CaWO_4$ crystals as absorber material. The detector structure can hold 33 modules of absorber which can be individually mounted and dismounted. Each module weights about 300 g. The detector operates at about 10 mK. The phonon signal from the $CaWO_4$ crystal is read by a superconducting tungsten TES thermometer and the photon signal by a separate but nearby cryogenic light detector, which consists of a silicon wafer with a tungsten TES thermometer. For an effective background discrimination the light detector has to be very efficient. This was achieved by applying an electric field to the silicon crystal leading to an amplification of the thermal signal due to the Neganov-Luke effect [122]. The time constant of the emission of scintillation photons from $CaWO_4$ at mK temperatures is of the order of ms, which requires a long thermal relaxation time for the light detector. The characteristics of the background rejection power depends on the knowledge of the quenching factor, which is the reduction factor of the light output of the nuclear recoil event relative to an electron event. These quenching factors were measured by the CRESST collaboration for various recoiling nuclei in $CaWO_4$ in a separate experiment [123]. The knowledge of these quenching factors would allow in principle to identify WIMP interactions with different nuclei in the $CaWO_4$ crystal. This method seems very promising, not only for identifying the background but also the quantum

numbers of the WIMP candidates. The three types of nuclei in $CaWO_4$ together with a low nuclear recoil energy threshold of 300 eV allowed CRESST II to extend the dark matter search with high sensitivity into a mass region below 10 GeV c^{-2} [124].

Besides the dark matter search the CRESST collaboration is developing a cryogenic detector to measure coherent neutrino nucleus scattering [125]. This process is predicted by the Standard Model (SM), but has been unobserved so far. If successful, a possible application could be the real time monitoring of nearby nuclear power plants. With a small size prototype cryogenic Saphire detector with a weight of 0.5 g a recoil energy threshold of 20 eV was achieved [126]. Nevertheless, the first detection of coherent neutrino scattering was reported from the COHERENT collaboration only recently [127]. They measured neutrino-induced recoils with conventional scintillating CsI(Na) crystals with a weight of 14.5 kg. Their experiment was located in a basement under the Oak Ridge National Laboratory Spallation Neutron Source.

The experimental results are usually presented as exclusion plots, which show the WIMP-nucleon cross section versus the WIMP mass. They are derived from the expected nuclear recoil spectrum for a given set of parameters [104]:

$$\frac{dR}{dE} = \frac{\sigma_0 \rho_\chi}{2\mu^2 m_\chi} F^2(E) \int_{\upsilon_{min}}^{\upsilon_{max}} \frac{f(\upsilon)}{\upsilon} d\upsilon \tag{19.17}$$

with m_χ the mass of the WIMP, μ the reduced mass of the WIMP-nucleus system, σ_0 the total elastic cross section at zero momentum transfer, $\rho_\chi = 0.3\,\mathrm{GeV\,cm^{-3}}$ the dark matter halo density in the solar neighborhood, $F(E)$ the nuclear form factor, $f(\upsilon)$ an assumed isothermal Maxwell-Boltzmann velocity distribution of the WIMPs in the halo, $\upsilon_{min} = \sqrt{E m_N / 2\mu^2}$ the minimum velocity which contributes to the recoil energy E, and $\upsilon_{max} = 575\,\mathrm{km\,s^{-1}}$ the escape velocity from the halo. The recoil energy E is given by $E = \mu^2 \upsilon^2 (1 - \cos\theta)/m_N$, with m_N the mass of the nucleus, υ the velocity of the WIMP, and θ the scattering angle in the centre of mass system. The expected nuclear recoil spectrum for interactions with WIMPs of a given mass will then be folded with the detector response, which was obtained experimentally from calibration measurements with neutron sources or in neutron beams. From a maximum likelihood analysis, an upper limit cross section value (90% C.L.) can be extracted for several different WIMP masses. Current limits for spin-independent WIMP interactions are depicted in Fig. 19.8, taken from [128]. The Figure includes only a selection of some of the most sensitive experiments. WIMP masses below 4 GeV c^{-2} are accessible by detectors like CRESST II [124], CDMSlite [115], EDELWEISS [118] and DAMIC [129] because of their low recoil energy thresholds and/or their light absorber nuclei. For WIMP masses above 6 GeV c^{-2} the best constraints are provided by experiments like XENON1T [130], LUX [131], PANDAX II [132], XENON 100 [133] and Dark Side 50 [134], which are based on massive dual phase (liquid and gas) Xenon or Argon detectors with time projection (TPC) read out. The DAMA experiment has observed

an annual modulation signal, which they claim is satisfying the requirements of a dark matter annual modulation signal [135]. Their detector is operated in the Gran Sasso Laboratory in Italy (LNGS) and is based on highly radiopure NaI (Tl) crystal scintillators. Similar results, but less significant, were reported by the CoGeNT experiment with a cryogenic Ge detector in the Soudan Underground Laboratory (SUL) [136]. Annual modulations have not been observed by other even more sensitive experiments and the interpretation of a WIMP signal is controversial. In order to better understand the origin of the observed modulation the SABRE experiment is planning to build twin detectors one of which will be placed in the northern hemisphere at the LNGS and the other in the southern hemisphere at the Stanwell Underground Physics Laboratory (SUPL) in Australia [137]. Both detectors will be identical and based on the same target material used in the DAMA experiment.

An extraction of the spin-dependent WIMP-nucleon cross section in a model independent way is not possible, since the nuclear and the SUSY degrees of freedom do not decouple from each other. Nevertheless, when using an "odd group" model which assumes that all the nuclear spin is carried by either the protons or the neutrons, whichever are unpaired, WIMP-nucleon cross sections can be deduced. The CDMSlite experiment [115] achieved constraints for spin dependent interactions below WIMP masses of 4 GeV c^{-2} complementary to LUX, PANDAX, XENON 100 and PICASSO [138].

Several experiments are planning to extend their sensitivity to a wide range of parameter space by operating multi tonnes of target material, reducing the energy thresholds and background level until the irreducible solar, earth and atmospheric neutrino background level is reached, Fig. 19.8. The EURECA (European Underground Rare Event Calorimeter) will bring together researchers from the CRESST

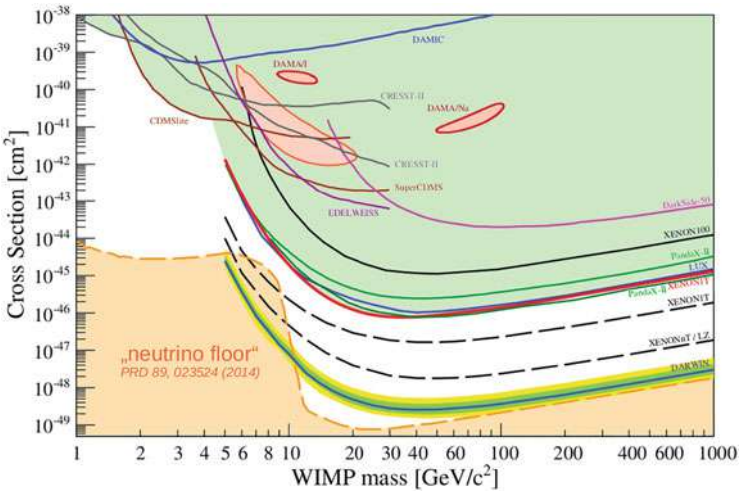

Fig. 19.8 Current limits for spin-independent WIMP interactions are shown

and EDELWEISS experiments to built a 1 ton cryogenic detector in the Modane Underground Laboratory in France [139]. SUPER CDMS will be operated in the Sudbury Neutrino Observatory (SNOLab) in Canada and is based on cryogenic Ge and Si absorber materials to increase their sensitivity for dark matter interaction cross sections to $10^{-43}\,\mathrm{cm}^2$ for masses down to $1\,\mathrm{GeV\ c}^{-2}$ [140]. The Dark Side 50 collaboration is planning to built a 23 ton dual phase liquid Argon TPC to be operated at the LNGS. The LUX-ZEPLIN (LZ) experiment is currently under construction in the Sanford Laboratory in South Dakota. It uses 10 ton of liquid XENON (dual phase) in a radio-poor double vessel cryostat [142]. The ultimate WIMP detector is proposed by the DARWIN collaboration at the LNGS [141]. It will be based on multi tonnes of liquid Xenon and will fill almost the entire parameter space for spin independent WIMP interaction cross sections down to the background level of neutrino interactions in the detector material as shown in Fig. 19.8. The ambitious project will also be sensitive to other rare interactions like solar axions, galactic axion like particles, neutrinoless double beta decay in ^{136}Xe and coherent neutrino nucleus scatterings.

19.5.2 Neutrino Mass Studies

Since the discovery of neutrino oscillations by the Kamiokande and Super-Kamiokande experiments [143] a new chapter in physics started. These findings showed that neutrinos have a mass and that there is new physics to be expected beyond the Standard Model (SM) in particle physics. Among the most pressing questions remain the absolute values of the neutrino masses, since from oscillation experiments only mass differences can be obtained [144], and the Dirac or Majorana type character of the neutrino. The main streams in this field focus upon the search for the neutrinoless double beta decay and the endpoint energy spectrum of beta active nuclei. Cryogenic detectors are particularly well suited for this type of research since they provide excellent energy resolutions, an effective background discrimination and a large choice of candidate nuclei.

19.5.2.1 Neutrinoless Double Beta Decay

Double beta decay was first suggested in 1935 by Maria Goeppert Mayer [145]. It is the spontaneous transition from a nucleus (A,Z) to its isobar (A,Z+2). This transition can proceed in two ways: (A,Z)⇒(A,Z+2) + 2 e^- + 2 $\bar{\nu}_e$ or (A,Z)⇒(A,Z+2) + 2 e^-. In the first channel, where two electrons and two antineutrinos are emitted, the lepton number is conserved. It is the second channel, the neutrinoless double beta decay ($0\nu\beta\beta$), where the lepton number is violated. In this case, with no neutrino in the final state, the energy spectrum of the decay would show in a peak which represents the energy sum of the two electrons. The experimental observation of this process would imply that neutrinos are Majorana particles, meaning that the

neutrino is not distinguishable from its antiparticle and that it has a non-vanishing mass. From the measured decay rate $(1/T^{0\nu}_{1/2})$ one can derive in principle its effective mass $< m_\nu >$ or a lower limit of it:

$$(1/T^{0\nu}_{1/2}) = G_{0\nu}(E_0, Z) \mid M_{0\nu} \mid^2 < m_\nu >^2 \tag{19.18}$$

where $G_{0\nu}(E_0, Z)$ is an accurately calculable phase space function and $M_{0\nu}$ is the nuclear matrix element, which is not very well known [146]. The calculated values of $M_{0\nu}$ can vary by factors up to two. Consequently the search for $0\nu\beta\beta$ should be made with several different nuclei in order to confirm an eventual discovery of this important process.

The Milano group has developed an experiment with the name CUORICINO to search for the neutrinoless double beta decay of ^{130}Te. The experiment is located in the Gran Sasso Underground Laboratory. The detector consists of an array of 62 TeO_2 crystals with the dimensions $5 \times 5 \times 5\,cm^3$ (44 crystals) and $3 \times 3 \times 3\,cm^3$ (18 crystals) and a total mass of 40.7 kg. The crystals are cooled to ~8 mK and attached to Ge NTD thermistors for phonon detection. Among other possible nuclear candidates (like for example $^{48}CaF_2$, ^{76}Ge, $^{100}MoPbO_4$, $^{116}CdWO_4$, $^{150}NdF_3$, $^{150}NdGaO_3$), $^{130}TeO_2$ was chosen because of its high transition energy of 2528.8 ± 1.3 keV and its large isotopic abundance of 33.8%. Published first results of the CUORICINO experiment [147] show no evidence for the $0\nu\beta\beta$ decay, but they set a lower limit on the half lifetime $T^{0\nu}_{1/2} \geq 1.8 \cdot 10^{24}$ yr (90% C.L.) corresponding to $< m_\nu > \leq$ 0.2 to 1.1 eV (depending on nuclear matrix elements). In a next step the collaboration developed CUORE-0 as prototype for a larger detector CUORE. Its basic components consist of 52 TeO_2 crystals with dimensions $5 \times 5 \times 5\,cm^3$ and a total weight of 39 kg corresponding to 10.9 kg ^{130}Te. CUORE-0 was operated in the CUORICINO cryostat at 12 mK. The data taken from 2013 to 2015 show no evidence for a neutrinoless double beta signal. Combined with the CUORICINO results a limit on the half lifetime $T^{0\nu}_{1/2} \geq 4 \cdot 10^{24}$ yr (90% C.L.) corresponding to $< m_\nu > \leq$ 270 to 760 meV (depending on nuclear matrix elements) was achieved [148]. CUORE, contains 19 CUORE-0 type towers with 988 TeO_2 crystals of a total mass of 741 kg corresponding to 206 kg of ^{130}Te [149, 150]. The array will be cooled in a large cryostat to 10 mK. It started commissioning early 2017 and aims for a sensitivity to reach limits of $T^{0\nu}_{1/2} \geq 9 \cdot 10^{25}$ yr in 5 years running time [150]. For the future the CUPID collaboration plans to develope a tonne-scale cryogenic detector which will be based on the experience gained with the CUORE experiment [151].

Several experiments investigated other nuclei and set stringent upper limits on the decay rates, for example: KamLand-Zen in ^{136}Xe [152], EXO-200 in ^{136}Xe [153], GERDA in ^{76}Ge [154], NEMO-3 in ^{100}Mo [155]. So far no neutrinoless double beta signal was seen. Currently a limit on the half lifetime $T^{0\nu}_{1/2} \geq 1.07 \cdot 10^{26}$ yr (90% C.L.) corresponding to $< m_\nu > \leq$ 60 to 165 meV (depending on nuclear matrix elements) was achieved by the KamLand-Zen experiment. An ambitious alternative approach in looking for Majorana versus Dirac type neutrinos is proposed by the

PTOLEMY experiment in studying the interaction of cosmic relic neutrinos with Tritium [156].

19.5.2.2 Direct Neutrino Mass Measurements

So far the best upper limit for the electron neutrino mass of 2.2 eV was obtained from the electron spectroscopy of the tritium decay $^3\mathrm{H} \Rightarrow {}^3\mathrm{He} + e^- + \bar{\nu}_e$, with a transition energy of 18.6 keV, by the Mainz and the Troitsk experiments [157]. In the near future the KATRIN experiment, which measures the same decay spectrum with a much improved electron spectrometer, will be in operation aiming for a neutrino mass sensitivity down to 0.2 eV [158].

One of the problems with experiments based on a spectroscopic measurement of the emitted electrons is that they yield negative values for the square of the neutrino mass when fitting the electron energy spectrum. This is mainly due to final state interactions (like tritium decays into excited atomic levels of ^{3}He), which lead to deviations from the expected energy spectrum of the electron. Low temperature calorimeters provide an alternative approach, since they measure the total energy including final state interactions, such as the de-excitation energy of excited atomic levels. However, in order to reach high sensitivity for low neutrino masses the detector has to have an excellent energy resolution and enough counting rate statistics at the beta endpoint energy. The Genoa group [159] pioneered this approach and studied the beta decay of $^{187}\mathrm{Re} \Rightarrow {}^{187}\mathrm{Os} + e^- + \bar{\nu}_e$ with a cryogenic micro-calorimeter. Their detector was a rhenium single crystal (2 mg) coupled to a Ge NTD thermistor. Rhenium is a super-conductor with a critical temperature of 1.7 K. Natural Re contains 62.8% of ^{187}Re with an endpoint energy of about 2.6 keV. The operating temperature of the detector was $T = 90$ mK. In their first attempt they obtained precise values for the beta endpoint energy and the half life of the ^{187}Re beta decay and were able to obtain an upper limit of the electron neutrino mass of 19 eV (90% CL) or 25 eV (95% CL) [160]. Following this approach the Milan group [161] has built an array of ten thermal detectors for a ^{187}Re neutrino mass experiment. The detectors were made from $AgReO_4$ crystals with masses between 250 and 350 µg. The crystals were coupled to Si implanted thermistors. Their average energy resolution (FWHM) at the beta endpoint was 28.3 eV, which was constantly monitored by means of fluorescence X-rays. The natural fraction of ^{187}Re in $AgReO_4$ yields a decay rate of $5.4 \cdot 10^{-4}$ Hz/µg. From a fit to the Curie plot of the ^{187}Re decay they obtained an upper limit for $m_{\bar{\nu}_e} \leq 15$ eV. Their measured value for the beta endpoint energy is 24653.3 ± 2.1 eV and for the half live is $(43.2 \pm 0.3) \cdot 10^9$ yr. A higher sensitivity to low neutrino masses may be achievable in the future, provided that the energy resolution and the statistics at the beta endpoint energy can be improved significantly. The latter may raise a problem for thermal phonon detectors, since their signals are rather slow and therefore limit the counting rate capability to several Hz.

With their Rhenium cryogenic micro-calorimeters the Genoa group [162, 163] and the Milano collaboration [164] were also able to measure interactions between

the emitted beta particle and its local environment, known as beta environmental fine structure (BEFS). The BESF signal originates from the interference of the outgoing beta electron wave and the reflected wave from the atoms in the neighbourhood. BEFS is similar to the well known Extended X-ray Absorption Fine Structure (EXAFS) method. Their results demonstrated that cryogenic micro-calorimeters may also offer complementary new ways for material sciences to study molecular and crystalline structures.

Currently several groups MARE [165, 166], ECHo [167], HOLMES [168], NUMECS [169], are investigating the possibility to measure the electron neutrino mass from the Electron Capture (EC) decay spectrum of Holmium (^{163}Ho) using cryogenic micro-calorimeters. This approach was originally suggested by A. de Rujula and M. Lusignoli in 1982 [170]. ^{163}Ho decays via EC into ^{163}Dy with a half life of 4570 years and a decay energy of 2.833 keV. It does not occur naturally and it is not commercially available. It has to be produced by neutron or proton irradiation. After purification the ^{163}Ho atoms have to be implanted into a suitable absorber material of the micro-calorimeter. In order to reach a neutrino mass sensitivity in the sub-eV region a total ^{163}Ho activity of several MBq is required. Since the activity of a single micro-calorimeter should not exceed 100 Bq the total ^{163}Ho activity has to be distributed over a large number of pixels (10^5). The groups are devoting much effort in developing micro-calorimeters with energy and time resolutions of the order of 1 eV and 1 µs respectively. Various thermal sensors, like TES, MMC and MKID, are considered. Multiplexing schemes have still to be invented to be able to read out the enormous number of pixels.

As already mentioned above, an upper limit for all neutrino masses of $\sum m_\nu \leq 0.234\,\mathrm{eV\,c^{-2}}$ was derived from cosmology. It will still take some efforts to reach or go below these limits in the near future with direct mass measurements. A review of direct neutrino mass searches can be found in [171].

19.5.3 Astrophysics

Modern astrophysics addresses a large list of topics: Formation of galaxies and galaxy clusters, the composition of the intergalactic medium, formation and evolution of black holes and their role in galaxy formation, matter under extreme conditions (matter in gravitational fields near black holes, matter inside neutron stars), supernovae remnants, accretion powered systems with white dwarfs, interstellar plasmas and cosmic microwave background radiation (CMB). The investigation of these topics requires optical instruments with broad band capability, high spectral resolving power, efficient photon counting and large area imaging properties. The radiation received from astrophysical objects spans from microwaves, in the case of CMB, to high energy gamma rays. The subjects discussed here can be divided into three categories, X-ray, optical/ultraviolet (O/UV) and CMB observations. In order to avoid the absorptive power of the earth atmosphere many of the instruments are operated in orbiting observatories, in sounding rockets or in balloons. Progress in

this rapidly growing field of science is constantly asking for new instrumentation and new technologies. Cryogenic detectors are playing a key role in these developments providing very broad-band, imaging spectrometers with high resolving power. They also feature high quantum efficiencies, single photon detection and timing capabilities. The observation of large-scale objects, however, needs spatial-spectral imaging devices with a wide field of view requiring cryogenic detectors to be produced in large pixel arrays. The fabrication and the readout of these arrays remains still a big challenge.

19.5.3.1 X-Ray Astrophysics

The orbital X-ray observatories Chandra and XMM-Newton contained CCD cameras for large field imaging and dispersive spectrometers for narrow field high spectral resolution. Cryogenic devices are able to combine both features in one instrument. Although they have not yet reached the imaging potential of the 2.5 megapixel CCD camera on XMM-Newton and the resolving power $E/\Delta E_{FWHM} = 1000$ at $E = 1\,\mathrm{keV}$ of the dispersive spectrometer on Chandra, their capabilities are in many ways complementary. For example, the resolving power of cryogenic devices increases with increasing energy and is above 2 keV better than the resolving power of dispersive and grating spectrometers, which decreases with increasing energy. Since the cryogenic pixel array provides a complete spectral image of the source at the focal plane its resolving power is independent of the source size. Cryogenic detectors also provide precise timing information for each photon allowing to observe rapidly varying sources such as pulsars, etc. They cover a wide range of photon energies (0.05–10 keV) with a quantum efficiency of nearly 100%, which is 5 times better than the quantum efficiency (20%) of CCDs.

The first space-borne cryogenic X-ray Quantum Calorimeter (XQC), a collaboration between the Universities of Wiskonsin, Maryland and the NASA Goddard Space Flight Center, was flown three times on a sounding rocket starting in 1995 [172]. The rockets achieved an altitude of 240 km providing 240 s observation above 165 km per flight. The XQC was equipped with a 2×18 micro-calorimeter array consisting of HgTe X-ray absorbers and doped silicon thermistors yielding an energy resolution of 9 eV across the spectral band. The pixel size was $1\,\mathrm{mm}^2$. The micro-calorimeters were operated at 60 mK. It is interesting to note that, when recovering the payload after each flight, the dewar still contained some liquid helium. The purpose of the mission was to study the soft X-ray emission in the band of 0.03–1 keV. The physics of the diffuse interstellar X-ray emission is not very well understood. It seems that a large component is due to collisional excitations of particles in an interstellar gas with temperatures of a few 10^6 K. A detailed spectral analysis would allow to determine the physical state and the composition of the gas. Since the interstellar gas occupies a large fraction of the volume within the galactic disk, it plays a major role in the formation of stars and the evolution of the galaxy. The results of this experiment and their implications are discussed in [172]. As a next step the Japan-USA collaboration has put an X-ray spectrometer

(XRS) on board of the Astro-E2 X-ray Suzaku satellite which was launched in July 2005. This instrument is equipped with 32 pixels of micro-calorimeters (HgTe) and semiconducting thermistors and is an improvement over the XQC spectrometer in terms of fabrication techniques, thermal noise, energy resolution of 7 eV across the operating band of 0.03–10 keV and observation time [173, 174]. The pixels are 0.624 mm^2 and arranged in a 6×6 array giving a field of view of 2.9×2.9 arcmin. The observatory is looking at the interstellar medium in our and neighboring galaxies as well as at supernovae remnants. The investigations include super massive black holes and the clocking of their spin rate.

The next step is to develop cryogenic detectors with increased pixel numbers (30×30) and energy resolutions of 2 eV, which would be able to replace dispersive spectrometers in future experiments. Superconducting micro-calorimeters with TES sensors have the potential to reach energy resolutions of 2 eV and are likely to replace semiconducting thermistors. One of the problems with cryogenic detectors is that they do not scale as well as CCDs, which are able to clock the charges from the center of the arrays to the edge using a serial read out. Cryogenic detectors rely on individual readout of each pixel. Another problem is the power dissipation in large arrays. To solve some of these problems, dissipation-free metallic magnetic calorimeters (MMC) and microwave kinetic inductance detectors (MKID) [80] are among the considered possibilities. Details of these new developments can be found in the proceedings of LTD. Large cryogenic detector arrays are planned for ATHENA (Advanced Telescope for High Energy Astrophysics), a future X-ray telescope of the European Space Agency. It is designed to investigate the formation and evolution of large scale galaxy clusters and the formation and grows of super massive black holes. The launch of ATHENA is planned for 2028.

19.5.3.2 Optical/UV and CMB Astrophysics

Since the first optical photon detection with STJ's in 1993 [175] and TES's in 1998 [176] a new detection concept was introduced in the field of Optical/UV astrophysics. Further developments demonstrated the potential of these single photon detection devices to combine spectral resolution, time resolution and imaging in a broad frequency band (near infrared to ultraviolet) with high quantum efficiency. The principle of these spectrophotometers is based on the fact that for a superconductor with a gap energy of typically 1 meV an optical photon of 1 eV represents a large amount of energy. Thus a photon impinging on a superconductor like for example Ta creates a large amount of quasi-particles leading to measurable tunnel current across a voltage biased junction. A first cryogenic camera S-Cam1 with a 6×6 array of Ta STJs (with a pixel size of $25 \times 25\,\mu m^2$) was developed by ESTEC/ESA and 1999 installed in the 4 m William Herschel telescope on La Palma (Spain) [75]. For a first proof of principle of this new technique the telescope was directed towards the Crab pulsar with an already known periodicity of 33 ms. The photon timing information was recorded with a 5 µs accuracy with respect to the GPS timing signals. Following the success of the demonstrator model S-Cam1

and of the improved model S-Cam2 a new camera S-Cam3 consisting of a 10×12 Ta STJ pixel array (with pixel dimension $33 \times 33\,\mu m^2$) was installed at the ESA 1m Optical Ground Station Telescope in Tenerife (Spain). The STJ structure is 100 nm Ta/30 nm Al//AlOx//30 nm Al/100 nm Ta. The camera covers a wavelength range of 340–740 nm with a wavelength resolution of 35 nm at $\lambda = 500$ nm and has a pulse decay time of 21 µs [177]. The Stanford-NIST (National Institute of Standards and Technology) collaboration has also developed a camera with an 8×8 pixel array based on tungsten TES sensors on a Si substrate [178]. Each pixel has a sensitive area of $24 \times 24\,\mu m^2$ and the array has a $36 \times 36\,\mu m$ center to center spacing. In order to improve the array fill factor a reflection mask is positioned over the inter-pixel gaps. For both STJ and TES spectro-photometers thermal infrared (IR) background radiation, which increases rapidly with wavelength above 2 µm, is of concern. Special IR blocking filters have to be employed in order to extend the wavelengths range of photons from 0.3 µm out to 1.7 µm. A 4 pixel prototype of the Stanford-NIST TES instrument was already mounted at the 2.7 m Smith Telescope at McDonald Observatory (U.S.A) and observed a number of sources including spin powered pulsars and accreting white dwarf systems. The Crab pulsar served as a source to calibrate and tune the system. The obtained data are published in [179]. Already these first results have shown that STJ or TES based spectrophotometers are in principal very promising instruments to study fast time variable sources like pulsars and black hole binaries as well as faint objects, like galaxies in their state of formation. However, in order to extend the observations from point sources to extended objects much larger pixel arrays are required. Future developments concentrate on a suitable multiplexing system in order to increase the number of pixels, which are presently limited by the wiring on the chip and the size of the readout electronics. SQUID multiplexing readout systems [180–182] as well as Distributed Read-out Imaging Devices (DROID) [74, 183], in which a single absorber strip is connected to two separate STJs on either side to provide imaging capabilties from the ratio of the two signal pulses, are under study. However, these devices are slower than small pixel devices and can handle only lower count rates. A much faster device is the superconducting MKID detector [80], which allows a simple frequency-domain approach to multiplexing and profits from the rapid advances in wireless communication electronics. A more detailed review can be found in [184, 185]. A camera, ACRONS, for optical and near infrared spectroscopy has been developed. The camera contains a 2024 pixel array of cryogenic MIKD detectors. The device is able to detect individual photons with a time resolution of 2 µs and simultaneous energy information [80]. The instrument has been used for optical observations of the Crab pulsar [186].

Large cryogenic pixel antennas have been developed for ground-based and space-borne CMB polarization measurements. These devices aim to be sensitive to the detection of the so-called primordial E- and B-modes, which would appear as curling patterns in the polarization measurements. E-modes arise from the density perturbations while B- modes are created by gravitational waves in the early universe. The two modes are distinguishable through their characteristic patterns. However, B-mode signals are expected to be an order of magnitude weaker than E-

modes. Nevertheless, B-modes are of particular interest since they would provide revealing insight into the inflationary scenario of the early universe signaling the effect of primordial gravity waves. There are several instruments, which use TES based cryogenic bolometers, in operation: EBEX [187] and SPIDER [188] are balloon-borne experiments, overflying the Antarktis. POLARBEAR [189] is an instrument, which is coupled to the HUAN TRAN Teleskope at the James Ax Observatory in Chile. BICEP2 and the Keck Array [190] are located at the Amundsen-Scott South Pole Station. The PLANCK Satellite [191] carried 48 cryogenic bolometers operating at 100 mK in outer space. In March 2014 the BICEP2 collaboration reported the detection of B-modes [190]. However the measurement was received with some skepticism and David Spergel argued that the observation could be the result of light scatterings off dust in our galaxy. In September 2014 the PLANCK team [191] concluded that their very accurate measurement of the dust is consistent with the signal reported by BICEP2. In 2015 a joint analysis of BICEP2 and PLANCK was published concluding that the signal could be entirely attributed to the dust in our galaxy [192].

19.6 Applications

Electron probe X-ray microanalysis (EPMA) is one of the most powerful methods applied in material sciences. It is based on the excitation of characteristic X-rays of target materials by high current electron beams in the energy range of several keV. EPMA finds its application in the analysis of contaminant particles and defects in semiconductor device production as well as in the failure analysis of mechanical parts. It is often used in the X-ray analysis of chemical shifts which are caused by changes in the electron binding due to chemical bonding as well as in many other sciences (material, geology, biology and ecology). The X-rays are conventionally measured by semiconducting detectors (Si-EDS), which are used as energy dispersive spectrometers covering a wide range of the X-ray spectrum, and/or by wavelength dispersive spectrometers (WDS). Both devices have complementary features. WDS, based on Bragg diffraction spectrometry, has a typical energy resolution of 2–15 eV (FWHM) over a large X-ray range. However, the diffraction limits the bandwidth of the X-rays through the spectrometer as well as the target size, which acts as a point source, and makes serial measurements necessary, which is rather time consuming. Contrary, the Si-EDS measures the entire X-ray spectrum from every location of the target simultaneously, but with a typical energy resolution of 130 eV. Si-EDS is therefore optimally suited for a quick but more qualitative analysis. Cryogenic micro-calorimeter EDS provides the ideal combination of the high resolution WDS and the broadband features of the energy dispersive EDS. The application of cryogenic detectors for EPMA was first introduced by Lesyna et al. [193]. The NIST group developed a prototype TES based micro-calorimeter which is suitable for industrial applications [42, 194, 195]. It consists of a Bi absorber and a Al-Ag or Cu-Mo bi-layer TES sensor. It covers

an area of $0.4 \times 0.4\,\mathrm{mm}^2$ and yields an energy resolution of 2 eV at 1.5 keV and 4.5 eV at 6 keV. The detector is cooled to 100 mK by a compact adiabatic demagnetization refrigerator and is mounted on a scanning electron microscope. Cryogenic refrigerators for this and various other applications are commercially available [196]. In spite of its excellent resolving power (see Fig. 19.2) the micro-calorimeter EDS has still two shortcomings. It has a limited counting rate capability (1 kHz compared to 100 kHz WDS and 25 kHz Si-EDS) and a small effective detector surface. To compensate for the latter the NIST group developed an X-ray focusing device using poly-capillary optics. The device consists of many fused tapered glass capillaries which focus the X-rays by means of internal reflection onto the micro-calorimeter increasing its effective area. Another solution under study is a multiplexed micro-calorimeter array with a possible loss in resolution due to the variability of individual detectors [182]. Nevertheless, TES based cryogenic micro-calorimeters EDS have already demonstrated major advances of EPMA in scientific and industrial applications.

Time of flight mass spectrometry (ToF-MS) of biological molecules using cryogenic detectors was first introduced by D. Twerenbold [197]. The main advantage of cryogenic calorimeters over traditionally employed micro-channel plates (MP) is that the former are recording the total kinetic energy of an accelerated molecule with high efficiency, independent of its mass, while the efficiency of the latter is decreasing with increasing mass due to the reduction of the ionization signal. ToF-MS equipped with MP lose rapidly in sensitivity for masses above 20 kDa (proton masses). The disadvantages of cryogenic detectors are: first, they cover only a rather small area of $\sim 1\,\mathrm{mm}^2$ while a MP with $4\,\mathrm{cm}^2$ will cover most of the beam spot size of a spectrometer; second, the timing signals of the cryogenic calorimeters are in the range of µs and therefore much slower than ns signals from MP, which degrades the flight time measurements and thus the accuracy of the molecular mass measurements. A good review of early developments can be found in [198]. After the early prototype experiences made with STJs [199, 200] and NIS (Normal-conductor/Isolator/Super-conductor) tunnel junctions [201], which provided only very small impact areas, super-conducting phase transition thermometers (SPT) with better time of flight resolutions and larger impact areas of $3 \times 3\,\mathrm{mm}^2$ were developed [202–204]. These devices consist of thin super-conducting Nb meanders, or super-conducting films in thermal contact with an absorber, which are current biased and locally driven to normal conducting upon impact of an ion. A voltage amplifier is used to measure the signal pulse.

Cryogenic detectors as high resolution γ-ray, α and neutron spectrometers also found applications in nuclear material analysis, as broad band micro-calorimeters in Electron Beam Ion Traps (EBITs) and in synchrotrons for fluorescence-detected X-ray absorption spectroscopy (XAS) [202]. They are also employed in nuclear and heavy ion physics [205].

19.7 Summary

Cryogenic detectors have been developed to explore new frontiers in astro and particle physics. Their main advantages over more conventional devices are their superior energy resolution and their sensitivity to very low energy transfers. However, most thermal detectors operate at mK temperatures requiring complex refrigeration systems and they have limited counting rate capabilities (1 Hz–1 kHz). Today the most frequently used thermometers for calorimeters operating in near equilibrium mode are doped semiconductors (thermistors), superconducting transition edge sensors (TES) and metallic paramagnets (MMC). Because they are easy to handle and commercially available thermistors are quite popular. They have, however, the disadvantage of having to deal with Joule heating introduced by their readout circuit. The most advanced technology is provided by the TES sensors in connection with an auto-biasing electrothermal feedback system. This system reduces the effect of Joule heating, stabilizes the operating temperature and is self-calibrating, which turns out to be advantageous also for the operation of large detector arrays. The main advantage of MMCs is their magnetic inductive readout, which does not dissipate power into the system. This feature makes MMC attractive for applications, where large detector arrays are required. Non-equilibrium detectors like superconducting tunnel junctions (STJ), superheated superconducting granules (SSG) and microwave kinetic inductance devices (MKID) are based on the production and detection of quasi-particles as a result of Cooper pair breaking in the superconductor. These devices are intrinsically faster, providing higher rate capabilities (10 kHz and more) and good timing properties suitable for coincident measurements with external detectors. Because of their sensitivity to low energy photons arrays of STJs are frequently employed in infrared and optical telescopes, but also efficiently used in x-ray spectroscopy. SSG detectors with inductive readout have the potential to reach very low energy thresholds (order of several eV) which would be advantageous for various applications like for example in neutrino physics (coherent neutrino scattering, etc.). But the practical realization of SSG detectors is still very challenging. MKIDs provide an elegant way to readout large detector arrays by coupling an array of many resonators with slightly different resonance frequencies to a common transmission line with a single signal amplifier. Due to this feature they are very suited for the future instrumentation of astrophysical observatories and other applications. Cryogenic calorimeters with a large detector mass for dark matter searches and neutrino physics as well as large detector arrays for astrophysical measurements and other practical applications are under intense developments. Despite the enormous progress made in the past their fabrication and readout remain still a challenge.

References

1. S.P. Langley, Proc. Am. Acad. Arts Sci. **16** (1881) 342.
2. P. Curie, A. Laborde, Compt.Rend. Hebd. Seances Acad. Sci. Paris **136** (1903) 673–675.
3. C.D. Ellis, A. Wooster, Proc. R. Soc. **117** (1927) 109–123.
4. W. Orthmann, Z. Phys. **60** (1930) 10; and L. Meitner, W. Orthmann, Z. Phy. **60** (1930) 143.
5. F. Simon, Nature **135** (1935) 763.
6. D.H. Andrews, R.D. Fowler, M.C. Williams, Phys.Rev. **76** (1949) 154.
7. G.H. Wood, B.L. White, Appl. Phys. Lett. **15** (1969) 237; and G.H. Wood, B.L. White, Can. J. Phys. **51** (1973) 2032.
8. H. Bernas et al., Phys. Lett. A **24** (1967) 721.
9. A. Drukier, C. Vallette, Nucl. Instrum. Meth. **105** (1972) 285.
10. A. Drukier, L. Stodolsky, Phys. Rev. D **30** (1984) 2295.
11. N. Coron, G. Dambier, J. Leblanc, in: *Infrared detector techniques for Space Research*, V. Manno, J. Ring (eds.), Reidel Dordrecht (1972), pp. 121–131.
12. T.O. Niinikoski, F. Udo, CERN NP Report 74–6 (1974).
13. E. Fiorini, T.O. Niinikoski, Nucl. Instrum. Meth. **224** (1984) 83.
14. D. McCammon, S.H. Moseley, J.C. Mather, R. Mushotzky, J. Appl. Physics **56**(5) (1984) 1263.
15. N. Coron et al., Nature **314** (1985) 75–76.
16. K. Pretzl, N. Schmitz, L. Stodolsky (eds.), Low-Temperature Detectors for Neutrinos and Dark Matter LTD1, Schloss Ringberg, Germany, Springer-Verlag (1987).
17. L. Gonzalez-Mestres, D. Perret-Gallix (eds.), Low-Temperature Detectors for Neutrinos and Dark Matter LTD2, Gif-sur-Yvette, France, Ed. Frontieres (1988).
18. L. Brogiato, D.V. Camin, E. Fiorini (eds.), Low-Temperature Detectors for Neutrinos and Dark Matter LTD3, Gif-sur-Yvette, France, Ed. Frontieres (1989).
19. N.E. Booth, G.L. Salmon (eds.), Low Temperature Detectors for Neutrinos and Dark Matter LTD4, Gif-sur-Yvette, France, Ed. Frontieres (1992).
20. S.E. Labov, B.A. Young (eds.), Proc. 5th Int. Workshop Low Temperature Detectors LTD5, Berkeley, CA, J. Low Temp. Phys. **93**(3/4) (1993) 185–858.
21. H.R. Ott, A. Zehnder (eds.), Proc. 6th Int. Workshop Low Temperature Detectors LTD6, Beatenberg/Interlaken, Switzerland, Nucl. Instrum. Meth. A **370** (1996) 1–284.
22. S. Cooper (ed.), Proc. 7th Int. Workshop Low Temperature Detectors LTD7, Max Planck Institut of Physics Munich, Germany, ISBN 3-00-002266-X, (1997).
23. P. deKorte, T. Peacock (eds.), Proc. 8th Int. Workshop Low Temperature Detectors LTD8, Dalfsen, Netherlands, Nucl. Instrum. Meth. A **444** (2000).
24. F. Scott Porter, D. MacCammon, M. Galeazzi, C. Stahle (eds.). Proc. 9th Int. Workshop Low Temperature Detectors LTD9, American Institute of Physics AIP Conf. Proc. **605** (2002).
25. F. Gatti (ed.), Proc. 10th Int. Workshop Low Temperature Detectors LTD10, Genova, Italy, Nucl. Instrum. Meth. A **520** (2004).
26. M. Ohkubo (ed.), Proc. 11th Int. Workshop Low Temperature Detectors LTD11, Tokyo, Japan, Nucl. Instrum. Meth. A **559** (2006).
27. M. Chapellier, G. Chardin (eds.), Proc. 12th Int. Workshop Low Temperature Detectors LTD12, Paris, France, J. Low Temp. Phys. **151**(1/2, 3/4) (2008).
28. B. Cabrerra, A. Miller, B.Young (eds.), Proc. 13th Int.Workshop Low Temperature Detectors LTD13, Stanford/SLAC, USA, American Inst. of Physics (March 2010).
29. Ch. Enss, A. Fleischmann, L. Gastaldo (eds.), Proc. 14th Int.Workshop Low Temperature Detectors LTD14, Heidelberg, Germany, J.Low Temp.Phys.**167**(5/6) (2012) 561–1196.
30. E. Skirokoff(ed.), Proc. 15th Int.Workshop Low Temperature Detectors LTD15, Pasadena Cal., USA, J. Low Temp. Phys.**176**(3/6)(2014) 131–1108.
31. Ph. Camus, A. Juillard, A. Monfardini (eds.), Proc. 16th Int.Workshop Low Temperature Detectors LTD16, Grenoble, France, J. Low Temp. Phys. **184**(1/4)(2016) 1–978.

32. A. Barone (ed.), Proc. Superconductive Particle Detectors, Torino, Oct. 26–29, 1987, World Scientific.
33. A. Barone, Nucl. Phys. B (Proc. Suppl.) **44** (1995) 645.
34. D. Twerenbold, Rep. Prog. Phy. **59** (1996) 349.
35. H. Kraus, Supercond. Sci. Technol. **9** (1996) 827.
36. N. Booth, B. Cabrera, E. Fiorini, Annu. Rev. Nucl. Part. Sci. **46** (1996) 471.
37. K. Pretzl, *Cryogenic calorimeters in astro and particle physics*, Nucl. Instrum. Meth. A **454** (2000) 114.
38. Ch. Enss (ed.), *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Springer Berlin, Heidelberg, New York (2005).
39. D. McCammon, *Thermal Equilibrium Calorimeters-An Introduction*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 1.
40. J.C. Mather, Appl. Opt. **21** (1982) 1125.
41. S.H. Moseley, J.C. Mather, D. McCammon, J. Appl. Phys. **56**(5) (1984) 1257–1262.
42. D.A. Wollmann et al., Nucl. Instrum. Meth. A **444** (2000) 145.
43. T.C.P. Chui et al., Phys. Rev. Lett. **69**(21) (1992) 3005.
44. B.I. Shklovskii, A.L. Efros, *Electronic Properties of Doped Semiconductors*, Springer-Verlag (1984).
45. P. Colling et al., Nucl. Instrum. Meth. A **354** (1995) 408.
46. K.D. Irwin, Appl. Phys. Lett. **66** (1995) 1998.
47. K.D. Irwin, G.C. Hilton, *Transition Edge Sensors*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 63.
48. B. Young et al., IEEE Trans. Magnetics **25** (1989) 1347.
49. K.D. Irwin, B. Cabrera, B. Tigner, S. Sethuraman, in: Proc. 4th Int. Workshop Low Temperature Detectors for Neutrinos and Dark Matter LTD4, N.E. Booth, G.L. Salmon (eds.), Gif-sur-Yvette, France, Ed. Frontieres (1992), p. 290.
50. P. Ferger et al., Nucl. Instrum. Meth. A **370** (1996) 157.
51. P. Colling et al., Nucl. Instrum. Meth. A **354** (1995) 408.
52. U. Nagel et al., J. Appl. Phys. **76** (1994) 4262.
53. J. Hohne et al., X-Ray Spectrom. **28** (1999) 396.
54. J. Martinis, G. Hilton, K. Irwin, D. Wollmann, Nucl. Instrum. Meth. A **444** (2000) 23.
55. G. Brammertz et al., Appl. Phys. Lett. **80** (2002) 2955.
56. C. Hunt et al., Proc. SPIE **4855** (2003) 318.
57. B. Young et al., Nucl. Instrum. Meth. A **520** (2004) 307.
58. M. Buehler, E. Umlauf, Europhys. Lett. **5** (1988) 297.
59. E. Umlauf, M.Buehler, in: Proc. Int. Workshop Low Temperature Detectors for Neutrinos and Dark Matter LTD4, N.E. Booth, G.L. Salmon (eds.), Gif-sur-Yvette, France, Ed. Frontieres (1992), p. 229.
60. S.R. Bandler et al., J. Low Temp. Phys. **93** (1993) 709.
61. A. Fleischmann et al., Nucl. Instrum. Meth. A **520** (2004) 27.
62. A. Fleischmann, C. Enss, G.M. Seidel, *Metallic Magnetic Calorimeters*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 196.
63. A. Barone, G. Paterno, Physics and Applications of the Josephson Effect, New York, Wiley-Interscience (1984).
64. D. Twerenbold, A. Zehnder, J. Appl. Phys. **61** (1987) 1.
65. H. Kraus et al., Phys. Lett. B **231** (1989) 195.
66. C.A. Mears, S. Labov, A.T. Barfknecht, Appl. Phys. Lett. **63** (21) (1993) 2961.
67. P. Lerch, A. Zehnder, *Quantum Giaever Detectors: STJ's*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 217.
68. N.E. Booth, Appl. Phys. Lett. **50** (1987) 293.

69. K.E. Gray, Appl. Phys. Lett. **32** (1978) 392.
70. I. Giaever, K. Megerle, Phys. Rev. **122**(4) (1961) 1101.
71. S.B. Kaplan et al., Phys. Rev. B **14**(11) (1976) 4854.
72. P.A.J. de Korte et al., Proc. SPIE **1743** (1992) 24.
73. G. Angloher et al., J. Appl. Phys. **89**(2) (2001) 1425.
74. P. Verhoeve et al., Proc. SPIE **6276** (2007) 41.
75. N. Rando et al., Rev. Sci. Instrum. **71**(12) (2000) 4582.
76. R. Christiano et al., Appl. Phys. Lett. **74** (1999) 3389.
77. M.P. Lissitski et al., Nucl. Instrum. Meth. A **520** (2004) 240.
78. E. Figueroa-Feliciano, Nucl. Instrum. Meth. A **520** (2004) 496.
79. P.K. Day et al., Nature **425** (2003) 871.
80. B. Mazin et al., Publ.Astro.Soc. of the pacific **125** (2013) 1348–1361.
81. K. Borer et al., Astroparticle Phys. **22** (2004) 199.
82. K. Borer, M. Furlan, Nucl. Instrum. Meth. A **365** (1995) 491.
83. A. Kotlicki et al., in: Low-Temperature Detectors for Neutrinos and Dark Matter LTD1, K. Pretzl, N. Schmitz, L. Stodolsky (eds.) Springer-Verlag (1987), p. 37.
84. R. Leoni et al., J. Low Temp. Phys. **93**(3/4) (1993) 503.
85. B. van den Brandt et al., Nucl. Phys. B (Proc. Suppl.) **70** (1999) 101.
86. M. Abplanalp et al., Nucl. Instrum. Meth. A **360** (1995) 616.
87. S. Janos et al., Nucl. Instrum. Meth. A **547** (2005) 359.
88. G. Meagher et al., J. Low Temp. Phys. **93**(3/4) (1993) 461.
89. C. Berger et al., J. Low Temp. Phys. **93**(3/4) (1993) 509.
90. S. Calatroni et al., Nucl. Instrum. Meth. A **444** (2000) 285.
91. S. Casalboni et al., Nucl. Instrum. Meth. A **459** (2001) 469.
92. S. Calatroni et al., Nucl. Instrum. Meth. A **559** (2006) 510.
93. M. Abplanalp, Nucl. Instrum. Meth. A **370** (1996) 11.
94. K. Pretzl, Particle World **1**(6) (1990) 153.
95. K. Pretzl, J. Low Temp. Phys. **93** (1993) 439.
96. F. Zwicky, Helv. Phys. Acta **6** (1933) 110.
97. S. Perlmutter et al., Astrophys. J. **483** (1997) 565.
98. A.G. Riess et al., Astronom. J. **116** (1998) 1009.
99. The Planck collaboration (P.Ade et al.), Astronom. Astrophys. **594**A (13) (2016) and arXiv: 1502.01589 (astro-phys.Co) (2016)
100. R.D. Peccei, H.R. Quinn, Phys. Rev. Lett. **38** (1977) 1440.
101. P.W. Graham et al., An.Rev.Nucl. and Particle Searches **65** (2015) 485–514 and arXiv: 1602.00039 (hep - ex) (2016)
102. F.Kahlhofer, Int.J.Mod.Phys. A **32** (2017) 1730006 and arXiv: 1702.02430 (hep - ph)
103. K. Pretzl, Space Science Reviews **100** (2002) 209.
104. G. Jungman, M. Kamionkowski, K. Griest, Phys. Rep. **267** (1996) 195.
105. B. Sadoulet, in: Low Temperature Detectors for Neutrinos and Dark Matter LTD1, K. Pretzl, L. Stodolsky, N. Schmitz (eds.), Springer-Verlag (1987), p. 86.
106. L. Gonzalez-Mestres, D. Perret-Gallix, Nucl. Instrum. Meth. A **279** (1989) 382.
107. S. Cebrian et al., Phys. Lett. B **563** (2003) 48.
108. P. Meunier et al., Appl. Phys. Lett. **75**(9) (1999) 1335.
109. D.S. Akerib et al., Phys. Rev. D **72** (2005) 052009.
110. Z.Ahmed et al., Phys. Rev. Lett. **106** (2011) 131302 and R.Agnese et al., Phys. Rev. D **92** (7) (2015) 072003 and arXiv:1504.05871 (hep-ex).
111. R.Agnese et al., Phys. Rev. Lett. **111** (2013) 251301 and arXiv:1304.4279 (hep-ex).
112. N.Luke, J. Appl. Phys. **64** (1988) 6858.
113. G. Wang, J. Appl. Phys. **107** (2010) 094504.
114. C. Isaila et al.,Phys. Lett. B **716** (2012) 160.
115. R. Agnese et al.,Phys. Rev. Lett. **116** (7) (2016) 071301, R.Agnese et al., arXiv:1707.01632 (2017) submitted to Phys. Rev. D.
116. V. Sanglard et al., Phys. Rev. D **71** (2005) 122002.

117. E.Armengaud et al.,Phys.Lett. B **702** (2011) 329–335 and arXiv:1103.4070 (astro-ph.Co) (2011)
118. L.Hehn et al., Eur. Phys. J. C **76** (10) (2016) 548 and arXiv: 1607.03367 (astro-ph.Co) (2016), E.Armengaud et al., arXiv 1706.01070 (physics.ins-det) (2017)
119. T. Shutt et al., Nucl. Instrum. Meth. A **444** (2000) 34.
120. D.S. Akerib et al., Phys. Rev. D **68** (2003) 082002.
121. G. Angloher et al., Astroparticle Physics **31** (2009) 270–276 and arXiv:0809.1829 [astro-ph].
122. C. Isaila et al., Nucl. Instrum. Meth. **559** (2006) 399; and C. Isaila et al., J. Low Temp. Phys. **151**(1/2) (2008) 394.
123. J. Ninković et al., Nucl. Instrum. Meth. A **564** (2006) 567.
124. G. Angloher et al., Eur.Phys. J.C **76** (1) (2016) 25 and arXiv: 1509.01515.
125. R. Strauss et al., Eur.Phys. J.C bf77 (2017) 506 and arXiv: 1704.04320 (physics.ins-det) (2017).
126. R. Strauss et al., Phys.Rev. D bf96 (2) (2017) 022009 and arXiv: 1704.04317.
127. D. Akimov et al., Science **357** (2017) 1123
128. Mark Schumann, Physics Departement University Freiburg Germany, private communication.
129. A. Aguilar-Arevalo et al., Phys. Rev. D **94** (2016) 082006 and arXiv: 1607.07410.
130. E. Aprile et al., arXiv: 1705.06655.
131. D. Akerib et al., Phys. Rev. Lett. **118** (2) (2017) 021303 and arXiv: 1608.07648.
132. A. Tan et al., Phys. Rev. Lett. **117** (12) (2016) 12133 and arXiv: 1708.06917.
133. E. Aprile et al., Phys. Rev. D **94** (12) (2016) 122001 and arXiv: 1609.06154.
134. P. Agnes et al., Phys. Lett. B **743** (2015) 456–466 and arXiv:1410.0653.
135. R.Bernabei et al., Eur. Phys.J.Web Conf **13** (2017) 60500 and arXiv:1612.01387 R.Bernabei et al. Eur. Phys. J. C **73** (2013) 2648.
136. C. E. Aalseth et al., Phys. Rev. Lett. **106** (13) (2011) 131301 and arXiv:1401.3295.
137. F.Froborg et al.,arXiv:1601.05307.
138. E. Behnke et al., Astropart. Phys. **90** (2017) 85–92.
139. G. Agloher et al.,Phys.Dark Univ. **3** (2014) 41–74.
140. R.Agnese et al.,Phys. Rev. D **95** (8) (2017) 082002 and arXiv: 1610.0006.
141. J.Aalbers et al., JCAP **11** (2016) 017 and arXiv: 1606.07001.
142. D.S.Akerib et al., Astropart. Phys. **96** (2017) 1–10
143. Y. Fukuda et al., Phys. Rev. Lett. **81**(8) (1998) 1562.
144. M. Maltroni, T. Schwetz, M. Tortola, J.W.F. Valle, New J. Phys. **6** (2004) 122.
145. M. Goeppert Mayer, Phys. Rev. **48** (1935) 512.
146. P. Vogel, *Double Beta decay: Theory, Experiment and Implications*, in: *Current aspects of Neutrino Physics*, D.O. Caldwell (ed.), Springer-Verlag (2001), p. 177.
147. C. Arnaboldi et al., Phys. Rev. Lett. **95** (2005) 142501.
148. K.Alfonso et al., Phys. Rev. Lett. **115** (10) (2015) 102502 and arXiv: 1504.02454 (nucl.-ex).
149. C.Alduino et al., J.INST **11** (7) (2016) P07009 and arXiv: 1604.05465 (phys.ins-det) (2016).
150. C.Alduino et al., Eur. Phys. J. C **77** (8) (2017) 532 and arXiv: 1705.10816 (phys.ins-det) (2017).
151. G.Wang et al., arXiv: 1504.03599 (phys. ins.-det) (2015)
152. A.Gaudo et al., Phys. Rev. Lett. **110** (2013) 062502 and arXiv: 1605.02889 (2016)
153. J.B. Albert et al., Nature **510** (2014) 229
154. M. Agostini et al., Nature **544** (2017) 5
155. R.Arnold et al., Phys. Rev. D **92** (2015) 072011
156. S.Betts et al., arXiv: 1307.4738 (astro-ph.IM) (2013)
157. V.M. Lobashov, Nucl. Phys. A **719** (2003) 153, and references therein.
158. KATRIN experiment, https//www.katrin.kit.edu.
159. F. Fontanelli, F. Gatti, A. Swift, S. Vitale, Nucl. Instrum. Meth. A **370** (1996) 247.
160. F. Gatti et al., Nucl. Phys. B **91** (2001) 293.

161. M. Sisti et al., Nucl. Instrum. Meth. A **520** (2004) 125.
162. F. Gatti et al., Nature **397** (1999) 137.
163. F. Gatti, F. Fontanelli, M. Galeazzi, S. Vitale, Nucl. Instrum. Meth. A **444** (2000) 88.
164. C. Arnaboldi et al., Phys. Rev. Lett. **96** (2006) 042503.
165. A. Nucciotti, J. Low Temp. Phys. **151**(3/4) (2008) 597.
166. E. Ferri et al.,J. Low Temp. Phys. **176** (5/6) (2014) 885–890.
167. L. Gastaldo et al.,J. Low Temp. Phys. **176** (5/6) (2014) 876–884.
168. B.Alperit et al., The European Physical Journal C **75** (2015) 112.
169. M.P. Croce et al., J.Low Temp. Phys. **184** (3/4) (2016) 958–968 and arXiv:1510.03874.
170. A. de Rujula and M. Lusignoli, Phys. Lett. B **118** (1982) 429434.
171. O. Dragoun and D. Vnos, Open Physics Journal **3** (2016) 73–113.
172. D. McCammon et al., Astrophys. J **576** (2002) 188.
173. C.K. Stahle et al., Nucl. Instrum. Meth. A **520** (2004) 466.
174. D.D.E. Martin et al., Nucl. Instrum. Meth. A **520** (2004) 512.
175. N. Rando et al., J. Low Temp. Phys. **93**(3/4) (1993) 659.
176. B. Cabrera et al., Appl. Phys. Lett. **73** (1998) 735.
177. D.D.E. Martin et al., Proc. SPIE **6269** (2006) 62690O-1.
178. J. Burney et al., Nucl. Instrum. Meth. A **559** (2006) 506.
179. R.W. Romani et al., Astrophys. J. **563** (2001) 221.
180. J.A. Chervanek et al., Appl. Phys. Lett. **44** (1999) 4043.
181. P.A.J. de Korte et al., Rev. Sci. Instrum. **74**(8) (2003) 3807.
182. W.B. Doriese et al., Nucl. Instrum. Meth. A **559** (2006) 808.
183. R.A. Hijemering et al., Nucl. Instrum. Meth. A **559** (2006) 689.
184. B. Cabrera, R. Romani, *Optical/UV Astrophysics Applications of Cryogenic detectors*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 416.
185. P. Verhoeve, J. Low Temp. Phys. **151**(3/4) (2008) 675.
186. M.J.Strader et al. Astrophys. J. Letters **779** (2013) L12.
187. Asad M. Aboobaker et al., arXiv: 1703.03847 (astro-phys.IM) (2017).
188. J.M.Nagy et al., Astrophysics J. **844** (2) (2017) 151 and arXiv:1704.0025.
189. P.Ade et al., The Astrophysical Journal **794** (171) (2014) 21.
190. P.A.R. Ade et al., Phys. Rev. Lett. **112** (24) (2014) 241101.
191. PLANCK Collaboration, Astronomy & Astrophysics **586** (2014) A133.
192. P.A.R. Ade et al. Phys. Rev. Lett. **114** (10) (2015) 101301 and arXiv:1502.00612.
193. L. Lesyna et al., J. Low Temp. Phys. **93** (1993) 779.
194. R. Ladbury, Physics Today (July 1998) 19.
195. D.E. Newbury et al., *Electron Probe Microanalysis with Cryogenic Detectors*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 267.
196. V. Shvarts et al., Nucl. Instrum. Meth. A **520** (2004) 631.
197. D. Twerenbold, Nucl. Instrum. Meth. A **370** (1996) 253.
198. M. Frank et al., Mass Spectrometry Reviews **18** (1999) 155.
199. D. Twerenbold et al., Proteomics **1** (2001) 66.
200. M. Frank et al., Rapid Commun. Mass Spectrom. **10**(15) (1996) 1946.
201. G.C. Hilton et al., Nature **391** (1998) 672.
202. J.N. Ullom, J. Low Temp. Phys. **151**(3/4) (2008) 746.
203. S. Rutzinger et al., Nucl. Instrum. Meth. A **520** (2004) 625.
204. P. Christ et al., Eur. Mass Spectrom. **10** (2004) 469.
205. P. Egelhof, S. Kraft-Bermuth, *Heavy Ion Physics*, in: *Cryogenic particle detection*, Topics in Applied Physics Vol. **99**, Ch. Enss (ed.), Springer Berlin, Heidelberg, New York (2005), p. 469.

Chapter 20
Detectors in Medicine and Biology

P. Lecoq

20.1 Dosimetry and Medical Imaging

The invention by Crookes at the end of the nineteenth century of a device called spinthariscope, which made use of the scintillating properties of Lead Sulfide allowed Rutherford to count α particles in an experiment, opening the way towards modern dosimetry. When at the same time Wilhelm C. Roentgen, also using a similar device, was able to record the first X-ray picture of his wife's hand 2 weeks only after the X-ray discovery, he initiated the first and fastest technology transfer between particle physics and medical imaging and the beginning of a long and common history.

Since that time, physics, and particularly particle physics has contributed to a significant amount to the development of instrumentation for research, diagnosis and therapy in the biomedical area. This has been a direct consequence, one century ago, of the recognition of the role of ionizing radiation for medical imaging as well as for therapy.

20.1.1 Radiotherapy and Dosimetry

The curative role of ionizing radiation for the treatment of skin cancers has been exploited in the beginning of the twentieth century through the pioneering work of some physicists and medical doctors in France and in Sweden. This activity has very much progressed with the spectacular developments in the field of accelerators,

P. Lecoq (✉)
CERN, Geneva, Switzerland
e-mail: Paul.Lecoq@cern.ch

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_20

beam control and radioisotope production. Today radiotherapy is an essential modality in the overall treatment of cancer for about 40% of all patients treated. Conventional radiotherapy (RT) with X-rays and electrons is used to treat around 20,000 patients per 10 million inhabitants each year.

The main aim of radiation therapy is to deliver a maximally effective dose of radiation to a designated tumour site while sparing the surrounding healthy tissues as much as possible. The most common approach, also called teletherapy, consists in bombarding the tumour tissue with ionizing radiation from the outside of the patient's body. Depending on the depth of the tumour, soft or hard X-rays or more penetrating γ-rays produced by a ^{60}Co source or by a linac electron accelerator are used. However, conventional X-ray or γ-ray radiation therapy is characterized by almost exponential attenuation and absorption, and consequently delivers the maximum energy near the beam entrance. It also continues to deposit significant energy at distances beyond the cancer target. To compensate for the disadvantageous depth-dose characteristics of X-rays and γ-rays and to better conform the radiation dose distribution to the shape of the tumour, the radiation oncologists use complex Conformal and Intensity Modulated techniques (IMRT) [1]. The patient is irradiated from different angles, the intensity of the source and the aperture of the collimators being optimized by a computer controlled irradiation plan in order to shape the tumour radiation field as precisely as possible.

Another way to spare as much as possible healthy tissues is to use short range ionizing radiation such as β or α particles produced by the decay of unstable isotopes directly injected into the tumour. This method, called brachytherapy, has been originally developed for the thyroid cancer with the injection of ^{153}I directly into the nodules of the thyroid gland. It is also used in other small organs such as prostate or saliva gland cancer.

Following these trends a new generation of minimally invasive surgical tools appears in hospitals, allowing to precisely access deep tumours from the exterior of the body (gamma-knife), or by using brachytherapy techniques, i.e. by injecting radioisotopes directly in the tumour (beta-knife, alpha-knife and perhaps soon Auger-knife). More recently the radio-immunotherapy method has been successfully developed: instead of being directly inserted in the tumour, the radioactive isotopes can be attached by bioengineering techniques to a selective vector, which will bind to specific antibody receptors on the membranes of the cells to be destroyed. Typical examples are the use of the α emitter ^{213}Bi for the treatment of leukaemia and of the β-emitter ^{90}Y for the treatment of glioblastoma.

In 1946 Robert Wilson, physicist and founder of Fermilab, proposed the use of hadron beams for cancer treatment. This idea was first applied at the Lawrence Berkeley Laboratory (LBL) where 30 patients were treated with protons between 1954–1957. Hadrontherapy is now a field in rapid progress with a number of ambitious projects in Europe, Japan and USA [2], exploiting the attractive property of protons and even more of light ions like carbon to release the major part of their kinetic energy in the so-called Bragg peak at the end of their range in matter (Fig. 20.1).

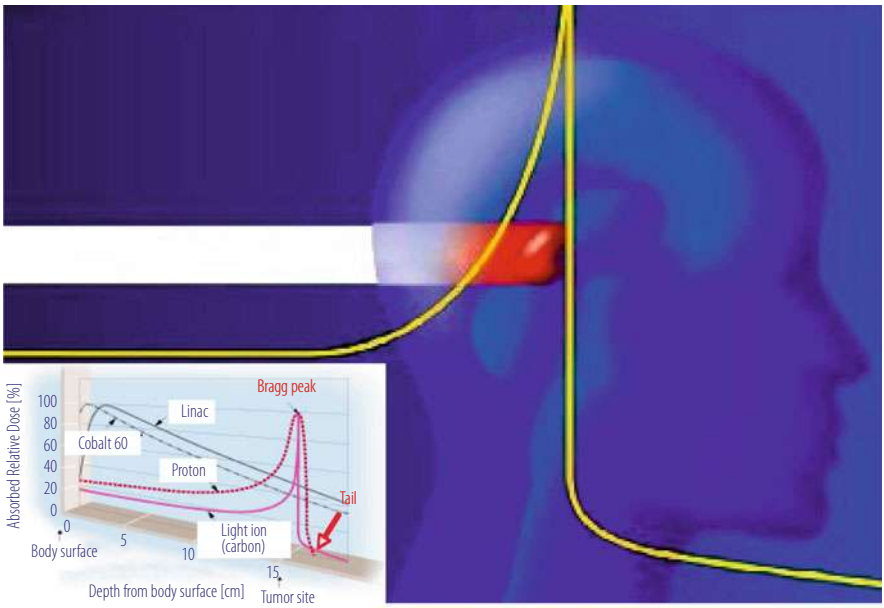

Fig. 20.1 Bragg peak for an ion beam in the brain of a patient. The insert shows the energy absorbed by tissues as a function of depth for different radiation sources (Courtesy U. Amaldi)

It is particularly important to treat the disease with the minimum harm for surrounding healthy tissues. In the last centimeters of their range the Linear Energy Transfer (LET) of protons or even more of carbon ions is much larger than the one of X-rays (low-LET radiations). The resulting DNA damages include more complex double strand breaks and lethal chromosomal aberrations, which cannot be repaired by the normal cellular mechanisms. The effects produced at the end of the range are therefore qualitatively different from those produced by X- or γ-rays and open the way to a strategy to overcome radio-resistance, often due to hypoxia of the tumour cells. For these reasons carbon ions with their higher relative biological effectiveness (RBE) at the end of their range, of around a factor of three higher than X-rays, can treat tumours that are normally resistant to X-rays and possibly protons. This treatment is particularly applicable to deep tumours in the brain or in the neck as well as ocular melanoma.

Whatever the detailed modality of the treatment planning, precise dosimetry is mandatory to develop an optimal arrangement of radiation portals to spare normal and radiosensitive tissues while applying a prescribed dose to the targeted disease volume. This involves the use of computerized treatment plan optimization tools achieving a better dose conformity and minimizing the total energy deposition to the normal tissues (Fig. 20.2). It requires a precise determination and simulation of the attenuation coefficients in the different tissues along the beam. These data are obtained from high performance anatomic imaging modalities such as X-ray computed tomography (CT) and magnetic resonance (MRI).

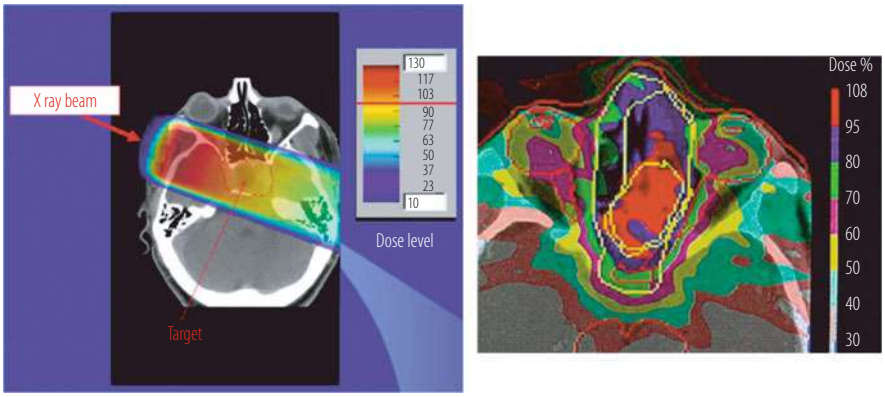

Fig. 20.2 Dosimetry for a brain tumour in the case of one (left) or nine crossed (right) X-ray beams. The treatment plan is based on a tumour irradiation of at least 90 Gy (Courtesy U. Amaldi)

For the particular case of hadrontherapy on-line dosimetry in the tissues is in principle possible. It relies on the production of positron emitter isotopes produced by beam spallation (^{10}C and ^{11}C for 12C beam) or target fragmentation during the irradiation treatment. The two 511 keV γ produced by the positron annihilation can be detected by an in-line positron emission tomography (PET) to precisely and quantitatively map the absorbed dose in the tumour and surrounding tissues. Although challenging because of the timing and high sensitivity requirements this approach is very promising and a number of groups are working on it worldwide [3].

20.1.2 Status of Medical Imaging

The field of medical imaging is in rapid evolution and is based on five different modalities: X-ray radiology (standard, digital and CT), isotopic imaging (positron emission tomography, PET, and single photon emission computed tomography SPECT), ultrasound (absorption, Doppler), magnetic resonance (MRI, spectroscopy, functional), and electrophysiology with electro- and magneto-encephalography (EEG and MEG). More recently, direct optical techniques like bioluminescence and infrared transmission are also emerging as powerful imaging tools for non-too-deep organs.

For a long time imaging has been anatomical and restricted to the visualization of the structure and morphology of tissues allowing the determination of morphometric parameters. With the advent of nuclear imaging modalities (PET and SPECT) and of the blood oxygen level dependent (BOLD) technique in magnetic resonance imaging (MRI) functional imaging became possible and medical doctors can now see organs at work. Functional parameters are now accessible in vivo and in real

time, such as vascular permeability, haemodynamics, tissue oxygenation or hypoxia, central nervous system activity, metabolites activity, just to cite a few.

In the current clinical practice medical imaging is aiming at the in-vivo anatomic and functional visualization of organs in a non- or minimally invasive way. Isotopic imaging, in particular PET, currently enjoys a spectacular development. Isotopic imaging consists in injecting into a patient a molecule involved in a specific metabolic function so that this molecule will preferentially be fixed on the organs or tumours where the function is at work. The molecule has been labeled beforehand with a radioisotope emitting gamma photons (Single Photon Emission Computed Tomography or SPECT) or with a positron emitting isotope (Positron Emission Tomography or PET). In the latter case, the positron annihilates very quickly on contact with ordinary matter, emitting two gamma photons located on the same axis called the line of response (LOR) but in opposite directions with a precise energy of 511 keV each. Analyzing enough of these gamma photons, either single for SPECT or in pairs for PET, makes it possible to reconstruct the image of the area (organ, tumour) where the tracer focused.

Since the beginning to the twenty-first century a new generation of machines became available, which combine anatomic and functional features: the PET/CT. This dual modality system allows the superposition of the high sensitivity functional image from the PET on the precise anatomic picture of the CT scanner. PET/CT has now become a standard in the majority of hospitals, particularly for oncology. This trends for multimodal imaging systems is increasing both for clinical and for research applications (Fig. 20.3).

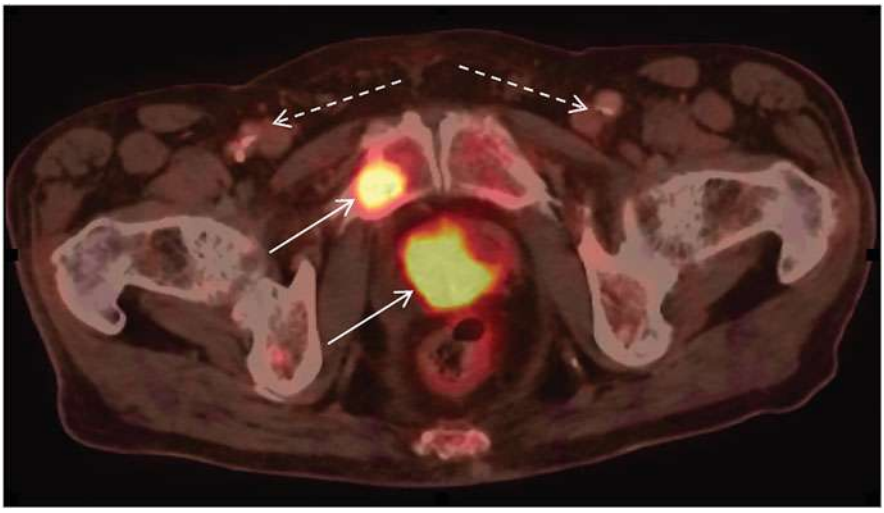

Fig. 20.3 Abdominal slice of a 78 year-old male, with biopsy-proven prostate adenocarcinoma and penile adenocarcinoma. Focal uptake in the prostate bed and in the penile shaft (full arrows). Multiple foci in the pelvis compatible with skeletal metastases (dashed arrows) (Courtesy D. Townsend)

Table 20.1 Comparison of the performances of four imaging modalities

Imaging modality	Type of imaging	Examination time	Spatial resolution
PET	Functional and molecular (picomolar sensitivity)	10–20′ (whole body)	3–5mm
SPECT	Functional and molecular	10–20′ (per 40 cm filed)	6–8 mm
MRI	Anatomical Functional (millimolar sensitivity)	30–60′	0.5 mm
CT	Anatomical	<1′ (whole body)	0.5 mm

The most frequently used positron emitters are ^{18}F, ^{11}C, ^{15}O, ^{13}N, the three last ones being isotopes of the nuclei of organic molecules.

As compared to other non-invasive imaging modalities isotopic imaging has a functional sensitivity at the picomolar level, which is several orders of magnitude better than magnetic resonance. It opens incredible perspectives for cell and molecular imaging, in particular for visualizing and quantifying genomic expression or tissue repair efficiency of stem cells. However, the detection efficiency compared to the dose injected to the patient, also called "sensitivity", is strongly limited by technical constraints, and the spatial resolution is still one order of magnitude worse than for CT or MRI (Table 20.1).

20.1.3 Towards In-Vivo Molecular Imaging

The challenge of future healthcare will be to capture enough information from each individual person to prevent disease at its earliest stage, to delineate disease parameters, such as aggressiveness or metastatic potential, to optimize the delivery of therapy based on the patient's current biologic system and to quickly evaluate the treatment therapeutic effectiveness.

New therapeutic strategies are entering the world of major diseases. They aim to acquire as fast as possible all the information on the pathological status of the patient in order to start adapted therapeutics and therefore to minimize the handicap. This applies to neurological and psychiatric diseases but also to the treatment of inflammatory diseases, such as rheumatologic inflammation, and of cancer. Moreover, the non-invasive determination of the molecular signature of cancers in the early stage of their development, or even before the tumour growth, will help to target the therapeutic strategy and to reduce considerably the number of unnecessary biopsies.

This trend is supported by the new paradigm of "personalized medicine" (also called precision medicine), which aims at delivering "the right treatment, to the right patient, at the right time". Personalised medicine refers to a medical model using

characterisation of individuals' phenotypes and genotypes (e.g. molecular profiling, medical imaging, lifestyle data) for tailoring the right therapeutic strategy for the right person at the right time, and/or to determine the predisposition to disease and/or to deliver timely and targeted prevention. In this new healthcare context, a radical shift is currently taking place in the way diseases are managed: from the present one-fits-all approach to one that delivers medical care tailored to the needs of individual patients. This includes the detection of disease predisposition, early diagnosis, prognosis assessment, measurement of drug efficacy and disease monitoring. To achieve this ambitious goal, there is an increased demand for simultaneous in-vivo quantitative and dynamic characterization of several biological processes at the molecular and genetic level. A new generation of whole body and organ-specific imaging devices is needed combining the excellent sensitivity and specificity of PET or SPECT with a high-spatial resolution imaging modality (CT, MR optical or US) providing additional functional, metabolic or molecular information.

For many years, physicians relied on the use of anatomical imaging to non-invasively detect tumours and follow up their growth. Functional imaging such as bone or thyroid scintigraphy and more recently PET using 18FDG for example, has provided more information for tumour staging. The next revolution being prepared will have to do with molecular imaging. The goal is in-vivo visual representation, characterization and quantification of biological processes at the cellular and sub-cellular level within living organisms. This is the challenge of modern biology: detect early transformations in a cell, which may lead to pathology (precancerous activity, modifications of neuronal activity as warning signs of Alzheimer or Parkinson disease). Besides early detection, assessment of prognosis and potential response to therapy will allow a better treatment selection through a precise delineation of molecular pathways from genes to disease. All aspects of gene expression will be addressed (genomics, proteomics, transcriptomics, enzymatic activity), but also the molecular signal transduction through cell membranes (a key to determine the efficacy of drugs) as well as the identification and quantification of specific cell receptors over-expressed in some pathological situations, such as dopamine receptors for epilepsy.

With the development of new imaging probes and "smart probes", imaging provides cellular protein and signal-pathway identification. There is an increasing amount of molecular probes dedicated to imaging but also to tumour therapy. The molecular phenotype of cells composing the tumour can lead to tailored therapies. This tumour phenotype can be determined ex-vivo on tissue samples. Molecular imaging should allow performance of an in-vivo tumour phenotyping by an appropriate use of specific imaging probes. This molecular profiling could already be envisioned in the very near future for some specific tumours overexpressing peptide hormone receptors such as breast and prostate cancers, and should become widely developed.

Therefore, it represents a major breakthrough to provide the medical community with an integrated "one-stop-shop" molecular profiling imaging device, which could detect tracers dedicated to Single Photon Emission Computed Tomography

(SPECT) or PET, as well as Magnetic Resonance Imaging (MRI), or X-ray Computed Tomography (CT) contrast agents.

Furthermore, since functional imaging allows the assessment of biochemical pathways, it will also provide accurate tools for experimental research. As an example, a large effort worldwide has recently allowed the precise mapping of the different genes in the DNA sequence but the mechanisms, by which these genes produce proteins, interact with each other, regulate their expression, are far from understood. In other terms we can say that the genomic alphabet has been decoded but its dynamic expression, its grammar, remains to be studied and understood. In-vivo molecular imaging of gene expression is now within reach through the development of ever more elaborated molecular probes as well as of sophisticated techniques which significantly improve the performances of modern imaging devices.

Drug development also takes advantage of technical progress in imaging technologies, like quantitative positron emission tomography in small animals, to determine drug pharmacokinetics and whole body targeting to tissues of interest. Moreover, the combination of functional imaging with a high resolution anatomical method such as MRI and/or X-ray CT will considerably enhance the possibility of determining the long term efficiency of a drug on basic pathological processes such as inflammation, blood flow, etc. In particular, the expected progress in sensitivity, timing and spatial resolution, coupled with a true multi-modality registration, will allow to explore the activity of a drug candidate or other essential pathophysiological processes of disease models in animals, like for instance cancer or adverse inflammatory effects.

This approach will require targeting cellular activity with specific contrast agents, but also a large effort on imaging instrumentation. Developments are needed for faster exams, correction of physiological movements during acquisition time (breathing, cardiac beating, digestive bolus), access to dynamic processes, quantification, true multimodality, dose reduction to the patient. This will require significant improvements in spatial and timing resolution, sensitivity and signal-to-noise ratio, all parameters very familiar to particle physicists. From the technologies already available, developed for instance for the LHC detectors or under development for the future linear collider, fast crystals, highly integrated fast and low noise electronics and ultrafast Geiger mode SiPMs open the way to time of flight (TOF) PET. These technologies are progressively being implemented in commercial PET's, resulting in an improvement of image signal/noise ratio with a corresponding sensitivity increase. Sensitivity to picomolar concentrations are within reach for whole body commercial PET scanners, which correspond to the molecular activity of a few hundreds of cells only.

20.2 X-Ray Radiography and Computed Tomography (CT)

20.2.1 Different X-Ray Imaging Modalities

X-ray radiology is the most popular imaging technique, which comprises X-ray Radiography, Computed Tomography (CT), also called Tomo-Densitometry (TDM), and Dual Energy X-ray Absorptiometry (DXA). For planar radiography the general trend is to progressively replace the film by digital devices, as already used for CT. The patient is exposed to an X-ray source, with its energy being adjusted as a function of the density of the tissues to be visualized. Present systems work in signal integration mode, although there is a trend towards photon counting devices, as will be explained in Sect. 20.2.4. In standard radiography, the projected image is recorded either on a photographic plate or on a digital device using a scintillation material coupled to a photosensitive array of Silicon diodes. Computed tomography or Tomo-Densitometry is based on the detection of X-ray attenuation profiles from different irradiation directions. Both the X-ray source and the detector (usually an array of scintillators coupled to a solid-state photodetector is rotating around the patient as the bed is moving through the scanner. This technique allows a 3D reconstruction of attenuation density within the human body. These density profiles can then be viewed from different directions and analyzed in a succession of slices allowing a full 3D reconstruction of the anatomical image (Fig. 20.4).

20.2.2 Detection System

The X-rays to be detected must be first converted into visible light in a scintillator or into electron-hole pairs in a semiconductor device, which are directly recorded. The X-rays are absorbed in a phosphor screen, in which they excite different luminescent centres depending on the nature of the phosphor. The visible light produced by these luminescent centres is recorded on an emulsion deposited on a film or a photographic plate or on a photodiode array in direct contact with the scintillating screen or through an optical relay lens system. For about one century, film has been the unique tool for X-ray radiography. There is a trade-off between the thickness of the phosphor screen, which has to be thick enough to efficiently absorb X-rays, but not too much in order to minimize the light spread and image blurring caused by the distance between the light emission point on the screen and the recording emulsion. To take advantage of the exponential absorption of X-rays in the screen causing a larger number of X-ray absorbed at the entrance of the phosphor screen, the film is generally placed in front of the screen in a so-called back screen configuration. Soft tissues, characterized by low X-ray absorption, are seen as bright areas on the phosphor screen because of the large number of X-rays reaching the screen. The visible light photons are absorbed in the emulsion where they convert (after development of the latent image) the silver halide grains into metallic silver. As a

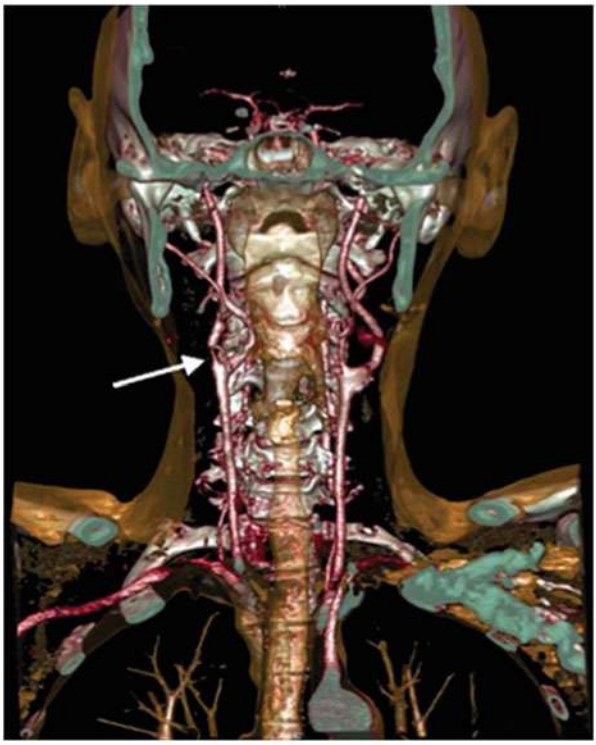

Fig. 20.4 64-Slice CT of the carotid arteries and circle of Willis of a patient. The arrow indicates a severe stenosis (Courtesy of Siemens Medical Solutions)

result, soft tissue produce black images whereas denser parts of the body like bones appear clear.

Digital radiography has progressively been replacing film-based radiography. Indeed, it offers a number of advantages such as better linearity, higher dynamic range, and most importantly, the possibility of distributed archiving systems. Besides direct conversion detectors like amorphous Silicon, CdTe or CdZnTe, which will be described in Sect. 20.2.4, scintillation materials are still the detectors of choice for modern X-ray detectors. For thin scintillation screens (0.1–0.2 mm thickness), which are well adapted to the lowest X-ray energies (for instance about 20 keV for X-ray mammography), ceramic phosphors are commonly used because they can be produced in any shape at a reasonable cost. On the other hand, for dental X-ray diagnostics (about 60 keV) and full body X-ray computed tomography (about 150 keV) the required stopping power would require much thicker screens and monocrystalline inorganic scintillators have been generally preferred up to now because of their much higher light transparency than ceramics. However, recent progress in producing more transparent ceramics based on nanopowders with low dispersion grain diameter may change this situation.

20.2.2.1 Scintillators for X-Ray Conversion

Detector elements of old CT scanners were prevalently implemented as ionization chambers filled with xenon at high pressure. Such detectors usually absorb 30–40% of the impinging photons and generate about 5500–6000 electrons per photon of 100 keV. Modern digital radiography devices and CT scanners use scintillator material arrays optically coupled to matching silicon p-i-n photodiode matrices. The scintillating material must be sufficiently thick to absorb close to 100% of the impinging photons, thus minimizing the patient dose required for a given image quality. Latest generation X-ray CT scanners are recording about 1000 projections (subject slices) per second. This imposes severe constrains on both the decay time and afterglow of the scintillating material. Afterglow is known to produce ghost images through a "memory effect" which deteriorates the quality of the images. The requirements for the scintillator material to be used in X-ray CT are:

- High absorption for X-rays in the energy range up to 140 keV. Absorption close to 100% for ~2 mm thick material layer is required to achieve an acceptable X-ray CT image to noise ratio. Indeed, the image quality is limited in low contrast regions by statistical fluctuations in the numbers of detected X-rays. A high detection efficiency allows to keep the patient dose exposure within reasonable limits for a given image quality. With last generation CT scanners, a whole body CT scan can now be achieved with a dose of less than 5 mSv, close to the level of 1 year natural radioactivity exposure.
- High light output, typically of the order or greater than 20,000 photons/MeV in order to reduce the image noise relative to (low) signal levels.
- Radioluminescence spectrum in the visible or near IR range to match the spectral sensitivity of the silicon photodetectors.
- Decay time in the range of 1–10 μs, in order to achieve sampling rates of the CT scanners in the $\geq$10 kHz range.
- Very low afterglow. Afterglow is generally caused by material imperfections (impurities, defects), causing delayed detrapping and carrier recombination with decay times in the range 100 ms to 10 s. An afterglow level of less than 0.1% is generally required 3 ms after the end of a continuous X-ray excitation. Afterglow causes blurring in the CT images.
- Good radiation hardness. The integrated exposure of the scintillators can reach several tens of kGy over the lifetime of a CT scanner. Changes in the light yield cause detector gain instability, resulting in image artifacts. Long-term changes of ~10% are acceptable, while only less than 0.1% short term changes during the daily operation (1000 R) can be tolerated without image quality degradation.
- Small temperature dependence of the light yield. The X-ray generation system usually dissipates a high amount of energy and the temperature of the detectors can change rapidly. A light output temperature coefficient within $\pm$0.1%/°C is desirable, which is a rather stringent requirement. Cadmium Tungstate (the most frequently used crystal in modern commercial CT scanners) has an acceptable temperature coefficient of $-$0.3%/°C [4].

- Good mechanical properties allowing micromachining of 2D scintillator arrays with pixel dimensions less than 1 mm.
- Affordable cost.

Table 20.2 summarizes the main characteristics of the scintillators used in medical CT imaging. During the last decade, there was a clear trend towards synthesized ceramic scintillators [5].

The only crystalline material still in use in medical and security systems CT scanners is cadmium tungstate, $CdWO_4$, also called CWO. Its main advantage over CsI(Tl) is a very low afterglow level of 0.05% 3 ms after the end of the X-ray exposure and a reasonable temperature coefficient of 0.3%/°C. In spite of their wide use CWO crystals are however not optimal for CT applications due to their brittleness and the toxicity of cadmium. Moreover, it is difficult to manufacture crystals with adequate uniformity. This has been an argument for the search of a new generation of CT scintillators. This search was initiated by General Electric and Siemens in the mid of the 1980s when they introduced the first polycrystalline ceramic scintillators. The host materials are yttrium and gadolinum oxides: Y_2O_3 and Gd_2O_3, which, after doping with Pr and Tb, demonstrate reasonable scintillation properties. However, their transmission is rather low, ceramics being more translucent than transparent. The additional Eu^{3+} activator efficiently traps electrons to form a transient Eu^{2+} state, allowing holes to form Pr^{4+} and Tb^{4+} and, therefore, competes with the intrinsic traps responsible for afterglow. This energy trapped on the Pr and Tb sites decays non-radiatively in presence of the Eu ions reducing therefore the level of afterglow [6].

Gadolinium oxide ceramic is now replaced by yttrium gadolinum oxide YGO [7], and gadolinum silicate GOS based ceramic materials [8]. When coupled to a silicon p-i-n photodiode they generate about 20 electrons per 1 keV of absorbed X-ray energy. However the long decay time of YGO (~1 ms) is a major concern and requires a complex algorithm of data deconvolution to suppress the effects of afterglow at the price of increased projection noise. Other ceramic materials proposed for CT applications are gadolinium gallium garnet, and lanthanum hafnate [9]. While ceramic materials are generally preferred to crystals because of their good performance and easy production in a variety of shapes, their low transparency requires the use of thin scintillators elements, with lower than optimal X-ray efficiency.

A large R&D effort is under way by several companies to produce flat panels for digital radiography. The standard scintillating crystal or ceramic pixels are replaced by detector arrays made of CsI(Tl) needles or small crystals (e.g. calcium tungstate CWO or YAP) directly coupled to photodiode arrays or segmented photomultipliers (see next section).

Table 20.2 Properties of scintillators used in X-ray CT imaging

Scintillator	Density (g/cm^3)	Thickness to stop 99% of 140 keV X-rays (mm)	Light yield (ph/MeV)/Temperature coefficient (%/°C)	Peak of emission band (nm)	Primary decay time (μs)	Afterglow (% at 3 ms)
CsI(Tl)	4.52	6.1	54,000/0.02	550	1	0.5
$CdWO_4$ (CWO)	7.9	2.6	28,000/−0.3	495	2, 15	0.05
Gd_2O_3:Eu^{+3}	7.55	2.6	–/–	610	–	–
$(Y,Gd)_2O_3$:Eu, Pr, Tb (YGO)	5.9	6.1	42,000/0.04	610	1000	5
Gd_2O_2S:Pr,Ce,F (GOS)	7.34	2.9	50,000/−0.6	520	2.4	<0.1
Gd_2O_2S:Tb(Ce) (GOS)	7.34	2.9	50,000/−0.6	550	600	0.6
La_2HfO_7:Ti	7.9	2.8	13,000/–	475	10	–
$Gd_3Ga_5O_{12}$:Cr,Ce	7.09	4.5	39,000/–	730	150	<0.1

20.2.2.2 Photodetectors

The visible or near infrared light produced by the X-ray absorption in the scintillating screen is converted into an electronic signal by a solid state photodetector usually in the form of an array of silicon p-i-n photodiodes. For CT applications the photodiode must satisfy the following requirements:

- High quantum and geometric efficiency to improve the signal statistics.
- High shunt resistance. This reduces the offset drift of the detector system due to the variations of the photodiode leakage current caused by temperature changes. Typically, these changes create image artifacts at high attenuation levels.
- Low capacitance to reduce the electronic noise.
- Ability to connect a large number of the photodiode pixels to the data acquisition system. This becomes increasingly difficult for 32 or 64 slice CT scanners.

The majority of the 16 slice scanners use conventional front-illuminated (FIP) silicon p-i-n photodiodes technology (Fig. 20.5a). However, it requires electrical strips on the front surface and electrical wirebonds from the edges of the silicon chip to the substrate. When the number of slices approaches 64, the increasing number of conductive strips the active area of the channels becomes unacceptably small, and the high density of the wirebonds cannot be handled by conventional

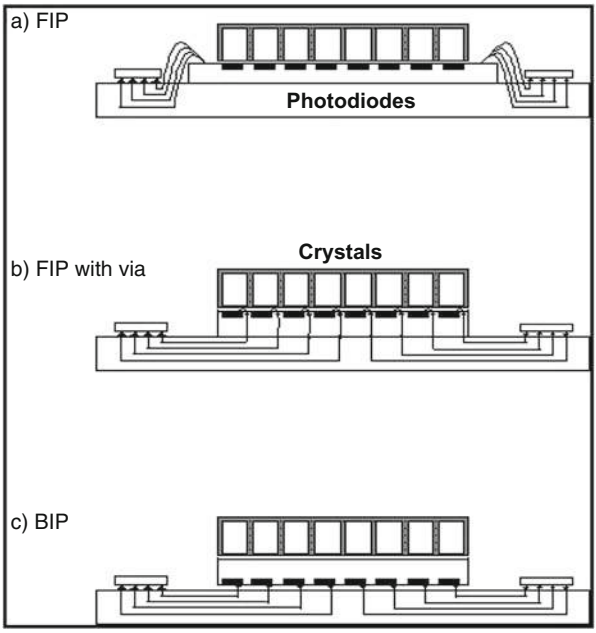

Fig. 20.5 Three types of photodiode arrays used in multislice CT. Front-illuminated FIP (**a**), Front-illuminated with via (**b**), and Back-illuminated BIP (**c**) (Courtesy R. Deych)

wirebond technology. In previous generation scanners, this limitation was addressed by combining the FIP technology with "vias" (electrically conducting feedthroughs) in the photodiode substrate [10]. The anodes of the photodiode elements were still wirebonded to via conductors, but the density of the wirebond was greatly reduced, because they were distributed over the area of the chip. The vias provided back-contacts for flip-chip connection to the detector board (Fig. 20.5b). Another approach is to use back-illuminated photodiodes (BIP) [11]. This solution solves the connectivity problem and the filling factor is almost 100% (Fig. 20.5c). It requires however very high resistivity silicon with large carrier lifetimes causing significant channel to channel cross-talk when standard silicon technology is used. In order to solve this problem BIPs are manufactured on 30 μm thick silicon wafer. Thinned BIP achieves almost 100% internal quantum efficiency in the spectral range 400–800 nm, and less than 1% cross-talk for 1×1 mm^2 pixels [12].

20.2.3 Scanner Geometry and Operating Conditions

20.2.3.1 Principle of Computed Tomography

The Computed Tomography principle introduced in the late 1960s by Allan MacLead Cormack and Sir Godfrey Newbold Hounsfield (Nobel Price laureates in Physiology and Medicine in 1979) marked a revolution in medical image reconstruction techniques. It is based on the relationship between the projections of a given parameter (X-ray attenuation for CT or radiotracer concentration for PET) integrated along line of responses (LOR's) at different angles through the patient and the Fourier transform of this parameter value distribution in the patient's body.

For the case of parallel beam illumination a projection at angle Φ is defined by the integration along all the parallel LORs of the parameter of interest as shown in Fig. 20.6 for a two-dimensional object f(x,y). The profile of all the LOR integrated values as a function of s, the radial distance from the centre of the projection, defines the projection at this angle Φ. The collection of all projections for $0 \leq \Phi < 2\pi$ forms a two-dimensional function of s and Φ. This function is called a sinogram because a fixed point in the f(x.y) object will trace a sinusoidal path in the projection space as shown in Fig. 20.6. A complex object will be represented in the projection space by the superposition of all the sinusoids associated to each individual point of the object. The line-integral transform of

$$f(x, y) \rightarrow p(s, \Phi)$$

is called the X-ray transform, also called the Radon transform for the two-dimensional case [13]. In this case, the projections are formed through a single transverse slice of the patient. By repeating this procedure through multiple axial slices, each displaced by a small increment in z, one can form a three-

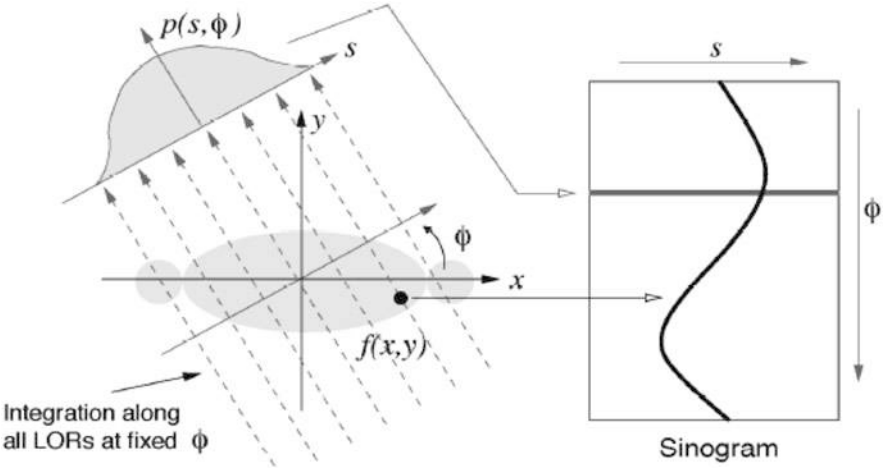

Fig. 20.6 Definition of the X-ray transform and the sinogram

dimensional image of a volumetric object f(x,y,z). It must be noticed that direct three-dimensional acquisition can be made by integrating LOR's not only in the transverse but also in oblique planes. Although more demanding in terms of computing power this fully three-dimensional approach is increasingly used in nuclear imaging (PET and SPECT), because it allows a significant gain in sensitivity.

The image reconstruction is based on the central-section theorem. This fundamental relationship in analytical image reconstruction states that the Fourier transform of a one-dimensional projection at angle Φ is equivalent to a section at the same angle through the centre of the two-dimensional Fourier transform of the object [14]. This is depicted in Fig. 20.7.

The image reconstruction process consists then in back-projecting and superposing all the data taken at all projection angles. However, to avoid oversampling in the centre of the Fourier transform (each projection will contribute to the central point, but increasingly less with increasing radial distance in the Fourier plane), the data are weighted, or filtered to correct for this oversampling. Basically, the Fourier transform of the back-projected image must be weighted by a cone filter

$$v = \sqrt{v_x^2 + v_y^2}$$

to decrease the values in the centre and increase them at the edges of the Fourier space.

However, in the specific case of X-ray CT the X-ray source is quasi-pointlike and the LOR's are not parallel. There is an important difference in the way parallel beam and divergent beam projections are back-projected. In a single view of divergent

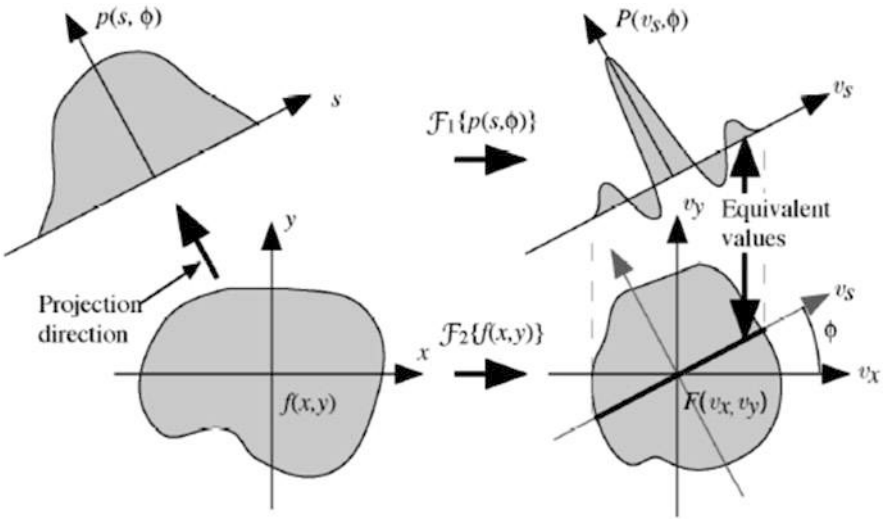

Fig. 20.7 Central-section theorem. $\mathcal{F}\{\ f\ (x, y)\ \}$ is the two-dimensional Fourier transform of the image and v_x is the Fourier conjugate of x

beam projections, the shift invariance of the image object is lost. As a consequence, equal weighting is not appropriate for back-projecting the measured divergent beam projections as it is in the parallel-beam cases. However, one can exploit the feature that in each single view, all the back-projections converge at the same X-ray focal spot. Therefore, the back-projection operation has a physical meaning only in a semi-infinite line: from the X-ray source position to infinity. Intuitively, an appropriate weight for the divergent beam back-projection operation should be a function of the distance from the X-ray source position to the back-projected point. If the distance from an X-ray source position $\vec{y}\,(t)$ to a back-projected point $\vec{x}$ is denoted as r, i.e. $r = \left|\vec{x} - \vec{y}\,(t)\right|$, then a weighting function $w(r)$ can be assigned for back-projecting the divergent beam projections. Using this general form of weighting function, a weighted back-projection can be defined as

$$G_d\left[\vec{x}, \vec{y}\,(t)\right] = w\left(r = \left|\vec{x} - \vec{y}\,(t)\right|\right) g_d\left[\hat{r} = \frac{\vec{x} - \vec{y}\,(t)}{\left|\vec{x} - \vec{y}\,(t)\right|}, \vec{y}\,(t)\right]$$

A measured projection value $g_d\left(\hat{r}, \vec{y}\right)$ is multiplied by a weight $w\left(r = \left|\vec{x} - \vec{y}\,(t)\right|\right)$ and then back-projected along the direction $\hat{r}$ to a point $\vec{x}$ with distance $r = \left|\vec{x} - \vec{y}\,(t)\right|$ as shown on Fig. 20.8.

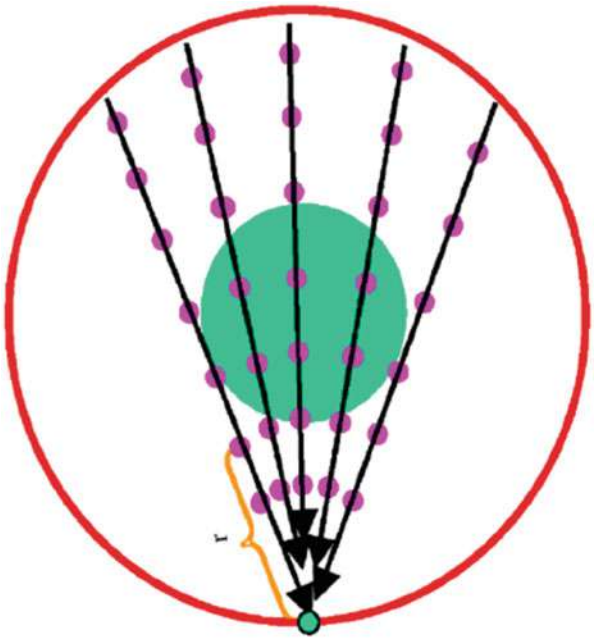

Fig. 20.8 Schematic representation of the divergent-beam weighted back-projection: The magenta points at seven different positions on each of the represented LOR's are visual guides to indicate how the sampling fraction varies as a function of the distance to the X-ray source, justifying the need for a distance-dependent weighting function in the backprojection algorithm (from [15])

20.2.3.2 Design of Modern CT Scanners

The most recent advance in CT scanning is the introduction of multi-slice helical scanning, sometimes known as spiral CT. A volume of tissue, e.g. the thorax or abdomen, is scanned by moving the patient continuously through the gantry of the scanner while the X-ray tube and detectors rotate continuously. The multi-slice systems offer the advantage over single-slice systems of being able to acquire information about the same volume in a shorter time, or alternatively to scan larger volumes in the same time or scan the same volume but obtaining thinner slices for better z-axis resolution. Helical CT has improved over the past years with faster gantry rotation, more powerful X-ray tubes, and improved interpolation algorithms [16].

The introduction of multi ring detectors and cone beam reconstruction algorithms have enabled the simultaneous acquisition of multiple slices: 4 slices in 1998, 16 slices 3 years later, 64 slices at the end of 2004, and up to 128 slices for the last generation scanners. Coupled with continuous increase of the gantry rotational speed (1.5 rotations per second in 1998, about 3 rotations per second in 2004) multislice acquisition is allowing shorter scan times (important for trauma patients, patients with limited ability to cooperate, pediatric cases and CT angiography),

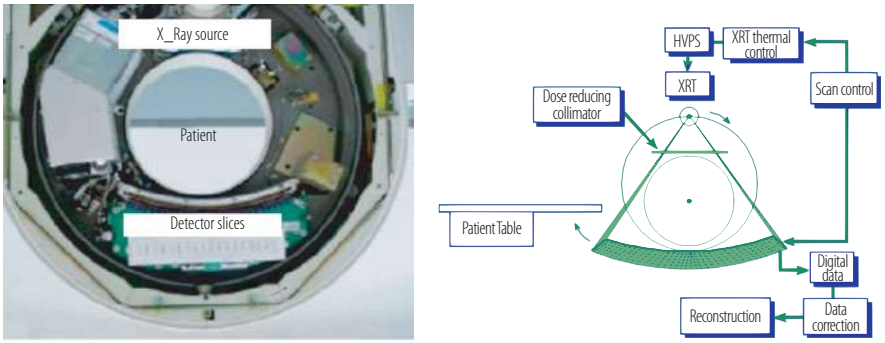

Fig. 20.9 Siemens Somaton 64 without cover and multislice CT block-diagram

extended longitudinal scan range (for combined chest abdomen scans, such as in oncological staging) and/or improved longitudinal resolution (typically 0.5 mm per slice). It has further improved the performance of the existing applications such as angiography and detection of lung and liver lesions as well as paved the way to the introduction of new ones, most notably in cardiology, where high quality images can be obtained in 10–20 heartbeats or in a single breath-hold only.

A third-generation multislice CT scanner and a block-diagram is shown in Fig. 20.9. The slip-ring technology was introduced at the end of 1980s and allowed the spiral scanning mode, when the X-ray tube and 2D arc of multislice detector system are rotated continuously around the patient while the scanning table is translated through the rotating gantry. The scan parameters, selected by the radiologist, define the X-ray tube protocol, X-ray collimation, patient table motion, data acquisition, and reconstruction parameters.

The scanner can be designed for a fixed slice collimation. It is however desirable, although more challenging, to design the detector in such a way as to meet the clinical requirement of different slice collimations adjustable to the diagnostic needs. There are basically two different approaches, the matrix detector with elements of a fixed size or the adaptive array principle (Fig. 20.10). As shown in the figure for a 16 slice array in an expanded way the cone beam geometry introduces a smearing over the field of view, which Increases the slice thickness on the edges of the cone compared to the centre.

There is therefore no need to have the same number of detector elements in the centre and in the periphery of the detector.

Most of the modern CT scanners have multiple, up to 128, detector rings, or slices. Typical CT scanners have a field of view FOV = 50 cm, and a spatial resolution of 0.5 mm in the middle of the FOV. Therefore, each detector ring houses more than 1000 detector elements per slice. The electronic channels amplify and filter the detector current and measure the filtered current at small time interval, called "view time" T_w, which is the time in which the disc rotates approximately one 1/4 to 1/3 of a degree. Since the rotational speed of modern CT scanners

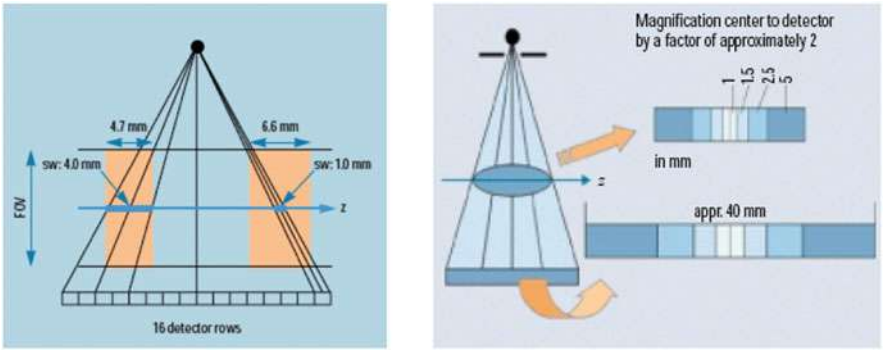

Fig. 20.10 Principle of the matrix detector (left) and of the Adaptive Array Detector (right)

can be as high as 3 or even 4 rotations per seconds, T_w can be as short as about 250 μs. At a typical focal spot-detector distance of 1000 mm, the exposure rate at the detector can reach ~0.1 Gy/s. Single detector channel detects up to 300,000 X-ray photons per sample at 3000 Hz acquisition rate. At such high photon fluxes, the detector cannot be operated in counting mode, and the majority of medical X-ray CT scanners operates in current measurement or integration mode.

One of the most widely used CT scanner, the Siemens SOMATOM Sensation 64 [17], is using a scintillating ceramics detector head. The X-ray focal spot is switched in two different positions during a view time, to reduce the aliasing of sampled data in the translational direction, i.e. along the z-axis. Consequently, the readout electronics must sample and measure the input signal two times in each view time. This machine uses an adaptive array detector with 40 detector rows in the longitudinal direction. The 32 central rows have a slice width of 0.6 mm in the centre of the field of view, whereas four detector rows on each side (in the penumbra of the collimated X-ray source) have a slice width of 1.2 mm. The slice widths being determined at the isocentre the actual detector size is about twice as large, due to the geometrical magnification (Fig. 20.11). Acquisition of 64 slices per rotation is possible through the use of a special X-ray tube with a flying spot capability (Fig. 20.12). The electron focal spot is wobbled between two different positions of a tilted anode plate by an variable electromagnetic field, resulting in a motion of the X-ray beam in the longitudinal direction. The amplitude of this periodic motion is adjusted in such a way that two subsequent readings are shifted by half a slice width in the longitudinal direction.

In general, it should be remembered that the performance factors of image quality, dose and speed can each only be improved at the expense of the other parameters. High contrast resolution in the final image is affected by noise, matrix size and contrast. Low contrast detection is affected by the size of the object, windowing and image noise. Image noise itself is affected by exposure factors, detection efficiency, slice width and, most critically, by the algorithms used in the reconstructions.

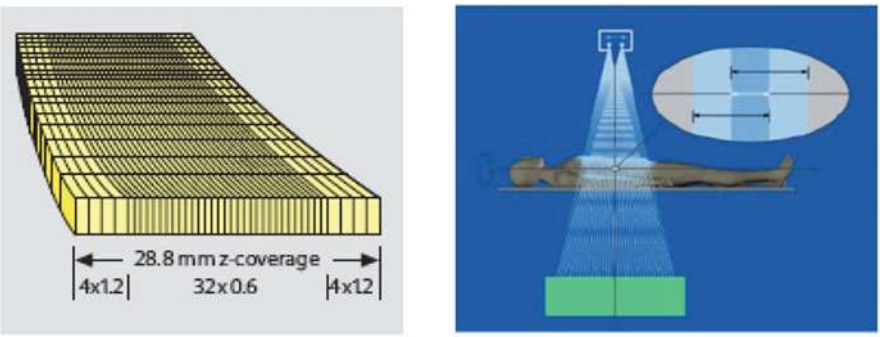

Fig. 20.11 Adaptive array detector with 32 slices of 0.6 mm in the central part, resulting in 64 slices with 0.3 mm sampling at the isocentre (Courtesy Siemens Medical Solutions)

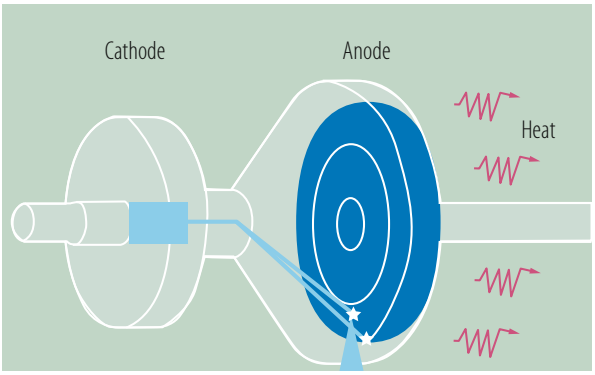

Fig. 20.12 Schematic drawing of a rotating envelope X-ray tube (Siemens STRATON, Forchheim, Germany) with z-flying spot technique (Courtesy Siemens Medical Solutions)

It must be noted that for modern scanners the X-ray tube operates with a higher duty cycle: heat output and heat dissipation are therefore a concern in the design of such multislice CT-scanners.

Another important trade-off is related to the radiation exposure of the patient. The continuing quest for better spatial resolution imposes ever smaller detector sizes. As the area of each detector cell decreases, the amount of X-rays incident on the detector decreases, leading to an increase in statistical noise. Retaining the original signal/noise ratio (and therefore the same level of contrast detection power) requires an increase in the number of X-rays and hence patient radiation dose. There is, therefore, a balance to be struck between radiation dose and resolution.

Recent developments are aiming at further decreasing the scan time and, most importantly, reducing the dose exposure to the patient. This is achieved by the introduction of the dual X-ray technology, with two X-ray sources of different energy (typically 80 kV and 140 kV) and selective photon shield for better spectral separation. Combined with two 128 slices detector panels and a

rotation speed of 0.28 s, the SOMATOM Definition Flash Spiral from Siemens achieves ultrafast image acquisition (not necessitating a breath hold) with an dose exposure approaching 1 mSv for a number of protocols, to be compared to 20 mSv 10 years ago (https://www.healthcare.siemens.com/computed-tomography/dual-source-ct/somatom-definition-flash/technical-specifications). A summary of the recent progress and new trends in CT imaging can be found in [18, 19].

20.2.4 Future of X-Ray Imaging

X-ray imaging is the historical imaging modality since the discovery of X-rays and the pioneering work of W. Roentgen in 1895. It is still the most widely used imaging diagnostic tool for physicians with nearly half a billion X-ray exams performed every year worldwide.

One major thrust for the future X-ray imaging devices is to obtain higher resolution data at a faster rate. For instance, cardiac applications would substantially benefit from CT scanners able to acquire heart images in one heartbeat or less so that motion artifacts can be minimized. One direction being pursued are scanners featuring multiple X-ray tube/data acquisition combinations operating simultaneously. Moreover, tubes incorporating a smaller focal spot are being introduced, enabling higher spatial resolution (up to around 25 line pairs per cm). Various other medical applications, such as surgical ones or the rapidly developing interventional radiography, would substantially benefit from CT scanners able to perform acquisitions at both normal and ultra-high spatial resolutions, as those required in fluoroscopic procedures. Ultra-high resolution can be achieved by substantially shrinking the physical size of both the focal spot and detector elements. Similar trends to reduce the pixel dimensions are observed for two-dimensional detector arrays used in digital radiography. This results in a considerable increase in the number of pixels and increases the complexity of the acquisition system in CT and planar digital radiography (see Sects. 20.2.4.1 and 20.2.4.2).

Present X-ray imaging only provides morphologic information but no information about the physiology of the organs under examination. However ongoing research suggests that information about the pathology of a tissue is conveyed not only by its overall X-ray attenuation, but also by its selective absorption at different X-ray beam wavelengths. This opens a new and exciting field: exploiting the new singly photon counting techniques for studying tissue pathologies with X-ray spectral images (see Sects. 20.2.4.3 and 20.2.4.4).

20.2.4.1 Indirect Detection with Phosphor Screen

The choice of the scintillating material is of course the key for a higher segmentation of a new generation of X-ray devices, as the pixel size is determined by mechanical properties of the crystal like hardness, cleavage, mechanical processing yield

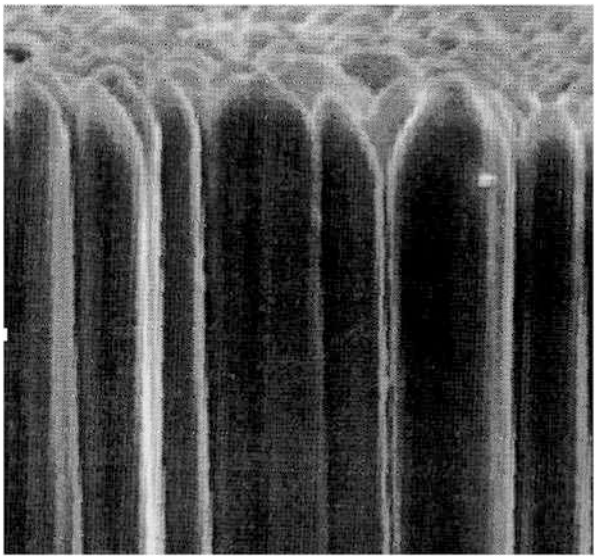

Fig. 20.13 Column structure of vapor deposited CsI(Tl). Columns have a typical diameter of 10 μm and a length of 500 μm

and cost. Large efforts have been devoted recently on specific technologies to develop a solid-state dynamic X-ray sensor with digital readout for matrixes manufacturing with sub-millimeter resolution. So called columnar structure screens were developed [20]. The rapid progress on position sensitive photomultipliers (PSPMT), Silicon photodiodes with different designs and Geiger mode Silicon photomultipliers (SiPMT) open attractive possibilities for pixel based arrays. The current design is based on large a-Si photodiodes (substrate) coupled to a CsI(Tl) layer. The scintillator layer growth is nucleated on the pattern substrate and transferred to a columnar system separated with grain boundaries as seen in Fig. 20.13. Each CsI(Tl) column is not only a scintillation pixel but also a light guide. This guide prevents or at least strongly suppresses the radial light spread and might be the way to obtain very high spatial resolution. Columnar structure growth technique allows to get 3–5 μm diameter columns and the pixel size is defined by the Si pad size as seen from Fig. 20.14. Currently, flat panels with dimensions up to 40×40 cm^2 are developed to image the human chest.

It should be noted that it is possible to use non-pixilated screens for low energy X-rays. If X-rays are absorbed in a very thin crystal layer, the angle of the emitted light is small (for the thin film) and the crosstalk to the neighbor photo-receiver is negligible maintaining therefore a good spatial resolution. The search for materials for such applications is now of very high importance.

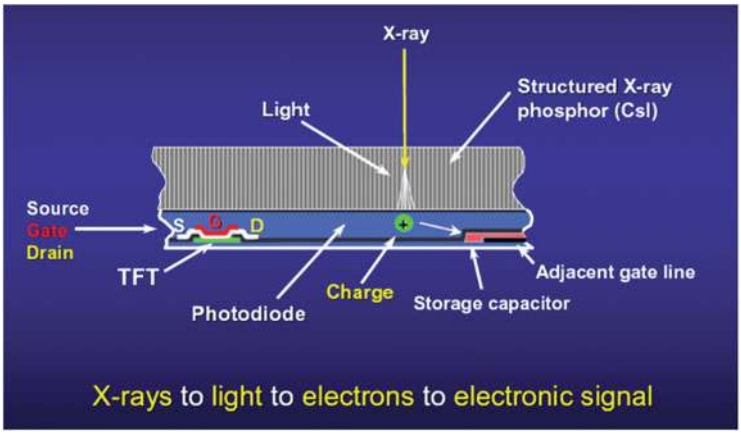

Fig. 20.14 Integrated columnar CsI(Tl) and a-Si photodiode readout for new generation X-ray flat panel (Courtesy J.A. Seibert, UC Davis Medical Centre, CA, USA)

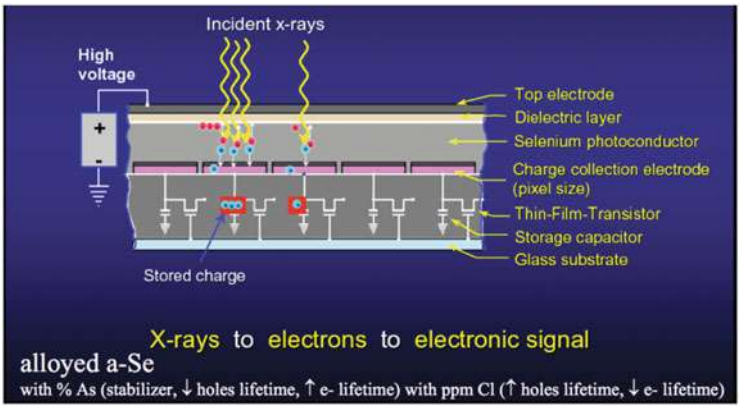

Fig. 20.15 Direct conversion detector for new generation X-ray flat panel (Courtesy J.A. Seibert, UC Davis Medical Centre, CA, USA)

20.2.4.2 Direct Conversion Screen

Another even more radical departure from the present X-ray detector technology may be the use of high-density room temperature semiconductors.

As shown in Fig. 20.15, direct detection flat panel technology is based on a uniform layer of X-ray sensitive photoconductor, e.g., amorphous selenium (a-Se) to directly convert incident X-rays to charge, which is subsequently electronically read out by a two-dimensional array of Thin-Film-Transistors (TFT). During readout, the scanning control circuit generates pulses to turn on the TFTs one row at a time, and transfers the image charge from the pixel to external charge sensitive amplifiers. They are shared by all the pixels in the same column. Each row of the detector

typically takes about 20 μs to read out. Hence a detector with 1000 × 1000 pixels can be read out in real-time (i.e., 30 frames/s).

A challenge for this approach is the practical implementation of the complex pixel design over a large area with consistent and uniform imaging performance. The problem of charge collection efficiency and speed for materials with high Z and sufficient thickness remains a major concern. Substantial technical problems must be resolved before these technologies will be implemented in commercial X-ray devices.

20.2.4.3 Single Photon Counting

The impressive CT images shown in the literature (Fig. 20.4) require several tens to 100 times higher X-ray exposures compared to standard radiography (typically 20–50 mSv as compared to 0.1 mSv). On the other hand, the enormous research effort on particle detectors has led to the development of digital X-ray detectors with very small pixels, based on silicon (Medipix) [21] and on gaseous detectors (GEM, Micromegas) [22]. The major and unique feature of these devices is their capability to work in single photon counting mode up to very high rates. Excellent high contrast images can therefore be obtained with X-ray doses up to 10 times smaller than for standard X-ray systems working in current mode.

There are a number of advantages of counting systems over current mode systems, such as:

- maximization of the contrast resolution, limited by the intrinsic Poisson statistics of the number of detected photons
- elimination of the excess noise resulting from the variance in the number of visible photons produced by the X-ray conversion in the phosphor screen, also called Swank factor [23]
- linear behavior over the whole dynamic range, which can be adapted to the specific application requirements
- possibility of implementing multiple thresholds for energy discriminating techniques, which can be used for instance for dual energy radiography, K-edge subtraction or Compton scattering discrimination
- no need for an energy dependent weighting factor as each event has equal weight whatever its energy.

This results in much better image contrast performance and significantly lower doses as shown in Fig. 20.16 in a comparative study of the signal-to-noise ratio for a 2 mm thick tumour as a function of the X-ray tube current for different mammography systems. The Medipix single photon counting device achieves the same image quality as the best commercial mammograph working in current mode for typically half the dose.

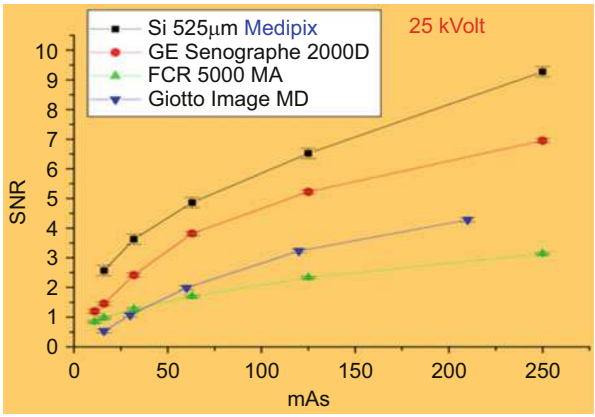

Fig. 20.16 Comparative study of the signal to noise ratio of different mammography systems as a function of the anodic current times exposure time, which is proportional to the X-ray fluence (from [24])

20.2.4.4 Spectral X-Ray Imaging

The introduction of hybrid pixel detectors in X-ray imaging**,** where the sensor array and the matching read-out chip are processed independently and are connected together only in the final step, has allowed high dynamic, noise free images to be recorded on the basis of single photon counting techniques [25]. Moreover, among the most promising recent developments in CT is the use of spectral information to improve contrast discrimination, by acquiring data with different energy thresholds. In traditional CT imaging, the overall attenuation of X-ray intensity is measured by the detector, but the detected X-rays are not spectrally resolved. This introduces a bias in the images because the absorption of X-rays by different materials depends on the X-ray energy. A significant amount of information can therefore be gained by including spectral data in the CT reconstruction process. Based on differences in X-ray absorption, different materials can be distinguished and quantified with a single spectral CT scan.

Two principal methods provide spectral CT data. The first method, dual-energy (DE) CT, uses X-ray sources with two different energy spectra and energy integrating X-ray detectors. The second method uses a single X-ray source but has energy-resolving detectors (photon counting detectors) that measure the energy of each detected photons. DE CT is currently used clinically and has been successful in improving imaging for a variety of applications.

While the soft tissue attenuation coefficient is rather wavelength independent, the photoelectric effect in high atomic Z materials strongly depends on X-ray energy. This feature can be exploited by using contrast agents containing such high Z elements. The attenuation coefficient of such substances (calcium in bone, iodine, gold) will show significant differences if the two energy spectra are recorded on either side of the K-edge for these heavy elements. The large increase in attenuation at energies above the K-edge leads to large signal differences between the two scans. By combining data from the two energy sets, these high Z materials can be distinguished and quantified. Figure 20.17 shows scans of a mice after injection

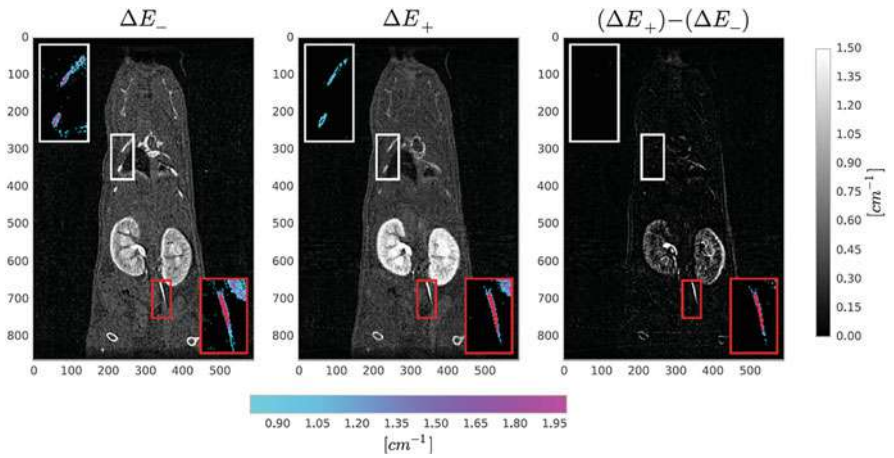

Fig. 20.17 Coronal slices of a sacrificed mouse after injection of an iodinated contrast agent. From left to right: slice reconstructed in the energy window just below the iodine K-edge, in the energy window just above the iodine K-edge and K-edge image (subtraction of the first slice from the second one). Zooms of two ribs and the ureter are also shown (from [26])

of a iodinated contrast agent. The difference between the scan taken in an energy window just below the iodine K-edge (left) and the scan taken in an energy window just above the K-edge (centre) shows the iodine concentration in some parts of the kidneys. More information can be found in [26].

20.3 Single Photon (SPECT) and Positron (PET) Emission Tomography

20.3.1 SPECT and PET Working Principle

Nuclear medicine relies on using radioactive molecules administered to a patient for diagnostic or therapeutic purposes. Radioactive molecules behave in vivo the same way as their non-radioactive "natural" equivalent involved in the metabolic or molecular processes under study. Nuclear medicine is used daily in oncology, cardiology, neurology, paediatrics, rheumatology or orthopaedics for diagnosis and therapy.

A new and recent concept is molecular imaging. It provides the ability to visualize and quantitatively measure in-vivo the activity of different biological and cellular processes activated or depressed in some pathologies.

The working principle of emission tomography is to image γ rays emitted by the radiotracers injected into the patient. Contrary to X-ray CT and standard nuclear magnetic resonance, which provide very precise images of the anatomy of

organs, nuclear molecular imaging modalities give in vivo access to the quantitative functioning of these organs.

20.3.1.1 SPECT

In Single Photon Emission Computed Tomography (SPECT) a molecule involved in the metabolism of the patient is labeled by a single photon emitter (usually ^{99}Tc emitting one 140 keV γ ray). After injection, this molecule concentrates preferentially in the organs or tumours where this metabolic function is active and allows their imaging through the reconstruction of the γ ray emitting points. The most popular technique is based on the "Anger logic", where γ rays are directed through a multi-hole collimator to a large slab of Sodium Iodide (NaI) or Cesium Iodide (CsI) scintillator. The coordinates of the interaction point are then determined by comparing the signals from a set of photomultipliers (PMT) coupled to the crystal, by the centre of gravity method (Fig. 20.18). This technique, called scintigraphy, is still largely used in many hospitals and medical imaging labs, but suffers from a relatively poor space resolution, of the order of a few centimeters.

More recent detector designs are based on discrete scintillating pixels coupled to multichannel photodetectors, such as multianode photomultipliers or avalanche photodiode matrices. But the most impressive progress has been made on the collimator, which is the main limiting factor for the spatial resolution and the sensitivity of SPECT devices. Several configurations have been studied:

- The parallel collimator, in which all the septa are perpendicular to the crystal surface
- The slanthole collimator, where the holes are parallel to each other but slanted, all in the same direction
- The fan beam collimator, where the holes are focused to a line
- The cone beam collimator, where the holes are focused to a point
- The pinhole collimator, at some distance from the crystal, where the field of view increases with the distance from the object

The best results so far are achieved with multi-pinhole configurations, for which sub-millimeter spatial resolution and sensitivities at the level of 1 cps/KBq have been obtained on small animal imaging SPECT cameras. The counterpart of using collimators, and particularly pinholes, is a reduction of the overall sensitivity of the SPECT camera. For clinical applications a compromise needs to be found between sensitivity and spatial resolution.

The spatial resolution is usually given by the Full Width at Half Maximum (FWHM) of the so-called point spread function (PSF):

$$FWHM = \frac{D}{L}(z_0 + L + B)$$

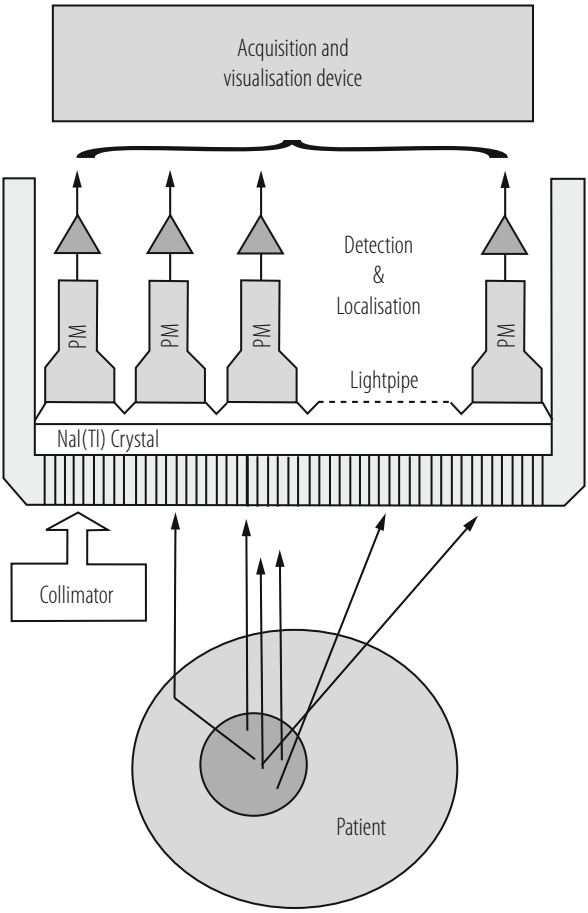

Fig. 20.18 Principle of an Anger camera

where D and L are the collimator hole diameter and length, z_0 is the distance from the γ-ray source to the collimator entrance and B is the distance between the collimator back face and the image plane in the crystal.

The collimator introduces an important loss of efficiency by a factor of 10^3 to 10^4. The resulting efficiency is given by the following formula:

$$\eta = \varepsilon \left(\frac{a_{hole}}{a_{cell}} \right) \left(\frac{a_{hole}}{4\pi L^2} \right)$$

where a_{hole} and a_{cell} are the collimator hole and cell area respectively and ε is the γ detection efficiency in the scintillator.

20.3.1.2 PET

In the case of Positron Emission Tomography (PET) the functional molecules are labeled with a β^+ emitter generally produced in a cyclotron. These PET tracers, injected into the patient, simulate natural sugars, proteins, water and oxygen presence, circulation and accumulation in the human body. Once fixed in the organ or the tumour, the molecule emits positrons, which annihilate very quickly on contact with the tissue, emitting two gamma photons located on the same axis—called the line of response (LOR)—but in opposite directions, with a precise energy of 511 keV each (Fig. 20.19).

The coincidence detection scheme introduces therefore an electronic collimation, which greatly enhances the background rejection as compared to SPECT. Moreover, the line of interaction being precisely determined by the two detectors hit in coincidence, there is no need for a collimator system, which severely reduces the sensitivity of SPECT cameras. In order to simplify and reduce the image reconstruction time the first generation PET scanners used septa to restrict the acquisition to transversal slices through the patient in a so called 2D acquisition mode. The slices were then combined off line for a 3D image reconstruction. Modern PET scanners benefit from the considerable progress in computer power and directly acquire data in 3D mode without septa (i.e. recording all the LORs independent of their direction relative to the scanner axis), which results in a significant gain in sensitivity.

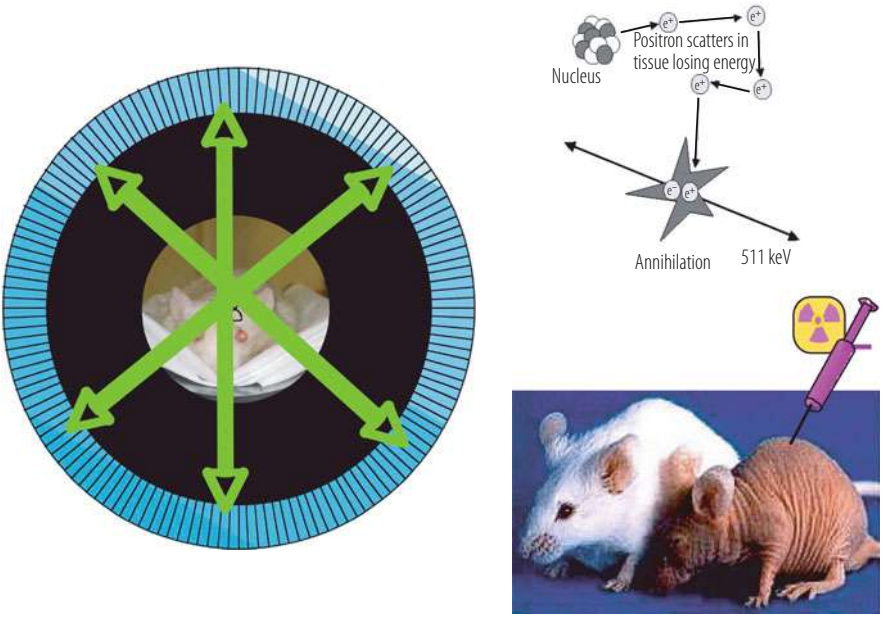

Fig. 20.19 Principle of a PET scanner

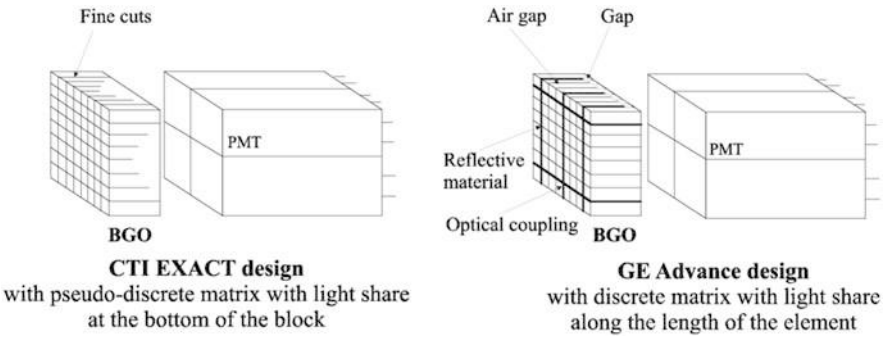

Fig. 20.20 The readout quadrant-sharing scheme for the CTI (now Siemens) and for the General Electrics PET scanners

Until recently, as a result of a compromise between performance and cost, PET scanners were using partially segmented BGO crystals readout by groups of four PMT's (quadrant-sharing scheme), allowing a reconstruction precision of the order of 4–5 mm (Fig. 20.20). Modern machines are going progressively to higher segmentation of the crystals and of the readout to achieve higher spatial resolution. Resolutions of the order of 1 mm have been reached at least for small dedicated machines, such as breast imaging devices or for small animals.

Positron Emission Tomography measures the uptake of the tracer in different organs or tumours and generates an image of cellular biological activities with a much higher sensitivity than any other functional imaging modality. The PET images can be used to quantitatively measure many processes, including sugar metabolism, blood flow and perfusion, oxygen consumption etc. Moreover, specialized PET scanners designed for experimental small animal studies (mouse, rat, rabbit) are powerful tools for fundamental research in disease models, new therapeutic approaches and pharmacological developments. The most commonly used radio-isotopes are ^{18}F with a lifetime of 109.8 min, ^{11}C (20.4 min), ^{13}N (10 min) and ^{15}O (2.1 min), the last three ones being among the basic building blocks of organic systems and therefore being easily introduced chemically in molecules involved in metabolic or pharmaceutical reactions. A typical example is FDG (^{18}F labeled fluorodeoxyglucose), which allows monitoring the energetic consumption of the cells in different parts of the body. FDG is a glucose analog, where a hydroxyl group has been substituted by a ^{18}F atom. Once phosphorylated by the hexokinase enzyme into FDG-6-phosphate, it remains trapped in the cell, where it accumulates. The interest in labeled glucose lies in the fact that tumour cells are characterized by an increase of glycolisis and expression of glucose transporters, such as GLUT-1, as compared to healthy tissues. This increase of FDG metabolism allows detecting tumours and related metastases through their abnormally high glucose concentration and therefore increased γ-ray activity.

PET has a very high sensitivity, at the picomolar level, which makes it one of the essential tools of molecular imaging with applications in many areas such as

expression and occupancy of therapeutic molecular targets, pharmacokinetics and pharmacodynamics, mechanisms of therapeutic action and functional response to therapy.

20.3.2 Detector Challenges for Modern Nuclear Medicine

The spectacular development of in-vivo molecular imaging will allow in the near future bridging the gap between post-genomics research and physiology and opening interesting perspectives for new diagnostic and therapeutic strategies for many diseases. Nuclear medicine and particularly PET imaging are already playing and will play an increasing role in molecular imaging [27, 28]. Constant progress in the medical and biological fields implies that imaging performances have to be continuously improved. In order to fulfill the needs of quantitative cell and molecular imaging, of dynamic studies over a certain time and of individualized therapy focusing on the patient's genotype, major technical improvements [29] will be necessary, comparable to those in large particle detectors, in order to deal with:

- integration of a very large number of increasingly compact measuring channels (several hundred thousands)
- data acquisition rates at the level of tens gigabytes/second
- several billions of events to reconstruct an image
- about 1000 gigabytes of data per image and commensurate computational power for the reconstruction
- integration of multiple technologies requiring pluridisciplinary competences for complex, compact and reliable systems.

The challenge for functional isotopic imaging lays in its capacity to identify the specific molecular pathways in action in a metabolic process and to quantitatively measure their relative metabolic activity. To achieve this, it is necessary to improve both the imaging system's spatial resolution, that is, its capacity to discriminate two separate objects, and the measurement's signal to noise ratio, that is, how precisely a metabolic agent's concentration in a body area can be determined. The precision of the concentration measurement depends mainly, but not only, on the imaging system's sensitivity, and therefore its capacity to accumulate the statistics needed to tomographically reconstruct the radiopharmaceutical tracer distribution. Moreover, the location of this metabolic activity must be precisely associated to the organs or parts of the organs under examination. This explains the increasing demand for combining functional and anatomical imaging devices.

The perspectives to develop isotopic imaging with multimodality and multifunctional capability revolve around three goals:

- improving sensitivity
- improving spatial resolution
- improving temporal resolution

20.3.2.1 Improving Sensitivity and Specificity

Sensitivity is defined as the ratio of the detected number of radioactive decays and the radioactivity injected into the patient and fixed on the organ under study. It reaches at best 10% in the case of PET scans on small animals, and a few percent only in the case of whole body PET. The main losses arise from poor geometrical acceptance, gaps between crystals, rejection of Compton events due to partial conversion of the γ-ray in the crystal and electronics dead time. Moreover, whole body PET scanners visualize only the patient's thorax in one acquisition run, which is sometimes a limiting factor, for instance in oncological studies of bone metastases in limbs. In the last few years, the use of faster scintillating crystals and electronics and improved geometrical acceptance has allowed to reach the above-mentioned sensitivity levels.

However, sensitivity has to be further improved for several reasons. First, examination durations have to be shortened. Today, a whole body scan lasts between 10 and 20 min. It would be desirable to reduce this time to a few minutes to improve the patients' comfort. It would also increase the exploitation of costly equipment and infrastructures with a significant impact on the cost of examinations. Shorter acquisition times would also improve the image quality because the impact of the patient's natural movements—breathing and cardiac activity, digestive bolus, etc.—would be reduced. Quicker metabolic processes could be followed, which are crucial for pharmacokinetic studies.

There is a strong interplay between sensitivity and spatial resolution, since the signal-to-noise ratio per voxel is the relevant image quality factor. Doubling the linear spatial resolution (i.e. reducing the volume by a factor 8), requires a 16 times (because it requires two detector pixels to identify one voxel) higher noise equivalent rate if the statistical quality of the image is to be maintained. The acquisition time of an image depends on many factors, which all influence the noise-equivalent measurement of the imaging system and of the radioactivity administered to the patient. Some of the relevant parameters are: the imaging system's geometrical acceptance; the efficiency and energy resolution, improving the energy selection of events and hence discriminating in a coincidence system the diffused events; the time resolution allowing to reduce the width of the coincidence window and rejecting random coincidences more efficiently; dead time of the detectors.

Increasingly, specific molecular signatures for the major diseases are being evaluated to devise individually targeted therapies adapted to the patient's genotype. This requires ever more performing equipment and more specific protocols. Indeed, it is highly desirable to study different molecular pathways simultaneously and to record the intensity, the range, the localization and the temporal development of various biochemical processes in their natural environment in the human body. In this way, the nature of the pathology can be established, at least partially, through molecular imaging, using an array of radiopharmaceuticals giving information on cell proliferation (FLT), on energetic metabolism (FDG) or on aminoacid synthesis (methionine) in the various tissues. For multitracer analysis of various biochemical or pathophysiological processes several radioactive tracers have to be administered.

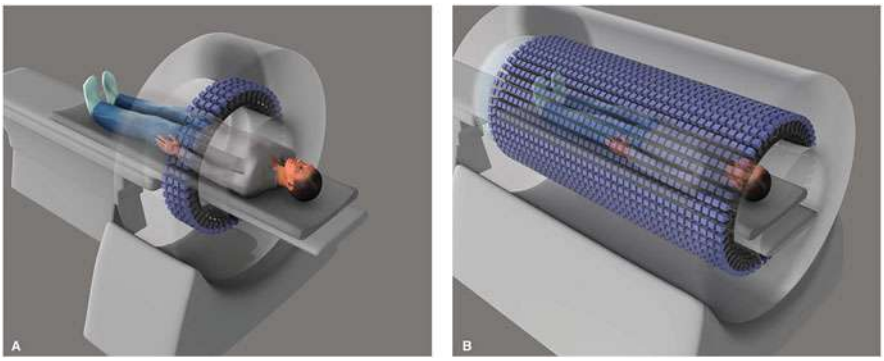

Fig. 20.21 Conventional PET scanner (**a**) and Total-Body PET (TB-PET) scanner (**b**). From [30]

In order to keep the doses tolerable for the patient, high-sensitivity PET scanners have to be developed, which would also open new prospects for young or pregnant women and for children.

The obvious approach for increasing the sensitivity is to increase the geometrical acceptance of the scanner. In some cases, developing dedicated equipment might be the right solution to study an organ (brain, breast, prostate) in a more efficient and optimized way. In this case the detectors can be placed closer to the organ under study, increasing therefore the geometrical acceptance of the events.

Until recently the length of the detector cylinder, or length of the system's sensitive volume ("field of view"), has remained essentially at the level of 15–18 cm (Fig. 20.21a), resulting in a very small geometric efficiency $\leq$0.2. Several static views need therefore to be acquired to perform a "whole body" scan, i.e. from the head to the pelvis. This procedure presents major limitations, as the acquisition time increases in proportion to the number of views. A "standard study" with the injection of 10 mCi of ^{18}F-FDG takes about 3 min per view and, therefore, approximately 20 min per entire study.

A few year ago, the EXPLORER consortium (http://explorer.ucdavis.edu) had launched the concept of a Total-Body PET scanner (TB-PET) to realize the full potential of PET—extending the FOV to cover the entire length of the body (Fig. 20.21b).

In TB-PET, the vast majority of the emitted photons could be captured. This step change in technological evolution would mean simultaneous coverage of all the tissues and organs in the body, with an overall >40-fold gain in effective sensitivity and a >6-fold increase in signal-to-noise ratio compared with whole-body imaging on current PET scanners. The challenge of funding the construction of the first prototype machine, an expensive novel device, was successfully overcome in September 2015 through funding from the NIH Transformative Research Award program, which recognizes high-risk, high-reward, paradigm-shifting innovative research [30].

To further improve sensitivity, we need denser and faster scintillating crystals or direct conversion materials, more compact and adaptable geometries, lower-noise and faster acquisition electronics, more parallelized acquisition architecture with integrated processing power, and at least partial use of the information included in the events diffused in the patient or in the detector. Potential progress in these fields are described in the next paragraphs.

Another promising way to significantly increase the sensitivity is to push the limits of time-of-flight PET scanners, as will be explained in Sect. 20.3.2.3.

20.3.2.2 Improving Spatial Resolution

Spatial resolution reaches 1.5–2 mm at the centre of the field of view of small animal PETs, but worsens off-axis. Modern technologies have allowed to reach a 1 mm resolution for small animals PETs or for scanners dedicated to specific organs, and a 4–5 mm resolution for whole body scanners.

Good spatial resolution is obviously of value for the study of small animals, but also for humans: increasingly smaller structures which are involved in specific metabolic processes can thus be visualized. Anatomical localization can also be more precise and combining CT or MRI information can be improved. But it is in the field of quantification that the improvement potential is likely to be the most significant. By reducing the blurring caused by insufficient spatial resolution (also called partial volume effect), the dynamic sensitivity of the radiotracer's concentration measurement, also called Standard Uptake Value (SUV), can be significantly improved.

There are four factors, which limit a PET camera's spatial resolution:

- the positron's mean free path: once the ligand has fixed itself on the organ or tumour being investigated, the radioisotope used to label it emits positrons with a kinetic energy depending on the isotope. As the annihilation probability of this positron is maximum when the positron has sufficiently slowed down, there is a difference between the positron emission point and its annihilation point. This difference is about 0.5 mm in the case of ^{18}F, but it can reach several mm for other isotopes—4.5 mm for ^{82}Rb, for instance. This blurring is often considered as an intrinsic limitation to the PET spatial resolution, but it can be significantly attenuated thanks to various electromagnetic artifices. For instance, the positron's trajectory usually revolves around the lines of a magnetic field—naturally present in the case of a combined PET-MRI camera—, which therefore reduces its conversion distance. This is however only effective in the plane perpendicular to the magnetic field and for new generation of MRI devices with high field (7T or more). It also has to be noted that positron annihilation probability as a function of its speed is a well-known function but it is not exploited today in image reconstruction algorithms.
- non-collinearity of the two gamma photons deriving from positron annihilation: momentum conservation implies that the two gamma rays resulting from the annihilation of a motionless positron are emitted on the same line of response

(LOR) in opposite directions. In practice the positron is not at rest when it is annihilated, which causes an average non-collinearity of the two gamma photons of about 0.25°. The error in the reconstruction of the emission point varies like the square of the scanner's radius. This error is reduced in equipment dedicated to the study of specific organs whose detectors can come as close as possible to the areas to be studied.

- size of the detection crystal (or pixel): it is the limiting factor in spatial resolution of commercially available scanners. Typically, the reconstruction error of each LOR is given by the half width of each pixel. The use of higher-density crystals and highly segmented photodetectors improves the spatial resolution. A significant increase in the number of channels, resulting from finer detector segmentation, implies that important efforts have to be made to develop cheap solutions for photodetectors and readout electronics. Difficult engineering problems have to be solved in order to integrate all the channels in a small volume and to keep the electronic equipment's thermal dissipation at an acceptable level.
- parallax effect: good spatial resolution has to be obtained not only on the axis, but also on the whole of the field of view (FOV). The depth of detecting crystals is limited by the density of the crystals and it cannot be reduced without altering the detector's sensitivity. If the conversion point of the gamma photon in the crystal is not known, spatial resolution deteriorates with increasing distance from the scanner axis. This error, which is known as parallax error, is increasing with decreasing the scanner's radius (Fig. 20.22). To limit this effect, one solution is to use several crystals in depth in a so-called phoswich configuration. If appropriate emission wavelength and decay time parameters are chosen for the crystals, the readout electronics can differentiate a conversion occurring in the front part or in the back part of the phoswich. Spatial resolution is therefore much more homogeneous on a larger field of view (Fig. 20.23). This scheme has been adopted in the ClearPET® small animal PET scanner with a combination of two 10 mm long LSO and LuAP crystals [31].

Another solution is to determine the γ conversion point in the crystal by means of a light sharing scheme with readout at both ends of the crystal. This solution has been chosen for 20 mm long LSO crystals in the ClearPEM®, a dedicated PET scanner for breast imaging [32].

20.3.2.3 Improving Time Resolution

Time-of-flight reconstruction can significantly reduce the signal to noise ratio in PET scanners, by constraining the annihilation point to a shorter segment on each LOR, with an uncertainty given by:

$$\Delta x = \frac{c}{2}\Delta t$$

where Δx is the position error, c is the speed of light, and Δt is the timing error.

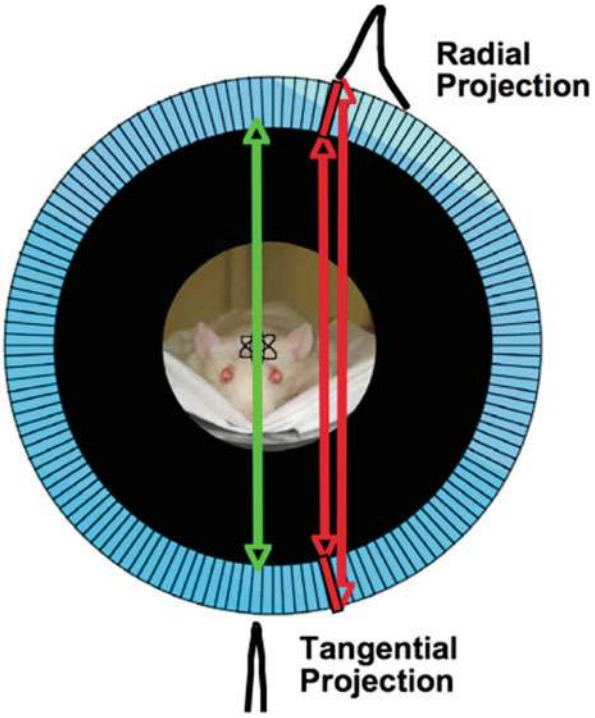

Fig. 20.22 Illustration of the parallax error for off centre events

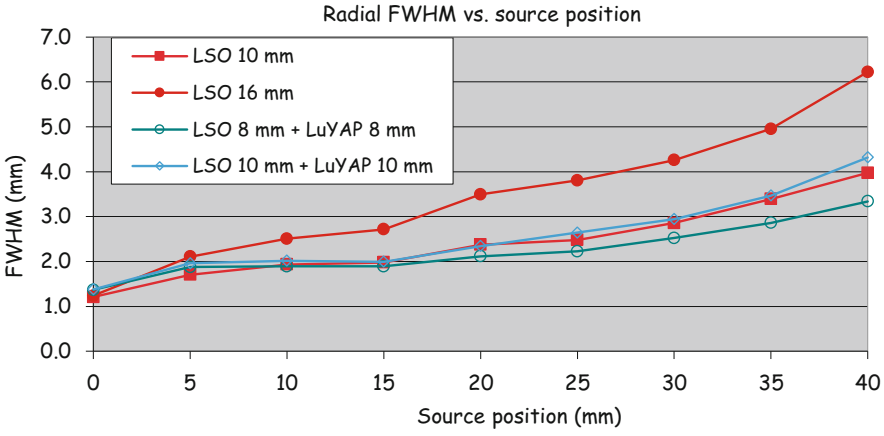

Fig. 20.23 Spatial resolution with and without phoswich for the ClearPET® (Courtesy Crystal Clear collaboration)

Until recently, PET scanners did not have any Time-of-Flight (TOF) capability to localize the position of the positron decay along the line of response (LOR) of the two γ-rays. Developments in fast scintillation crystals, photodetectors and electronics have open the way to TOFPET scanners with coincidence time resolution (CTR) improving progressively from 500–600 ps to 249 ps as recently announced by Siemens for their Biograph-Vision scanner. Pushing the limits of TOFPET techniques is motivated by the perspective for a significant improvement in the image signal-to-noise ratio (SNR), resulting in a corresponding clinical sensitivity increase and dose reduction, as given by the following equation:

$$SNR_{TOF} \Big/ SNR_{NONTOF} = \sqrt{\frac{2D}{c.CTR}}$$

where

D is the diameter of the Field of View (FOV),
c is the speed of light in vacuum
CTR is the Coincidence Time Resolution

Breaking significantly the 100 ps barrier, would dramatically improve the SNR (Fig. 20.24) and significantly remove artefacts affecting tomographic reconstruction in the case of partial angular coverage. This will open the field to a larger variety of organ-specific imaging devices as well as to imaging-assisted minimally invasive endoscopy.

Ultimately, a time resolution of 10 ps would lead to an uncertainty of only 1.5 mm for a given positron disintegration along the corresponding line of response (LOR). This is the order of accuracy achieved in today's very best small animal or organ

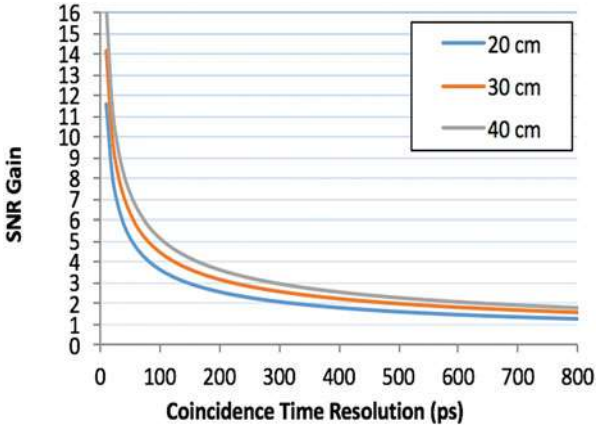

Fig. 20.24 Signal to Noise Ratio improvement as compared to non TOFPET as a function of the Coincidence Time Resolution for three different diameters of the Field of View (FOV)

specific PETs. The processing time of tomographic back-projection or iterative reconstruction algorithms would be considerably reduced, as true 3D information would be directly available for each decay event [33]. The possibility to see in real time the accumulation of the events during the acquisition could introduce a paradigm shift in routine clinical protocols, allowing in particular adapting the acquisition time to what is really observed and not to some predetermined evaluation. Moreover, such a timing resolution would allow recording the full sequence of all γ-ray interactions inside the scanner, including Compton interactions, like in a 3D movie, opening the way to the integration of at least a fraction of the Compton events in the image reconstruction, further improving the sensitivity.

To improve time resolution scintillating crystals with short decay time and fast treatment and acquisition electronics are needed. This has a double impact on image quality:

- as the width of the coincidence window is reduced, the number of isolated events decreases linearly. The proportion of random coincidences, which increase the detector's dead time and introduce noise into the image, is therefore reduced as the square of the single event rate. Images are less noisy and require less filtering, increasing spatial resolution and contrast.
- the use of time of flight information along the line of response (LOR) eliminates many random coincidences and reduces significantly the image noise.

In PET, the random event rate for an individual LOR is given by:

$$R = 2R_1R_2\Delta T$$

where R is the random event rate for that chord, R_1 and R_2 are the single event rates for two detector elements that form that chord, and ΔT is the width of the coincidence gate. The total number of random events in the image is the sum over all the chords, thus is proportional to ΔT. The mean contribution to the image from random events can be measured and subtracted, but the noise resulting from the statistical variations in this rate remains. The residual noise from random coincidences is usually estimated using the noise equivalent count rate (NECR) [34], a common figure of merit for comparing tomograph performance. The NECR is given by:

$$NECR = \frac{T^2}{T + S + 2R}$$

where *NECR* is the noise equivalent count rate, T is the true coincidence event rate, S is the scattered event rate, and R is random event rate. The noise equivalent count (NEC) metric is designed to obey counting statistics; that is, the NEC variance is equal to the NEC. Although the magnitude of the NECR is very sensitive to the source and camera geometries, this formalism is useful for predicting how changes in the trues, randoms, and scatters affect the image quality. Figure 20.25 shows such examples of NEC curves measured with a 20 cm long 20 cm diameter phantom

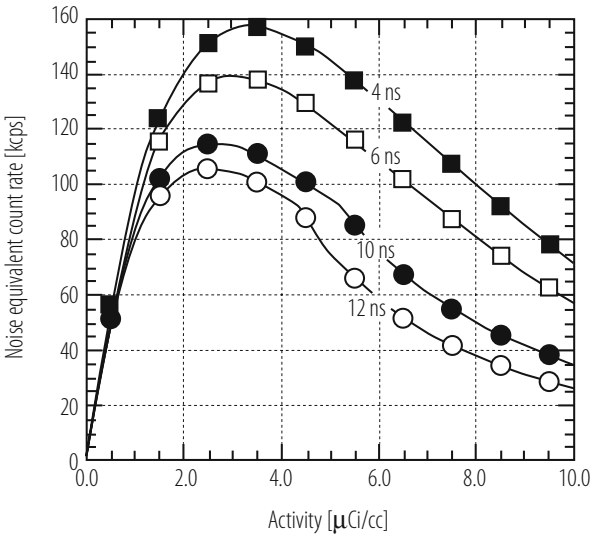

Fig. 20.25 NECR curves as a function of coincidence window width. The object imaged was a uniform 20 cm diameter cylinder and the camera had an 82 cm detector ring diameter and 15 cm axial extent (ECAT EXACT HR from Siemens) (Courtesy W.W. Moses, LBNL)

for a commercial PET scanner (ECAT EXAT HR from Siemens). The NEC value first increases linearly with the injected activity. It then progressively saturates and slowly decreases when the electronics dead time becomes significant as compared to the event rate. The importance of reducing the coincidence gate is evident from these plots.

20.3.3 Current and Future Technical Approaches

In the last few years, there have been noticeable improvements in commercial imaging equipment, with increased level of pixellization, better angular coverage, faster crystals, higher degree of integration of electronics with increased built-in functionality, more efficient reconstruction algorithms. Further progress is expected if medical imaging, and particularly nuclear imaging, can take advantage of significant technological advances in other fields like telecommunications or particle detectors.

Developments proceed along the following lines:

- new denser and faster scintillating crystals or direct conversion materials
- highly segmented and compact photodetectors
- low noise and highly integrated front end electronics
- data acquisition systems based on highly parallelized architectures
- efficient data filtering algorithms

- modern and modular simulation software based on universally recognized standards
- high performance image reconstruction and analysis algorithms

20.3.3.1 Conversion Materials and Metamaterials

The scintillating crystals used in PET scanners have to be dense, with a high atomic number, so as to optimize detection efficiency, and fast, in order to reduce dead time. The previous generation of PET scanners were using BGO crystal arrays, which have the advantage of being very dense (7.1 g/cm^3) and of having the highest atomic number known to this day for a scintillator (75), and therefore a high photoelectric conversion efficiency. Their main flaw is a slow decay time (300 ns) of the scintillating light. As a result, these scanners work with a limited sensitivity of about 1000 kcps/mCi/ml with a coincidence window of about 10–12 ns and a proportion of diffused events of more than 30%.

A new generation of scanners is now using LSO (Lutetium oxyorthosilicate) crystals [35], about 10 times faster than BGO and in some cases, the capability of determining the interaction depth in the crystals thanks to phoswich technology or double readout schemes. Combining these developments with progress in readout electronics and data acquisition, a gain in sensitivity by about one order of magnitude and in spatial resolution by a factor 2 or 3 has been achieved.

In the last 10 years, many groups, among them the Crystal Clear collaboration [36], have devoted a large effort on pluridisciplinary work to develop new scintillating materials meeting the demands for increasingly efficient detectors in physics and medical imaging. The most attractive scintillating crystals currently available or being developed for nuclear medicine are presented in Table 20.3. Cadmium Tungstate (CWO) and two ceramics compositions used for CT scanners are also mentioned for comparison. Figure 20.26 shows some pictures of the growth of LuAP ingots and pixel production developed in this context [37] for the preparation of phoswich pixels in combination with LSO for the ClearPET small animal PET scanner.

Other attractive crystals presently being developed are from the Lathanum halide group [38]. $LaBr_3$:Ce for instance has a higher light yield than CsI:Tl with more than 60,000 photons/MeV. The combination of high scintillation efficiency and good low energy linearity gives this crystal an unprecedented energy resolution (about 3% measured with avalanche photodiodes for 511 keV photons) and excellent timing properties.

Contrary to scintillators, semi-conductors convert the energy of the gamma photons to electric charge carriers (electrons and holes), which are directly collected on electrodes. However, most of the semi-conducting materials known today and used industrially, such as silicon, are not dense enough and do not have sufficient stopping power for 511 keV gammas (density 2.33 g/cm^3 and atomic number 14, as compared, for instance, to BGO density, 7.13 g/cm^3, and average atomic number,

Table 20.3 Scintillators already used or in development for medical imaging

Scintillator	Type	Density (g/cm^3)	Light yield (Ph/MeV)	Emission wavelength (nm)	Decay time (ns)	Hygroscopic
NaI:Tl	Crystal	3.67	38,000	415	230	Yes
CsI:Tl	Crystal	4.51	**54,000**	550	1000	Lightly
CWO	Crystal	7.9	28,000	470/540	20,000/5000	No
$(Y,Gd)_2O_3$:Eu	Ceramics	5.9	19,000	610	1000	No
Gd_2O_2S:Pr,Ce,F	Céramics	**7.34**	21,000	520	3000	No
BGO	Crystal	**7.13**	9000	480	300	No
GSO:Ce	Crystal	6.7	12,500	440	60	No
LSO:Ce	Crystal	**7.4**	**27,000**	420	**40**	No
LuAP:Ce	Crystal	**8.34**	10,000	365	**17**	No
$LaBr_3$:Ce	Crystal	5.29	**61,000**	358	35	Very

Particularly attractive parameters are marked in bold

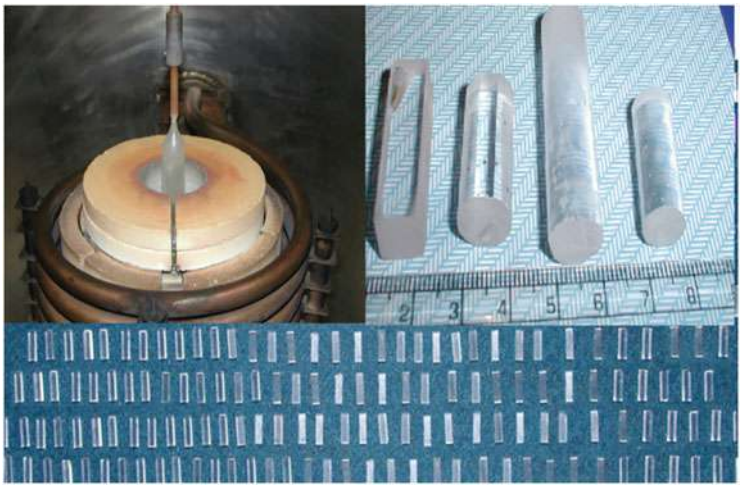

Fig. 20.26 LuYAP crystals produced in Bogoroditsk, Russia (Courtesy Crystal Clear, CERN)

75). This technique is nevertheless used in single photon X-ray imaging and makes the acquisition of high resolution and high contrast digital images possible [39].

For gamma imaging, multi-layer systems could be considered, but to this day integrating a huge number of channels in these conditions has not been solved, especially in terms of connectivity. Yet interesting solutions to these problems have recently become available through recent developments on pixel detectors for tracking devices. For example, using bump-bonding techniques semiconductors are coupled directly to their readout electronics. It has also become possible to integrate the semiconductor directly with ASIC readout chips and to read a large number of channels on a very small surface quickly and with low noise.

New semiconducting materials, denser than silicon, are also being developed: Gallium Arsenide (GaAs) [40], with a density of 5.32 g/cm^3 and an average atomic number of 31, Cadmium Telluride (CdTe), with a density of 5.85 g/cm^3 or Cadmium and Zinc Telluride (CdZnTe, or CZT) [41], with a density of 5.78 g/cm^3 but whose atomic number is higher, 49 instead of 32. One of these materials is particularly attractive because of its density and high atomic number: Mercuric Iodide (HgI_2). With a density of 6.4 g/cm^3 and an average atomic number of 62, it nearly equals the stopping power of the best scintillating crystals (BGO, LSO and LuAP). It is unfortunately very difficult to grow in reasonable size and consistent quality.

One remaining problem for these materials is the limited charge collection speed and efficiency, which requires well designed geometries with small drift regions and optimized, cost effective production technologies with a very good control of charge carriers traps.

As shown in [33], the ambitious target of a few tens of picoseconds Time-of-Flight resolution can only be met with scintillators exhibiting a very fast rise time in the scintillation process and the possibility to combine standard scintillation processes with a few hundreds of prompt photons generated by another mechanism. One of the very attractive mechanisms for the production of sub-ns scintillation processes is related to quantum confinement in nanocrystals, as explained in [42, 43].

The challenge is to optimize the design of a metamaterial combining the high density and stopping power (small radiation and interaction lengths, small Moliere radius, high photoelectric fraction) of already well known scintillators (LSO, L(Y,G)SO, PWO, BGO, LuAG, YAG, GGAG, Lu(Y)AP, etc . . .) to the ultrafast (<1 ns) light emission of nanocrystals. Different solutions are presently under study for an optimal combination of these two classes of materials, solving at the same time the problem of light transport by the use of photonic fibers, as proposed in [44].

20.3.3.2 Photodetectors

In nuclear medicine the basic technique to detect ionizing radiation uses scintillators to convert X- or gamma-rays into light and then into an electric signal by a photodetector. Until recently, the standard commercial imaging cameras were equipped with photomultiplier tubes (PMT) used as light sensors. However, these technologically mature products are approaching their limits in terms of dimension, efficiency and cost. However, their sensitivity to magnetic fields prevents their use in combined PET/MRI devices. The trend toward larger numbers of scintillating pixels of increasingly smaller size will limit their use in the future.

New compact photodetectors have been developed over the last two decades, for instance hybrid photodetectors (HPD), photodiodes and avalanche photodiodes, thanks to which sensitivity, spatial resolution and immunity to magnetic field could be significantly improved. Arrays of avalanche photodiodes have been considered

for hybrid PET/MRI scanners and several prototypes dedicated to brain and breast imaging have been built.

Avalanche photodiodes suffer however, a number of drawbacks, such as a limited gain of a few hundred, a large excess noise factor and relatively poor timing characteristics preventing their use in PET Time-of-Flight systems. For these reasons the intense ongoing R&D activity on multipixel Geiger mode avalanche photodiodes (also called Silicon Photomultipliers or SiPM) is followed with particular attention. Their working principle is based on the segmentation of the large coupling area with the scintillating crystal into a large number of small avalanche photodiode cells working in Geiger mode and connected in parallel via individual quenching resistors. The first devices of this type were developed in the late 1990s in Russia and since then several designs have been realized [44, 45].

The cells in a SiPM are all identical with dimensions ranging from 7×7 to $70 \times 70\ \mu m^2$. Each cell operates as an independent photon counter in Geiger-mode when the bias voltage is 10–20% higher than the breakdown voltage and behaves as a binary device since the signal from a cell always has the same shape and amplitude.

The gain is similar to the one of a photomultiplier, in the range of 10^5 to 10^7. Since each cell acts as a digital single photon counter the excess noise factor is very small. The light yield is directly given by the number of fired cells. This assumption is of course valid only, if crosstalk between individual cells can be eliminated, which has been solved by the use of trenches between the pixels.

Present SiPMs have a dead-time per cell of the order of several μs. For the device to be linear to the light response of fast scintillators having a decay time in the ns range, the number of cells must be larger than the maximum number of photons per event. This requires SiPM to have a cell density above 1000 cells/mm^2. SiPMs with 100–10,000 cells/mm^2 are currently available.

The overall efficiency of the device depends on the quantum efficiency of each individual cell, which is wavelength dependent but is now reaching 60–65% at the emission wavelength of LSO, and with the geometrical acceptance due to the dead space between the cells, which ranges between 20% and 70% depending on the design.

One of the most promising features of SiPMs for medical imaging applications is related to their excellent timing resolution. The active layer of silicon in a SiPM is very thin (2–4 μm), the avalanche breakdown process is fast and the signal amplitude is large. Impressive timing resolution of about 10 ps for single photons have been reported. For a system of a SiPM and a LYSO crystal with dimensions of $2 \times 2 \times 3$ mm^3, $2 \times 2 \times 10$ mm^3 and $2 \times 2 \times 20$ mm^3, a time resolution of 73 ps, 100 ps and 122 ps FWHM respectively has been measured for 511 keV X-rays [46].

20.3.3.3 Highly Integrated Low Noise Front-End Electronics

The large number of readout channels requires highly integrated low-noise and high-speed readout electronics, typically using integrated circuits of the VLSI CMOS type [47]. Institutes of particles physics are experienced in designing and integrating

large numbers of multichannel and multifunction low noise and fast electronics into complex detector systems.

Medical imaging should also benefit from the progress in large-scale integration of electronic channels with complex functions and highly segmented sensors. The concept of a hybrid detector, in which each pixel is integrated directly to its readout electronics opens totally new perspectives in the conception and architecture of new imaging systems.

20.3.3.4 Highly Parallelized and Intelligent Data Acquisition System (DAQ)

The data acquisition system has a double function: first, it has to discriminate between the interesting events—coming from real X- or γ-ray interactions—and the various types of background, and, secondly, it has to transfer data to the computer(s) where these events will be processed. In medical imaging, most of the "real events" are triggered by the time and amplitude analysis of each sensor. In this acquisition system, data are selected, standardized, organized, corrected, processed with more or less complex algorithms and finally presented as an image file.

In data acquisition system with conventional architecture, data are treated sequentially. A new event is only accepted, once the processing of the previous event is completed; pile-up is thus avoided, but dead time occurs at each treatment stage affecting the data collection efficiency and coming from three main sources: first, the sensor and the electronic pulse generation system, second, the analog-to-digital conversion of this signal, and third, the logic treatment (in general, the major one).

Although pipeline architectures could be seen as more complex and more expensive than conventional ones, improvements in electronics (ASIC and FPGA) in terms of integration and cost suggest that in a very close future medical imaging devices entirely based on this concept might be designed. Similarly, progress in data transfer speed between the electronic system and the analysis system is no more a limiting factor for data transfer at the 1 Gbit/s level. Finally, with the parallelization of processors in cheap PC clusters (processor farms), adequate processing power is now available.

In the future, data acquisition system will no longer be a limiting factor in medical imaging.

20.3.3.5 Simulation Software

Monte Carlo simulation methods are an essential tool for developing new detectors in medical imaging.

Versatile generic simulation tools have been developed for particles physics, for instance Geant4 and FLUKA at CERN and INFN, EGS4 at SLAC or MCNP at Los Alamos National Laboratory. More recently, the development of Geant4

has made it possible to include efficient geometrical modelling and visualisation tools. GATE [48] was developed on this basis to simulate PET and SPECT imaging devices; it is a simulation platform written specifically to model imaging systems, through which time-dependent phenomena—detector motion, decrease in isotope radioactivity, dead time phenomena—can be followed. This new simulation tool, developed, validated and documented by the OpenGATE collaboration regrouping about 20 laboratories of medical and particles physics, is freely available on the Internet and is currently used by a community of more than 200 scientists and industry worldwide.

20.3.4 Image Reconstruction Algorithms

The data provided by transmission and emission tomography make it possible to reconstruct projections of the image, which are then combined to images thanks to tomographic reconstruction methods [49]. There are two possible approaches to deal with tomographic reconstruction problems. The first one is analytical and consists in treating the measured projections as if they were perfect mathematical projections. In this case, it is necessary to make a number of hypotheses on linearity and continuity, but the data, often incomplete and noisy, will not always fulfil these requirements and give rise to artefacts. The second approach is phenomenological. It consists in modelling the measuring process using a probability matrix, which has to be inverted through iterative algebraic techniques. Such algorithms iterate a process aiming at optimising an objective function, for instance the verisimilitude function coming from the Poissonian nature of the data recorded by the tomograph. This is the case for instance of the maximum likelihood expectation maximization (MLEM) algorithm and its variant using ordered subsets expectation maximization (OSEM). These iteration methods are less demanding on the geometry of the detector, and do not require a complete set of projections. On the other hand, they require high calculation power. Fortunately, in the near future the Grid or Cloud will provide considerable and massively distributed computing resources. The quality of the results of an iterative process are more difficult to evaluate than the results of an analytical one. Monte Carlo simulation is often a key element to study and optimise these algorithms.

20.4 Multimodality

20.4.1 Need for a Multimodal Approach

SPECT or PET scanners allow localizing radiotracers uptakes in the human body and are, as such, very powerful tools for basic research in cognitive sciences, for

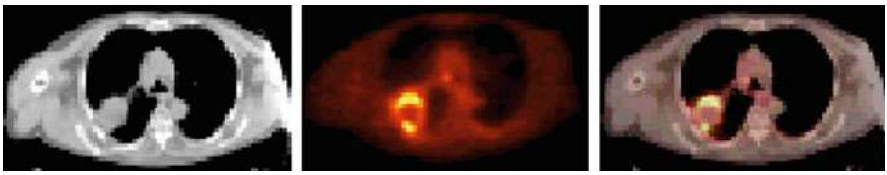

Fig. 20.27 Primary lung cancer imaged with a PET/CT scanner. A large lung tumour, which appears on CT as a uniformly attenuating hypo dense mass (left), has a rim of FDG activity and a necrotic centre revealed by PET (middle). The combined image (right) allows a precise localization of the active parts of the tumour (Courtesy Dave Townsend, University of Tennessee)

clinical oncology and cardiology and for kinetic pharmaceutical studies. However, they do not deliver precise anatomical images, like MRI or X-ray CT for instance. Whilst effective software image fusion techniques exist there is a great deal of interest in performing functional and anatomical studies as simultaneously as possible in order to improve registration accuracy and to resolve the logistical problems associated with software registration.

Modern scanners combine the very high sensitivity of PET for metabolic imaging with the high spatial resolution anatomic information delivered by X-ray CT or another anatomical modality. Indeed, the majority of PET and SPECT systems currently being installed now incorporate a CT scanner in the same gantry, so that functional and anatomical images can be performed in rapid succession. Features identified with PET or SPECT can then be accurately localized via the CT scan. The CT data can also be used to determine the photon attenuation correction, an advantage, in particular for overweight patients. These machines provide impressive images giving the very precise localization of the metabolic activity of organs and tumours (Fig. 20.27).

The development of bimodal acquisition systems, for metabolic, functional and anatomical data, like PET/CT combined scanners [27]—several thousand machines exist as of today—, is radically modifying the patients care thanks to increased precision in diagnosis. This approach also decreases significantly the number of imaging scans for a patient.

Combined PET/CT imaging brings additional benefits in the planning of radiotherapy, which is a promising area for research and clinical applications. The principle of radiation therapies is to modulate the intensity according to the spatial distribution of the area to be treated (Intensity-Modulated Radiation Therapy, IMRT). The combination of PET images, which provide information on the metabolic extension and heterogeneity of tumoural tissues, and CT images, which provide precise structural information and location of the tumour, helps defining an irradiation map to focus the therapy on the particularly active areas of the tumour.

Multimodality has become an intensive research area, the challenge being to take the best advantage in the combination of anatomic and functional information by optimizing the choice of the imaging modalities as a function of the application domain. The progress in SiPM technology has opened the way for a vibrant

field of development for PET/MRI systems, in particular for brain studies, taking advantage of the high functional sensitivity of the PET and of the very good soft tissue contrast capability of the MRI. Similarly, coupling optical fluorescence methods and PET or SPECT is very attractive for studying biologic processes on small animals. In another domain a Cerimed [50] collaboration has designed a PET/SPECT/Ultrasound dedicated breast imaging camera allowing to simultaneously access a variety of parameters on breast tumours such as energetic metabolism of cells, response to specific hormonal ligands (herceptin, bombesin), echogenecity, elastometry, Doppler, and to correlate them in order to optimize the treatment plan of the patients.

Finally, the combination of PET and SPECT imaging associated to the labeling of various ligands signaling different molecular pathways opens the way to multifunctional imaging. Such an approach could prove useful for identifying the molecular profiling of a tumour in a single exam. It would also allow the simultaneous recording of the expression of several neurotransmitters under specific stimuli, a very powerful approach in cognitive sciences.

20.4.2 Outlook: Towards Integrated Morphologic and Functional Imaging

Biological systems are so complex that there is an important need to develop imaging modalities capable of simultaneously recording different molecular pathways in a quantitative and dynamic manner. Helping to address this issue, multi-parametric molecular imaging involves combining the excellent sensitivity and specificity of molecular imaging (PET or SPECT) with a complementary high-spatial resolution imaging modality (CT, MRI or ultrasound).

The most frequently used equipment combines a PET scanner and an X-ray CT. At present, these combined systems are large, consisting of independent scanners mounted in-line in a common gantry, not generally mounted on the same rotary holder. This results in some imprecision in the image fusion process due to external and internal movements of the patient. CT data provide crucial information for the correction of the unavoidable attenuation factors from the patient's body in PET images and for improving image quality by decreasing the influence of artefacts. Partial volume effects are caused by PET's limited spatial resolution, which dilutes information from small hot spots onto a larger area because of the blurring of the image. CT information, which provides much more precise anatomical information, helps to correct, at least partially, these negative effects. This correction is difficult if both imaging systems acquire data in distinct, poorly correlated spaces. In the case of attenuation and partial volume correction, it is crucial to record both data sets as simultaneously as possible so as to guarantee good image superposition. A major challenge is to further integrate the readout of X-rays and γ-rays. Simultaneous recording of anatomical (CT) and functional (PET and/or SPECT) information by

the same reading head is in principle possible thanks to progress in microelectronics. The large functionality of modern ASIC's makes it possible to develop electronic readout channels able to count each individual event, well suited for CT, PET and SPECT signal treatment. This is a particularly interesting perspective, because it would make it possible to correct attenuation and partial volume parameters more precisely.

Another way of obtaining images associating anatomical and functional information is to merge MRI and PET images. Again both data sets have to be acquired as simultaneously as possible, even if a universal acquisition system cannot be considered here because of the fundamental differences between these two modalities. The PET/MRI approach is particularly promising for brain studies. Indeed, MRI gives much better images than CT for the soft brain tissues behind the skull.

Besides, BOLD contrast MRI, which relies on the variation of blood oxygenation level, is promising as a tracer of neuronal activity in functional MRI imaging. Combining this approach with PET functional imaging using various ligands (dopamine, serotonine, acetylcholine, glutamates, opiates, etc.) opens the way to a better understanding of fundamental neurotransmission mechanisms in the brain.

However, a number of significant technological problems arise from recording almost simultaneously MRI and PET images; these problems are mainly caused by the presence of powerful magnetic fields in MRI, with a high homogeneity requirement. To combine the two systems, innovative technologies are needed. The development of SiPM matrices has solved the problem of photodetectors having to be immune to the magnetic field. Moreover, they are extremely compact and require an operating voltage of only a few hundred volts. However, a number of other constraints remain: in a PET/MRI system conducting or ferromagnetic materials must also be carefully avoided because they would alter the homogeneity of the MRI magnetic field. Other technical difficulties, which have to be solved are linked to gradient coils and to MRI's radiofrequency fields, which require effective shielding for PET parts against Eddie currents and electromagnetic noise.

Thanks to multiparametric molecular imaging, a radical shift is currently taking place in the way diseases are managed: from the present onefitsall approach to one that delivers medical care tailored to the needs of individual patients. This includes the detection of disease predisposition, early diagnosis, prognosis assessment, measurement of drug efficacy and disease monitoring. Thus, the introduction of personalized medicine requires an unprecedented effort to develop new technologies in fields of diagnostic and image-guided therapeutic medicine (theranostics) including pathology and imaging. Such imaging tools should characterize diseases and assess treatment efficacy, with the added advantage of non-invasive monitoring at multiple time points. The recent explosion of molecular biology and imaging technologies is now allowing simultaneous quantitative and dynamic characterization of several biological processes inside the body at the molecular and genetic level. This exciting new field will transform the future of medicine on a massive scale and will have an enormous impact on the advancement of targeted therapies for personalized medicine.

References

1. C.C. Ling, J. L. Humm, S. M. Larson, H. Amols, Z. Fuks, S. Leibel, J. A. Koutcher, Towards multidimensional radiotherapy (MDCRT): biological imaging and biological conformality. *Int. J. Radiat. Oncol. Biol. Phys.* 2000; 47:551–560.
2. R. Orecchia, A. Zurlo, A. Loasses, M. Krengli, G. Tosi, P. Zucali, S. Zurrida, and U. Veronesi, *Particle Beam Therapy (Hadrontherapy): Basis for Interest and Clinical Experience*, European Journal of Cancer 34: 456 –468 (1998).
3. W. Enghardt, P. Crespo, F. Fiedler, R. Hinz, K. Parodi, J. Pawelke, F. P.onisch, Charged hadron tumour therapy monitoring by means of PET, NIMA 525, 284-288 (2004).
4. Deych D, Dobbs J, Marcovici S, Tuval B (1996), Cadmium tungstate detector for computed temography. In: Dorenbos P, van Eijk CWE (eds). Inorganic scintillators and their application. Delft University Press, pp 36–39.
5. C. Greskovich, S. Duclos, Annu. Rev. Mater. Sci., 27(1997)69.
6. Kostler W, Winnacker A, Rossner W, Grabmaier BC (1993), Effect of Pr-codoping on the X-ray induced afterglow of $(Y,Gd)_2O_3$:Eu, J Phys Chem Solids 56: 907–913.
7. C. Greskovich, D. Cusano, D. Hoffman, R. Riedner, American Ceramic Society Bull., 71(1992)1120.
8. E. Gorokhova, V. Demidenko, O. Khristich, S. Mikhrin, P. Rodnyi, J. Opt. Technology, 70(2003)693.
9. Y. Ji, J. Shi, J. Mater. Res., 20(2005)567.
10. S. Sekine, T. Yanada, U.S. Patent No. 6,876,086 B2, 2005.
11. R. Luhta, R. Mattson, N. Taneja, P. Bui, R. Vasbo, in: Medical Imaging 2003: Physics of Medical Imaging, Proc. Of SPIE 5030(2003)235.
12. A. Goushcha, A. Popp, E. Bartley, R. Metzler, C. Hicks, in: Medical Imaging 2004: Physics of Medical Imaging, Proc. Of SPIE 5368(2004)586.
13. F. Natterer, The Mathematics of Computerized Tomography, New York: Wiley, 1986.
14. A.C. Kak, M. Slaney, Principles of Computerized Tomographic Imaging, New York: IEEE Press, 1988.
15. Guang-Hong Chen et al., A novel extension of the parallel-beam projection-slice theorem to divergent fan-beam and cone-beam projections, Med. Phys. 32 (3), March 2005.
16. Crawford, C. R., King, K. F.: Computed tomography scanning with simultaneous patient translation. Med. Phys. 1990; 17:967-82.
17. T. G. Flohr, K Stierstorfer, S. Ulzheimer, H. Bruder, A. N. Primak, C. H. McCollough, Med. Phys., 32 (2005) 2536.
18. N.J. Pelc, Recent and future directions in CT imaging, Ann. Biomed Eng. 2014 Feb: 42(2): 260-268.
19. D. Fornel, Technology Improvements in Current Generation CT systems, ITN September 2015, https://www.itnonline.com/article/technology-improvements-current-generation-ct-systems.
20. Wieczorek H., Frings G., Quadfield P., et al.(1995) CsI:Tl for solid state X-ray detectors, Proc. In Dorenbos P, van Eijk CWE (eds). Inorganic Scintillators and Their Applications, Delft University Press, 547-554.
21. X. Llopart et al., Timepix, a 65k programmable pixel readout chip for arrival time, energy and/or photon counting measurements, Nucl. Intr. And Meth. A (2007), doi:https://doi.org/10.1016/j.nima.2007.06.054.
22. M. Titov, New developments and future perspectives of gaseous detectors, Nucl. Intr. And Meth. A (2007), doi:https://doi.org/10.1016/j.nima.2007.07.022.
23. R.K. Swank, Absorption and noise in X-ray phosphors, J. Appl. Phys. 44, 4199-4203 (1973).
24. M.G. Bisogni et al., NIMA 546, 14 (2005).
25. P. Delpierre, A history of hybrid pixel detectors, from high energy physics to medical imaging JINST 9 C05059, 2014.
26. F. Cassol et al., Characterization of the imaging performance of a micro-CT system based on the photon counting XPAD3/Si hybrid pixel detectors, Biomed. Phys. Eng. Express 2 (2016) 025003.

27. H. Schöder et al., PET/CT: a new imaging technology in nuclear medicine, Eur J Nucl Med Mol Imaging (2003) 30:1419-1437.
28. T. Jones, Molecular imaging with PET – the future challenges, The British Journal of Radiology, 75 (2002), S6–S15.
29. European Society of Radiology (ESR), Medical imaging in personalised medicine: a white paper of he research committee of the European Society of Radiology (ESR), Insight Imaging, Apr. 2015, 6(2): 141-155.
30. T. Jones, D. Townsend, History and future technical innovation in positron emission tomography, J. Med. Imag. 4(1), 011013 (2017).
31. P. Sempere Roldan et al., Performance Evaluation of Raytest ClearPET®, a PET Scanner for Small and Medium Size Animals, Conference records of the IEEE NSS/MIC conference, Oct. 27-Nov. 3, 2007, Hawaï, 2859-2864.
32. M. C. Abreu et al., Clear-PEM: A PET imaging system dedicated to breast cancer diagnostics, NIMA 571, 81-84 (2007).
33. P. Lecoq, Pushing the limits in Time-of-Flight PET imaging, IEEE Transactions on Radiation and Plasma Medical Sciences, Vol. 1, N°6, November 2017.
34. S. C. Strother, M. E. Casey, and E. J. Hoffman, "Measuring PET scanner sensitivity: Relating count rates to image signal-to-noise ratios using noise equivalent counts," *IEEE Trans. Nucl. Sci.*, vol. NS-37, pp.783-788, Apr. 1990.
35. C.L. Melcher et al., A promising new scintillator: cerium doped lutetium oxyorthosilicate, Nuclear Instruments and Methods in Physics Research A 314 (1992) 212-214.
36. Crystal Clear Collaboration, RD18, CERN/DRDC/P27/91-15.
37. C. Kuntner et al., Advances in the scintillation performance of LuYAP:Ce single crystals, Proceedings of the 7th conference on Inorganic Scintillators and their Use in Scientific and Industrial Applications, Valencia, Spain, Sept 2003, Nuclear Instruments and Methods in Physics Research A 537 (2005) 295-301.
38. V.D. van Loef et al., Scintillation properties of $K_2LaX_5:Ce^{3+}$ (X=Cl, Br, I), Proceedings of the 7th conference on Inorganic Scintillators and their Use in Scientific and Industrial Applications, Valencia, Spain, Sept 2003, Nuclear Instruments and Methods in Physics Research A 537 (2005) 232-236.
39. JC Clemens et al., PIXSCAN: CT-Scanner for Small Animal Imaging Based on Hybrid Pixel Detectors. To be published in conf rec 7th International Workshop on Radiation Imaging Detectors, IWORID-7, July 4-7, 2005 in Grenoble, France.
40. J. C. Bourgoin, A new GaAs material for X-ray imaging,. Nuclear Instruments and Methods in Physics Research A460 (2001), 159-164.
41. A. Owens et al., The X-ray response of CdZnTe, Nuclear Instruments and Methods in Physics Research A484 (2002), 242-250.
42. R.M. Turtos et al., "Timing performance of ZnO:Ga nanopowder composite scintillators", Phys. Status Solidi RRL 10, No. 11, 843–847 (2016).
43. R.M. Turtos et al., "Ultrafast emission from colloïdal nanocrystals under pulsed X-ray excitation", *JINST_068P_06*.
44. P. Lecoq, "Metamaterials for novel X- or γ -ray detector designs," in Proc. IEEE Nucl. Sci. Symp. Conf. Rec., Dresden, Germany, 2008, pp. 680–684.
45. I. Britvitch et al., Development of scintillation detectors based on avalanche microchannel photodiodes, Proceedings of the 1st international conference on Molecular Imaging Technology, Marseilles, France, May 9-12, 2006, Nuclear Instruments and Methods in Physics Research A 571 (2007) 317-320.
46. S. Gundacker et al., "State of the art timing in TOF-PET detectors with LuAG, GAGG and L(Y)SO scintillators of various sizes coupled to FBK-SiPMs", *2016 JINST 11 P08008.*
47. E.H.M. Heijne, Future semiconductor detectors using advanced microelectronics with post-processing, hybridization and packaging technology, Nuclear Instruments and Methods in Physics Research A 541 (2005) 274-285.

48. D. Strul et al., “GATE (Geant4 Application for Tomographic Emission): a PET/SPECT general-purpose simulation platform,” Nucl. Phys. B (Proc. Suppl.) 125C (2003) 75-79. (http://www.opengatecollaboration.org)
49. B. Bendriem and D.W. Townsend, The theory and Practice of 3D PET, Kluwer Academic Publishers, 1998, ISBN 0-7923-5108-8.
50. Cerimed, European Centre for Research in Medical Imaging, based in Marseille, France. http://www.cern.ch/cerimed/.

Chapter 21
Solid State Detectors for High Radiation Environments

Gregor Kramberger

21.1 Introduction

The solid state particle detectors emerged in 1950 [1]. Initially Si and Ge detectors operated as junction diodes were used for charged particle detection and γ spectroscopy measurements (Chap. 5). Although these detectors are superior to gaseous detectors in many respects, being a crystalline medium meant that they are susceptible to radiation damage. Unlike in gaseous detectors where the detection media can be exchanged the semiconductor crystals have to retain their detection properties over the entire envisaged period of operation. The particle detection capabilities and the energy resolution degrade gradually with irradiation, which limits their lifetime.

A large majority of present high energy experiments uses position sensitive silicon detectors which became widely available after the introduction of planar process in 1980 [2]. Their goal is achieving desired position resolution with as few read out channels as possible, while keeping detection efficiency close to 100%. At the present and particularly future experiments high particle rates close to the interaction point require very fine segmentation and high position resolution of detectors in order to be able to associate hits with tracks.

In the future a precise timing information associated with a track and even with each sensor hit may be required to cope with large multiplicity of tracks. The sensor hits and associated tracks will be therefore separated not only spatially, but also in time allowing easier assignation of tracks to different collisions occurring within each colliding particles bunch crossing.

G. Kramberger (✉)
Experimental Particle Physics, Jožef Stefan Institute, Ljubljana, Slovenia
e-mail: gregor.kramberger@ijs.si

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_21

High particle rates cause radiation effects. The most important is the damage of the crystal lattice which leads to the degradation of the measured charge after passage of ionizing radiation. At the same time the noise may increase for various reasons thus significantly reducing the signal-to-noise ratio. Consequently the detection efficiency, energy, and position resolution may degrade to the level where the detectors become unusable. Extensive research was made in the last decades to understand the damage in silicon detectors and to manipulate the properties of silicon aiming at radiation-harder detectors. The research was not only limited to silicon but alternative semiconductor materials were considered.

It is not only the bulk crystal that is affected by irradiation, but also the surface. The radiation effects at the silicon—silicon oxide interface not only change the performance of silicon detectors, but are the main reason for radiation damage of electronics. The latter was often considered a bigger problem than the radiation damage of detectors, particularly in environments where the ionization dose was large (e.g. synchrotron radiation). With the advent of deep sub-micron CMOS processes, electronics was thought to became intrinsically radiation hard and no special radiation hardening processes would be required. An important contribution to the radiation hard electronics was also introduction of radiation-tolerate design rules. However, for very small feature sizes, e.g. very deep sub-micron processes, such as 0.130, 0.065 μm, radiation hardness of electronics, rather than sensors, could become a limiting factor at harshest radiation environments.

On the other hand the effects of radiation damage were exploited for dose measurements. Active dosimeters appeared for both measurements of ionizing and non-ionizing energy losses in silicon crystal such as $p-i-n$ diodes [3] and radiation sensitive field effect transistors [4].

21.2 High Radiation Environments

The radiation environments differ in composition and energies of the particles producing the radiation damage. Although the particles that are to be detected contribute largely to the damage it is often the background particles that dominate. As will be described later the damage depends on the type of the particle. While X-rays alter the properties of the detector surface they can not displace the semiconductor atoms from the lattice. On the other hand neutron irradiation affects only the lattice and energetic charged hadrons and leptons damage both the lattice and the surface. The difference in damage creation and its effects to detector operation will be discussed later. First we review the radiation environments where particle detectors are employed.

Collider Experiments In general there are three major types of accelerators with respect to collision particles: hadron ($p-p$, $\bar{p}-p$, heavy ions), lepton (e^+-e^-, eventually $\mu^+-\mu^-$) and lepton-hadron ($e^{+,-}-p$). The flux of particles traversing

the detectors is given by the particles originating from the collisions (ϕ_{coll}) and secondary radiation that originates from the spectrometer or the accelerator (ϕ_{sec})

$$\phi = \phi_{coll} + \phi_{sec}, \tag{21.1}$$

The flux of particles crossing the detectors is much larger at hadron colliders than at lepton colliders, owing to a difference in total cross-section σ_{tot} of colliding particles. The radiation environment at lepton colliders is dominated by $e^{\pm}$ from Bhaba scattering. Consequently the radiation damage of detectors at hadron colliders is much more severe than at lepton colliders.

A significant secondary irradiation, particularly at high luminosity colliders, can arise from back-scattered neutrons originating in breakup of nuclei in calorimeters and other parts of spectrometers after interaction with highly energetic hadrons. The secondary radiation originating from the accelerator such as synchrotron radiation, beam-gas interactions or halo particles scraping the collimators should be small but can represent in case of an accident a significant contribution to the total fluence Φ (integral of flux $\Phi = \int \phi \, dt$) of particles traversing the detectors.

The required radiation tolerance/hardness of vertex detectors at different colliders is given in the Table 21.1. Placing of the detectors in the spectrometer determines their exposure. The ϕ_{coll} decreases quadratically with the distance from the interaction point. The large cross-section for soft collisions result in larger ϕ_{coll} at small angles with respect to beam. Large ϕ_{coll} at small angles is also characteristic for asymmetrical beams (energy, particle) or fixed target experiments. A particle fluence profile for ATLAS experiment [5] at the Large Hadron Collider (LHC) is shown in Fig. 21.1. The dominating particles are at small radii mainly pions and protons originating from collisions and "albedo" neutrons from the calorimeters for $R > 20\,\text{cm}$.

Table 21.1 The review of basic parameters of some accelerators and required radiation hardness of the most exposed detectors for the entire operation period

Accelerator	Type	σ_{tot} [barn]	$\mathcal{L}$ [$\text{cm}^{-2}\,\text{s}^{-1}$]	$\sim \int \phi dt$ [$\text{n}_{eq}\,\text{cm}^{-2}$]	Dose in Si [Gy]
Super KEK-B	e^+-e^- (8,3.5 GeV)	4n	$5.0 \cdot 10^{35}$	$<2 \cdot 10^{12}\,\text{cm}^{-2}$	<10 k
ILC	e^+-e^- (250,250 GeV)	3p 3p	few 10^{34}	$\sim 10^{10}$	few k
HERA	$e^{+,-}$-p (27.5, 920 GeV)	10^{-3} ($Q^2 < 100\,\text{GeV}$)	$7 \cdot 10^{31}$	$<10^{13}$	<2 k
Tevatron	$\bar{p}$-p (0.98,0.98 TeV)	70 m	$1.7 \cdot 10^{32}$	$<10^{13}$	<30 k
LHC HL-LHC (>2026)	p-p (7,7 TeV)	100 m	$10^{33} - 10^{34}$ $5 - 7.5 \cdot 10^{34}$	up to $5 \cdot 10^{15}$ up to $2 \cdot 10^{16}$	~2.5 M ~10 M
FCC	p-p (50, 50 TeV) foreseen >2040	100 m	$5 - 30 \cdot 10^{34}$	up to $6 \cdot 10^{17}$	~400 M

The total cross-section without Bhaba scattering is given for $e^+ - e^-$ accelerators

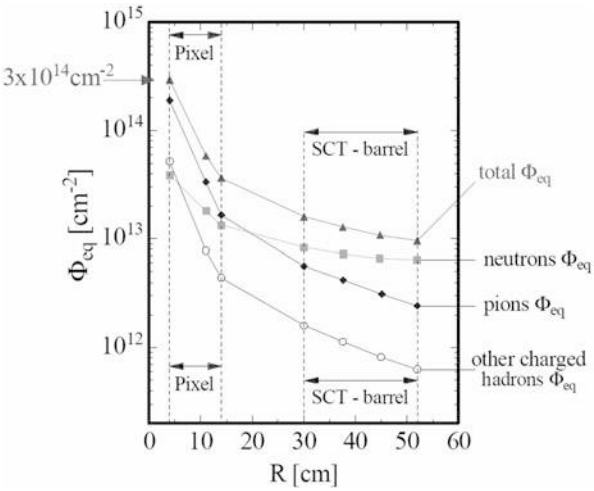

Fig. 21.1 Yearly fluence profile in ATLAS experiment at LHC design luminosity. The radiation damage caused by different particles was used to normalize the fluences (see next section for Φ_{eq}). The arrows denote the location of the pixel and strip detectors (SCT)

The choice of detector technology at a given radius depends on the ability to retain the detection efficiency and position resolution at required levels. At the same time the material budget should be kept low in order not to spoil the tracking performance. At many experiments the most exposed detectors are beam position/condition and radiation monitors (Chap. 18).

Space Applications

Particle detectors are an important constituent of many space missions. They are mainly used as spectrometers, visible light detectors and charged particle trackers. The radiation fields are far less severe than that at accelerators experiments, but the detectors and the information that they provide can be far more susceptible to the radiation effects (e.g. CCD, DEPFet, Si-drift detectors). The origin of radiation in space comes from three sources:

- **Galactic cosmic radiation;** Consists primarily of nuclei (85% protons, 14% Helium, 1% heavier ions among which Fe and C are most abundant ones). The relevant particles for damage creation have energies between 1–20 GeV. The fluxes of cosmic particles are shown in Fig 21.2a. The flux depends on the activity of the sun through interaction with solar wind (a continuous stream of high ionized plasma emerging from the sun). Interactions of highly energetic particles with nuclei in the earth's atmosphere or space-vessel produce showers of ionizing particles which increase the intensity of the radiation.
- **Solar particles;** The sun is also a sporadic source of lower energy charged particles (solar particles) accelerated during certain solar flares and/or in the

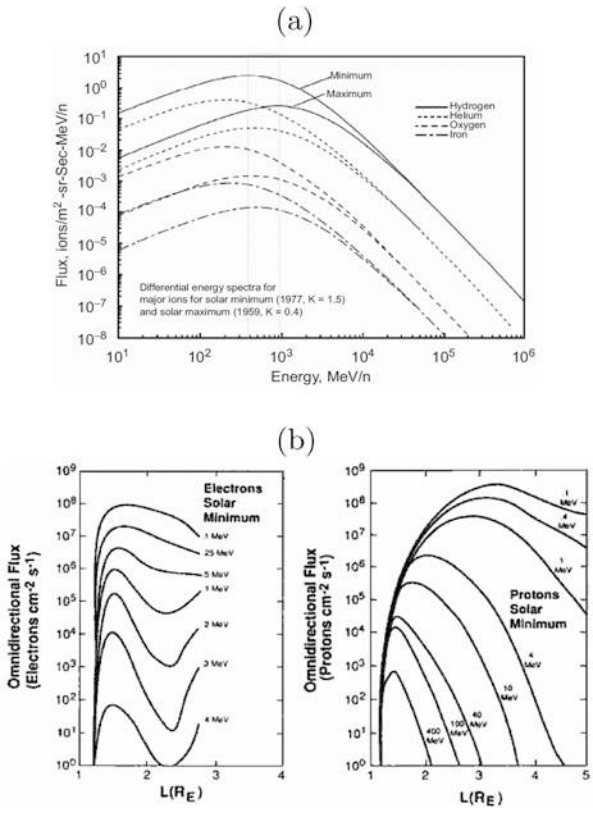

Fig. 21.2 (**a**) Galactic cosmic ray particle spectra and their modification by solar activity [6]. (**b**) Equatorial electron and proton flux vs. the distance from the Earth's center. Each curve gives the total flux above the specified threshold [7]

subsequent coronal mass ejections. These solar particles comprise both protons and heavier ions with variable composition from event to event. Energies typically range up to several hundred MeV and occasional events produce particles of several GeV. Although such events are rare, typically one per month and lasting several hours to days, the flux integrals as large as $10^{10}\,\mathrm{cm}^{-2}$ for protons with energy >1 MeV were measured.

- **Radiation belts;** The charged particles trapped in the Earth's magnetic field form so called Van Allen's belts. The inner belt extends to 2.5 Earth radii and comprises protons up to 600 MeV and electrons up to several MeV. The outer belt extends to 10 Earth radii where there are mainly electrons and soft protons (0.1–100 MeV). The fluxes of electrons and protons trapped in the radiation belts are shown in Fig. 21.2b. The sharp fall of flux at high energies makes shielding very effective.

Environmental Applications

- **Medical application;** The most widely used source of radiation are X-rays, Linacs and radio active isotopes used for cancer treatment. The energy of photons used is: up to 100 keV for X-rays, below 1 MeV for isotopes and up to 25 MeV for Linacs.
- **Fusion in fission reactors and nuclear waste managements;** The main damage comes from neutrons and γ rays, both with energies up to few MeV. Fusion reactors of TOKAMAK type require plasma, fuel impurity and fusion products monitoring instrumentation close to the first wall. The foreseen neutron fluences to which the sensors (e.g. silicon sensors for X-ray spectroscopy) and electronics will be exposed at International Thermonuclear Experimental Reactor (ITER) are comparable with that of the HL-LHC, up to few $10^{16}\,\mathrm{cm}^{-2}$.

21.3 Damage Mechanism in Solid State Detectors and Electronics

As radiation (photons, leptons, hadrons) passes through material, it loses energy by interaction with the electrons and nuclei of the material atoms. The effects produced in the material are dependent on the energy-loss processes and the details of the material structure. The damage in semiconductor detectors can be divided into bulk and surface damage.

21.3.1 Bulk Damage

The interaction with electrons results in creation of electron-hole pairs (ionizing energy loss, ionizing dose) that does not affect the lattice and causes no bulk damage. The bulk damage in crystalline and poly-crystalline material is a consequence of displacement of lattice atoms by impinging particles, due to elastic scattering on a nuclei and nuclear reactions. In order to produce Primary Knocked off Atom (PKA) the transfer of kinetic energy should be sufficient. Approximately 25 eV of recoil energy is required for example in silicon. The displaced atom may come to rest in a interstitial position (I), leaving a vacancy (V) at its original location. If the kinetic energy of the recoiling atom is sufficient ($\sim$5 keV in Si [8]) it can displace further atoms, creating a dense agglomeration of defects at the end of the primary PKA track. Such disordered regions are referred to as defect clusters.

Most of the resulting vacancies and interstitials recombine while others diffuse away and eventually create stable defects with impurity atoms and other vacancies or interstitials. Those defects disturb the lattice periodicity and give rise to energy levels in the band-gap, which alter the properties of the semiconductor. In most semiconductor materials the cross-section for nuclear reaction is much smaller than

Table 21.2 Material properties of some semiconductors used as ionizing particle detectors

Property	Si	Diamond	GaAs	GaN	4H-SiC	a-Si(H)
Z	14	6	31/33	31/7	14/6	14
E_g [eV]	1.12	5.5	1.4	3.39	3.3	1.7
E_{bd} [MV/cm]	0.5	10			2.2–4	
μ_e [cm^2/Vs]	1350	∼2000	≤8500	1000	800–1000	1–10
μ_h [cm^2/Vs]	450	∼1400	≤400	30	30–115	0.01–0.005
$v_{sat,e}$ [cm/s]	$2 \cdot 10^7$	$2.7 \cdot 10^7$	$1.2 \cdot 10^7$		$2 \cdot 10^7$	
ϵ	11.9	5.5	0.4		9.7	
e–h energy [eV]	3.6	13	4.3	8.9	7.8	4–4.8
e-h/μm for m.i.p.	90	36			51	75
Density [g/cm^3]	2.3	3.5	5.3	6.2	3.2	2.3
Displacement [eV]	25	43	10	Ga-20 N-10		

for elastic scattering, hence the creation rate of defects, resulting from nuclear reactions, is usually more than two orders of magnitude lower when compared to creation rates of defects originating from displaced silicon atoms.

The energy E_p required for an incoming particle of mass m_p to produce PKAs and clusters with a creation threshold E_{th} can be calculated from non-relativistic collision kinematics as

$$E_p = E_{th} \frac{(m_p + m_l)^2}{4\, m_p\, m_l}, \tag{21.2}$$

where the lattice atom has a mass m_l. In silicon a neutron needs at least 175 eV to produce a PKA and 35 keV to form a cluster. For an electron a relativistic kinematics should be used giving 260 keV and 8 MeV. It should be noted that the radiation damage caused by γ-rays from radioactive decays is primarily due to the interaction of Compton electrons with a maximum energy well below the one required for cluster production. The bulk damage is therefore exclusively due to point defects. As the thresholds are of the same order also in other semiconductor materials (see Table 21.2) similar conclusions are valid.

A part of vacancies and interstitials formed immediately after irradiation can recombine, while others diffuse away and eventually recombine or react with other defects or impurities. The defects can evolve in time. They can either dissociate or react with each other and form new defects. The evolution of defects is described by first order dynamics in case of dissociation (Eq. (21.3)) or second order dynamics for reactions of two defects (Eq. (21.4)):

$$X \rightarrow Y \quad , \quad \frac{dN_Y}{dt} = -\frac{dN_X}{dt} = k_1^Y N_Y \tag{21.3}$$

$$X_A + X_B \rightarrow Y \quad , \quad \frac{dN_Y}{dt} = -\frac{dN_{X_A}}{dt} = -\frac{dN_{X_B}}{dt} = k_2^Y N_{X_A} N_{X_B}, \tag{21.4}$$

where $k_{1,2}^Y$ denotes the reaction constants. The Eq. (21.4) turns into a first order process in cases when one type of the reacting defects is present in much larger quantities than the other. The solution of Eq. (21.3) is exponential with

$$N_Y = N_X^0\left(1-\exp\left(-\frac{t}{\tau_1^Y}\right)\right) \quad , \quad N_X = N_X^0 \exp\left(-\frac{t}{\tau_1^Y}\right), \quad \tau_1^Y = \frac{1}{k_1^Y} \tag{21.5}$$

with N_X^0 denoting the initial concentration of defects X proportional to the fluence. The solution of the Eq. (21.4) for $(N_{X_A}^0 > N_{X_B}^0)$ is given by

$$N_Y(t) = N_{X_B}^0 \frac{1 - e^{-k_2^Y t (N_{X_A}^0 - N_{X_B}^0)}}{1 - (N_{X_B}^0/N_{X_A}^0) e^{-k_2^Y t (N_{X_A}^0 - N_{X_B}^0)}} \, . \tag{21.6}$$

In the case of two defects with similar initial concentrations $N_{X_A}^0 = N_{X_B}^0 = N_X^0$ or a reaction between defects of the same type one obtains

$$N_X(t) = \frac{N_X^0}{1 + N_X^0 k_2^Y t} = \frac{N_X^0}{1 + t/\tau_2^Y} \quad , \tau_2^Y = \frac{1}{k_2^Y \, N_X^0} \tag{21.7}$$

$$N_Y(t) = N_X^0 - N_X(t) = N_X^0 (1 - \frac{1}{1 + t/\tau_2^Y}). \tag{21.8}$$

From Eqs. (21.3) and (21.4) it can be seen that for first order reactions, the rate depends linearly on defect concentration while for second order reactions the dependence is quadratic.

Since the energy needed for breaking up the defect (dissociation) or forming a new defect is supplied by the lattice vibrations, the reaction constant is strongly temperature dependent. The lattice atom vibration energy is governed by the Maxwell-Boltzmann distribution. The probability of sufficient energy transfer from lattice vibration to the defect is therefore exponential with temperature (T). If the reaction rate given by the Arrhenius relation is known at T_0 then the rate at T_1 is calculated as:

$$k_{1,2}^Y \propto \exp\left(-\frac{E_a}{k_B \, T}\right) \quad \Longrightarrow \quad \frac{\tau_{1,2}^Y(T_0)}{\tau_{1,2}^Y(T_1)} = \frac{k_{1,2}^Y(T_1)}{k_{1,2}^Y(T_0)} = \exp\left[\frac{E_a}{k_B}\left(\frac{1}{T_0} - \frac{1}{T_1}\right)\right], \tag{21.9}$$

where E_a is the energy required for defect dissociation or formation.

21.3.1.1 Non-Ionizing-Energy-Loss Hypothesis of Damage Effects

The energy loss of impinging particles suffered in a process of displacing lattice atoms is called non-ionizing energy loss—NIEL. First experimental findings have led to the assumption that damage effects produced in the semiconductor bulk by energetic particles may be described as being proportional to non-ionizing energy loss, which is referred to as the NIEL-scaling hypothesis. According to it any displacement damage induced change in the material properties scales with the amount of energy imparted in displacing collisions, irrespective of the spatial distribution of the defects in a PKA cascade and irrespective of the various annealing sequences taking place after the initial damage [10].

The non-ionizing energy deposit in a unit cell of the target nuclei (ρ_{dis}) exposed to the fluence of particles with energy E can be calculated as

$$\rho_{dis} = D(E) \cdot \Phi, \tag{21.10}$$

where $D(E)$ [9] is so-called displacement damage function, sometimes also referred to as damage cross-section. For a spectrum of particles the contributions to the ρ_{dis} for each energy should be summed:

$$\rho_{dis} = \int_0^\infty \frac{d\Phi(E)}{dE} D(E) dE. \tag{21.11}$$

According to NIEL hypothesis ρ_{dis} determines the damage effects. The damage efficiency of any particle spectrum $d\Phi/dE$ can therefore be expressed as that of an equivalent 1 MeV neutron fluence. The equivalent fluence of 1 MeV neutrons Φ_{eq} is calculated as

$$\Phi_{eq} = \kappa \ \Phi = \frac{\rho_{dis}}{D_n(1 \text{ MeV})} \tag{21.12}$$

$$\kappa = \frac{1}{D_n(1 \text{ MeV})} \cdot \frac{\int_0^\infty D(E) \frac{d\Phi}{dE}(E) \, dE}{\int_0^\infty \frac{d\Phi}{dE}(E) \, dE}, \tag{21.13}$$

where κ is so called hardness factor for that particle spectrum and $D_n(1 \text{ MeV})$ the D for 1 MeV neutrons, 95 MeV mb for Si and 10 MeV mb for diamond [12]. The displacement damage cross-section for pions, protons, electrons and neutrons in silicon is shown in Fig. 21.3. The hardness factors for most commonly used irradiation facilities are given in the Table 21.3.

The NIEL hypothesis is violated in silicon for highly energetic charged hadrons. In addition to the hard core nuclear interactions, being dominant for neutrons, charged hadron reactions are also subjected to Coulomb interactions leading to low energy recoils below the threshold for cluster creation. In this case the damage is a mixture of homogeneously distributed point defects and clusters. This distinct

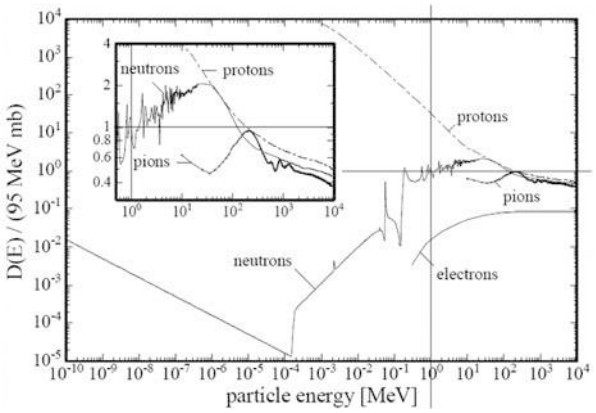

Fig. 21.3 Non Ionizing Energy Loss NIEL for different particles in silicon [11]. The insert shows magnified $D(E)$ for most damaging particles at LHC

Table 21.3 Measured hardness factors of commonly used irradiation particles

	26 MeV[a] protons	70 MeV[b] protons	800 MeV[c] protons	23 GeV[d] protons	200 MeV[e] pions	Reactor [f] neutrons
κ	1.85	1.43	0.71	0.62	1.14	0.92

[a] KIT, Germany and University of Birmingham, UK
[b] CYRIC, Japan
[c] LANL, USA
[d] CERN, Switzerland
[e] PSI, Switzerland
[f] JSI, Slovenia

difference between neutron and proton induced damage is depicted in Fig. 21.4. Different impurities (e.g. O,C) are homogeneously distributed over the volume and the probability for such an impurity to form a defect complex with vacancy or interstitial is much larger if the latter are also homogeneously distributed. Hence, the defects formed after irradiation and consequently the lattice properties can be different for various irradiation particles at equal NIEL. It should be emphasized again that the NIEL scaling can only be regarded as a rough approximation as it disregards the specific effects resulting from the energy distribution of the respective recoils.

21.3.1.2 Impact on Bulk Damage on Detector Performance

As already mentioned the defects in the semiconductor lattice give rise to energy levels (states) in the band gap affecting the operation of semiconductor detector mainly in three ways as shown in Fig. 21.5.

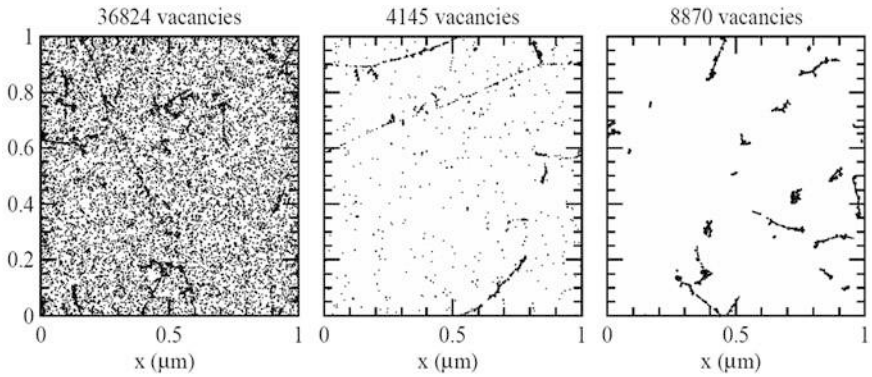

Fig. 21.4 Initial distribution of vacancies produced by 10 MeV protons (left), 23 GeV protons (middle) and 1 MeV neutrons (right). The plots are projections over 1 μm of depth (z) and correspond to a fluence of 10^{14} cm^{-2} [10]

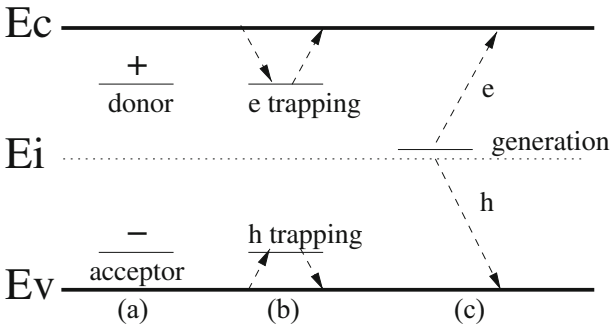

Fig. 21.5 Consequences of deep energy levels to operation of semiconductor detectors: (**a**) charged defects alter the space charge and therefore the electric field, (**b**) defects can trap and detrap free carriers and (**c**) defects act as generation-recombination centers. Electrons and holes are denoted by e and h

- Some of the defects can be charged which leads to (Chap. 5) (Fig. 21.5a) changes in the electric field. For semiconductor detectors this may result in loss of the depleted (active) region requiring an increase of the applied bias. The bias voltage is however limited by the device break down. The space charge is calculated as a difference in concentration of charged donors and charged acceptors,

$$N_{eff} = \sum_{donors} N_t\,(1 - P_t) - \sum_{acceptors} N_t\,P_t, \tag{21.14}$$

where N_t denotes the concentration of deep traps and P_t the probability of a trap being occupied by an electron. The traps continuously emit and capture

carriers. The difference in emission and capture rate is called the excess rate. In a stationary state the occupation probability is constant, therefore excess rates of holes and electrons for a given trap have to be equal. The derivation of occupation probability from this condition can be found in any solid state physics text book. As the P_t is needed for calculation of detector properties we will just state the result:

$$P_t = \left[\frac{c_p\, p + \epsilon_n}{c_n\, n + \epsilon_p} + 1\right]^{-1}, \tag{21.15}$$

$$c_{n,p} = v_{th_{e,h}}\, \sigma_{t_{e,h}}, \epsilon_{n,p} = n_i\, c_{n,p}\, \exp\left(\pm\frac{E_t - E_i}{k_B\, T}\right). \tag{21.16}$$

where $c_{n,p}$ is the capture coefficient and $\epsilon_{n,p}$ emission rate of electrons and holes, respectively. The concentration of free electrons and holes is denoted by n and p and their thermal velocity by $v_{th_{e,h}}$. The capture coefficients and emission rates depend on trap and semiconductor properties. The carrier capture cross-section is given by $\sigma_{t_{e,h}}$ and the level in the band gap by E_t. The Fermi level and free carrier concentration in intrinsic semiconductor are denoted by E_i and n_i. They occupation probability depends on temperature only for levels close to middle of the band-gap. The exponential term in Eq. (21.15) prevails once E_t is few $k_B\, T$ away from the E_i. If follows from here that only donors in the upper part of the band gap and acceptors in lower part of the band gap contribute to the space charge.

- The states can act as trapping centers for the drifting charge generated by the particles we want to detect (Fig. 21.5b). If trapped charges remain trapped and do not complete the drift within the integration time of the read-out electronics they are lost for the measurement, which leads to smaller signal.

  The probability for electrons and holes to be trapped at the trap t can be calculated as

$$\frac{1}{\tau^t_{tr_e}} = c_n\,(1 - P_t)\, N_t, \quad \frac{1}{\tau^t_{tr_h}} = c_p\, P_t\, N_t. \tag{21.17}$$

  The trapping time $\tau^t_{tr_{e,h}}$ represents the mean time that a free carrier spends in the part of the detector before being trapped by t. According to Eq. (21.17) electron traps have energy levels in the upper part of the band gap ($P_t \approx 0$), while hole traps have energy levels in the lower part of the band gap ($P_t \approx 1$).

  To get the effective trapping probability $1/\tau_{eff_{e,h}}$ for electrons and holes one has to sum over the trapping probabilities of all traps with emission times $(1/\epsilon_{n,p})$ longer than integration time of the electronics:

$$\frac{1}{\tau_{eff_e}} = \sum_t^{defects} c_n\,(1 - P_t)\, N_t, \tag{21.18}$$

$$\frac{1}{\tau_{eff_h}} = \sum_{t}^{defects} c_p \, P_t \, N_t. \tag{21.19}$$

The emission times decrease with distance from the mid-gap and become at certain energy level short enough not to be included in the Eq. (21.19). The traps close to the mid-gap have therefore a dominant contribution to the effective trapping times.

- States close to the mid-gap region also act as generation-recombination centers (Fig. 21.5c). The thermally generated electron hole pairs are separated in the electric field before they can recombine, which gives rise to the bulk generation current. The increase of current leads to the increase of noise and power dissipation.

  The generation current can be calculated with the assumption of equal generation rates $G_t = G_n = G_p$ of electrons and holes in thermal equilibrium:

$$G_t = N_t \, P_t \, \epsilon_n = N_t \, \frac{\epsilon_n \, (\epsilon_p + c_n \, n)}{\epsilon_n + \epsilon_p + c_p \, p + c_n \, n} \tag{21.20}$$

$$G_t = N_t \, \frac{1}{1/\epsilon_n + 1/\epsilon_p} \quad \text{for} \quad n, \, p \approx 0. \tag{21.21}$$

  Both carrier types generated in the active volume drift to the opposite electrodes. The current density, albeit different for holes and electrons, is constant everywhere in the detector. The measured current is therefore calculated as

$$I = e_0 \, w \, S \sum_{t}^{defects} G_t \tag{21.22}$$

  where w denotes the active thickness and S the active surface of the detector. It follows from Eq. (21.21) that only the levels close to mid-gap $E_i \sim E_t$ contribute significantly to the current. If traps are far from the mid-gap, emission times are either very long or very short.

Apart from the changes in the depletion region, the properties of the non-depleted silicon bulk are also affected by irradiation. The resistivity of the bulk increases. The increase depends on both initial dopant concentration as well as on irradiation fluence. The minority carrier lifetime also decreases as $1/\tau_r \propto \Phi$ and reaches values of few tens ns at $\Phi = 10^{14}\,\text{cm}^{-2}$ and below ns at $\Phi > 10^{16}\,\text{cm}^{-2}$ [13].

Recent measurements [14] also show that mobility of free carriers is affected by radiation. The concentration of defects, not only electrically active, is high enough to affect the low field mobility. A significant decrease of low field mobility was observed at fluences of $\Phi_{eq} > 5 \cdot 10^{15}\,\text{cm}^{-2}$.

Although silicon detectors are by far the most widely used there are other semiconductor detectors which can be used in high radiation fields and have a higher

PKA displacement energy. The material properties of different semiconductors used as particle detectors are summarized in Table 21.2.

Effects of irradiation on detector performance strongly depend on the choice of material. In wide band gap semiconductors for example the rate of thermally generated carriers will be small even if states close to mid-gap are present in abundant concentrations due to small intrinsic carrier concentration. Thus the leakage current increase is negligible. If the drift velocity is large and charge collection time is short then the increase of trapping probability will be less important. The small dielectric constant reduces the capacitance of a detector leading to lower noise, which can partially compensate for larger e-h pair creation energy. The choice of the semiconductor detector for a specific application is often governed by a compromise in semiconductor properties. Also availability, reliability and experience play an important role. In this respect diamond is the choice of detector material next to silicon.

21.3.1.3 Most Important Defects in Silicon

A lot of effort was invested over the R&D phases of LHC/HL-LHC in identifying the defects responsible for changes in performance of silicon detectors. A comprehensive list of defects identified by so called "microscopic" techniques such as Deep Level Transient Spectroscopy (DLTS) or Thermally Stimulated Current (TSC) can be found in [15]. The summary plot with the most important defects is shown in Fig. 21.6. The effects for which they are mainly responsible will be addressed in

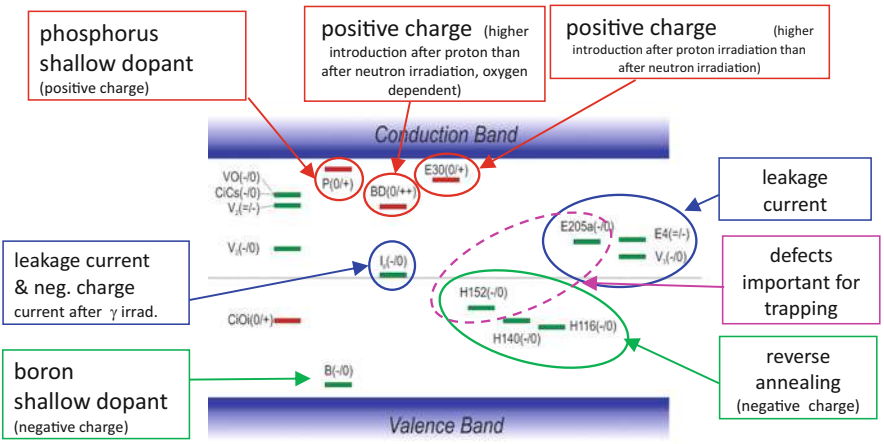

Fig. 21.6 A schematic view of known defects and their main effects on the detector performance. The defect charge state is given in brackets. For the defects with unknown chemical composition the temperature at which electron -E or hole-H traps were identified with DLTS/TSC techniques is used. The near mid-gap H levels are likely multivacancy complexes

the following sections. Note, that for only few identified energy levels the chemical composition of the corresponding defects is known.

21.3.2 Surface Damage

The semiconductor detector bulk needs to form a contact with readout electronics. The contacts used, either Ohmic or Schottky, as well as the rest of the surface are prone to changes due to irradiation. The description of surface radiation damage given here will be focused on the border of silicon bulk and oxide (Chap. 5). The surface damage affects the electrical properties of the detectors such as inter-electrode resistance, inter-electrode capacitance and dark current. It is particularly important for sensors where charge flow is close to the surface, such as 3D-Si detectors, CCDs, Active CMOS Pixel Detectors and MOS-FET transistors.

The surface of particle detectors is usually passivated by thermal oxidation [16]. The oxide isolates and stabilizes the crystal surface with respect to chemical and electrical reactivity. The cross-section of the device surface is generally divided into silicon/oxide interface and oxide bulk depicted in Fig. 21.7. The border region between oxide and silicon crystal is characterized by a large defect density due to bond stress. In general surface defects can be caused by growth and irradiation. According to their position in the oxide the traps are divided in the oxide bulk traps (OT), border traps (BT) and interface traps (IT). The latter two are located close to the interface and can exchange charges with underlying silicon (switching traps). The oxide traps are mostly donors, which is the reason that net oxide charge

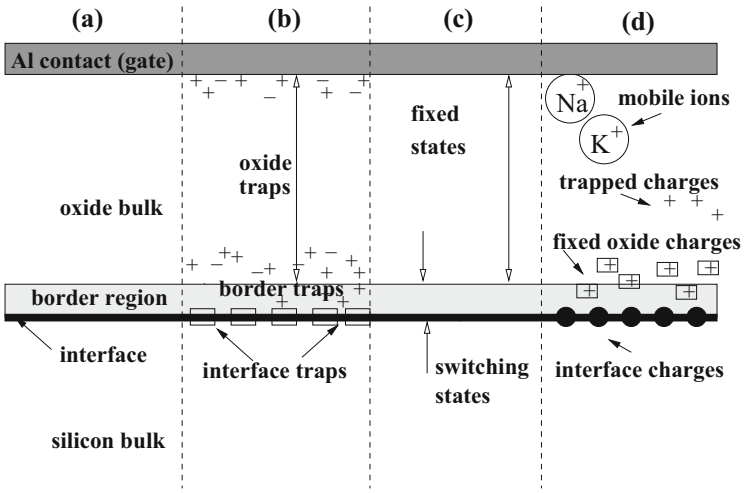

Fig. 21.7 Schematic view of the surface of a silicon detector according to [17]; (**a**) surface regions (**b**) trap locations (**c**) states (**d**) oxide charges

density is always positive. The most important oxide defects are trivalent Si (≡ Si·, donor), interstitial oxygen (O_I, donor) and non-bridging oxygen (≡ Si–O·, acceptor). Other important defects include hydrogen related defects (all donors) [18]. Hydrogen is particularly important since it passivates the dangling bonds by attaching to them. The build-up of interface traps is not fully understand yet and there are different models explaining it [18, 19]. The bulk and interface traps formed during processing of the oxide can be passivated by annealing (350–500 °C) in hydrogen rich environment.

If the creation of e–h pair in the silicon bulk is completely reversible process, it is not in SiO_2 and at the interface. Ionizing radiation has a significant impact on the defect generation and activation. The damage mainly manifests itself as a regeneration process of already present but deactivated defects. Hence the processing of the oxide, preparation and temperature treatments (annealing) impacts the performance after irradiation.

Although the underlying physics of formation is not yet fully understood, it is assumed that radiation ionizes oxide bulk defects that remain charged

$$\mathrm{Si}\cdot + \mathrm{radiation} \rightarrow \mathrm{Si}^{+} + \mathrm{e}^{-} \tag{21.23}$$

or free holes are trapped by passivated defects

$$\mathrm{O_I} + \mathrm{h} \rightarrow \mathrm{O_I}^{+}. \tag{21.24}$$

Similarly to oxide bulk damage the interface state density also increases with irradiation. After [20] the interface states are generated by breaking up the bonds between surface silicon atoms (Si_s) and hydrogen, due to hole trapping at the interface (Si_s-H+h→Si_s·+H^+; Si_s-H+h→Si_s^++H· followed by Si_s^++e^- →Si_s·). The dangling bonds enable surface silicon atoms to react with the underlying silicon and induce different states in the silicon band-gap. The state build-up can continue over a long period of time after exposure to radiation.

The electrons are much more mobile in the oxide ($\mu_e(20\,°\mathrm{C}) \sim 20\,\mathrm{cm}^2/\mathrm{Vs}$) and are in the presence of electric field promptly swept away, while holes ($\mu_h(20\,°\mathrm{C}) = 10^{-4}-10^{-11}\,\mathrm{cm}^2/\mathrm{Vs}$) slowly drift to the interface. The absence of electric field in the oxide is therefore beneficial as the recombination can take place in the oxide bulk as well as at the interface.

21.3.2.1 Impact of Surface Damage on Device Properties

Positive Oxide Charge

As shown by many experiments the exposure to ionizing radiation causes an increase of positive space charge. The different contributions to the oxide charge are shown in Fig. 21.7. Apart from the oxide traps and mobile ion impurities also trapped holes at interface states contribute to the positive oxide charge. An

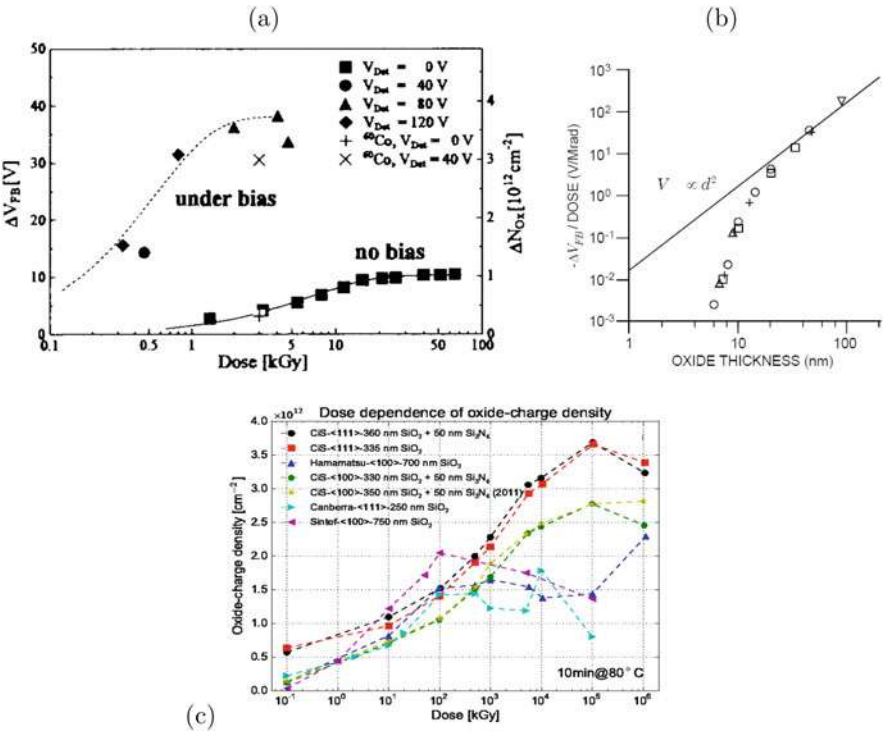

Fig. 21.8 (**a**) Oxide charge measured from a change in flat band voltage for silicon gated diodes [21] after irradiation with 20 keV electron and γ-rays from ^{60}Co. (**b**) Dependence of flat band voltage on oxide thickness [22]. (**c**) Recent measurements to very large doses for samples with different producer/orientation/oxide thickness [23].

effective net sheet charge (surface density) in the oxide N_{ox} is calculated as the sum of all contributions. It has been shown that under bias the oxide charge density increases with irradiation up to few kGy where it starts to exhibit saturation. In an unbiased devices saturation occurs at significantly larger doses up to few 10 kGy (see Fig. 21.8) [21]. The saturation sheet charge depends on thickness of the oxide and is of order $N_{ox} = 10^{12}\,\mathrm{cm}^{-2}$. Latest measurements show an increase of oxide charge, although at a much slower rate, up to the doses of 1 GGy (see Fig. 21.8c).

The positive oxide charge attracts electrons which can form a conductive layer underneath the surface. The resistivity between the nearby n^+ contacts can therefore decrease producing a short circuit. A p^+ implant is therefore commonly used to cut these conductive paths. A more novel approach is to use a moderate p implant over the whole surface (p-spray [24]). The p-spray dose must be sufficiently high ($\approx 10^{11}-10^{13}$ ions/cm^2; the same order as N_{ox}) to prevent decrease of inter-strip resistivity and not too high to cause early breakdowns. Very often both methods are used together.

In very thin oxides the tunneling of electrons from nearby electrodes occurs. The oxide traps get passivated, by reversing the reactions described by Eqs. (21.23), (21.24). Thinning down the oxide therefore reduces the N_{ox} (see Fig. 21.8b) [22], which makes the device more radiation hard. The flat band voltage which should follow the $V_{FB} \propto d^2$, if the oxide charge is uniform, shows a steep decrease in thin oxide films $<$20 nm. The importance of this effect will be discussed in section on radiation hard electronics.

Surface Generation Current

Interface states act as charge carrier generation centers. As soon as the silicon surface is depleted, the thermally generated carries are separated in electric field and contribute to the dark current of a nearby $p-n$ junction or a MOS transistor. This current is called interface generation current and is calculated as

$$I_{ox} = e_0 \, n_i \, S_s \, v_{surf} \tag{21.25}$$

where v_{surf} is the surface recombination velocity and S_s the depleted silicon surface area. The surface recombination velocity is directly proportional to the density of interface states. The density of states rather than discrete states is used as experimentally it is impossible to distinguish between different trap levels [25]. The increase of surface current and surface recombination velocity with irradiation is shown in Fig. 21.9.

Trapping

The interface states act as trapping centers for the charge drifting close to silicon surface in analogous way to trapping of drifting carries in the bulk. Equation 21.17 is multiplied with an exponential term $\exp\left(\frac{e_0 < \psi >}{k_B T}\right)$ to take into account the average band bending $< \psi >$ close to the surface.

21.4 Detector Technologies

21.4.1 Design Considerations

The design of the detector should minimize the radiation effects most crucial for the successful operation of the detector while retaining the required functionality. The material and operational conditions determine to a large extent the radiation hardness of a detector. However, some of the radiation effects can be reduced by

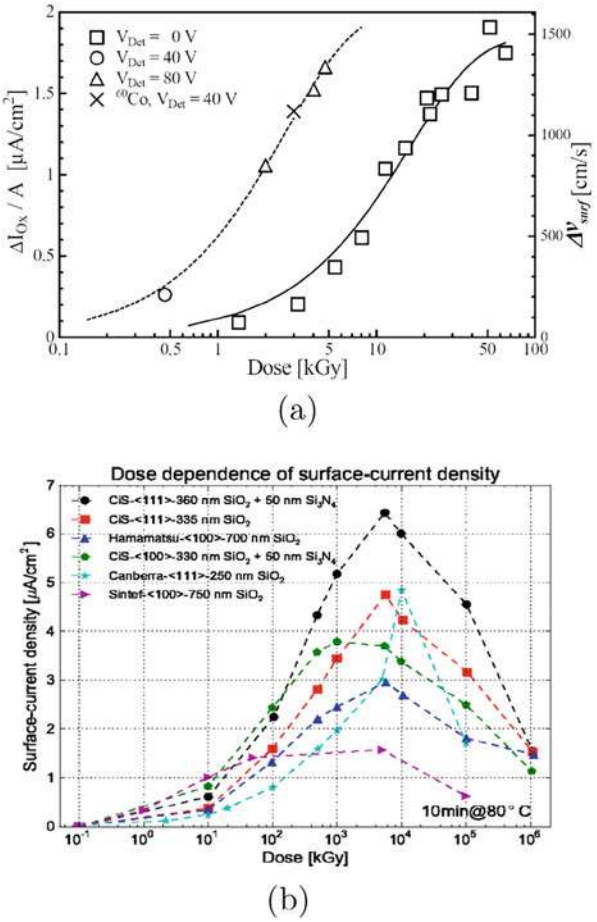

Fig. 21.9 (**a**) The increase of surface current density (surface recombination velocity) after 20 keV electron and ^{60}Co irradiations for biased and unbiased gate. (**b**) Surface current density after 12 keV X-rays irradiations of different samples to very high doses [23]

a choice of the read-out electrodes and detector geometry. At the new accelerator experiments the largest obstacle is the radiation-provoked decrease of measured charge and increase of noise. The consequent degradation of signal-to-noise ratio can lead to the loss of detection efficiency up to the level where successful operation of the detectors is no longer possible.

In terms of charge collection the radiation hard detector design follows directly from the calculation of the induced charge Q. The current induced (I) by a motion of charge q in the detector is given by Shockley-Ramo's theorem [26] and is discussed in the section on signal processing. The charge induced in the electrodes is given

by the difference in the weighting potential (U_w) traversed by the drifting charge (Chap. 10, Eq. 10.2):

$$Q(t) = q[U_w(\vec{r}(t)) - U_w(\vec{r_0})], \tag{21.26}$$

where $\vec{r}_0$ and $\vec{r}$ denote position at the both ends of the traversed path. The distinct difference in weighting potential for a pixel detector and simple pad detector is shown in Fig. 21.10 and discussed in section 6.2.2.

For an electron hole pair the induced charge is a sum of both contributions $Q_{e-h} = Q_e + Q_h$. A track of an ionizing particle therefore induces the charge Q^t

$$Q^t = \sum_{all\ pairs} Q_e + Q_h = Q_e^t + Q_h^t, \tag{21.27}$$

$$Q_{e,h}^t = \mp e_0 \sum_i^{e,h} U_w(\vec{r}_i) - U_w(\vec{r}_{i,0}). \tag{21.28}$$

If all carries complete the drift on the sensing electrode $U_w(\vec{r}_i) = 1$ if on non-sensing $U_w(\vec{r}_i) = 0$. In the absence of trapping and homogeneous ionization the sum in Eq. (21.28) becomes integral which can be easily calculated. For the track through the center of the pixel shown in Fig. 21.10 the contribution of electrons drifting to sensing electrode is $Q_e^t/Q^t = 0.82$, which is significantly larger than $Q_e^t/Q^t = 0.5$ for pad detectors. The fact that in segmented devices one carrier type contributes more to the total induced charge, can have important consequences after irradiation if the difference in mobility or/and trapping probability is large for electrons and holes.

If carriers are trapped and not released in time to finish the drift within the integration time of the amplifier (t_{int}) then $U_w(r_i) \neq 1, 0$. Using $\vec{v_{e,h}} = \mu_{e,h}\vec{E}$ and $q = e_0 \exp\left(\frac{-t}{\tau_{eff,e,h}}\right)$ the Eqs. (21.28), Eq. 1 (Section 6) turn to

$$Q_{e,h}^t = \mp e_0 \sum_i^{e,h} \int_0^{t_{int}} \exp\left(\frac{-t}{\tau_{eff,e,h}}\right) \mu_{e,h}(E) \left[\vec{E}(\vec{r}_i) \cdot \vec{E_w}(\vec{r}_i)\right] dt, \tag{21.29}$$

where $\mu_{e,h}$ represents carrier mobility. Three conclusions can be drawn without actually solving the Eq. (21.29) for a given detector and charge particle track:

- A better charge collection efficiency CCE (ratio of measured and generated charge) of the hit electrode is achieved when it collects the carriers with larger $\mu \cdot \tau_{eff}$. They contribute a larger part to Q^t and hence reduce the effect of the trapping.
- If the electric field can not be established in the entire detector (e.g. partial depletion or polarization of detector) it is important to have the region with

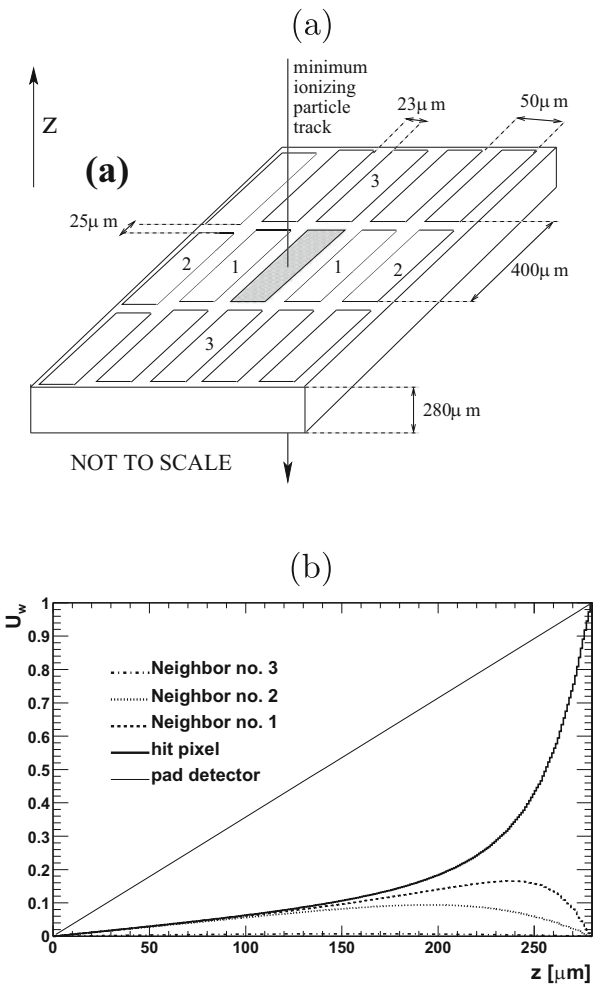

Fig. 21.10 (**a**) A schematic picture of the ATLAS pixel detector with pixel dimensions of $400 \times 50\,\mu\text{m}^2$. The hit pixel for which the U_w was calculated is shaded. Neighbors are denoted by the corresponding numbers. (**b**) The weighting potential along the axis through the center of the hit pixel and through the center of the three closest neighbors. For comparison U_w of a pad detector is also shown

electric field around the read-out electrodes, where E_w is large (large $\vec{E} \cdot \vec{E_w}$). Operation of partially depleted detectors therefore requires that the junction grows from the segmented side. Growth of depletion region from the back of the detector, shown in Fig. 21.10, would result in smaller induced charge in hit pixel than expected from the thickness of the active region.

- A detector design where the number of generated e–h pairs is disentangled from their drift time is optimized for large induced charge.

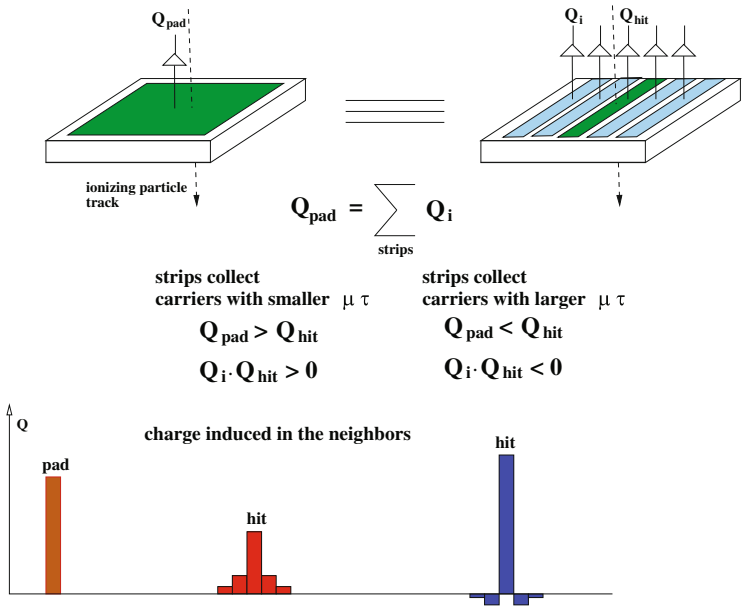

Fig. 21.11 Explanation of trapping induced charge sharing

As U_w depends on the geometry only it is obvious that it is possible to optimize the electrode design for maximum signal. However, as the paramount parameter for any detector is its signal-over-noise ratio, the optimization should also include the inter-strip capacitance and leakage current of electrode both affecting the noise.

Charge collection in segmented devices leads to "signal cross-talk" as described in the section on Signal processing 6.2.1. The bi-polar current pulses in the neighboring electrodes (see e.g. U_w in Fig. 21.10) yield zero net charge for integration time larger than the drift time (see Signal processing Fig. 6.2). In irradiated detector some of the carriers are trapped and do not complete their drift. Therefore the integrals of the bipolar pulses do not vanish. A significant amount of charge can appear in the neighbors adding to the usual charge shared by diffusion (see explanation in Fig. 21.11). Unlike diffusion, where the polarity of the induced charge is equal for all electrodes, the trapping can result in charges of both polarities. If electrodes collect carriers with smaller $\mu \cdot \tau_{eff}$, the polarity of the charge is the same for all electrodes. Otherwise the polarity of the charge induced in the neighbors is of opposite sign compared to the hit electrode [27, 28]

The effect can be used to enhance the spatial resolution due to larger charge sharing at the expense of smaller charge collection efficiency or vice versa.

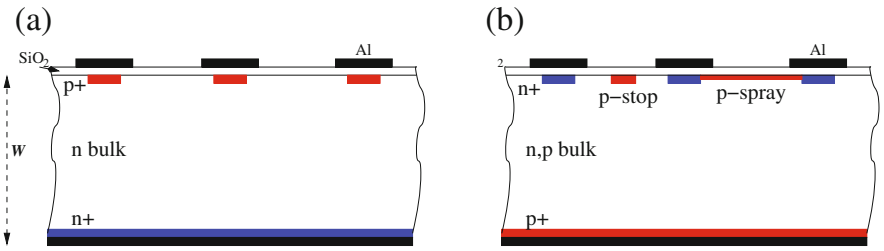

Fig. 21.12 Schematic view of (**a**) $p^+ - n - n^+$ and (**b**) $n^+ - n - p^+$, $n^+ - p - p^+$ strip detectors (AC coupled)

21.4.2 Silicon Detectors

Silicon is by far the most widely used semiconductor detector material. A large majority of silicon particle detectors exploit the asymmetric $p - n$ junction bias in the reverse mode as a basic element. Up to recently the detector grade silicon was produced by the so called float zone (FZ) technique, where concentration of impurities and dopants can be precisely controlled to very low values ($\sim 10^{11}$ cm^{-3}). The step further in radiation hardening of silicon detectors was the enrichment of the float zone silicon through oxygen diffusion (DOFZ). Recently, detectors were processed on Czochralski[1] and epitaxially grown silicon and are in some respects radiation harder than float zone detectors.

Most of the detectors used up to now were made on n-type silicon with p^+ readout electrodes (see Fig. 21.12a), which collect holes. Electrons have larger $\mu\,\tau_{eff}$ in silicon, hence n^+ readout electrodes are more appropriate for high radiation environments where the loss of charge collection efficiency is the major problem. They are mostly realized by segmentation of n^+ side of the n-type bulk (see Fig. 21.12b), which however requires more complex processing on both detector sides. The double sided processing can be avoided by using p-type bulk material with n^+ electrodes [29]. This is the preferred choice silicon detector type at HL-LHC.

21.4.2.1 Effective Doping Concentration

The defects produced by irradiation lead to change of the effective doping concentration. The main radiation induced defects responsible for the change of effective dopant concentration can be found in Fig. 21.6 and consist of both donors and acceptors.

[1]If magnetic field is used to control the melt flow in crucible the process is called Magnetic-Czochralski.

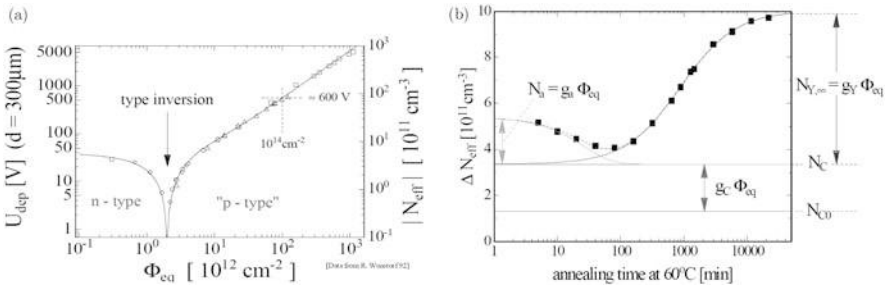

Fig. 21.13 (**a**) Effective doping concentration in standard silicon, measured immediately after neutron irradiation [30] (**b**) Evolution of ΔN_{eff} evolution with time after irradiation [31]

It is a well established, that irradiation by any particle introduces effectively negative space charge in detectors processed on float zone silicon, which is most commonly used. The change in effective doping concentration is reflected in the full depletion voltage V_{fd}, needed to establish the electric field in the entire detector:

$$V_{fd} = \frac{e_0 \, |N_{eff}| \, W^2}{2\epsilon_0 \epsilon}. \tag{21.30}$$

The V_{fd} of initially n-type detectors ($p^+ - n - n^+$), therefore decreases to the point, where the negative space charge prevails, so called space charge sign inversion point (SCSI). The N_{eff} turns to negative and depleted region grows from the n^+ contact at the back. The V_{fd} thereafter continues to increase with fluence beyond any tolerable value, which is usually set by the breakdown of a device (see Fig. 21.13a). The space charge of p-type detectors ($n^+ - p - p^+$) remains negative with irradiation so that the main junction stays always at the front $n^+ - p$ contact.

For both detector types not only deep radiation induced defects are created, but also initial shallow dopants are electrically deactivated (removed)—so called initial dopant removal. The initial dopant removal impacts to large extent the performance of some detector technologies such as Low Gain Avalanche Detectors and depleted CMOS detectors, which will be reviewed later.

Evolution of Effective Dopant Concentration—Hamburg Model

After the irradiation the defects responsible for space charge evolve with time according to defect dynamics described by Eqs. (21.3), (21.4). The time scale of these processes varies from days to years already at close to room temperatures which makes the annealing studies lengthy procedures. At elevated temperature the underlying defect kinetics can be accelerated, and thus the simulation of the damage investigation at real experiments spanning several years is possible in weeks.

The radiation induced change in the effective doping concentration is due to historical reasons defined as $\Delta N_{eff} = N_{eff,0} - N_{eff}(t)$, where $N_{eff,0}$ denotes the initial doping concentration. The fact that the radiation introduced space charge is negative means that ΔN_{eff} is positive. The evolution of N_{eff} after irradiation is shown in Fig. 21.13b. ΔN_{eff} initially decreases, reaches its minimum and then starts to increase. The measured evolution can be described by a so called Hamburg model, which assumes three defects [31] all of them obeying first order kinetics (see Eq. (21.3)). The initial decrease of ΔN_{eff} is associated with decay of effective acceptors (N_a). After a few days at room temperature a plateau, determined by defects stable in time (N_c), is reached. At late stages of annealing effective acceptors are formed again (N_Y) over approximately a year at room temperature. The corresponding equations are:

$$\Delta N_{eff} = N_{eff,0} - N_{eff} = N_a(\Phi, t) + N_c + N_y(\Phi, t) \tag{21.31}$$

$$\Delta N_{eff} = g_a \Phi_{eq} \exp\left(-\frac{t}{\tau_a}\right) + N_c + g_Y \Phi_{eq}(1 - \exp\left(-\frac{t}{\tau_{ra}}\right)) \tag{21.32}$$

$$N_c = \pm N_{id}(1 - \eta(1 - \exp(-c \cdot \Phi_{eq}))) + g_c \Phi_{eq}, \tag{21.33}$$

where g_a, g_c and g_Y describe the introduction rates of defects responsible for the corresponding part of the damage and τ_a and τ_{ra} the time constants of initial and late stages of annealing.

The **stable part** of the damage incorporates also **initial dopant removal**, where $\pm N_{id}$ (negative/positive sign for donors/acceptors) denotes the concentration of initial dopants, η fraction of removed dopants and c the removal constant. Displacement of the initial dopant from the lattice site, deactivates it. Once in the interstitial position, initial dopants (mainly boron and phosphorous) can react with other defects leading to possibly new electrically active defects. The new defects formed can also be charged, hence the removal can be partial, i.e. $N_{id} \neq N_{eff,0}$ [32, 33]. For example, the interstitial boron can undergo different reactions with impurities forming both donor and acceptor like defects [32]. As the reactions can take place also with impurities the removal rate depends on their concentration.

The initial donor (phosphorous) removal was intensively studied for high resistivity $p^+ - n - n^+$ detectors [34], where initial donor removal is attributed to formation of electrically inactive Vacancy-Phosphorous (V-P) complex. The rate of removal was found to depend on initial concentration with $N_{id} \times c \approx 0.008\,\mathrm{cm}^{-1}$. The reason for such relation is unclear. It was observed that donor removal is complete for charge hadron irradiated detectors while around half of the initial donors remain effectively active after neutron irradiations ($\eta \sim 0.45 - 0.7$).

The initial acceptor (boron) removal was much less studied in the $n^+ - p - p^+$ particle detectors, more for solar cells [35]. The required radiation hardness of p-type detectors for HL-LHC is such that deep acceptors exceed the concentration of

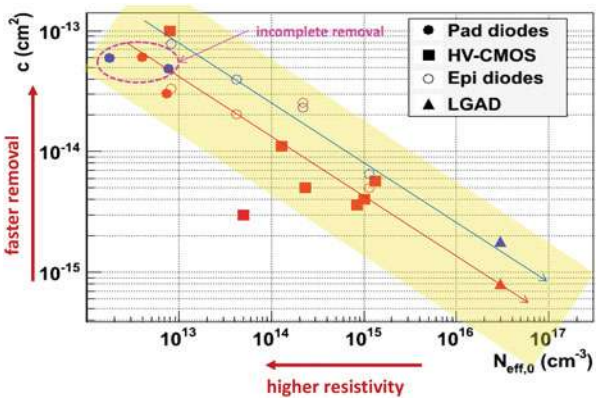

Fig. 21.14 Initial acceptor removal rate dependence on initial dopant concentration. The data were obtained from measurements with different detectors/technology: pad diodes (float zone and epitaxial), depleted (HV) CMOS and LGADs. The red markers show neutron irradiations and the blue markers show fast charged hadron irradiations. The red and blue arrows guide the eye. Data from Refs. [36–41]

Table 21.4 The survey of Hamburg model parameters for standard and diffusion oxygenated float zone detectors

	Standard FZ		Diffusion Oxygenated FZ	
	Neutrons	Charged hadrons	Neutrons	Charged hadrons
g_a [cm^{-1}]	0.018	–	0.014	–
τ_a [h at 20 °C]	55	–	70	–
g_c [cm^{-1}]	0.015	0.019	0.02	0.0053
g_Y [cm^{-1}]	0.052	0.066	0.048	0.028
τ_{ra} [days at 20 °C]	480	500	800	950

The uncertainty in the parameters is of order 10% and mainly comes from variation of silicon materials used

initial ones by far, hence their removal was not in focus. However, new detector technologies (LGAD, depleted CMOS) with significant/dominant concentration of initial dopants also after foreseen fluences, triggered extensive studies of initial acceptor removal. Similarly to donor removal c was found to depend on initial concentration as shown in Fig. 21.14. The rate of removal is around two times larger for fast charged hadrons and only for large initial dopant concentrations the removal is complete ($\eta \approx 1$).

The **parameters of the Hamburg model** related to radiation induced defects (deep traps) are given in the Table 21.4 and are valid for p- and n-type silicon detectors. For reasons that will be explained later, the model parameters are also shown for FZ detectors which were deliberately enriched by oxygen.

The time constants of initial (τ_a) and late stage annealing (τ_{ra}) can be scaled to different annealing temperatures by using Eq. (21.9). The activation energies for initial and long term annealing are $E_{ra} \approx 1.31\,\text{eV}$ and $E_a \approx 1.1\,\text{eV}$ [34].

After around 80 min annealing at 60 °C $N_a, N_y \ll N_c$ and ΔN_{eff} is almost entirely due to stable defects. If the initial dopant removal is complete or initial dopant concentration is small (with respect to to deep defects) the effective doping concentration is given by a simple relation $|N_{eff}| \approx g_c \cdot \Phi_{eq}$.

Often the irradiations follow the planned operation scenario. For example at LHC the detectors are operated at $T \approx -10°\text{C}$ for 1/3 of the year then stored for few weeks at close to room temperature and the rest of the year at $T \approx -10°\text{C}$. The corresponding temperature history of a whole year can be compressed roughly to 4 min at 80°C. The whole period of operation therefore consists of multiple irradiation and annealing steps, which is also referred to as CERN scenario [34].

The parameters of Hamburg model are used to predict the evolution of full depletion voltage of silicon pixel ($n^+ - n - p^+$) and strip detectors ($p^+ - n - n^+$) at LHC experiments. The agreement of predictions with measurements during LHC operation was good, as shown on few examples in Fig. 21.15.

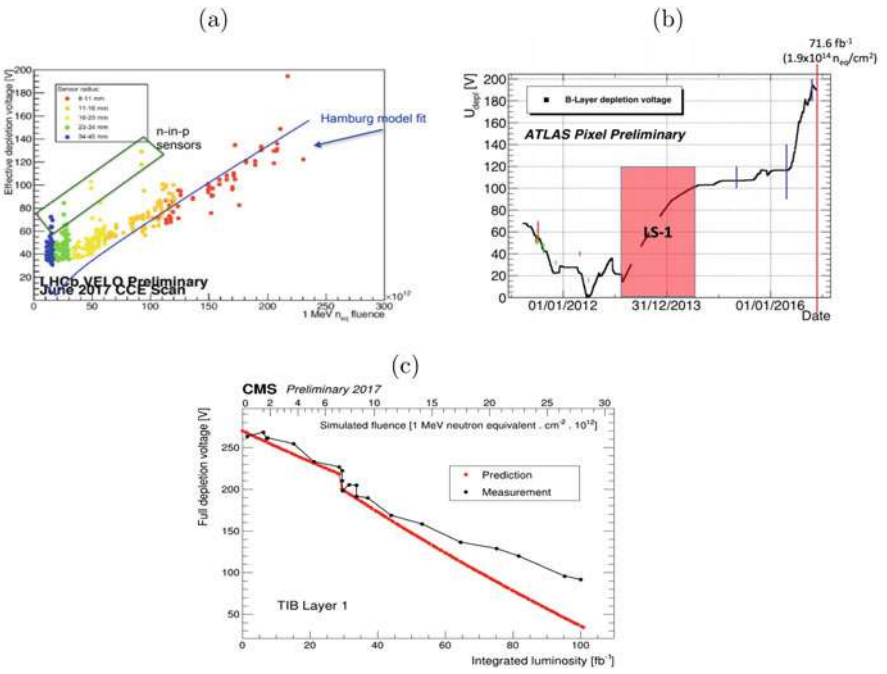

Fig. 21.15 The agreement of predicted and measured V_{fd} for (**a**) LHCb Velo detector [42], (**b**) ATLAS-Insertable B layer pixel detector [43] (**c**) CMS—strip detectors in the outer region [44]. For (**b**) the prediction is denoted by black dots and measurements as bars with different colors

As can be seen in Figs. 21.15, the agreement of Hamburg model with measurements is reasonable and allows for predictions of operation up to the end of their lifetime at LHC. It is evident that careful planning of maintenance and technical stops is required to keep V_{fd} as low as possible. Even though oxygen rich silicon was used for ATLAS pixel detectors, they will be operated under-depleted at least for some time at the end of LHC operation. The depleted region after space charge inversion grows from the pixel side and for $V_{bias} < V_{fd}$ the detector performance is similar to that of somewhat thinner detector, still providing efficient tracking.

On the other hand irradiated strip detectors at LHC ($p^+ - n - n^+$) require at all times $V_{bias} > V_{fd}$ as the region around the strips needs to be depleted for achieving sufficient charge collection efficiency. The maximum bias voltage for e.g. ATLAS strip detectors is set to 450 V, which is sufficient for full depletion over the entire operation program before the HL-LHC upgrade. Standard float zone detectors are used for the fact that the larger fraction of damage is coming from neutrons and oxygenated detectors would therefore offer no significant advantage.

Defect Engineering

The radiation tolerance of silicon can be improved by adequate defect engineering. Defect engineering involves the deliberate addition of impurities in order to reduce the radiation induced formation of electrically active defects or to manipulate the defect kinetics in such a way that less harmful defects are finally created. It has been established that enhanced concentration of oxygen in FZ detectors reduces the introduction rate of stable defects by factor of $\sim$3 after charged hadron irradiations (see Table 21.4). The most likely explanation is that oxygen acts like a trap for vacancies (formation of an uncharged V-O complex) and therefore prevents formation of charged multi-vacancy complexes. In addition, Oxygen is also related to formation of deep donors (see Fig. 21.6).

On the opposite carbon enhances the concentration of vacancies as it traps interstitial silicon atoms and reduces the recombination. Since the concentration of oxygen is not high enough in the disordered regions-clusters, it has little or no effect after neutron irradiations. Different stable damage in neutron and charged hadron irradiated detectors at equal NIEL is an evidence of NIEL hypothesis violation. The diffusion oxygenated float zone detectors are used for the inner-most tracking detectors at LHC, where significant reduction of V_{fd} is required as shown in Fig. 21.15.

The oxygen concentration in DOFZ detectors is around $2 \cdot 10^{17}\,\mathrm{cm}^{-3}$, which is up to an order of magnitude lower than the oxygen concentration in Czochralski (Cz) silicon. They have only recently become available as detector grade material with resistivity (>1 kΩcm) high enough to allow production of 300 μm thick detectors [45]. The increase of V_{fd} after irradiation was found to be smaller or equal to that of DOFZ detectors as shown in Fig. 21.16a. Moreover, for n-type Cz detectors (less evident in p-type Cz) stable donors ($g_c \sim -5 \cdot 10^{-3}\,\mathrm{cm}^{-1}$) are introduced instead of acceptors after fast charged hadron and γ-ray irradiations. The oxygen in form

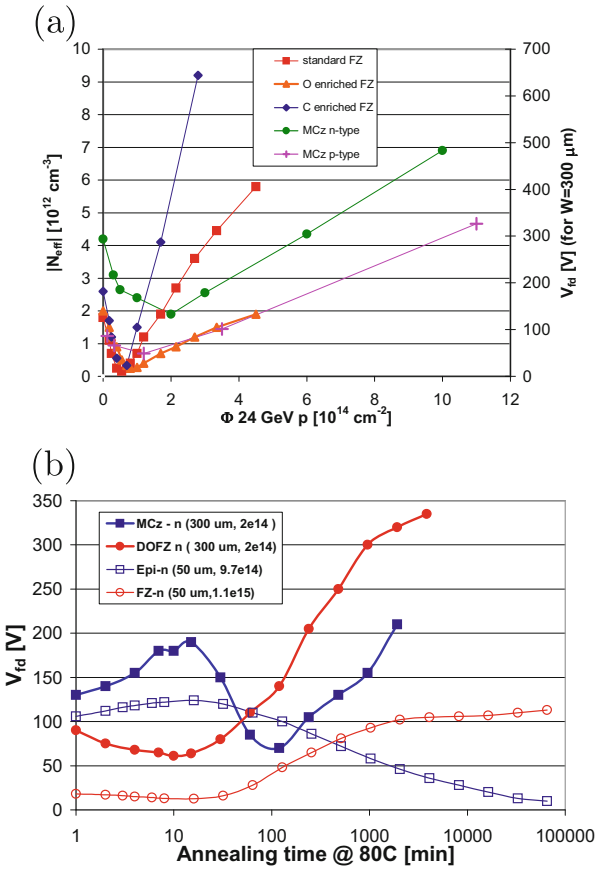

Fig. 21.16 (**a**) Influence of carbon and oxygen enrichment and wafer growth on the change of N_{eff} as function of fluence. (**b**) Annealing of the Magnetic Cz-*n* type (MCz) and diffusion oxygenated samples after $2 \cdot 10^{14}\,\mathrm{cm}^{-2}$. Also shown are thin epitaxial and standard FZ detectors irradiated to fluences around $10^{15}\,\mathrm{cm}^{-2}$. Note the typical behavior of detectors with positive space charge for epitaxial and MCz detectors

of a dimer $[O_{2i}]$, which is more abundant in Cz than FZ detectors, is likely to be responsible. It is a precursor for formation of radiation induced shallow donors (thermal donors) [46]. The reverse annealing in Cz detectors has approximately the same amplitude as in FZ but is delayed to such extent that may not even play an important role at future experiments. The different sign of g_c and g_Y produce a different shape of N_{eff} annealing curve (see Fig. 21.16). During the short term annealing the V_{fd} increases and then starts to decrease as acceptors formed during late stages of annealing compensate the stable donors. Eventually the acceptors prevail and the V_{fd} starts to increase again.

Another interesting material is epitaxial silicon grown on low resistivity Cz substrate [47]. Stable donors are introduced after charge hadron irradiation with

rates depending on the thickness of the epitaxial layer ($g_c = -4 \cdot 10^{-3}$to $-2 \cdot 10^{-2}\,\mathrm{cm}^{-1}$, for thickness of 150–25 μm). They exhibit also the smallest increase of $|N_{eff}|$ after neutron irradiations, but are only available in thicknesses up to 150 μm.

Control of Space Charge

The opposite sign of g_c and g_Y and $|g_Y| > |g_c|$ opens a possibility to control V_{fd} with a proper operation scenario and to keep it low enough to assure good charge collection (see Fig. 21.16b).

This has been demonstrated with thin epitaxial detectors which were irradiated in steps to $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$ and annealed for 50 min at 80°C during the steps which is roughly equivalent to room temperature storage during non-operation periods at LHC or HL-LHC (see Fig. 21.17) [48]. The compensation of stable donors by acceptors activated during the irradiation steps resulted in lower V_{fd} after $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$ than the initial V_{fd}. Allowing detectors to anneal at room temperature during non-operation periods has also a beneficial effect on leakage current and trapping probability as will be shown later.

The use of silicon material with opposite sign of stable damage for neutrons and charged hadrons can be beneficial in radiation fields with both neutron and charged hadron content. The stable acceptors introduced by neutron irradiation compensate stable donors from charged hadron irradiations and lead to reduction of V_{fd} as demonstrated in [49]. An example is shown in Fig. 21.17b for MCz n-type pad detectors which were irradiated by 23 GeV protons (open symbols) and then by neutrons (solid symbols). The additional neutron irradiation decreases the V_{fd}.

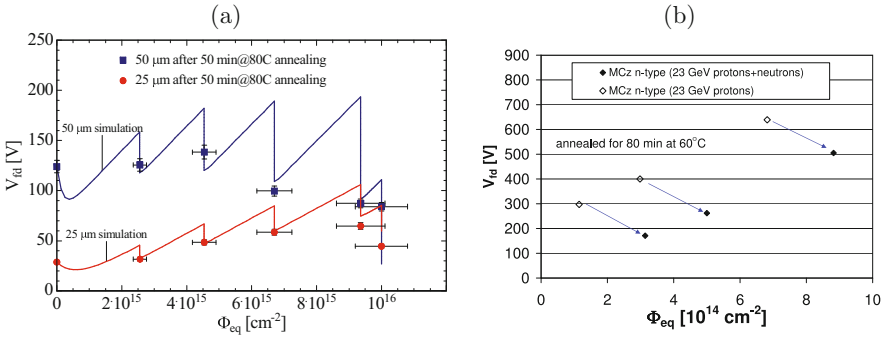

Fig. 21.17 (**a**) An example of space charge compensation through annealing in a thin epitaxial detector irradiated with 23 GeV protons to $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$. The lines denote the Hamburg model prediction. (**b**) Beneficial effect of irradiations by protons and neutrons on V_{fd} for MCz n-type detectors

21.4.2.2 Electric Field

The occupation probability (Eq. (21.15)) of a deep level is determined by its position in the band gap, temperature and concentration of free carriers. The occupancy of initial shallow dopants is largely unaffected by p, n, T and N_{eff} is constant over the entire bulk. The irradiation introduces deep levels which act as generation centers. Thermally generated carriers drift in the electric field to opposite sides (bulk generation current). The concentration of holes is thus larger at the p^+ contact and of electrons at the n^+ contact. Some of these carriers are trapped and alter the space charge i.e. steady state P_t in Eq. (21.14). As a result the N_{eff} is no longer uniform, but shows a spatial dependence, with more positive space charge at p^+ and more negative at n^+ contact Such a space charge distribution leads to an electric field profile different from linear.

The electric field profile can be probed by measuring the current induced by the motion of carriers generated close to an electrode (so called Transient Current Technique). They drift over the entire thickness of detector. The measured induced current at time t after the injection, is then proportional to the electric field, at the position of the drifting charge at time t according to equation $i = -q\vec{E}_w \cdot \vec{v}$. An example of such a measurement can be seen in Fig. 21.18b, where carriers at the back of the detector (n^+ contact) are generated close to electrode by a short pulse of red light. The shape of the current depends on the voltage and temperature. At lower voltages and higher temperatures the electric field shows two peaks, which can only be explained by the space charge of different signs at both contacts. This is usually referred to as "double junction" profile [50, 51], the name indicating that the profile is such as if there were two different junctions at both contacts ($p^+ - n\ p - n^+$ structure). This is evident for under-depleted detectors where both junctions are separated by an un-depleted bulk. Usually one of the regions dominates spatially

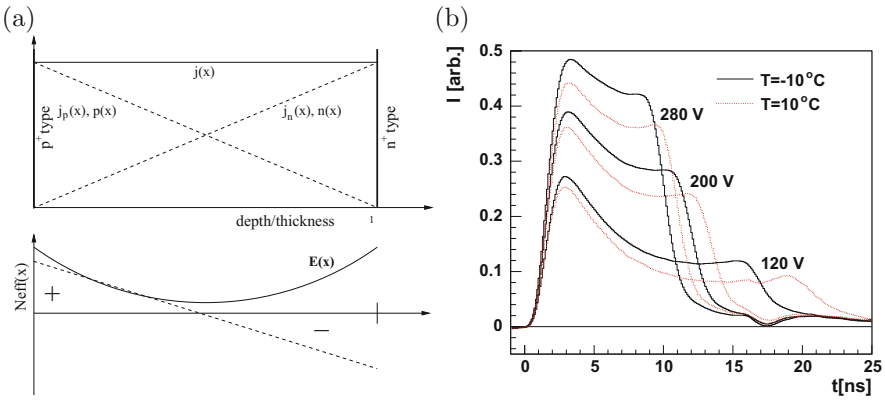

Fig. 21.18 (**a**) Illustration of mechanism leading to non-uniform N_{eff}. (**b**) Induced current due to drift of holes from n^+ side to p^+ side in 300 μm thick oxygenated detector irradiated with 23 GeV protons to $2 \cdot 10^{14}$ cm^{-2}

(also called the "main junction") which determines the predominant sign of the space charge and annealing properties. The space charge profile depends on the balance between the deep levels which occupation depends on n, p and shallow defects mostly unaffected by n, p.

Apart from thermally generated carriers the non-equilibrium carriers which modify the electric field can also be generated by ionizing particles or continuous illumination of detector by light [52].

Modeling of the Field

Even more precise insight in electric field, particularly for heavily irradiated detector ($\Phi_{eq} > 10^{15}\,\mathrm{cm}^{-2}$), is obtained by a more elaborate technique called Edge-TCT [53] shown in Fig. 21.19, where the polished edge of the silicon strip detector is illuminated by narrow beam of infra-red light. The induced current measured promptly after light injection is proportional to the sum of the drift velocities of electrons and holes at a given depth of injection. The drift velocity profile of an detector is hence obtained by scanning over the edge of the detector at different depths. The profiles of heavily irradiated silicon detectors are shown in Fig. 21.20.

The velocity profile in detector moderately irradiated with neutrons (Fig. 21.20a) deviates only slightly from simple model of constant N_{eff} inside the bulk, while at higher fluence (Fig. 21.20b) the electric field shows typical "double junction" behavior, with some remarkable features:

- the main junction penetrates deeper than expected using g_c measured at low fluences
- the high field region at the back extends deep into the detector
- the electric field is present in the whole bulk even at very modest voltages

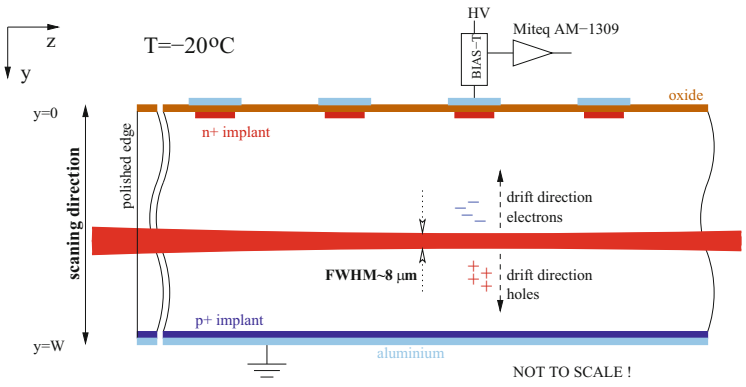

Fig. 21.19 The principle of the Edge-TCT technique

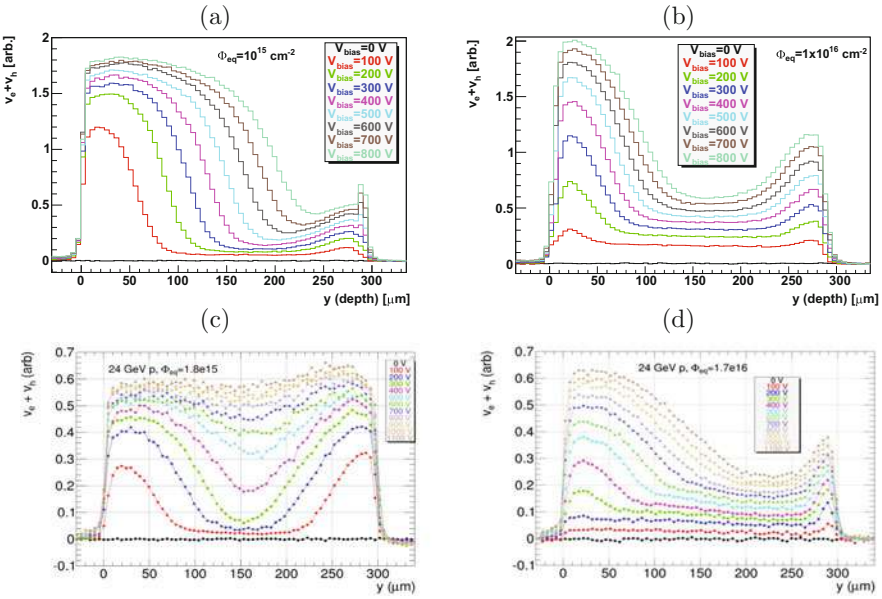

Fig. 21.20 The velocity profiles of neutron irradiated detectors to (**a**) $\Phi_{eq} = 10^{15}\,\mathrm{cm}^{-2}$, (**b**) $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$ and 23 GeV proton irradiates detectors to (**c**) $\Phi_{eq} = 1.8 \cdot 10^{15}\,\mathrm{cm}^{-2}$ and (**d**) $\Phi_{eq} = 1.7 \cdot 10^{16}\,\mathrm{cm}^{-2}$. The measurements were performed with 300 μm thick ATLAS-07 prototype strip detectors with 100 μm pitch and 20 μm implant width at −20 °C. Strips are at $y = 0$ μm

- the velocity in the neutral bulk is very high reaching almost a third of the saturation velocity at very high bias voltages

The appearance of the electric field in the neutral bulk can be explained by the increase of undepleted bulk resistivity and increase of generation current. As both increase also higher field is required for transport of thermally generated carriers across the detector in a steady state.

The electric field in charged hadron irradiated detectors is almost symmetrical at lower fluence (Fig. 21.20c) and becomes similar to neutron irradiated ones only at very high bias voltages (Fig. 21.20d). Already at 500 V the detector is fully active after receiving $\Phi_{eq} = 1.8 \cdot 10^{15}\,\mathrm{cm}^{-2}$. The reason for such behavior is not clear, but points to higher oxygen content of the silicon wafers and different energy levels associated with changes of N_{eff} with respect to the neutron irradiated detectors.

Extraction of electric field from velocity profile is not straightforward [53], due to large uncertainties arising from saturation of drift velocity with the electric field. Instead of precisely modeling $N_{eff}(y)$ several key parameters can be extracted from the measured velocity profiles which can be used to constrain/anchor any electric field model, either effective or calculated from known defects. These parameters are

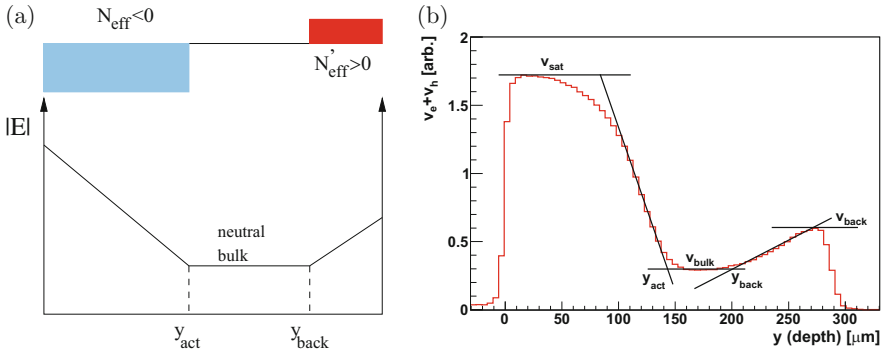

Fig. 21.21 (**a**) Simplest effective space charge and electric field model in irradiated strip detectors. (**b**) Extraction of key parameters determining electric field from the measured velocity profile

shown in Fig. 21.21 and are:

- depth of active region with negative space charge extending from the electrode side y_{act}
- velocity in undepleted bulk v_{bulk}
- depth of positive space charge region at the back of the detector $W - y_{back}$
- velocity at the back of the detector v_{back}

The parameters extracted for neutron irradiated detectors are shown in Fig. 21.22. The change of active region depth y_{act} with voltage is compatible with g_c up to the fluence of $\Phi_{eq} < 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$, while a three times lower g_c was extracted at $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$. Drift velocity in neutral bulk increases both with fluence and voltage, while the depth of the active region at the back is less dependent on fluence.

It is clear that in heavily irradiated detectors ($\Phi_{eq} > 1 - 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$) the V_{fd} doesn't serve as a relevant parameter determining the active thickness as the whole detector becomes active with irradiation.

21.4.2.3 Charge Multiplication

The increase of N_{eff} with irradiation and high applied bias voltages lead to very high electric fields close to electrodes. They can become high enough so that the electrons gain enough energy in its free path to create new e–h pairs, a process called impact ionization. After drifting over the distance dx the number of free carriers increases by

$$dN_{e,h} = \alpha_{e,h}\, N_{e,h}\, dx \tag{21.34}$$

where $\alpha_{e,h}$ are the impact ionization coefficients for electrons and holes [54, 55]. Charge multiplication through impact ionization is a well known process and widely

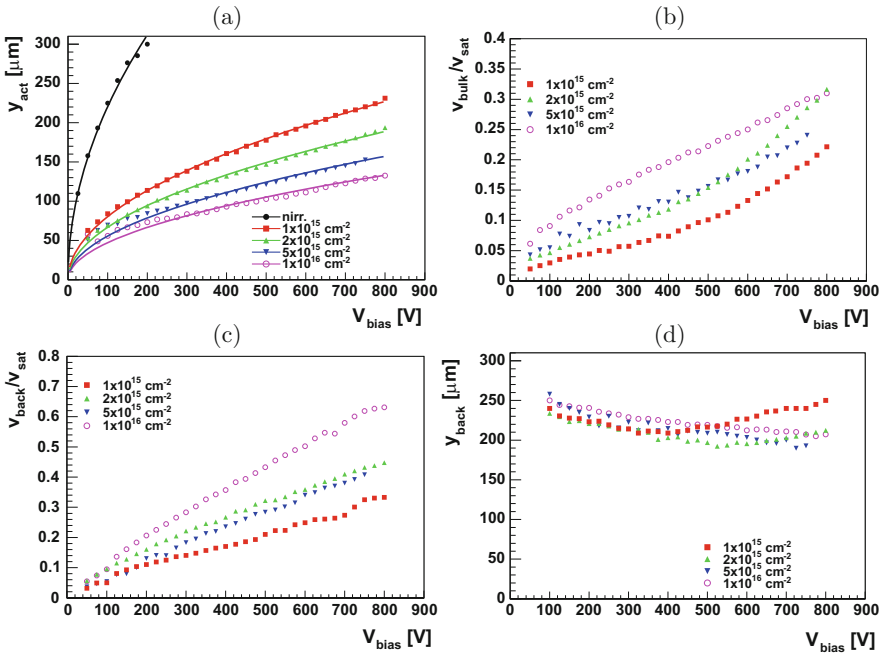

Fig. 21.22 The relevant parameters of the electric field in the neutron irradiated silicon detector—see Fig. 21.21 for explanation

exploited in Avalanche Photo Diodes and Si-Photo-multipliers. It was however not observed directly in irradiated silicon detectors. Prediction of detector performance a decade ago based on extrapolation of damage parameters to fluences well above $\Phi_{eq} > 10^{15}\,\mathrm{cm}^{-2}$ greatly underestimated the charge collection and detection efficiency.

Part of this, better than expected, performance can be attributed to favorable electric field profile, part to smaller trapping (discussed later) and part to charge multiplication. A key factor was improved high voltage tolerance of detectors which allowed application of bias voltages exceeding 1 kV.

Charge multiplication has since been undoubtedly observed with charge collection efficiency $CCE > 1$ in pad detectors [56], 3D detectors [57] and mostly strip detectors [58, 59] (see Fig. 21.23a). Another direct evidence came from TCT measurements where the drift of holes produced in multiplication was clearly observed as shown in Fig. 21.23b. There are several aspects of charge multiplication that make it difficult to control and master:

- Charge multiplication is geometry/process dependent; fields between 15–25 V/μm are required to produce sizable gain (∼1 e_0/μm). To achieve high gains the shape of implant and segmentation of electrodes (pitch and implant width) are very important. Strong field focusing close to implant edges leads to

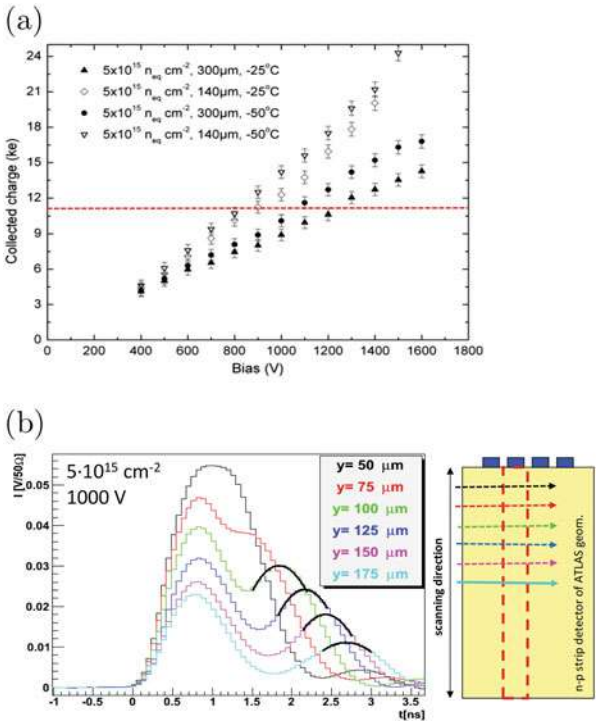

Fig. 21.23 (**a**) Measured charge collection dependence on voltage for 140 and 300 μm thick strip detectors. The red line denotes the charge measured in non-irradiated 140 μm thick detector. (**b**) Induced current pulses in strip detector for different depths of Edge-TCT injection. The second peak in the induced current pulses is due to multiplied holes drift

higher gains. This is also the reasons why larger gains were observed in highly segmented detectors.

- Charge gain depends on the hit position within the electrode. In highly irradiated strip detectors higher gain was observed for tracks few μm away from the implant, where the electric fields are highest [60].
- The holes produced in multiplication are trapped by deep defects (change of free hole concentration, p, in Eq. (21.15)) which reduce the negative space charge—act as a feedback. Therefore gain increases moderately with voltage and is usually limited to factors below <10.
- Gain can vary on time scale of days when detector is under bias [61].
- It is difficult to parametrize the field and reliably simulate the operation.

Annealing Performance of Highly Irradiated *p*-type Detectors Active bulk and charge multiplication have an important impact on performance of p-type detectors after annealing. Increase of N_{eff} with time and consequent increase of electric field increases gain. On the other hand smaller high field region near the electrodes affects less the performance due to significant field in the neutral bulk. A typical annealing

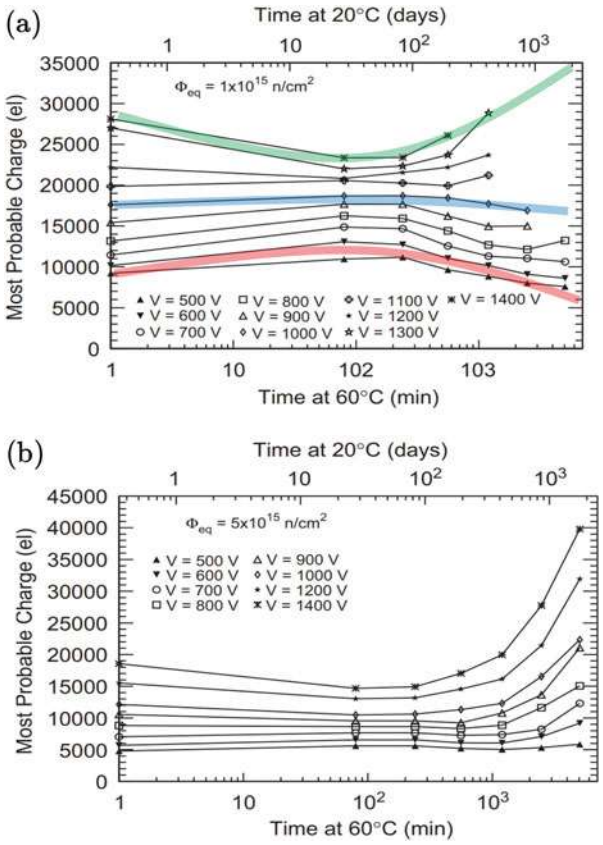

Fig. 21.24 Dependence of charge collection on annealing time at 60 °C at different bias voltages at (**a**) $\Phi_{eq} = 1 \cdot 10^{15}$ cm^{-2} and (**b**) $\Phi_{eq} = 5 \cdot 10^{15}$ cm^{-2} [62]

performance is shown in Fig. 21.24a. At lower voltages charge collection increases during short term and decreases during long term annealing (red band), which is in agreement with evolution of effective doping concentration. At higher voltages the charge multiplication compensates the decrease of active region (blue band) and at highest voltages overcompensates it, resulting in smallest charge collection for completed short term annealing (green band). At higher fluences and voltages shown in Fig. 21.24b the beneficial effect of long term annealing is even more pronounced.

Noise The increase of noise due to multiplication can diminish the benefits or even deteriorate the performance in terms of signal/noise ratio. The details about the noise in multiplication mode will be discussed at in the section on electronics.

21.4.2.4 Charge Trapping

The decrease of charge collection efficiency is determined by the trapping term and the product $\vec{E} \cdot \vec{E_w}$ in Eq. (21.29). At fluences beyond that at LHC the trapping term dominates and ultimately sets the limit of efficient operation. The influence of trapping on charge collection can be clearly seen for a fully depleted detector, where the degradation of the induced charge is exclusively due to trapping. The collected charge degrades with fluence as shown in Fig. 21.25a. The degradation is severe and around half the charge in non-irradiated detector (12000 e_0) are measured at V_{fd} for $\Phi_{eq} \sim 10^{15}\,\mathrm{cm}^{-2}$. The induced charge increases further for bias voltages larger than V_{fd}. Higher electric field reduces the drift time and by that the influence of trapping term.

If the deep levels responsible for trapping are constant in time or change with a first order process (see Eq. (21.5)), then at any time after irradiation their concentration is linearly proportional to the fluence. Under this assumption Eq. (21.19) can be rewritten as

$$\frac{1}{\tau_{eff_{e,h}}} = \frac{1}{\tau_{eff0_{e,h}}} + \beta_{e,h}(t,T)\,\Phi_{\mathrm{eq}}, \tag{21.35}$$

where $\beta_{e,h}$ is called effective electron and hole trapping damage constant which depends on temperature, time after irradiation and irradiation particle. In detector grade silicon the effective trapping probability of a non-irradiated detector $\frac{1}{\tau_{eff0_{e,h}}}$ is negligible and is usually omitted from Eq. (21.35). Alternatively the trapping distance can be defined as

$$\lambda_{e,h} = \mu_{e,h}\,\tau_{eff_{e,h}}\,E \tag{21.36}$$

measuring the distance the carriers drift before being trapped.

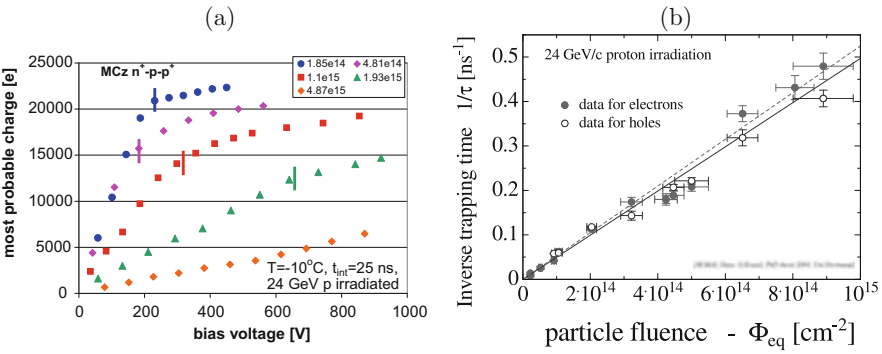

Fig. 21.25 (**a**) Dependence of induced charge on voltage for MCz p-type pad detector irradiated to different fluences. The V_{fd} for each measurement is denoted by vertical bar. (**b**) Effective trapping times of electrons and holes as found in Ref. [64]

The trapping times in silicon were systematically measured with Transient Current Technique [63]. The trapping probabilities for 23 GeV protons are shown in Fig. 21.25b. At $\Phi_{eq} \sim 10^{15}\,\mathrm{cm}^{-2}$ the effective trapping times are around few ns.

The trapping damage constant was studied as a function of different material properties: resistivity, oxygen concentration, carbon concentration, wafer production (MCz, FZ, epi-Si) and type of silicon (p-type or n-type). It was found, within the error margin, not to depend on any, thus being universal for silicon. The average values of β for neutrons and charged hadrons are given in the Table 21.5 [65]. It shows that the trapping probability for electrons is smaller than for holes. The NIEL hypothesis is slightly violated as charged hadrons produce more damage than reactor neutrons.

The evolution of trapping probability with time after irradiation is described in the simplest model by the decay of the dominant trap to another dominant trap (Eq. (21.3)) or a model with two traps one constant in time and one that decays. Both models can be described by the following equation [63]

$$\beta_{e,h}(t) = \beta_{0_{e,h}} \cdot e^{-\frac{t}{\tau_{ta,e,h}}} + \beta_{\infty_{e,h}} \cdot (1 - e^{-\frac{t}{\tau_{ta,e,h}}}) \tag{21.37}$$

with $\beta_{0_{e,h}}$ and $\beta_{\infty_{e,h}}$ the trapping rates at early and late annealing times, respectively. For the annealing temperatures of interest β_0 is very close to β measured at the end of short term annealing ($\beta(t_{min})$) given in Table 21.5. There is a distinctive difference between annealing of effective trapping times for holes and electrons. The trapping probability of holes increases with annealing time and that of electrons decreases (see Fig. 21.26) irrespective of material properties and type of irradiation

Table 21.5 Trapping time damage constants for neutron and fast charged hadron irradiated silicon detectors measured after the end of short term annealing [65]

t_{min}, $T = -10°\mathrm{C}$	β_h [$10^{-16}\,\mathrm{cm}^{-2}$/ns]	β_e [$10^{-16}\,\mathrm{cm}^{-2}$/ns]
Reactor neutrons	4.7 ± 1.2	3.5 ± 0.6
Fast charged hadrons	6.6 ± 1.1	5.3 ± 0.5

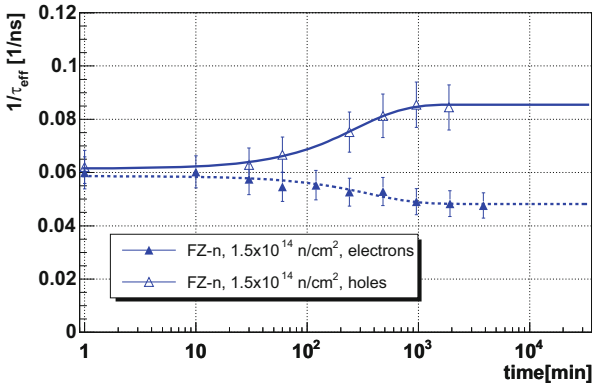

Fig. 21.26 Annealing of $1/\tau_{eff,e,h}$ for a detector irradiated with neutrons to $\Phi_{eq} = 1.5 \cdot 10^{14}\,\mathrm{cm}^{-2}$

Table 21.6 Parameters used to model annealing of effective trapping times

	τ_{ta} [min at 60 °C]	$(\beta_0 - \beta_\infty)/\beta_0)$	E_{ta} [eV]
Electrons	650±250	0.35±0.15	1.06 ± 0.1
Holes	530±250	0.4±0.2	0.98 ± 0.1

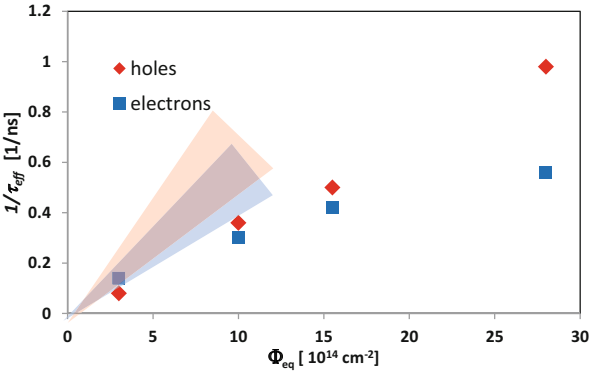

Fig. 21.27 Effective trapping probability measured at high fluences of charged hadrons [66]. The red and blue bands indicate the predictions of trapping probability of holes and electrons from Table 21.5

particle. The parameters describing annealing of effective trapping probabilities are shown in Table 21.6. The activation energy E_{ta} should be used in Eq. (21.9) for scaling τ_{ta} to different temperatures. The $\beta_{e,h}$ depends only moderately on temperature [63]. At temperatures of interest for most applications the trapping probabilities for both holes and electrons decrease with temperature by around 10–20% if the temperature changes from −20° to 20 °C.

The linear relation of Eq. (21.35) breaks down at equivalent fluences higher than $\sim 10^{15}\,\mathrm{cm}^{-2}$, where it starts to exhibit saturation. Unfortunately the TCT can not be directly used to measure trapping probabilities and values have be extracted by combining both TCT and CCE measurements with simulations. The study performed by CMS collaboration is shown in Fig. 21.27 [66]. It can be seen that already at few times $10^{15}\,\mathrm{cm}^{-2}$ the effective trapping probabilities deviate significantly from linear. Recently studies [67] showed that at extreme fluences of $\sim 10^{17}\,\mathrm{cm}^{-2}$ the trapping probability is around an order of magnitude smaller than predicted from the low fluence measurements.

21.4.2.5 Generation Current

The defects influencing the generation current (Eq. (21.22)) were found to either dissociate or are constant in time. The bulk damage-induced increase of the reverse current (ΔI) exhibits therefore a simple dependence on particle equivalent fluence at any time after irradiation

$$\Delta I_{gen} = \alpha(t, T)\, V\, \Phi_{eq}, \tag{21.38}$$

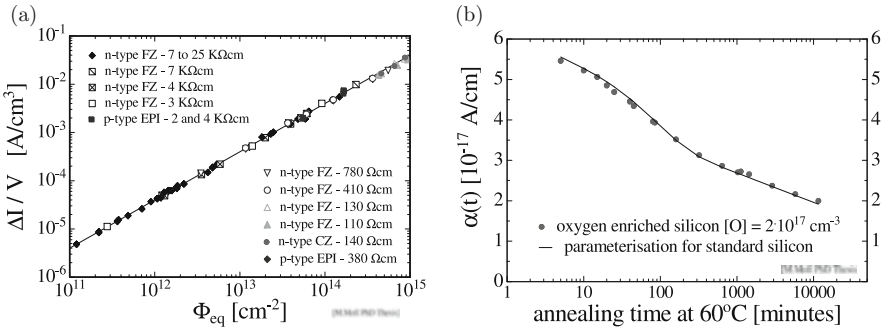

Fig. 21.28 (**a**) Dependence of bulk generation current on fluence for different detectors after 80 min storage at 60 °C. (**b**) Annealing of leakage current damage constant (after [31])

where V is the active volume ($V = S\,w$) and α the leakage current damage constant. The bulk generation current scales with NIEL, hence the leakage current damage constant is independent of the silicon properties and irradiation particle type as shown in Fig. 21.28a [68]. The measured value of the leakage current depends exponentially on the operating temperature as (see terms in Eq. (21.22))

$$I_{gen}(T) \propto T^2 \exp\left(-E_g/2k_BT\right), \tag{21.39}$$

and accordingly all α-values can be scaled to any temperature.

The damage induced bulk current undergoes also a temperature dependent beneficial annealing, described by

$$\alpha(t) = \alpha_1 \exp\left(-\frac{t}{\tau_\alpha}\right) + \alpha_0 - \alpha_2 \ln\left(\frac{t}{t_{norm}}\right), \tag{21.40}$$

with $\alpha_0 = 5.03 \cdot 10^{-17}$ A/cm, $\alpha_1 = 1.01 \cdot 10^{-17}$ A/cm, $\alpha_2 = 3.34 \cdot 10^{-18}$ A/cm, $\tau_\alpha =$ 93 min and $t_{norm} = 1$ min all measured at 60 °C. The first term in the Eq. (21.40) describes the decay of the defect and the second contribution of the defects constant in time. The last term is associated with the decay of the cluster, a conclusion based on its absence in ^{60}Co irradiations [68]. The leakage current annealing can be seen in Fig. 21.28b. Universality of the annealing described by Eq. (21.40) can be used to reliably monitor the equivalent fluence of particle sources even in cases of wide energy distributions. As a standard α(80 min at 60°C, 20°C) $= 4 \cdot 10^{-17}$ A cm^{-1} is used.

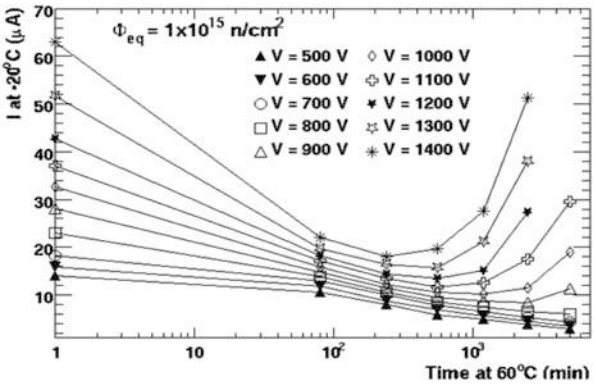

Fig. 21.29 Dependence of leakage current on annealing time at different voltages. The increase of leakage current with annealing is due to charge multiplication (from Ref. [62])

Leakage Current in Presence of Charge Multiplication

For devices with gain the leakage current is given by the current gain M[2] and generation current $I = M \cdot I_{gen}$. An example of the leakage current increase at high bias voltages during annealing is shown in Fig. 21.29. One should however be careful as the increase of leakage current at high bias voltages can also be attributed to other effects such as the onset of thermal runaway or rise of the surface current, however without clear increase of the collected charge.

21.4.2.6 Alternative Ways of Operation

The key reason for changes in performance of an irradiated detector are deep traps. The manipulation of their occupancy therefore has an influence on the detector properties. Variation of the operation temperature and/or concentration of free carriers can be used to change the occupancy of deep traps. The first observation of charge collection efficiency recovery after gradually cooling down the heavily irradiated silicon detector from room temperature to cryogenic temperatures (see Fig. 21.30a) was reported in [69] and referred to as "Lazarus effect". However the operation of silicon detectors under reverse bias turned out to be very sensitive to previous biasing conditions and ionizing particle rates. The signal varies with time after exposure to ionizing particles as shown in Fig. 21.30b. The trapping of the drifting carriers enhances the space charge of different signs at both detector contacts (see Sect. 21.4.2.2) to the point where the applied voltage is insufficient to establish the electric field in the entire detector. As a consequence the charge

[2]Current and charge gains can be in principle different, but have been so far observed to be very similar.

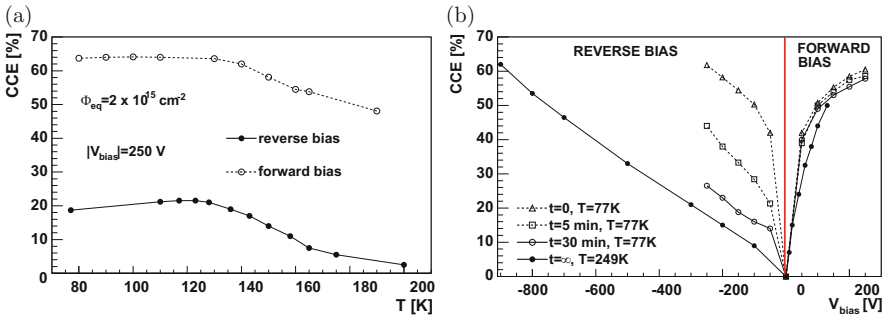

Fig. 21.30 (**a**) Charge collection efficiency in 400 μm thick detector irradiated to 10^{15} cm^{-2} in forward and reverse direction. (**b**) The dependence of CCE on voltage at $T = 77$ K in both forward and reverse direction of a detector irradiated to $2 \cdot 10^{15}$ cm^{-2}

collection efficiency is reduced. The phenomena of polarization of the detector by trapped charge is not unique to silicon and is present also in other semiconductors. Since emission times depend on $E_g/(2\,k_B\,T)$, silicon at cryogenic temperatures behaves similarly as wide band gap semiconductors at room temperature.

At cryogenic temperatures a more stable operation is achieved with detectors biased in forward direction [70] (see Fig. 21.30a,b). The resistivity of the bulk increases with irradiation and it effectively becomes a heavily doped insulator. Applied bias in forward direction injects carriers in the detector. These are trapped at deep levels and affect the electric field. The predominately negative space charge is naturally compensated by injection of holes. The electric field grows from $E \approx 0$ at the injection point towards the other contact with the square root of the distance x from the injecting junction [71]

$$E(x) = \frac{3\,V}{2\,W}\sqrt{\frac{x}{W}}. \tag{21.41}$$

The electric field extends through the entire detector thickness regardless of the applied voltage or concentration of the deep levels. This is an important advantage over the biasing of detectors in reverse polarity. The drawback of forward bias operation is the increased current, requiring intensive cooling. The current dependence on voltage is quadratic ($I \propto V^2$), followed by a sharp rise at threshold voltage V_T as shown in Fig. 21.31. It happens when the space charge saturates due to filling all the traps and current can not be limited by increasing the concentration of the trapped carriers, therefore $V_T \propto \Phi_{eq}$. An important feature of this mode of operation is the fact that the current at a given voltage progressively decreases with fluence (see Fig. 21.31), approximately as $I(\Phi_{eq}) \propto \Phi_{\mathrm{eq}}^{-1.5}$. The larger the concentration of traps the smaller is the current which is needed to adjust the electric field. Nevertheless, it is still larger than in reverse direction.

In principle, a $p^+ - n - n^+$ structure should inject holes and electrons, which would not produce the aforementioned properties. However it turns out that at n^+

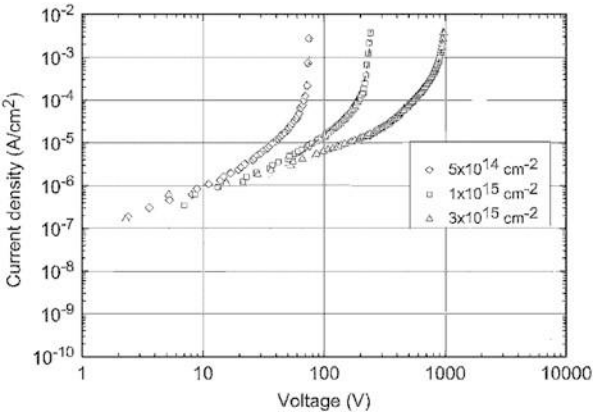

Fig. 21.31 Leakage current-voltage characteristics in forward mode of operation

contacts electrons are not injected [71]. The symmetric structure $p^+ - n - p^+$, where only holes are injected, has the same properties pointing to the same underlying physics process. The same condition of carrier injection can be also achieved in reverse bias mode by continuous illumination of one side by light of short penetration depth [72]. The injected carries establish the same condition as under forward bias and the Eq. (21.41) applies.

The filling of deep levels affects effective trapping probabilities of electrons and holes. Measurements have shown that the same charge collection efficiency is achieved as for a fully depleted detector at few times smaller bias voltage [70, 73] (see Fig. 21.30b). Smaller bias results in smaller average electric field and therefore longer collection times. As the reduction of charge collection efficiency depends in first approximation on the ratio of the drift time to the trapping time of the carriers, the latter must be longer than under the reverse bias.

It is obvious that forward bias operation mode becomes usable once the detectors are already heavily irradiated. There are two ways of how to use detectors in real experiments. With read-out electronics sensitive to both polarities detectors can be first used in reverse and later in forward direction or the detectors are irradiated before being used. In general the use of the forward bias means replacing the problem of the high voltage required for the reverse bias operation by the problem of a high dark current. Therefore detectors with small element size (i.e. pixels) are more suitable for this mode of operation.

21.4.2.7 3D Detectors—A Radiation Harder Detector Design

One approach to address the issue of radiation damage are optimized detector geometries. A good example of radiation hard detector design are so called 3D detectors. An schematic view of such detector is shown in Fig. 21.32 [74]. The

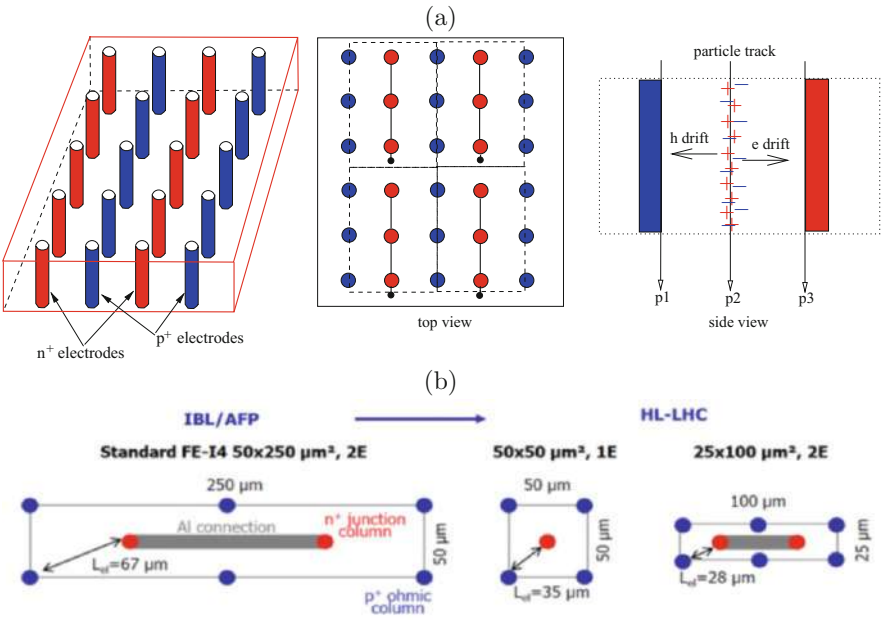

Fig. 21.32 (**a**) The schematic view of the 3D detector (left). The view of the detector surface (middle); gray n^+ electrodes, dark gray p^+ electrodes, black metal line, black dot the bump-bond. The dashed line marks the pixel cell with three columns. Q_e^t/Q^t for different tracks: p1=1, p2=0.5 in p3=0 (right). (**b**) Layout of a single cell/pixel of an IBL 3D detector (2 electrode configuration—2E) and of HL-LHC detector with both options 1E and 2E. The maximum drift length of carriers is indicated

electrodes in such detectors are perpendicular to the surface. Such placement of electrodes has two beneficiary effects for heavily irradiated detectors. The small distance between the electrodes effectively reduces the full depletion voltage. Even more importantly, the drift length of carriers is reduced and therefore the probability of drifting carriers to get trapped ($\tau_{eff,e,h} \gg t_{drift}$). As the signal (number of e-h pairs in Eq. (21.29)) is determined by the detector thickness, vertical electrode configuration ensures good charge collection at moderate voltages. Several columns can be connected together to form pixel cells or strips (Fig. 21.32b). The thickness of the detector is limited by the deep reactive ion etching process used to produce holes. The standard aspect ratio (hole length/hole diameter) is around 24. Apart from a more complex processing, which can be simplified by electrodes not penetrating fully the detector [75, 76], there are some drawbacks of the 3D design:

- Reducing the inter-column spacing results in higher inter-electrode capacitance
- Columnar electrodes are a non-active part of the detector volume and can lead to particle detection inefficiency; most of the tracks in experiments are, however, inclined which mitigates the problem.

- Induced charge depends on the hit position of the particle track. Unlike in planar detectors, the ratio Q_e^t/Q^t varies between 0 and 1 across the detector and can affect the position resolution and efficiency (see Fig. 21.32c).

Nevertheless, these detectors are often first choice for tracking detectors at highest fluences. ATLAS pixel detector (Insertable B Layer—IBL) [77] saw the first application of 3D detectors for tracking in high energy experiments, covering 25% of the total IBL surface at both sides of the staves. The 3D technology is improving with different ways of processing the detectors with single-sided process or more elaborate double sided processing with possibility of active/slim edges reducing the inactive part at the detector border. Efficient charge collection was achieved also for sensors where columns don't penetrate the whole depth. Such a design improves the yield of sensor production. The latter remains one of the main concerns for 3D technology reaching around 50–60% for the IBL module production [78].

At HL-LHC the 3D detectors are planned for the first pixel layer. A small cell size will have a single junction column (cell $50 \times 50\,\mu m^2$) or two columns (cell $25 \times 100\,\mu m^2$), where the maximum drift distances will be reduced to mere 37 and 28 μm making these detectors extremely radiation hard. The first beam tests with such 230 μm thick detectors showed [79] 97% detection efficiency for perpendicular tracks after extreme fluences of $2.5 \cdot 10^{16}\,cm^{-2}$ at >200 V using IBL readout electronics (FE-I4) [80].

21.4.2.8 Timing Detectors

At HL-LHC coping with large particle fluxes emerging from collisions will be an enormous challenge. On average 200 p-p collisions will occur every 25 ns, with collision points distributed normally along the beam with $\sigma_z = 5$ cm and in time with $\sigma_t = 180$ ps. Resulting track and jet densities in the detector complicate the analysis of the underlying reactions that took place. A way to cope with that problem is separation of individual collisions also in terms of time of occurrence within each bunch crossing. This is particularly important for tracks/jets in forward direction for which the position resolution of primary vertex is much worse (∼1 mm). If tracks are not resolved in time, this can lead to false vertex merging. A timing resolution of around 30 ps with respect to the HL-LHC clock is required to successfully cope with pileup. Such an outstanding single particle timing resolution was up to recently impossible with silicon detectors.

Three factors determine the timing resolution of each sensor: time walk which is a consequence of non-homogeneous charge deposition by an impinging particle, noise jitter ($\sigma_{jitter} = t_{rise}/(S/N)$) and resolution of time-to-digital conversion. Standard silicon detectors of 300 μm are not appropriate for precise timing measurement as the integration time to collected all the charge and consequent rise time t_{rise} are large, hence the jitter. In addition fluctuations, not only of the amount of the

charge (time walk correctable by e.g. constant fraction discrimination), but also of the deposition pattern (non-correctable time walk)—so called Landau fluctuations—ultimately limit the time resolution to effectively >100 ps [81]. High enough signal-to-noise S/N in thin detectors can be achieved by using so called Low Gain Avalanche Detectors (LGAD) [82].

They are based on a $n^{++} - p^{+} - p - p^{++}$ structure where an appropriate doping of the multiplication layer (p^{+}) leads to high enough electric fields for impact ionization (see Fig. 21.33) [82]. Gain factors in charge of few tens significantly improve the resolution of timing measurements, particularly for thin detectors. The main obstacle for their operation is the decrease of gain with irradiation, attributed to effective acceptor removal in the gain layer [41]. A comprehensive review of time measurements with LGADs is given in Ref. [83].

The most probable charge in 50 μm and 80 μm thick pad devices before and after irradiation is shown in Fig. 21.34a. As soon as multiplication layer is depleted the gain appears. At lower fluences the gain degradation at the depletion of multiplication layer (around 40 V) can be clearly seen. At higher fluences the gain appears at high bias voltages where over-depletion ensures that high enough electrical fields are reached; above $\Phi_{eq} > 10^{15}\,\mathrm{cm}^{-2}$ the onset of multiplication is observed only at highest voltages of around 700 V. Such voltages correspond to very high average fields of 15 V/μm. At fluences $\Phi_{eq} > 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$ the beneficial effect of multiplication layer is gone. The devices of the same design without multiplication layer show similar behavior as LGADs. The time resolution of LGADs was extensively measured in the test beams [84] and with ^{90}Sr electrons. It is shown in Fig. 21.34b for the 50 μm thick non-irradiated devices.

At very large fluences of $\Phi_{eq} > 2 \cdot 10^{15}\,\mathrm{cm}^{-2}$ the gain, although lower than the initial, appears due to deep traps (see section on charge multiplication)

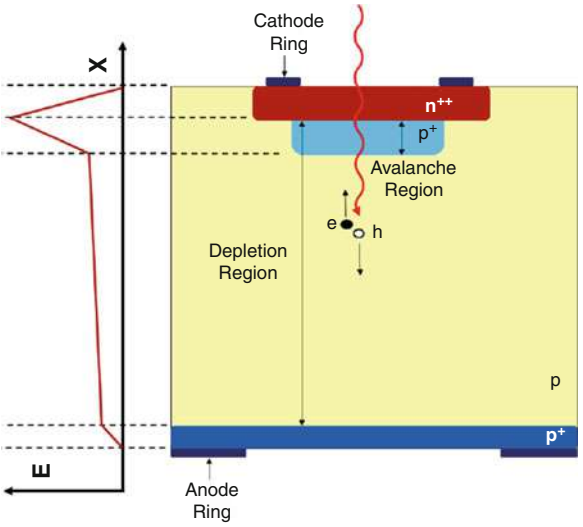

Fig. 21.33 Schematic view of the Low Gain Avalanche Detector

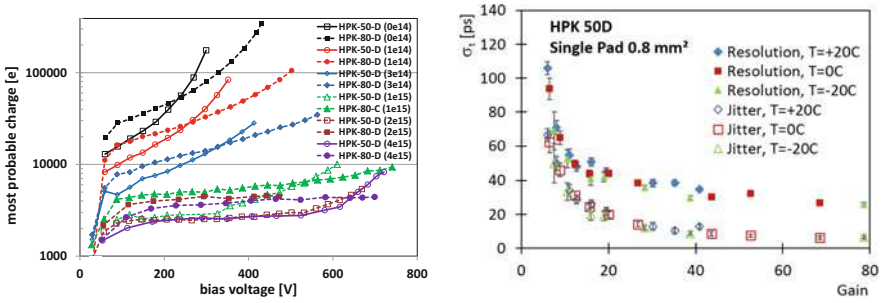

Fig. 21.34 Dependence of most probable charge for irradiated LGAD devices on voltage for different thickness (50 and 80 μm). Fluences in the brackets are in [cm^{-2}] [85]. Around 3000 e is expected for a 50 μm device without gain layer. (**b**) Time resolution and its noise jitter contribution measured for the non-irradiated 50 μm detector at different temperatures [86]

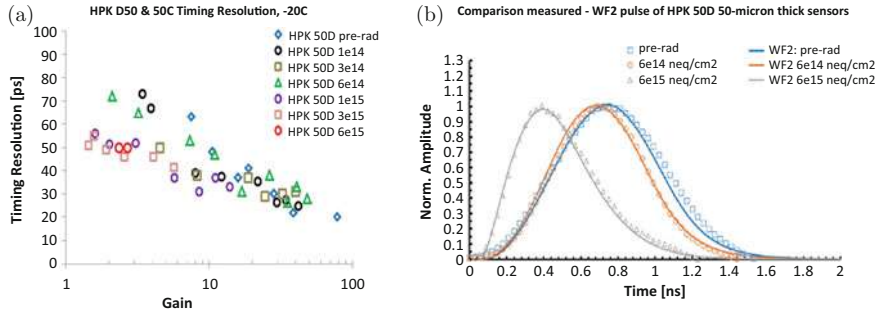

Fig. 21.35 (**a**) Time resolution of irradiated LGAD detectors at different gains and fluences ([cm^{-2}]). (**b**) Measured and simulated induced current pulse shape at different irradiation levels. Note that amplitudes were normalized to one

and the timing resolution degrades only moderately (see Fig. 21.35a). Moreover, multiplication in larger volume of the bulk results in faster rise time of the induced current which at given gain leads to better timing resolution (see Fig. 21.35b).

The leakage current in LGADs follows the same equation as discussed in section on charge multiplication. Hence, the gain can be calculated from measurement of leakage current and calculated generation current [85, 86].

A lot of effort was spend in recent years to increase the radiation hardness of LGADs by mitigating the acceptor removal. The efforts concentrated to use of co-implantation of carbon [87] to multiplication layer aiming to reduce the removal constant or replacing boron with gallium, which should be more difficult to displace [87, 88].

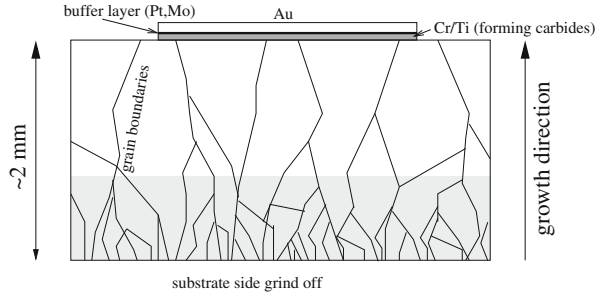

Fig. 21.36 Schematic view of the pCVD diamond detector

21.4.3 Diamond Detectors

Although the specific ionization in diamond detectors is around three times smaller than in silicon, larger detector thickness, small dielectric constant, high break down voltage and negligible leakage current make them the most viable replacement for silicon in the highest radiation fields.

The intrinsic concentration of carriers in diamond is extremely low (good insulator, $\rho > 10^{16}$ Ωcm). The detectors are therefore made from intrinsic diamond metallized at the back and the front (see Fig. 21.36) to form ohmic contacts. Most of the diamond detectors are made from poly-crystalline diamond grown with chemical vapor deposition technique (CVD). Recently also single crystalline (scCVD) detectors have become available. The quality of the poly-crystalline (pCVD) diamond as a particle detector depends on the grain size. The grains in this material are columnar, being smallest on the substrate side, and increasing in size approximately linearly with film thickness.[3] Crystal faults at the boundaries between the grains give rise to states in the band gap acting like trapping centers.

A widely used figure of merit for diamond is its charge collection distance (CCD), which is defined as

$$CCD = \frac{Q_t}{\rho_{e-h}}. \tag{21.42}$$

The CCD represents the average distance over which carriers drift. If $CCD \ll W$, it is equivalent to the trapping distance $\lambda_e + \lambda_h$.[4]. After irradiation, and for pCVD detectors also before irradiation, the CCD depends on electric field, due to reduced probability for charge trapping at larger drift velocity. Only for non-irradiated scCVD detectors $\lambda_{e,h} \rightarrow \infty$ and the $CCD = W$ regardless of the bias voltage

[3]For this reason, many detectors have the substrate side etched or polished away.

[4]The exact relation between the charge collection and the trapping distance is: $CCD = \lambda_e\,[1 - \frac{\lambda_e}{W}(1 - exp(-\frac{W}{\lambda_e}))] + \lambda_h\,[1 - \frac{\lambda_h}{W}(1 - exp(-\frac{W}{\lambda_h}))]$.

applied. Most commonly, the CCD is defined at $E = 1\,\mathrm{V/\mu m}$ or $E = 2\,\mathrm{V/\mu m}$, although sometimes also CCD at saturated drift velocity is stated.

The CCD of typically 500 μm thick diamond detectors has improved tremendously over the last 20 years. Current state of the art pCVD detectors reach up to 300 μm at 2 V/μm and are available from 6 inch wafers.

21.4.3.1 Radiation Hardness

In pCVD detector the leakage current does not increase with irradiation; moreover it may even decrease, which is explained by passivation of defects at grain boundaries. The current density in high quality pCVD diamond is of order 1 pA/cm^2, a value strongly dependent on the quality of metallized contacts.

Irradiation decreases the CCD for both scCVD and pCVD diamonds with similar rate [89], pointing to the in-grain defects being responsible. It has been observed that exposing such an irradiated detector to ionizing radiation (10^{10} minimum ionizing particles/cm^2) improves the charge collection efficiency of pCVD detectors by few 10%. This process is often called "pumping" or priming. The ionizing radiation fills the traps. The occupied traps become inactive, hence the effective trapping probability decreases. The traps can remain occupied for months due to large emission rates if kept in the dark at room temperatures. Once detectors are under bias the ionizing radiation leads to polarization of detectors, in the same way as in silicon, but with the polarization persisting over much longer times. The measurements of charge collection can therefore depend on previous biasing condition and relatively long times are needed to reach steady state of operation.

The irradiation decreases the trapping distance of electrons and holes proportionally to the fluence. The relation can be derived by inserting the effective trapping time (Eq. (21.35)) in the expression for the trapping distance (Eq. (21.36)):

$$\frac{1}{\lambda_{e,h}} = \frac{1}{\lambda_{0_{e,h}}} + K_{e,h} \cdot \Phi, \tag{21.43}$$

where λ_0 denotes the trapping distance of an unirradiated detector and $K_{e,h}$ the damage constant. Assuming $\lambda_e \approx \lambda_h$ and $\lambda_e + \lambda_h \ll W$ for simplicity reasons, CCD dependence on fluence can be calculated as:

$$\frac{1}{CCD} \approx \frac{1}{CCD_0} + K\,\Phi. \tag{21.44}$$

Although only approximate the Eq. (21.44) fits the measurements well over a large fluence range as shown in Fig. 21.37a. The extracted damage constant K ($\sim 1/2\,K_{e,h}$) from source and test beam data for particles of different energy and spectrum are gathered in Table 21.7. For high fluences the second term in Eq. (21.44) prevails and the scCVD and pCVD diamonds perform similarly. At $\Phi = 2 \cdot 10^{16}\,\mathrm{cm}^{-2}$ of 23 GeV protons the $CCD \approx 75$ μm which corresponds

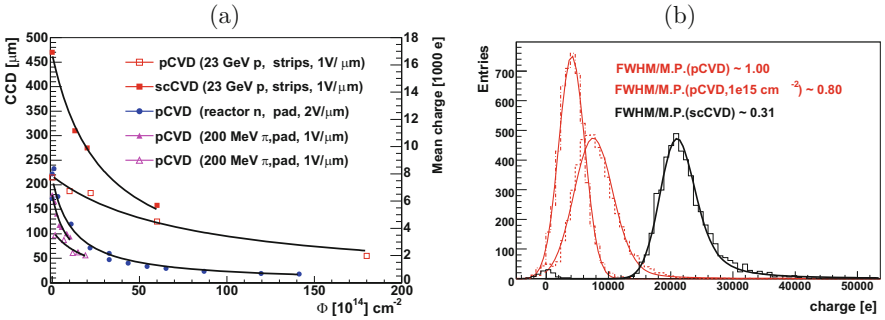

Fig. 21.37 (**a**) CCD vs. fluence of different particles. Detectors were 500 μm thick. The Eq. (21.44) is fitted to the data [89–91]. The irradiation particle, electrode geometry and electric field is given in brackets. (**b**) Energy loss distribution in CVD pad detectors. The value of FWHM, corrected for electronics noise, over most probable energy loss is shown

Table 21.7 Charge collection distance degradation parameter for different irradiation particles [89–92]

	70 MeV	800 MeV	23 GeV p	200 MeV π	Reactor neutrons
$K[10^{-18}\,\mu m^{-1}cm^{-2}]$	1.76	1.21	0.65	~3.5	~3 – 4

to mean charge of 2770 e_0. At lower fluences the first term dominates and for $CCD_0 \sim 200\ \mu m$ CCD only decreases by 15% after 10^{15} cm^{-2} of 23 GeV protons.

The homogeneity of the response over the detector surface, which is one of the drawbacks of pCVD detectors, improves with fluence for pCVD as the collection distance becomes smaller than grain size. For the same reason also the distribution of energy loss in pCVD detector, initially wider than in scCVD, becomes narrower (Fig. 21.37b). The energy loss distribution in pCVD diamond is Gaussian, due to convolution of energy loss distributions (Landau) in grains of different sizes.

One of the main advantages of the diamond is the fact that at close to room temperatures no annealing or reverse-annealing effects were observed.

The drawbacks of grains in pCVD detectors can be largely overcome by using **3D diamond detectors** [93], who share the same concept with silicon detectors (see Fig. 21.38). The vertical electrodes are produced by focused laser light which graphitizes the diamond. Whether the vertical electrode serves as cathode or anode depends on metal bias grid on the surface of the detector. Very narrow electrodes of ~2 μm diameter can be made along 500 μm thick device with low enough resistivity to allow good contacts and doesn't increase the noise. Such a good aspect ratio allows even smaller cell sizes than in silicon.

The first tests showed >75% charge collection efficiency in 500 μm pCVD diamond detector of ganged 150 × 150 μm^2 cells with bias voltages of only few tens Volts [94]. A much better homogeneity of charge collection over the surface (columns are parallel to grain boundaries) and narrower distributions of collected charge were obtained than in planar diamond detectors.

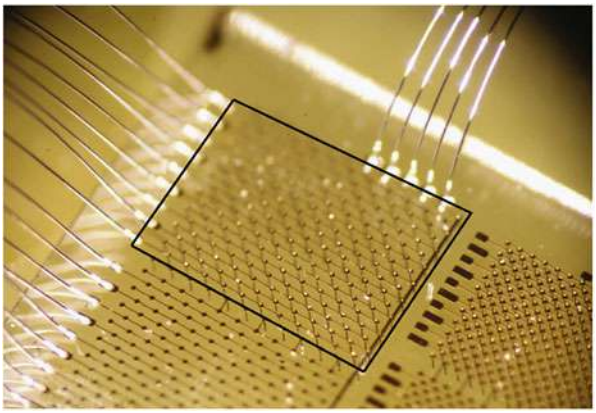

Fig. 21.38 Photograph of 3D diamond strip detector in black rectangle with a square cell of 150 μm size. Strip detector of the same geometry with planar electrodes is shown below (from [93])

The main problem with diamond 3D detectors is the rate of production in particular for large area as even if laser beam is powerful enough and is split into several parallel beams. Currently the rate is limited to roughly ten thousand holes a day.

The diamond detectors are used also outside particle physics for particle detection such as for fusion monitoring where neutrons are detected, for alpha particle detection, for determination of energy and temporal distribution of proton beams and in detection of ions during the teleradiology. They are also exploited for soft X-ray detection, where the solar-blindness and fast response of diamond detectors are the keys of their success.

21.4.4 Other Semiconductor Materials

Silicon is in many respects far superior to any other semiconductor material in terms of collected charge, homogeneity of the response and industrial availability. Other semiconductor materials can only compete in niche applications where at least one of their properties is considerably superior or where the existing silicon detectors cannot be used. For example, if low mass is needed or active cooling can not be provided, high leakage current in heavily irradiated silicon detectors is intolerable and other semiconductor detectors must be used.

The growth of compound semiconductors is prone to growth defects which are frequently unmanageable and determine the properties of detectors before and after irradiation. If a high enough resistivity can be achieved, the detector structure can be made with ohmic contacts. However, it is more often that either a Schottky contact or a rectifying junction is used to deplete the detector of free carriers. Only

a few compound semiconductors have reached a development adequate for particle detectors. They are listed in the Table 21.2. For particle physics application some other very high-Z semiconductor such as CdZnTe or HgI_2 are inappropriate due to large radiation length.

Silicon Carbide was one of the first alternatives to silicon proposed in [95, 96]. It is grown as epitaxial layer or as bulk material. Even though at present the latter exhibits a lot of dislocations (inclusions, voids and particularly micro-pipes) in the growth and the former is limited to thicknesses around 50 μm, both growth techniques are developing rapidly and wafers are available in large diameters (10 inch). Due to the properties similar to diamond the same considerations apply as for diamond with an important advantage of 1.4 larger specific ionization (55 e-h/μm).

Presently the best performing detectors are produced by using slightly n-doped epitaxial layers of ≈50 μm forming a Schottky junction.They exhibit 100% charge collection efficiency after full depletion and negligible leakage current [97]. Also detectors processed on semi-insulating bulk (resistivities $\sim 10^{11}$ Ωcm) with the ohmic contacts show CCD up to 40 μm [96] at 1 V/μm for few hundred μm thick material.

After irradiation with hadrons the charge collection deteriorates more than in silicon or in diamond. For epitaxially grown SiC the degradation of CCD is substantial with $K_e \approx 20 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ and $K_h \approx 9 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ for reactor neutron and 23 GeV p irradiated samples at high electric fields of 10 V/μm [97, 98]. The leakage current is unaffected by irradiation or it even decreases [98].

GaAs The resistivity of GaAs wafers is not high enough for the operation with ohmic contacts and detectors need to be depleted of free carriers, which is achieved by Schottky contact or a $p-n$ junction. GaAs detectors were shown to be radiation hard for γ-rays (^{60}Co) up to 1 MGy [99]. As a high Z material these detectors are very suitable for detection X and γ rays.

Their tolerance to hadron fluences is however limited by loss of charge collection efficiency, which is entirely due to trapping of holes and electrons. The V_{fd} decreases with fluence [100, 101] which is explained by removal or compensation of as grown defects by irradiation. Although larger before the irradiation, the trapping distance of electrons shows a larger decrease with fluence than the trapping distance of holes. The degradation of charge collection distance at an average field of 1 V/μm (close to saturation velocity) in 200 μm thick detectors is very large $K_{e,\pi} \approx 30 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ and $K_{h,\pi} \approx 150 \cdot 10^{-18}\,\mathrm{cm^2/\mu m}$ [100]. One should however take into account that specific ionization in GaAs is four times larger than in diamond.

The leakage current increases moderately with fluence up to few $10\,\mathrm{nA/cm^2}$, much less than in silicon, and starts to saturate at fluences of around $10^{14}\,\mathrm{cm^{-2}}$ [100, 101]. The GaAs exhibit no beneficial nor reverse annealing of any detector property at near to room temperatures.

GaN The GaN detectors produced on few μm thin epitaxial layer shown charge collection degradation which is much larger than in Si [102]. Further developments in crystal growth may reveal the potential of material.

21.4.5 Comparison of Charge Collection for Different Detectors

The key parameter relevant to all the semiconductor particle detectors is the measured induced charge after passage of minimum ionizing particles. The charge collection dependence on fluence in different semiconductor pad detectors is shown in Fig. 21.39a. A 3D pad detector (all columnar electrodes connected together) shows best performance, while smallest charge is induced in SiC and pCVD diamond detectors. The induced charge decreases with fluence and at most few thousand e_0 can be expected at $\Phi_{eq} > 10^{16}\,\mathrm{cm}^{-2}$.

Although pad detectors are suitable for material comparison the effect of segmentation and choice of the type of the read-out electrodes determine to a large extent the performance of the detectors. The superior charge collection performance of segmented silicon planar detectors with n^+ electrodes to pad detectors can be seen in Fig. 21.39b. A signal of around 7000 e_0 is induced in epitaxial p-type and Fz p-type strip detectors at $\Phi_{eq} = 10^{16}\,\mathrm{cm}^{-2}$. At the highest fluence shown the signal in a silicon pad detector is only half of that in a strip detector. On the other hand a device with p^+ readout performs worst of all.

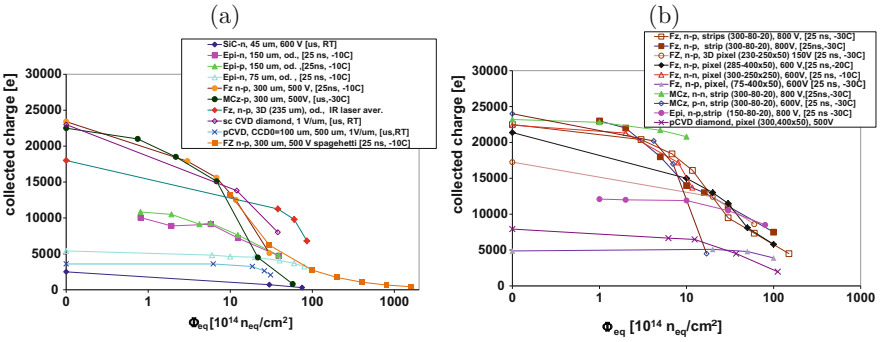

Fig. 21.39 (**a**) Comparison of charge collection in different detectors and materials; given are material, thickness, voltage, [shaping time of electronics and temperature]. "od." means at $V_{bias} > V_{fd}$. All detectors were irradiated with 23 GeV protons, except 75 μm epi-Si and 300 μm thick "spaghetti" diode which where irradiated with reactor neutrons. For diamond detectors the mean, not the most probable, charge is shown. (**b**) Charge collection in different segmented devices; the segmentation is denoted for strips (thickness-pitch-width) and pixels (thickness-cell size)]. Solid markers denote neutron irradiated and open 23 GeV proton irradiated samples

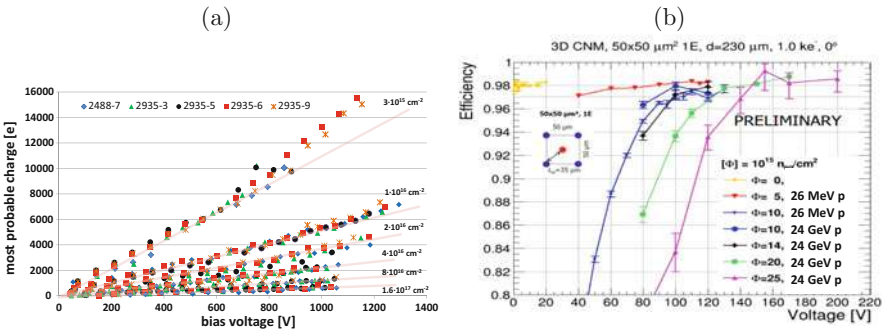

Fig. 21.40 (**a**) Dependence of collected charge in different planar silicon detectors on voltage up to the extreme fluences at −10 °C. The color bands are to guide the eye. (**b**) Test beam (120 GeV π) measurement of detection efficiency in heavily irradiated 3D detector with single electrode cell of $50 \times 50 \mu m^2$ shown in the inset [79]

The detection efficiency, however, depends on signal-to-noise ratio, which should be maximized. A choice of material, electrode geometry and thickness determine the electrode capacitance which influences the noise of the connected amplifier (Chap. 10). At given pixel/strip geometry the highest electrode capacitance has a 3D silicon detector, followed by a planar detector with n^+ electrodes, due to required p-spray or p-stop isolation which increases the inter-electrode capacitance. Even smaller is the capacitance of p^+ electrodes which is of 1 pF/cm order for strip detectors. The smallest capacitance is reached for diamond detectors owing to small dielectric constant.

21.4.5.1 Operation at Extreme Fluences

A combination of trapping times saturation, active neutral bulk and charge multiplication allows silicon detectors to be efficient in radiation environments even harsher than that of HL-LHC, approaching those of FCC. The operation of silicon detectors was tested up to $\Phi_{eq} = 1.6 \cdot 10^{17}\,\mathrm{cm}^{-2}$ and is shown in Fig. 21.40a for short strip detectors with ganged electrodes ("spaghetti" diode). Detectors remained operational and most probable charge of around 1000 e was measured in 300 µm thick detectors at 1000 V. At high fluences ($>2 \cdot 10^{15}\,\mathrm{cm}^{-2}$) the collected charge is linearly proportional with o bias voltage in whole range of applicable voltages and the dependence of charge on voltage and fluence can be parametrized with only two free parameters [103],

$$Q(V, \Phi_{eq}) = k \cdot V \cdot \left(\frac{\Phi_{eq}}{10^{15}\ \mathrm{cm}^{-2}}\right)^{b}, \tag{21.45}$$

where $b = -0.683$ and $k = 26.4$ e/V for 300 µm thick detectors.

A small cell size 3D detector ($50 \times 50\mu m^2$, 1E) irradiated with charged hadrons to $\Phi_{eq} = 2.5 \cdot 10^{16}\,cm^{-2}$ was recently found to be fully efficient at voltages even below 200 V (see Fig. 21.40b) [79]. A rough simulation of collected charge in such a device based on known data predicts collected charge $>3000\ e_0$ after the fluence of $10^{17}\,cm^{-2}$, which may be already enough also for successful tracking.

21.4.6 Radiation Damage of Monolithic Pixel Detectors

The monolithic pixel detectors, which combine active element and at least first amplification stage on the same die, are widely used in x-ray and visible imaging applications. Their use as particle detectors is limited for applications where radiation environments are less severe (space applications, $e^+ - e^-$ colliders), either because of small hadron fluences or because of radiation fields dominated by leptons and photons (see Fig. 21.3). The CCD is the most mature technology while CMOS active pixel sensors were successfully used for particle detection in STAR experiment at Relativistic Heavy Ion Collider over the last decade. These detectors are more susceptible to radiation damage due to their charge collection mechanism and the readout cycle. Recently several CMOS foundries offered a possibility to apply high voltage which can be used for depletion of substrate on which CMOS circuitry resides thereby enabling fast charge collection by drift. This greatly enhanced both radiation hardness of CMOS detectors and their speed.

The principles of operation of these detectors were addressed in section on Solid state detectors. Here on only the aspects of radiation hardness of aforementioned detectors will be addressed.

21.4.6.1 CCDs

The CCD[5] is intrinsically radiation soft. The transfer of the charge through the potential wells of the parallel and serial register is very much affected by the charge loss. At each transfer the fraction of the charge is lost. The charge collection efficiency is therefore calculated as $CCE = (1 - CTI_p)^n \times (1 - CTI_s)^m$, where CTI_s and CTI_p denote the charge collection inefficiency of each transfer in serial and parallel register. An obvious way of improving the CCE is a reduction of the number of transfers (m and n). Applications requiring high speed such as ILC, where the readout of the entire detector ($n \sim 2500$) within 50 μs is neeeded, the serial register is even omitted ($m = 0$) and each column is read-out separately (column parallel CCD [104]). The CCDs suffer from both surface and bulk damage.

[5]CCD is often not considered to be monolithic devices.

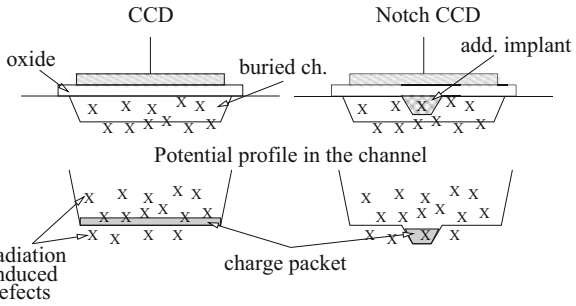

Fig. 21.41 The principle of notch CCD. An additional n^+ implant creates the minimum in potential

The increase of CTI is a consequence of bulk and interface traps. There are several methods to improve the CTI:

- The transport of the charge takes place several hundred nm away from the surface by using a n^+ implant (buried channel), which shifts the potential minimum. The transport is less affected by trapping/detrapping process than at the interface traps.
- Operation of CCDs at low temperatures leads to filling of the traps with carriers—electrons. Since emission times are long (Eq. (21.16)) the amount of active traps is reduced.
- If the density of signal electrons (n_s) is larger than the trap concentration only a limited amount of electrons can be trapped, thus $CTI \propto N_t/n_s$. An additional n^+ implant can be used for buried channel CCDs to squeeze the potential minimum to much smaller volume (see Fig. 21.41).
- The CTI depends on the charge transfer timing, i.e. on the clock shapes. The transfer of the charge from one pixel to another should be as fast as possible to reduce the trapping. The choice of the clock frequency, number of the phases (2 or 3) and shape of the pulses, which all affect the CTI, is a matter of optimization (see Fig. 21.42). The transfer time from one well to another can be enhanced by an implant profile which establishes gradient of the electric field.
- If traps are already filled upon arrival of the signal charge they are inactive and CTI decreases. The effect can be achieved either by deliberate injection of charges (dark charge) or by exploiting the leakage current. In the same way also the pixel occupancy affects the CTI.

The radiation affects also operation of detectors due to surface effects. The surface generation current which is a consequence of interface traps is in most of the applications the dominant source of current in modern CCDs. Very rarely the bulk damage is so high that the bulk generation current dominates. The surface dark current can be greatly suppressed by inverse biasing of the Si-SiO_2 interface [105].

The voltage shift due to oxide charge requires proper adjustment of the amplitude of the gate drive voltages. However the supply current and power dissipation of the gate drivers can exceed the maximum one as they both depend on the square of the voltage amplitude.

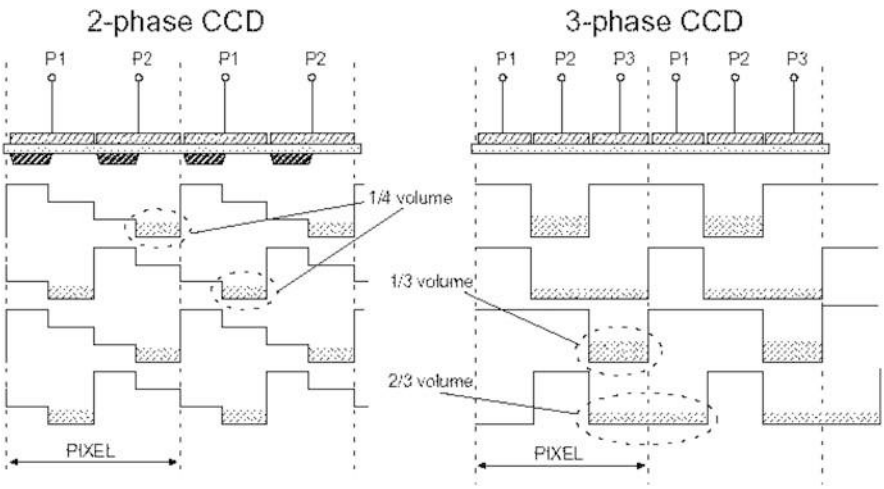

Fig. 21.42 Comparison of charge transfer of 2 and 3 phase CCD (P1,P2,P3 denote gate drive voltages). Note that the potential well occupies 1/4 of the pixel volume for 2 phase CCD and up to 2/3 for 3-phase CCD. The signal charge is shaded

The CCDs can probably not sustain radiation fields larger than at the ILC (see Table 21.1), particularly because of the bulk damage caused by neutrons and high energy leptons.

21.4.6.2 Active CMOS Pixels

In conventional monolithic active CMOS pixel sensors (Chap. 5) [106, 107] the n^+ well collects electron hole pairs generated by an ionizing particle in the p doped epitaxial layer (see Fig. 21.43). The built-in depletion around the n^+ well is formed enabling the drift of the carriers. In the major part of the detector the charge is collected from epitaxial p-type silicon through the diffusion. The charge collection process depends on epitaxial layer thickness and takes tens of ns. Above 90% of the cluster charge is induced within ~100 ns for 15 μm thick epitaxial layer [108]. Since the n^+ wells are used as collection electrodes, only nMOS transistors can be used for the signal processing circuit. The level of complexity of signal processing after the first stage depends on the CMOS technology used (number of metal layers, feature size).

The charge collection by thermal diffusion is very sensitive to electrons lifetime, which decreases due to the recombination at deep levels. The loss of collection efficiency and consequently smaller signal-to-noise ratio is the key limitation for their use. The way to increase the radiation tolerance is therefore the reduction of diffusion paths. This can be achieved by using many n^+ collection diodes per pixel area, which improves the charge collection efficiency. The price for that is a larger

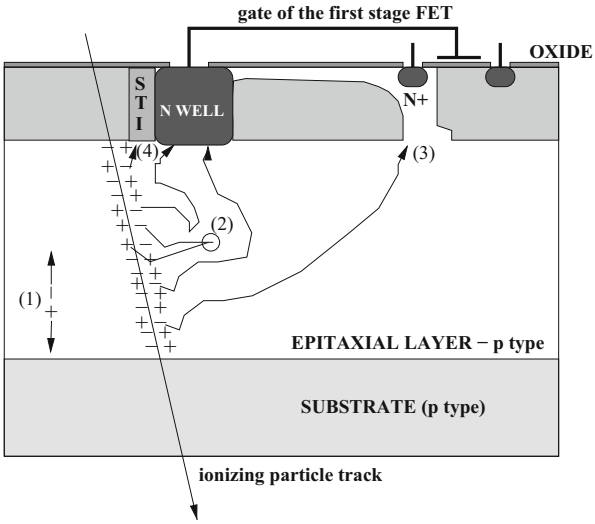

Fig. 21.43 Schematic view of the radiation effects in active CMOS Pixel Detector: (1) generation of carriers—leakage current; (2) recombination of diffusing carriers; (3) positive oxide charge buildup leads to punch trough the p well; (4) charge trapping at shallow trench isolation structure

capacitance and leakage current of the pixel. An increase of the epitaxial layer will increase the fraction of recombined charge, but the absolute collected charge will nevertheless be larger. The reduction of the collection time can be achieved by a gradual change of epitaxial layer doping concentration which establishes electric field.

The generation-recombination centers give rise to the current and cooling is needed to suppress it. It increases the noise and requires more frequent reset of the pixel.

The active pixel detectors were proven to achieve detection efficiencies of >95% at $\Phi_{eq} = 2 \cdot 10^{12}\,\mathrm{cm}^{-2}$ [109], suggesting an upper limit of radiation tolerance to hadron fluences of $\Phi_{eq} \leq \cdot 10^{13}\,\mathrm{cm}^{-2}$.

The active pixel sensors are CMOS circuits and therefore susceptible to surface damage effects. Apart from the damage to transistor circuitry which is discussed in next section in some CMOS processes the n-well is isolated from p-well by shallow trench SiO_2 isolation. The radiation induced interface states serve as trapping centers and reduce the signal. The active pixel sensors were shown to be tolerant to ionizing radiation doses of up to 10 kGy [110]. The damage effects discussed above are shown in Fig. 21.43.

Depleted CMOS

In recent years a so called depleted CMOS or high voltage CMOS (HV-CMOS) process has become available by different foundries. These processes allow application of high voltage to the p substrate which becomes depleted. Charge collection by drift significantly improves both speed and radiation tolerance of these devices.

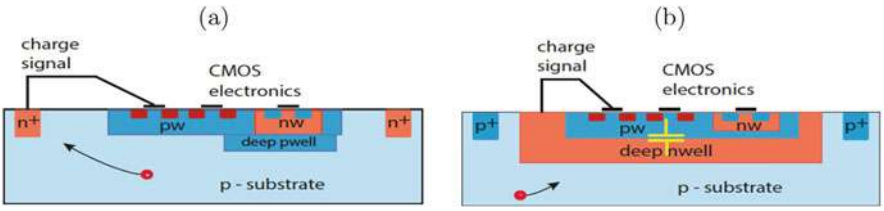

Fig. 21.44 Schematic view of (**a**) small electrode and (**b**) large electrode HV-CMOS detectors [111]

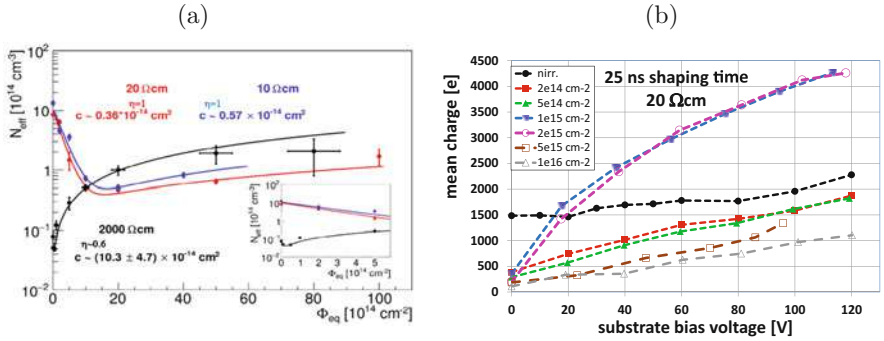

Fig. 21.45 (**a**) Dependence of substrate N_{eff} on fluence of devices produced by two different foundries on different substrate resistivities. Fit of Hamburg model to the data is shown with initial acceptor removal parameters left free [37]. (**b**) Charge collection in irradiated passive HV-CMOS diode array connected to LHC speed electronics [36]

The devices differ mostly in the way the collection electrode is realized. A small n^+ collection electrode is beneficial (see. Fig. 21.44a) for its small capacitance. If, however, a n^+ electrode is inside a large n-well the capacitance is determined by the size of n-well (see. Fig. 21.44b), but the charge collection is faster and more homogeneous. The optimum design therefore depends on the application (see Ref. [112]). Both options are under consideration for the upgrade of pixel sub-detectors at HL-LHC.

Relatively high doping concentration of the substrate (from N_{eff} =few 10^{12} to 10^{15} cm^{-3}) emphasizes the importance of effective acceptor removal with irradiation. For low resistivity substrates the removal of shallow acceptors dominates over the creation of deep ones and the effective doping concentration initially decreases with irradiation. The active/depleted thickness at given voltage increases resulting in larger collected charge. After the initial acceptors are removed the deep acceptors determine the depleted thickness regardless of the choice of initial substrate. The dependence of N_{eff} on fluence for different initial substrate resistivities/doping is shown in Fig. 21.45a [37]. The increase of active thickness is reflected also in charge collection measurements shown in Fig. 21.45b for a low resistivity, 20 Ωcm, device. Note that after the irradiations the contribution from the charges diffusing

from the undepleted substrate to the depleted region vanishes and almost no charge is measured without bias. The contribution of carriers diffusing from the undepleted substrate disappears already after $\Phi \sim 10^{14}\,\mathrm{cm}^{-2}$ [36] which is the reason for initial drop of charge collection efficiency.

Recent studies of pixelated devices established the need for metallization their backside and/or thinning them down [113]. In most processes the high voltage for depletion of the substrate is applied from the contact on top of the device (see Fig. 21.44). After irradiation the increase of resistivity of undepleted bulk can have a large impact on fraction of weighting potential traversed by the carriers and therefore induced charge. Low impedance biasing electrode (HV bias) relatively far away from the sensing electrode and long lateral drift paths of carriers in devices without back side biasing can result in smaller induced charge than expected from the active thickness.

21.5 Electronics

The front-end electronics is an essential part of any detector system. The application specific integrated circuits (ASIC) are composed of analog and digital parts. The analog part usually consists of a preamplifier and a shaping amplifier, while the digital part controls the ASIC and its communication with readout chain. The fundamental building block of the circuit, transistors, can be either bipolar or field-effect devices.

The benefits of either bipolar or field-effect transistors as the first amplifying stage are comprehensively discussed in [114] (see Chapter 6). The equivalent noise charge of the analog front end is given by

$$ENC^2 \approx 2\,e_0\,(I_{nm} + I_m M^2 F)\,\tau_{sh} + \frac{4\,k_B\,T}{g_m\,\tau_{sh}}(C_d + C_c)^2 \tag{21.46}$$

where τ_{sh} is shaping/integration time of the amplifier, I_{nm} and I_m the non-multiplied and multiplied currents flowing in the control electrode of the transistor, M current multiplication factor and F excess noise factor, C_d detector capacitance, C_c capacitance of the transistor control electrode and g_m the transconductance of the transistor. The transconductance measures the ratio of the change in the transistor output current vs. the change in the input voltage. The first term is also called current (parallel) noise while the second term is called voltage (series) noise [115]. The radiation of particle detectors therefore increases both parallel noise through I_{nm}, I_m and series noise through C_d.

The excess noise factor is determined by the gain and effective ratio of hole and electron ionization coefficients k_{eff} [116]

$$F = k_{eff}\,M + (1 - k_{eff})(2 - 1/M) \tag{21.47}$$

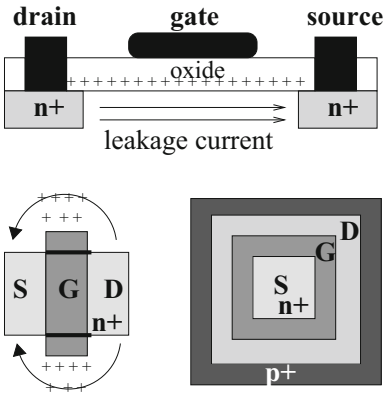

Fig. 21.46 Schematic view of the MOSFET leakage current (top). The standard FET design (bottom left) with parasitic current paths and enclosed transistor design (bottom right)

For moderate gains of $M \approx 10$ and $k_{eff} < 0.01$ usual for silicon tracking detectors $F \sim 2$. It follows from Eq. (21.46) that a voltage noise should dominate the current noise if charge multiplication should increase the signal/noise ratio. If this in not true the current noise increases faster than the signal with bias voltage. Therefore, integration time and electrode size and design should be carefully optimized.

The bipolar transistors are susceptible to both bulk and surface damage while field effect transistors suffer predominately from surface effects. It is the transconductance that is affected most by irradiation.

MOSFET

Advances in integrated electronics circuitry development lead to reduction of feature size to the deep sub-micron level in CMOS technology. The channel current in these transistors is modulated by the gate voltage. The accumulation of positive oxide charge due to ionizing radiation influences the transistor threshold voltage V_{th} (See Fig. 21.46). For nMOS transistors the channel may therefore always be open and for pMOS always closed after high doses. This is particularly problematic for the digital part of the ASIC leading to the device failure. The operation points in analog circuits can be adjusted to some extent to accommodate the voltage shifts. The threshold voltage depends on the square of the oxide thickness and with thick oxides typical for MOS technologies in the previous decades (>100 nm) the radiation hardness was limited to few 100 Gy. At oxide thickness approaching 20 nm the relation $V_{th} \propto d^2$ breaks down (see Fig. 21.8b) as explained in Sect. 21.3.2.1. The deep sub-micron CMOS processes employing such thin oxides are therefore intrinsically radiation hard. The weak dependence of V_{th} on dose for deep sub-micron CMOS processes is shown in Fig. 21.47a. The interface states introducing the leakage current are largely deactivated (see Sect. 21.3.2) in deep sub-micron CMOS transistors, leading to almost negligible surface current (Fig. 21.47b). Also mobility changes less than 10% up to 300 kGy.

Even with transistor parameters not severely affected by radiation the use of so called enclosed transistor layout (ELT) [117] is sometimes required to eliminate the

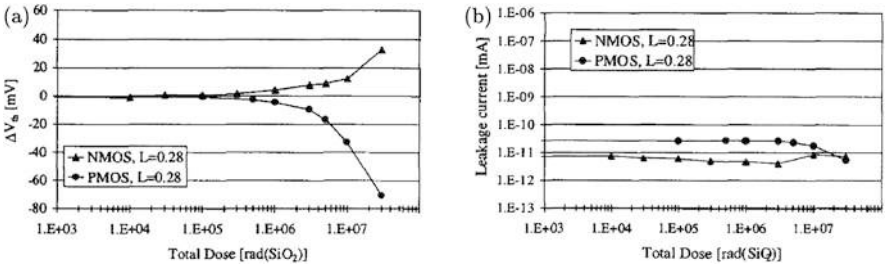

Fig. 21.47 (**a**) Threshold voltage shift of enclosed nMOS and standard pMOS transistors as a function of the total dose for a 0.25 μm technology. (**b**) Leakage current for the same transistors [117].

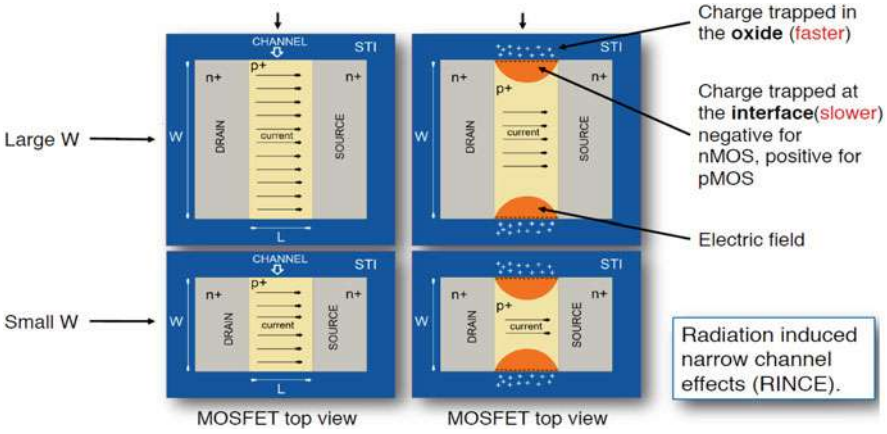

Fig. 21.48 Schematic view of Radiation-Induced Narrow Channel Effect

radiation effects on large arrays of transistors. Radiation induces transistor leakage through the formation of an inversion layer underneath the field oxide or at the edge of the active area (see Fig. 21.46). This leads to source-to-drain and inter-transistor leakage current between neighboring n^+ implants. The former can be avoided by forcing all source-to-drain currents to run under the gate oxide by using a closed gate. The inter-transistor leakage is eliminated by implementing p^+ guard rings.

Development of dedicated libraries to implement enclosed transistors for each deep-sub micron process is often too demanding or the functionality required for a given surface doesn't allow enclosed transistors. In such cases a so called Radiation-Induced Narrow Channel Effect (**RINCE**) shown in Fig. 21.48 can occur.

The positive charge trapped in the lateral shallow trench isolation oxide (STI) attracts electrons and opens a conductive channel through which leakage current can flow between source and drain. This current is usually small and [119] compared to the current that can flow in the main transistor and it only influences the subtreshold region of the transistor I–V curve. Even if the functionality of the chip is preserved

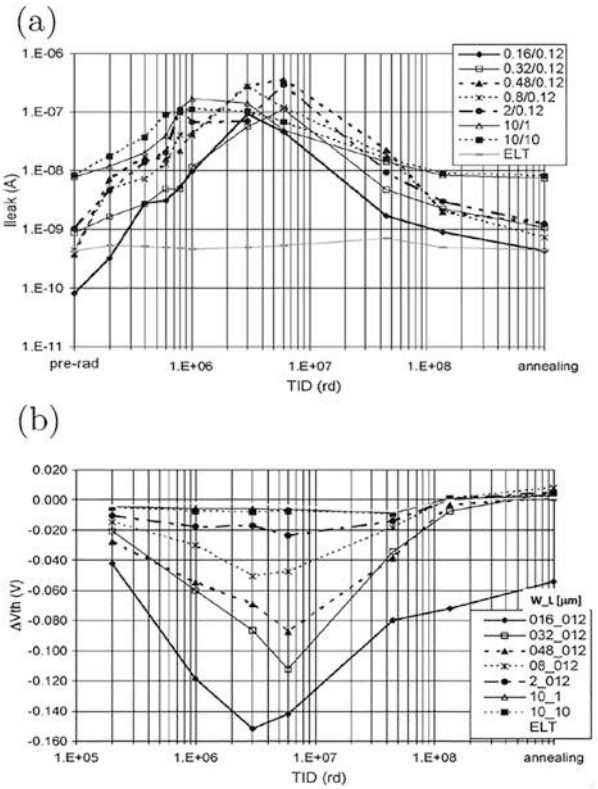

Fig. 21.49 (**a**) Evolution of the leakage current with TID for different NMOS transistor sizes (width/length in μm), up to 1.36 MGy. The last point refers to full annealing at 100 °C. The first point to the left is the pre-rad value (**b**) Same as (**a**) but showing transistor threshold voltage shift [119].

this impacts power consumption and thermal performance of the chip. At higher doses the negative charge trapped at the interface states compensates the positive space charge (NMOS transistors) and leakage current decreases (see Fig. 21.49a). Both processes of positive oxide charge and negative charge build-up at the interface states are highly dependable on dose rate, process and annealing. The increase of the transistor leakage current affected the operation of ATLAS-IBL detector [77].

Apart from parasitic leakage current the trapped oxide charges can also moderate electric field in the transistor channel particularly for narrow channel transistor where relatively larger part of the transistor is affected. If the change in threshold voltage for NMOS is small (see Fig. 21.49b), RINCE can be a bigger problem for PMOS transistors. There, positive trapped charge (holes) at the interface states adds to the positive oxide charge. As a consequence the threshold voltage and the required current to turn transistor on change significantly. As shown in Figs. 21.49 only marginal annealing effects were observed.

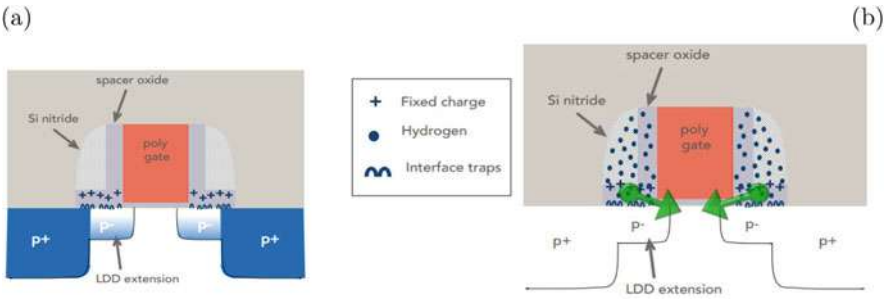

Fig. 21.50 (**a**) Radiation-Induced Short Channel Effect—charging of spacer oxide modifies free carrier concentration in LDD (p^-) layer. (**b**) Annealing releases protons/hydrogen atoms/molecules to the gate oxide [120]

Radiation Induced Short Channel Effect (RISCE) appears in both PMOS and NMOS transistors with very short channel (for both ELT and open-layout transistors) and is a consequence of transistor design with so called spacer oxide shown in Fig. 21.50a. This oxide charges and affects the amount of carriers in Low Drain Doping (LDD) extension of the transistors leading to a decrease of the transistor-on current during exposure. The radiation and temperature/annealing frees protons, neutral hydrogen atoms/molecules from spacer oxide that can reach the nearby gate oxide. There they depassivate Si-H bonds and by that change the threshold voltage and transistor-on current. This process is strongly dependent on (annealing) temperature. It can be avoided at low temperature operation ($T < 0°C$) and by switching off biasing at high temperatures [120].

The constant reduction of feature size in modern CMOS processes going from 0.35, 0.25, 0.13, 0.065, 0.045 µm have also other beneficial consequences. Ever shorter transistor channel lengths result in higher speed of the devices which consumes also less power particularly in the digital part. Larger transistor densities allow more complex signal processing while retaining the die size. Unfortunately at given power constraints, the basic noise parameters of bipolar and field-effect front-end transistors will not improve with the reduction of feature size [115].

Bipolar Transistor The main origin of damage in bipolar transistors is the reduction of minority carrier life time in the base, due to recombination processes at radiation-induced deep levels. The transistor amplification factor $\beta = I_c/I_b$ (common emitter) decreases according to $1/\beta = 1/\beta_0 + k\Phi$. The pre-irradiation value is denoted by β_0 and damage constant dependent on particle and energy by k. Since $g_m \propto \beta$, degradation of β leads to larger noise and smaller gain of the transistor. Thinner base regions are less susceptible to radiation damage, so faster transistors tend to be better.

The choice of base dopant plays an important role. A boron doped base of a silicon transistor is not appropriate for large thermal neutron radiation fields due to the large cross-section for neutron capture (3840 barns). The kinetic energy released (2.3 MeV) to Li atoms and α particles is sufficient to cause large bulk damage [118].

Single Event Effects (SEE) Unlike the bulk and surface damage, the single event effects are not cumulative. They are caused by the ionization produced by particles hitting certain elements in the circuit. According to the effect they have on operation they can be:

- **transient**; Spurious signals propagating in the circuit, due to electrostatic discharge.
- **static**; The state of memory or register bits is changed (Single Event Upset). In case of altering the bits controlling the circuit, they can disturb functionality or prevent circuits from operating (Single-event Functional Interrupt).
- **permanent**; The can destroy the circuit permanently (Single Event Latchup).

The SEE become a bigger problem with reduction of the feature size, as relatively smaller amount of ionization is required to change properties. The radiation hardening involves the use of static-RAM instead of dynamic-RAM and processing of electronics on SOI instead on silicon bulk (physical hardening). The logical hardening incorporates redundant logical elements and voting logic. With this technique, a single latch does not effect a change in bit state; rather, several identical latches are queried, and the state will only change if the majority of latches are in agreement. Thus, a single latch error will not change the bit.

21.6 Conclusions

The radiation damage of crystal lattice and the surface structure of the solid state particle detectors significantly impacts their performance. The atoms knocked-off from their lattice site by the impinging radiation and vacancies remaining in the lattice interact with themselves or impurity atoms in the crystal forming defects which give rise to the energy levels in the semiconductor band-gap. The energy levels affect the operation of any detector in mainly three ways. Charge levels alter the space charge and the electric field, the levels act as generation-recombination and trapping centers leading to increase of leakage current and trapping probability for the drifting charge. The magnitude of these effects, which all affect the signal-to-noise ratio, depends on the semiconductor material used as well as on the operation conditions.

Although the silicon, by far most widely used semiconductor detector material, is affected by all three, silicon detectors still exhibit charge collection properties superior to other semiconductors. Other semiconductors (e.g. SiC, GaN, GaAa, a-Si) can compete in applications requiring certain material properties (e.g. cross-section for incoming radiation, capacitance) and/or the crucial properties are less affected by the radiation (e.g. leakage current and associated power dissipation). Radiation effects in silicon detectors were thoroughly studied and allow for reliable prediction of the detector performance over the time in different irradiation fields.

The state of the art silicon strip and pixel detectors used at experiments at LHC retain close to 100% detection efficiency for minimum ionizing particles at hadron fluences in excess of $10^{15}\,\mathrm{cm}^{-2}$ and ionization doses of 1 MGy. The foreseen upgrade of Large Hadron Collider require hardness to even an order of magnitude larger fluences, which presently set the ultimate benchmark for operation of semiconductor particle detectors. The efforts for improving the silicon detection properties in order to meet these demanding requirements include defect engineering by adding impurity atoms, mainly oxygen, to the crystal, operation at cryogenic temperatures and placement electrodes perpendicularly to the detector surface—3D detectors.

The increase of effective doping concentration with fluence together with high voltage operation lead to charge multiplication in heavily irradiated silicon detectors which in combination with electric field in the neutral bulk and saturation of effective trapping probabilities result in efficient operation in radiation environments even harsher than that at the HL-LHC.

In recent years new detector technologies appeared, such as Low Gain Avalanche Detectors which offer along with position also time resolution and depleted CMOS monolithic detectors. The latter offer for the first time fully monolithic devices with fast response and sufficient radiation hardness. The initial dopant removal plays a crucial role in performance of both LGADs and depleted CMOS detectors. Among other semiconductors diamond is the most viable substitute for silicon in harsh radiation fields, particularly with the advent of 3D diamond detectors.

The silicon detector employed in less severe environments e.g. monolithic active pixels, charge coupled devices, silicon drift detectors are optimized for the required position and/or energy resolution and the radiation effects can be well pronounced and even become the limiting factor already at much lower doses. Longer drift and/or charge integration times increase the significance of leakage current, charge trapping and carrier recombination.

The silicon-silicon oxide border and the oxide covering the surface of silicon detectors and electronics is susceptible to ionizing radiation. The positive charge accumulates in the oxide and the concentration of interface states, acting as trapping and generation-recombination centers, increases. These effects can be effectively reduced in silicon detectors by proper processing techniques. Thin oxides ($<$20 nm) allow tunneling of electrons from the gate electrode through the oxide. They can recombine with positive charges in the oxide and also passivate interface traps. Deep sub-micron CMOS processes which utilize oxides of such thicknesses are therefore intrinsically radiation hard especially if proper design rules are used. In very deep-sub micron processes where often the use of special design rules is not possible two effects Radiation-Induced Narrow Channel Effect and Radiation Induced Short Channel Effect appear which require special adjustments in operation scenarios.

References

1. K. McKay, Phys. Rev. 76 (1949) 1537.
2. J. Kemmer, Nucl. Instr. Meth. A 169 (1980) 499.
3. M. Swartz, M. Thurston, J. Appl. Phys., 37(2) (1966) 745.
4. A.G. Holmes-Siedle and L. Adams, Radiat. Phys. Chem. 28(2) (1986) 235.
5. ATLAS Inner Detector Technical design report, CERN/LHCC/97-16, ISBN 92-9083-102-2.
6. G.D. Badhwar, Rad. Res. 148 (1997) 3.
7. W.N. Spjeldvik and P.L. Rothwell, The Earth's Radiation Belts. In: Environmental Research Paper No. 584, Air Force Geophysics Laboratory, U.S. Department of the Air Force, AFGL-TR-83-0240, Massachusetts (1983).
8. V.A.J. van Lint, T.M. Flanagan, R.E.Leadon, J.A. Naber, V.C.Rogers, *Mechanism of Radiation Effects in Electronic Materials*, John Wiley & Sons, 1980.
9. T.F. Luera et al., IEEE Trans. NS 34(6) (1987) 1557.
10. M. Huhtinen, Nucl. Instr. and Meth. A 491 (2002) 194.
11. G. Lindström, Radiation Damage in Silicon Detectors, Nucl. Instr. and Meth. A 512, 30 (2003).
12. W. de Boer, Phys. Stat. Sol. (a) 204, No. 9 (2007) 3004.
13. E. Gaubas et al., Mat. Sc. Sem. Proc. 75 (2018) 157165.
14. M. Mikuž et al., "Extreme Radiation Tolerant Sensor Technologies" presented at 26th Vertex conference, Las Caldas, Spain, September, 2017..
15. R. Radu et al., Journal Of Applied Physics Vol. 117 (16) (2015) 164503.
16. M.M. Atalla, E. Tannenbaum and E.J.Scheibner, Bell. Syst. Techn. J. 38 (1959) 749.
17. D.M. Fleetwood, J. Appl. Phys. 73 (10) (1993) 5058.
18. C.T. Sah, IEEE Trans. NS 23 (6) (1976).
19. A. Goetzberger, V. Heine, E.H. Nicollian, Appl. Phys. Lett. 12 (1968) 95.
20. A.G. Revesz, IEEE Trans. Electron Dev. ED-12 (1965) 97.
21. R. Wunstorf et al., Nucl. Instr. and Meth. A 377 (1996) 290.
22. N.S. Saks, M. G. Ancona and J. A. Modolo, IEEE Trans. NS 31(6) (1984) 1249.
23. J Zhang et al., JINST 7 (2012) C12012.
24. R.H. Richter et al., Nucl. Instr. Meth. A 377 (1996) 412.
25. W. Füsel et al., Nucl. Instr. and Meth. A 377 (1996) 177.
26. S. Ramo, Proc. I.R.E. 27 (1939) 584.
27. G. Kramberger et al., IEEE Trans. NS 49(4) (2002) 1717.
28. T.J. Brodbeck et al., Nucl. Instr. and Meth. A 395 (1997) 29.
29. G. Casse et al., Nucl. Instr. and Meth. A 487 (2002) 465.
30. R. Wunstorf, Ph.D. thesis, Hamburg University 1992, DESY FH1K-92-01 (October 1992).
31. Michael Moll, Ph.D. thesis, Hamburg University 1999, DESY-THESIS-1999-040, ISSN-1435-8085.
32. J. Adey, PhD Thesis, Univeristy of Exceter, 2004.
33. J. Adey et al., Physica B 340342 (2003) 505508.
34. G. Lindström et al.(RD48), Nucl. Instr. and Meth. A 466 (2001) 308.
35. A. Khana et al., Solar Energy Materials & Solar Cells 75 (2003) 271.
36. A. Affolder et al., JINST Vol. 11 (2016) P04007.
37. I. Mandić et al., JINST 12 P02021 2017.
38. E. Cavallaro et al., JINST 12 C01074 2017.
39. B. Hiti et al., JINST 12 P10020 2017.
40. P. Dias de Almeida et al., "Measurement of the acceptor removal rate in silicon pad diodes", 30^{th} CERN-RD50 Workshop, Krakow, 2017.
41. G. Kramberger et al., JINST Vol. 10 (2015) P07006.
42. E. Buchanan for LHCb Velo collaboration, "The LHCb VELO & ST Operational Performance Run II", PoS (Vertex 2017) 016.

43. M. Kocian for ATLAS collaboration, "Operational Experience of ATLAS SCT and Pixel Detector", PoS(Vertex 2017) 017.
44. C. Barth for CMS collaboration, "CMS pixel and strip rad damage measurements", 31^{st} CERN-RD50 Workshop, Geneve, 2017.
45. J. Härkönen, Nucl. Instr. and Meth. A 518 (2004) 346.
46. I. Pintilie, et al., Meth. Instr. and Meth. A 514 (2003) 18.
47. G. Kramberger et al., Nucl. Instr. and Meth. A 515, 665 (2003).
48. G.Lindström et al., Nucl. Instr. and Meth. A 568 (2006) 66.
49. G. Kramberger et al., Nucl. Instr. and Meth. A 609 (2009) 142.
50. V. Eremin et al., Nucl. Instr. and Meth. A 360 (1995) 458.
51. V. Eremin et al., Nucl. Instr. and Meth. A 476 (2002) 556.
52. G. Kramberger et al., Nucl. Instr. and Meth. A 497 (2003) 440.
53. G. Kramberger et al.,"Investigation of Irradiated Silicon Detectors by Edge-TCT ", IEEE Trans. Nucl. Sci. Vol. 57(4), 2010, p. 2294.
54. R. van Overstraetem and H.de Man, Solid-State Electronics 13(1970),583–608.
55. W. Maes, K. de Meyer, R. van Overstraetem, Solid-State Electronics 33(1990),705–718.
56. J. Lange et al.,"Charge Multiplication Properties in Highly-Irradiated Epitaxial Silicon Detectors", PoS(Vertex 2010) 025.
57. M. Koehler, IEEE Trans. NS 58 (2011) 3370.
58. I. Mandić et al. Nucl. Instr. and Meth. A603 (2009) 263.
59. G. Casse et al. Nucl. Instr. and Meth. A624 (2010) 401.
60. I. Mandić et al., JINST 8 (2013) P04016.
61. R. Mori et al., Nucl. Instr. and Meth. A796 (2015) 131.
62. I. Mandić et al.,Nucl. Instr. and Meth. 629 (2011) 101.
63. G. Kramberger et al., Nucl. Instr. and Meth. A 481 (2002) 297.
64. O. Krasel et al., IEEE Trans. NS 51(1) (2004) 3055.
65. RD50 Status Report 2008, CERN-LHCC-2010-012 and LHCC-SR-003 (2010).
66. W. Adam et al., JINST Vol. 11 (2016) P04023.
67. M. Mikuž et al.,"Extreme Radiation Tolerant Sensor Technologies", presented at 26^{th} Vertex Conference, Las Caldas, 2017.
68. M. Moll, et al., Nucl. Instr. and Meth. A 426 (1999) 87.
69. V. Palmieri et al., Nucl. Instr. and Meth. A 413 (1998) 475.
70. K. Borer et al., Nucl. Instr. and Meth. A 440 (2000) 5.
71. V. Eremin et al., Nucl. Instr. and Meth. A 583 (2007) 91.
72. E. Verbitskaya et al., IEEE Trans. NS 49(1) (2002) 258.
73. A. Chilingarov et al., Nucl. Instr. and Meth. A 399 (1997) 35.
74. S.I. Parker et al., Nucl. Instr. and Meth. A395 (1997) 328.
75. A. Zoboli et al., IEEE Trans. NS 55(5) (2008) 2775.
76. G. Pellegrini et al, Nucl. Instr. and Meth. A 592 (2008) 38.
77. M. Backhaus et al., Nucl. Instr. and Meth. A831 (2016) 65.
78. G. Darbo et al., JINST 10 (2015) C05001.
79. J Lange et al., JINST Vol. 13 (2018) P09009.
80. M. Garcia-Sciveres et al. Nucl. Instr. and Meth. A 636 (2011) 155.
81. H. Sadrozinski et al., Nucl. Instr. and Meth. A831 (2016) 18.
82. G. Pellegrini et al., Nucl. Inst. Meth. A765 (2014) 14.
83. N. Cartigli, H. Sadrozinski and A. Seiden, REPORTS ON PROGRESS IN PHYSICS Vol. 81 (2018) 026101.
84. N. Cartiglia et al., Nucl. Instr. and Meth. A850 (2017) 83.
85. G. Kramberger et al., Nucl. Instr. and Meth. A891 (2018) 68.
86. H. Sadrozinski et al., "Properties of HPK UFSD after neutron irradiation up to 6e15 n/cm^2 " to appear in Nucl. Instr. and Meth. A (2018).
87. M. Ferrero et al., Nucl. Instr. and Meth. A919 (2019) 16.
88. G. Kramberger et al., Nucl. Instr. and Meth. A898 (2018) 53.
89. W. Adam et al., Nucl. Instr. and Meth. A447 (2000) 244.

90. H. Kagan et al., Nucl. Instr. and Meth. A582 (2007) 824.
91. M. Mikuž et al., "Study of polycrystalline and single crystal diamond detectors irradiated with pions and neutrons up to 3 × 1015 cm-2", IEEE Nucl. Sci. Symp. Conference Record, San Juan (2007) N44-5.
92. N. Venturi et al., "Results on Radiation Tolerance of Diamond Detectors", presented at 11^{th} International Hiroshima Symposium on the Development and Application of Semiconductor Tracking Detectors, Okinawa, Japan, (2017), to appear in Nucl. Instr. and Meth. A.
93. F.Bachmair et al., Nucl. Instr. and Meth. A786 (2015) 97.
94. N. Venturi et al., JINST 11 (2016) C12062.
95. F. Nava et al., Nucl. Instr. and Meth. A437 (1999) 354.
96. M. Rogala et al., Nucl. Phys. B 78 (1999) 516.
97. S. Sciortino et al., Nucl. Instr. and Meth. A552 (2005) 138.
98. F. Moscatelli et al., IEEE Trans. NS 53(4) (2006) 1557.
99. S.P. Beaumont et al., Nucl. Instr. and Meth. A322 (1992) 472.
100. R.L. Bates et al., Nucl. Instr. and Meth. A395 (1997) 54.
101. M. Rogala et al., Nucl. Instr. and Meth. A410 (1998) 41.
102. J. Grant et al., Nucl. Instr. and Meth. A576 (2007) 60.
103. G. Kramberger et al., JINST 8 (2013) P08004.
104. C.J.S. Damerell et al., Nucl. Instr. and Meth. A512 (2003) 289.
105. N.S. Saks et al., IEEE Trans. NS 27 (1980) 1727.
106. G. Claus et al., Nucl. Instr. and Meth. A465 (2001) 120.
107. R. Turcheta et al., Nucl. Instr. and Meth. A458 (2001) 677.
108. G. Deptuch et al., Nucl. Instr. and Meth. A465 (2001) 92.
109. M. Deveaux et al., Nucl. Instr. and Meth. A583 (2007) 134.
110. M. Deveaux et al., Nucl. Instr. and Meth. A552 (2005) 118.
111. N. Wermes and H. Kolanoski, *Teilchendetektoren: Grundlagen und Anwendungen*, Springer Spektrum (2016), ISBN 978-3-662-45349-0.
112. H. Pernegger et al., JINST 12 (2017) P06008.
113. I. Mandić et al., Nucl. Instr. and Meth. A903 (2018) 126.
114. V. Radeka et al., Ann. Rev. Nucl. Part. Sci. 37 (1988) 217.
115. H. Spieler, *Semiconductor Detector Systems*, Oxford University Press, Oxford (2005) ISBN 0-19-852784-5 TK9180. S68 2005.
116. R. J. MaIntyre, IEEE Trans. Elec. Devices Vol. 13 (1966) 164.
117. P. Jarron et al., Nucl. Phys. B 78 (1999) 625.
118. I. Mandić et al., Nucl. Instr. and Meth. A518 (2004) 474.
119. F. Faccio and G. Cervelli, IEEE Trans. Nucl. Sci. Vol. 57 (2005) 2413.
120. F. Faccio et al., IEEE Trans. Nucl. Sci. Vol. 65 (2018) 164.

Chapter 22
Future Developments of Detectors

Ties Behnke, Karsten Buesser, and Andreas Mussgiller

22.1 Introduction

Large scale detectors in particle physics take many years to plan and to build. The last generation of large particle physics detectors for the energy frontier, ATLAS and CMS, have been operating for more than 10 years, and upgrades for them are now being done. Studies for the next generation of experimental facilities have been ongoing for a number of years. In this section future directions in integrated detector design are discussed, as they were visible at the time of writing this report.

At the moment the biggest approved project in particle physics is the upgrade of the Large Hadron Collider (LHC) towards high luminosity running. This project is scheduled to be completed by 2027, and major upgrades to the two main collider detectors ATLAS and CMS are planned. Beyond the LHC, an electron-positron collider has been discussed for many years, to fully explore the Higgs and the top sector and to complement the discovery reach of a hadron machine at the energy frontier with a high precision program.

The requirements as far as detectors are concerned are very different for these two types of projects: for the LHC luminosity upgrade fundamental changes to the underlying philosophy of the existing detectors are not possible, but significant technological development is needed to meet the challenges of extreme radiation environments and high event rates. For a yet not existing electron-positron collider a detector can be designed from ground up, optimised to meet the ambitious physics agenda of such a facility.

T. Behnke (✉) · K. Buesser · A. Mussgiller
DESY, Hamburg, Germany
e-mail: ties.behnke@desy.de

C. W. Fabjan, H. Schopper (eds.), *Particle Physics Reference Library*,
https://doi.org/10.1007/978-3-030-35318-6_22

Several strategy discussions at national and international levels have consistently put a high energy electron positron collider far up on the list of future projects in the field [1–4]. Such a facility should serve as a Higgs factory, run at least at an energy of 250 GeV, but should also provide an upgrade path towards the top threshold and beyond. With the results from the current run of the LHC which show no indications of direct signs for new physics, the role of ultimate precision especially at the Higgs production threshold has been much strengthened [5].

The International Linear Collider, ILC, is a mature project to build an electron-positron collider, which could eventually push into the TeV regime, realised as a linear accelerator. The facility is described in the Technical Design Report from 2012 [6], and targets an initial energy of 250 GeV, upgradable to 1 TeV. To reach energies in the multi-TeV range in an electron-positron collider, another technology will be needed. The CLIC technology, developed mostly at CERN, is a promising candidate for such a machine [7, 8].

With the strong emphasis on precision Higgs physics, circular machines have become again a subject of study. A circular collider like the FCC-ee project, pursued at CERN [9], could reach the Higgs and possibly the top pair threshold in a ring of around 100 km circumference. Such a facility could also be used for the next large hadron collider, reaching energies of up to 100 TeV [10]. A similar project, CEPC/SppC, is under discussion in China [11–13].

A number of smaller projects are currently pursued in the field of experimental high-energy physics as well, for example, the B-factory at KEK, or long baseline neutrino experiments like Dune.

22.2 Challenges at Future Colliding Beam Facilities

The Large Hadron Collider, LHC, has seen first beams in 2008. Until 2018, a spectacular physics harvest has taken place, with the undisputed highlight the discovery of the long-sought-after Higgs particle in 2012. The energy of the collider has reached its design value, and the collider will continue to run in this configuration for another approximately 5 years, until 2024.

Already now the LHC has exceeded its design luminosity of $10^{34}\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and is expected to accumulate a total integrated luminosity of $300\,\mathrm{fb}^{-1}$ in the first running phase ("Phase-I") that extends to 2024. This will result in significant new insights into the physics of the electroweak symmetry breaking, and significant new information on physics beyond the standard model.

During the Phase-II, starting around 2027 and extending to 2035 or beyond, the LHC will increase its luminosity by about a factor of 10. ATLAS and CMS will extend their physics reach [15] significantly with this upgrade. The discovery reach for supersymmetric particles for example will be extended by some 20–30%, access to rare decay modes e.g. of the Higgs Boson will be improved, and flavour changing neutral currents through top decays might become accessible. Many other measurements will profit from this improvement as well. However, the increased

luminosity is payed for with more severe background conditions, with a much larger number of events per beam crossing, and a resulting challenge to the sub-detectors. In particular the innermost detectors will need major upgrades, together with the readout and data acquisition systems, to handle the new conditions.

It should be noted here that also studies have been initiated for detectors of possible future very large hadron colliders that could succeed the LHC and explore energy ranges of up to 100 TeV. One such concept of a hekto-TeV hadron collider is discussed within the framework of the FCC-hh study at CERN [10], another, SPPC, is part of the CEPC study in China [12]. The requirements for the detectors of such machines are just being explored and are far from being fully understood. The main challenges are related to the large jet energies and boosted event topologies, that require very large magnetic fields, large detector dimensions, and highly segmented detectors. In addition, the radiation environment is harsh and requires very radiation hard detectors.

An electron-positron collider like the ILC or FCC-ee poses different but unique challenges to its detectors. It puts a premium on precision physics, particularly on the precision reconstruction of jet masses. The experimental environment is benign by LHC standards, which allows one to consider technologies and solutions which have not been possible during the development of the LHC detectors.

To reach high precision in the overall reconstruction of event properties, each sub-system must reach excellent precision by itself. In addition, however, in the combination of sub-systems into a complete detector extreme care has to be taken to be able to fully utilise the precision of the sub-detectors. Among the most relevant parameters is the amount of dead material, in particular for the inner tracking detectors, and its radiation hardness. Low mass detectors are a key requirement, and add a major challenge to the system. High readout-speed is another ingredient, without which the high luminosity of the collider can not be exploited fully.

An experiment at an electron-positron collider has to be designed to extract maximum information from the event, and utilise the available luminosity as much as possible. It has to be able to reconstruct as many different topologies and final states as possible. This implies that the focus of the development has to be the reconstruction of hadronic final states, which are by far the most numerous ones in nearly all reactions of interest. A typical event topology is a multi-jet final state, with typical jet energies of order of 50–100 GeV. In contrast to the LHC, where many collisions occur in one bunch crossing, typically only one event of interest takes place at the linear collider, even at very high luminosities. With well below 100 particles per jet the total number of particles in the final state is comparatively small. This makes it possible to attempt the reconstruction of every single particle, neutral or charged, in the event. A major focus of the detector development therefore will be the capability of the detector to identify individual particles as efficiently as possible, and to reconstruct the properties of each particle as precisely as possible. This has large implications for the overall design of the detector.

Even though the event topology at an electron-positron collider is intrinsically clean, and there are no underlying events nor multiple interactions, as they are present in a hadron collider, backgrounds nevertheless do play a role. In particular

for the innermost and the most forward systems, beam induced backgrounds are significant. Electron-positron pairs created in the interaction of the two highly charged bunches add significant background to the event, and detectors close to the beam need to be able to cope with these. This background is particularly relevant at linear colliders, which, due to the smaller repetition rate of the interactions, need to focus their beams very strongly at the interaction region to reach the luminosity goals. Circular electron-positron colliders on the other hand can operate with less strongly focussed beams, since they re-use the beams after each turn, operating at much larger repetition rates.

22.3 Hadron Colliders

The LHC and its envisaged upgrade to the HL-LHC provides a physics program well until the middle of the 2030s. As discussed above, plans for the next colliders at the energy frontier are being made already now. A possible far-future option is a very large hadron collider. Recently, the conceptual design report for the Future Circular Collider (FCC), a ≈100 km long storage ring proposed for CERN, has been published. The proposal foresees to start with an e^+e^- collider for Higgs precision studies (FCC-ee [9]) that could be replaced by a hadron collider, the FCC-hh [10], at a later stage, probably not before the 2060s. Table 22.1 summarises the basic parameters of HL-LHC and FCC-hh in comparison to the LHC.

The LHC detectors are operating since quite some time now and are very well understood. This experience helped to design the upgrades that are required to cope with the challenges of the oncoming LHC luminosity upgrade, as will be discussed in the next Sect. 22.3.1. The FCC-hh challenges to the detectors are quite different; first concepts for detectors are under discussion and will be presented in Sect. 22.3.2.

Table 22.1 Some basic design parameters of the LHC, HL-LHC and FCC-hh (nominal) [10]

Parameter	Unit	LHC	HL-LHC	FCC-hh
Center-of-mass energy	TeV	14	14	100
Peak luminosity/IP	10^{34} cm^{-2} s^{-1}	1	5	30
Number of bunches	#	2808	2808	10,400
Bunch population	10^{11}	1.15	2.2	1.0
Time between bunches	ns	25	25	25
Beam spot size at IP	µm	16.7	7.1	3.5
Bunch length	cm	7.55	7.55	8
Accelerator length	km	27	27	97.75
Peak pile-up events/bunch crossing	#	25	130	950
Pile-up line density	mm^{-1}	0.2	1.0	8.1
Pile-up time density	ps^{-1}	0.1	0.29	2.43
Total ionising dose at $r = 2.5$ cm	MGy	1.3	13	270

22.3.1 Detector Upgrades for the High-Luminosity-LHC

The two major colliding beam experiments at the LHC, ATLAS and CMS, have recorded large data sets starting in 2010. The currently installed innermost detectors were designed to cope with track densities and to withstand the radiation doses expected during the LHC Phase-I running that extends until 2024. For the high-luminosity operation phase of the LHC both experiments will replace their inner tracking detector with completely new systems.

The tracking detectors of both large LHC experiments are mostly based on silicon technology detectors. Over the past years, an intense R&D effort has taken place, to re-design and re-optimise the inner detectors for both ATLAS and CMS. Fundamentally no changes in technology will take place, both detectors will rely on an all-silicon solution for the tracking. In addition, ATLAS will remove the transition radiation detector from its system, and extend its silicon tracker to larger radii. Owing to the track trigger concept, CMS completely re-designs its tracker and utilises novel detector modules that allow for an on-module p_T discrimination of charged particle tracks. Both Phase-II trackers will again follow a classical barrel and end cap design. However, compared to the Phase-I trackers, ATLAS will use wedge-shaped sensor modules in its end caps of the tracker, whereas CMS will rely on rectangular modules in this part of the detector. Both future trackers will have substantially increased granularity to cope with the expected pile up of up to 200 events per bunch crossing, and very much improved radiation tolerance, which will significantly go beyond the one of the Phase-I detectors and suffice for operation throughout the Phase-II era.

The amount of insensitive material is a significant performance limiting factor of the current trackers, both at ATLAS and at CMS. The large amount of material in the present trackers not only reduces the performance of the trackers themselves, but also has a negative impact on the performance of the electromagnetic calorimeters directly outside of the tracking systems. The reduction of material is therefore another important goal of the tracker upgrades. CMS will use 320 μm thick sensors with an active thickness of 200 μm, as compared to 500 μm thickness in the present detector, novel structural materials, and novel powering and cooling schemes will make this goal achievable.

For the innermost layers of the future trackers radiation tolerance will be of even larger importance than today. Current technologies are not able to withstand the anticipated rates for longer periods. A number of novel technologies are under consideration, 3D Silicon pixel sensors or diamond tracking detectors. Even solutions which do not involve Silicon—like Micromegas trackers—are being discussed.

The higher rates at the upgraded LHC will not only challenge the hardware of the tracker, but also put large demands on the readout and the trigger system. In particular, the latter will have to be significantly upgraded to handle the anticipated rates without a loss of sensitivity. The tracker might well play a central role here, as the early trigger on track-like objects already in the level-1 trigger will significantly

reduce the trigger rate. Triggering on tracks rather than just increasing the trigger thresholds will maintain a much better sensitivity to a broad range of signals, in particular for the much sought-after new physics signals.

The final layout is based on the concept of a "long pixel" detector. In this Ansatz the pixel size is increased compared to current pixels to something like $100\,\mu\text{m} \times 2\,\text{mm}$. It appears possible to keep the power per pixel constant compared to current pixel readouts, thus resulting in a tracker which has a channel count larger by two orders of magnitude than the current strip trackers, but a similar overall power consumption.

Although the tracking detectors are most affected by the increased luminosity, other detectors will be affected as well. The calorimeters will see much increased backgrounds in the forward direction, which might necessitate upgrades or significant changes. A serious problem might be that the ATLAS liquid argon calorimeter in the forward direction heats up under the backgrounds to a point where is will no longer function. In this case—which will only be known once operational experience under real LHC conditions is available—the replacement by a warm forward calorimeter might be necessary. CMS intends to make major changes to its calorimeter system, replacing the hadronic section and part of the electromagnetic section with a highly granular calorimeter, using technology which has been developed and will be described later in the section on detectors at electron positron colliders. For all detectors the capability to handle larger rates will be needed, and might make updates and replacements of the readout electronics necessary. This even applies to parts of the muon system, again primarily in the forward direction. ATLAS e.g. is considering to replace the drift tubes in the forward direction with ones of smaller diameter, to limit the occupancy. In any case upgrades to the trigger and the data acquisition are needed.

22.3.1.1 Novel Powering Schemes

The minimisation of power consumption will play a central role in the upgrades of the LHC tracker detectors for the LHC Phase-II. Traditionally readout electronics are the main generators of heat in the detectors, which needs to be cooled away. Both ATLAS and CMS employ sophisticated liquid cooling systems, operating at pressures below the atmospheric pressure, to cool away approximately 33 kW from the tracking detector alone. Power is brought to the electronics at low voltages, typical for semiconductor operation. The resulting large cross sections of conductors add significantly to the overall material of the detector.

Several alternative schemes are under consideration, to limit the material and volume needed by the power lines. In one approach, called DC-DC, a large voltage is provided at the frontend. For the same power delivered a significant reduction of the amount of copper needed can be obtained. On the front-end the larger voltage is then transformed to the needed lower voltage. An optimised method to transform

the voltages without large power loss, and without large and bulky circuitry, is the subject of intense R&D.

An alternative option is serial powering. Here as well power is supplied to the front-end at a high potential. By putting several readout circuits in series, the power is reduced at each chip to the needed level. This approach promises reduced power loss and less material at the detector, but presents the experimenter with problems of proper grounding of the detector elements. By putting systems in series potentially a correlation between chips due to changing power consumption levels may be introduced. This method as well is the subject of intense R&D.

22.3.1.2 Novel Mechanical Structures and Cooling

The all-silicon trackers developed for the ATLAS and CMS operation at LHC Phase-II conditions rely on sophisticated mechanical systems, which are light-weight and at the same time provide the necessary precision and services to the detector modules. They need to be able to operate at low temperatures, and withstand thermal cycles with a temperature differential of up to 50°.

In contrast to previous designs, where cooling and positioning of modules was achieved via separate features of the mechanical structures, the new designs will combine these functionalities in single features with the goal of substantially reducing the amount of passive materials in the tracker volume. In addition, bi-phase evaporative CO2 cooling will be used as coolant, which not only has a larger radiation length X_0 compared to conventional coolants, but also allows to use pipes with smaller diameters and wall thickness, which even further reduces the material budget. However, smaller pipe diameters require significant improvements in the type of heat spreaders that are used to transport the heat from the source to the coolant. Due to their thermal properties carbon foams are widely used for this purpose. They provide a relatively large thermal conductivity at low mass. Moreover, carbon foams can be tuned to the specific needs of an application, by adjusting the pore-size and the amount of carbon deposited on the cell structure, which defines both the density of the foam and its thermal conductivity. Figure 22.1 shows a microscopic image of a stainless steel cooling pipe embedded in a block of carbon foam. In the sample shown the heat transfer between foam and cooling pipe is established via a layer of Boron Nitride doped glue that is pushed into the open-pore foam.

Support structures for silicon tracking devices are typically made of carbon fibre reinforced polymer (CFRP), which—due the demand of high stiffness rather than high strength—employ high or even ultra-high modulus carbon fibres. These fibres have the positive side effect that carbon fibres with a high Young modulus typically also have a large thermal conductivity in fibre direction, which is beneficial for cooling the detector or is even actively used for cooling. As the HL-LHC trackers are designed for an integrated luminosity of up to $4000\,\mathrm{fb}^{-1}$ over a operation time of 12 years without maintenance and several thermo cycles, longevity and in particular moisture uptake is a concern for the mechanical support systems.

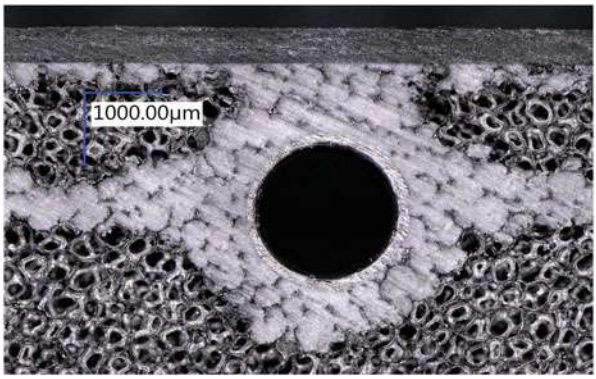

Fig. 22.1 Stainless steel cooling pipe embedded in a block of Carbon foam (credit DESY)

CFRPs with cyanate ester based resin systems are known for their low moisture uptake, however, recent industrial developments show that epoxy based systems have similar behaviour with the advantage of longer shelf life times and thus easier use of the raw material.

In general machining of CFRP with the precision required for e.g. positioning of the sensitive detector modules is not feasible, especially for layouts with small number of layers. The designs of tracker support structure therefore often follow the paradigm of "precision by glueing". The positioning elements requiring high precision machining and placement are made from e.g. Aluminium or PEEK plastic and placed on a jig prior to the assembly. The CFRP parts are then glued to these positioning elements resulting in a stiff and precise support structure. With this design and production method the tolerances on the machining and production of the used CFRP can be relaxed, which eases the production process and reduces cost while maintaining the quality of the final support structure.

22.3.2 Emerging Detector Concepts for the FCC-hh

FCC-hh will pose new challenges to the detectors [10]. A 100 TeV proton collider has not only discovery potential, given by the increased energy compared to LHC, but will also provide precision measurements as the cross sections for Standard Model (SM) processes in combination with the high luminosity lead to large event samples [13]. The envisaged detector concepts must therefore be able to measure multi-TeV jets, leptons, and photons from heavy resonances as well as Standard Model processes with high precision. As the established SM particles are small in mass, compared to the 100 TeV CMS energy of the collider, event topologies will be heavily boosted into the forward directions. A further challenge are the expected simultaneous pp collisions in one bunch crossing ('pile-up') that are expected to

reach numbers of 1000 at the FCC-hh, significantly above what is seen at LHC (60) and expected for HL-LHC (200). In particular, the anticipated separation between vertices of pile-up events is of the same order as the multiple scattering effect on the tracker vertex resolution, which renders resolving pile-up with classical 3D tracking nearly impossible. A promising approach to overcome this problem is to use 4D tracking by adding precise timing information to the tracker hits and exploiting the time structure of the pile-up events. For its HL-LHC operation the CMS experiment is foreseeing this approach already by introducing of the so-called MIP Timing Detector (MTD) that will be installed directly after the future tracker and provide timing information with a resolution of about 30 ps [14].

A reference detector for FCC-hh has been defined that at this time does not represent a specific choice for the final implementation, but rather serves as a concept for the study of physics potential and subsystem studies [10]. Figure 22.2 shows a rendering of the reference detector together with a quadrant view that shows the coverages in $|\eta|$. The detector has an overall length of 50 m and a diameter of 20 m. The central detector covers the regions of $|\eta| \leq 2.5$. Two forward

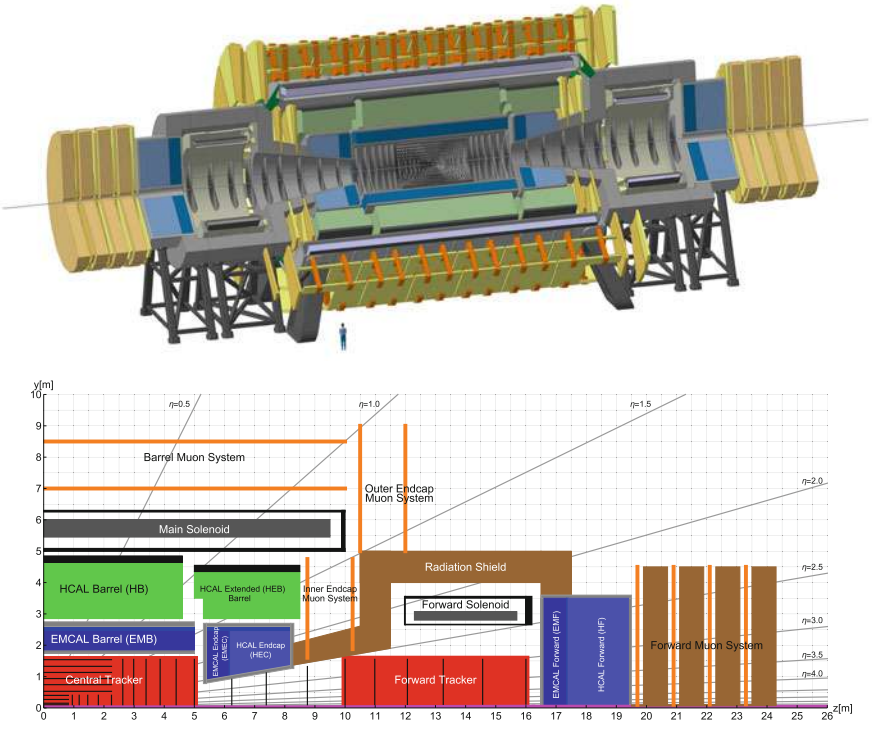

Fig. 22.2 The FCC-hh reference detector (top) has an overall length of 50 m and a diameter of 20 m. The quadrant view (bottom) shows the main detector elements and the coverage in $|\eta|$. Both figures from [10] (credit CERN/CC BY 4.0)

spectrometers cover rapidity regions of up to $|\eta| \approx 4$. A central detector solenoid with an inner bore of 10 m delivers a field of 4 T for the central regions. Two options are under study for the forward spectrometer magnets, either solenoids or dipoles. No iron return yokes are foresee, as the necessary amount of iron would be very heavy and expensive. As a consequence, the magnetic stray fields in the detector cavern will be significant which raises the need for separate service caverns some distance away.

The central tracker extends to a radius of 1.6 m. The calorimeter system consists of a LAr electromagnetic calorimeter with a thickness of 30 radiation lengths and a scintillator-iron based hadronic calorimeter of 10.5 nuclear interaction lengths. A muon system is foreseen for the outer and forward parts of the detector.

A significant challenge for the FCC-hh detector will be the radiation levels. Figure 22.3 (top) shows the expected total ionising dose rate in the detector components after a total luminosity of 30 ab^{-1} has been integrated. It is expected that the total rate for the inner tracking layers would accumulate to about 300 MGy. The radiation levels in the hadronic calorimeters would be at about 6–8 kGy,

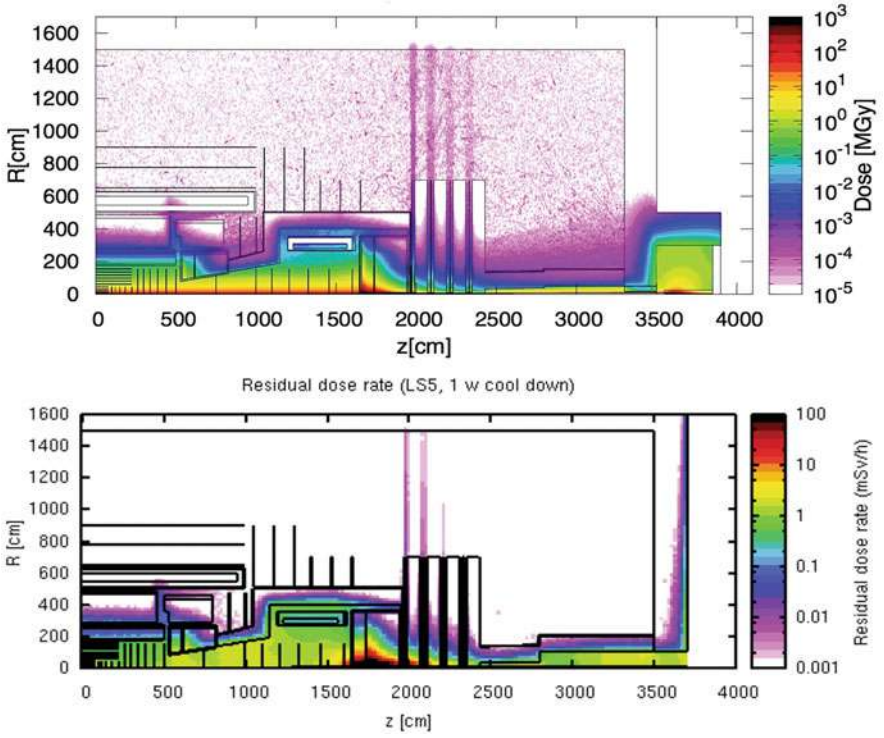

Fig. 22.3 Top: Total ionising dose for 30 ab^{-1} of integrated luminosity. Bottom: Radiation dose after one week of cool-down towards the end of the FCC-hh operation [10] (credit CERN/CC BY 4.0)

which is below the limiting number for the use of organic scintillators. Figure 22.3 (bottom) shows the radiation dose rate after one week of cool-down time towards the end of FCC-hh operations. The resulting dose rates of about 1 mSv/h in the tracker volume put limitations on person access for maintenance purposes.

22.4 Electron-Positron Colliders

The realisation of high energy electron-positron collisions has been the subject of many studies over the last years. Two fundamentally different options exist: a large circular collider, as e.g. proposed in form of the FCC-ee at CERN, or a linear collider. Due to synchroton radiation losses a circular collider is limited in its energy reach. The FCC proposal, with a ring of about 100 km in circumference, could reach with acceptable losses a final energy enough to reach the top-pair production threshold. It is economically not very sensible to go beyond this energy stage.

A linear accelerator on the other hand is intrinsically capable to reach higher energies, by extending the length of the accelerator. Over the last 20 years several technologies have been developed which promise to reach a centre-of-mass energy of 1 TeV. The international linear collider, ILC, uses superconducting cavities, a by now well established and mature technology. A fully costed design has been published in 2012 [16]. With the successful completion of the construction of the European XFEL, a large system based on the same technology has been build and successfully commissioned, providing a solid basis for estimating both costs and risks associated with this technology. An artist's drawing of the ILC facility is shown in Fig. 22.4.

To reach even higher energies the superconducting technology is not very well suited, as the achievable accelerating gradients are limited and, thus, the systems

Fig. 22.4 Artist's view of the ILC tunnel in Japan. Credit: Rey Hori/KEK

will become too large. An option based on normal conducting cavities and an innovative two-beam acceleration scheme is under development at CERN in the context of the CLIC collaboration. Even more ambitious projects like plasma accelerators are being discussed as well, but are far from being available for large scale systems [17].

Politically, Japan has been discussing to come forward and host the ILC. At the time of writing this report, no final decision has been reached.

At the core of the ILC are superconducting radio-frequency cavities, made from Niobium, which accelerate the beams. After many years of intense research and development the Tesla Technology Collaboration (TTC, [18]) has developed these cavities and industrialised their production. About 800 such cavities are used in the European Free Electron Laser, built at DESY, the European XFEL [19]. Here an average acceleration gradient of 23.5 MV/m has been reached routinely, with most cavities exceeding the design value by far, almost reaching ILC design requirements. For the ILC a gradient of 31.5 MV/m is anticipated, which, however, at the time of writing this report seems to be in reach, but has not yet been realised for large numbers of cavities nor in an industrial type series production environment. Recently, an intense R&D efforts has been started to further increase the reachable gradients in superconducting RF structures. Nitrogen doping, discovered at Fermilab [20], is one subject of study, as are alternative shapes of the cavities, optimisation of the preparation of the Niobium material, and other ideas. It is hoped that the results from this R&D, which is however not the subject of this review, will significantly reduce the cost of the ILC project.

The ILC facility poses many additional challenges to the accelerator builder, which are being attacked in an intense and long-term research and development (R&D) program. The preparation of low emittance beams, the production of high intensity polarised positron beams, and the final focus of the high energy beams down to nanometer spot sizes are just some of these [21].

The key parameters of the proposed ILC facility are summarised in Table 22.2. With the current knowledge from the LHC, the importance of a high luminosity run at the Higgs threshold is strongly stressed, which led to the re-definition of the first stage of the ILC as a 250 GeV collider [24]. This also results in a significant cost saving for this first stage, an important consideration for the political discussions taking place in Japan and elsewhere. Such a collider could be realised in a tunnel infrastructure of about 20 km length. In Japan a promising site in the north of the country has been identified, which is under close scrutiny at the moment. However, it should be noted that no official decision has been reached by Japan neither on hosting the ILC, nor on its location within Japan.

The CLIC accelerator is based on normal conducting cavities, operated at 12 GHz, reaching gradients of between 80 and 120 MV/m. It is based on a novel 2-beam acceleration scheme, where one high power, low energy beam is used to produce the radio frequency needed to accelerate the high energy beam. The feasibility of this technology is investigated at CERN at the CLIC Test Facility. Over

Table 22.2 Some basic design parameters of the ILC (250 and 500 GeV options [22, 24]), CLIC (3 TeV option) [8] and FCC-ee (240 GeV parameters) [9]

Parameter	Unit	ILC-250	ILC-500	CLIC	FCC-ee
Center-of-mass energy	GeV	250	500	3000	240
Peak luminosity/IP	$cm^{-2}s^{-1}$	1.35×10^{34}	1.79×10^{34}	5.9×10^{34}	8.5×10^{34}
Pulse rate	Hz	5	5	50	
Pulse length	μs	727	727	0.24	
Number of bunches/pulse	#	1312	1312	312	328/beam
Time between bunches	ns	554	554	0.5	994
Beam size (horizontal) at IP	nm	516	474	40	13,748
Beam size (vertical) at IP	nm	7.7	5.9	1	36
Bunch length at IP	μm	300	300	44	3150
Electron polarisation	%	>80	>80	>80	0
Positron polarisation (optional)	%	>30	>30	>30	0
Accelerator length	km	20.5	33.4	≈50	97.756
Total site AC power	MW	129	164	589	308

the last year significant progress was made on demonstrating the CLIC technology (see [8] and references therein). However a major limitation remains the lack of a significant demonstration setup, which would allow full system tests in a sizeable installation.

In recent years efforts to study a circular collider option have intensified. Both at CERN and in China designs are being developed for a circular collider, based in a tunnel of about 100km in circumference, which could host an electron positron collider. The technology for such a collider is available and does not provide unsurmountable challenges, apart from the scale of the project. A design study is currently ongoing, led by CERN, to develop a conceptual design report for a such a collider hosted in the Geneva area [9]. A similar study, CEPC, is led by IHEP in Beijing, about hosting this collider in China [11]. A circular collider at the Higgs threshold would be able to deliver integrated luminosities which are—at the Higgs production threshold—higher by a factor >5 for the same running time and one interaction region, as a linear collider. It could also serve more than one interaction region simultaneously in the recirculating beams, adding up the integrated luminosities of each installed experiment. This is a big advantage over a linear collider, where the colliding beams are used only once and disposed off in beam dumps after the collision. On the other hand, the infrastructure for a 100 km installation becomes very challenging, and the energy reach of a circular machine is limited due to the losses by synchrotron radiation. It is clear that any electron-positron collider that goes beyond 350 GeV has to be linear. In that respect, linear colliders do scale with energy while circular colliders do not.

For the experimenter however the challenges at any of the proposed electron-positron collider facilities are similar. The biggest difference between the proposals is the distance between bunches. At the ILC and FCC-ee (at the Higgs threshold) this time difference is with a few 100 ns very benign. At CLIC bunch distances at sub-ns

level are anticipated, and pose additional challenges to the experiment. Nevertheless, the goals for all facilities are the same: the experiment should be able to do precision physics, even for hadronic final states, should allow the precise reconstruction of charged and neutral particles, of secondary vertices. It has to function with the very large luminosity proposed for these machines, including significant backgrounds from beam-beam interactions.

22.4.1 Physics at an LC in a Nutshell

The design of a detector at a large facility like the ILC or CLIC can not be described nor understood without some comprehension about the type of measurements which will be done at this facility. A comprehensive review of the proposed physics program at the ILC facility can be found in [23, 25, 26], a review of CLIC physics is available under [7, 8].

The discussion in this section concentrates on the physics which can be done at a facility with an energy below 1 TeV. In recent years, the physics reach of a facility operating at around 250 GeV has been closely scrutinised, both at ILC and at CLIC (which is proposed to run in an initial energy stage at 380 GeV). Earlier studies have looked at the science case for a 500 GeV machine, and have explored the additional measurements possible if an energy upgrade up to 1 TeV might be possible.

At a center-of-mass energy of 250 GeV the ILC will be able to create Higgs bosons in large numbers, mostly in the so called Higgs-Strahlungs process. Here a Higgs boson is produced associated to a Z boson. The great power of this process is that by reconstructing the Z, and knowing the initial beam energies, one can reconstruct the properties of the Higgs boson without ever looking at the Higgs boson itself. Thus a model independent and decay mode-blind study of the Higgs particle will become possible. In addition through the reconstruction of exclusive final states for the Higgs particle, high precision measurements of the branching ratios will be possible. On its own, precisions on the most relevant branching ratios of around 1% will be possible. Combined with the results from the LHC, this precision can be pushed to well below the percent level. Samples of the heavy electroweak bosons, W and Z, a focus of the program at the LEP collider, will in addition be present in large numbers, and might still present some surprises if studied in detail.

If the energy of the facility can be increased to above 350 GeV, top-quark pairs can be produced thus turning the ILC into a top factory. Again, due to the cleanliness of the initial and the final states, high precision reconstruction of the top and its parameters will become possible. A precision scan of the top pair production threshold would determine the top mass with a statistical error of 27 MeV [27], which relates to a relative precision of $\approx$0.015%, far better than what can be done at LHC.

Operating at 500 GeV or slightly above, the ILC will gain access to the measurement of the top-Higgs coupling, and start to become sensitive to a measurement

of the Higgs self coupling. This latter experiment might provide evidence for the existence of this interaction at 500 GeV, but would vastly profit from even higher energies. At 1 TeV the Higgs self coupling could be measured to within 10%, which allows for reconstructing the Higgs potential and, thus, testing a cornerstone of the predictions of the standard model and the Higgs sector. Together these measurements will allow a complete test of the Higgs sector, and thus a in-depth probe of the standard model in this unexplored region.

There are good reasons to assume that the standard model is only an effective low-energy theory of a more complex and rich theory. A very popular extension of the standard model is supersymmetry, which predicts many new states of matter. Even though the LHC sofar has not found any evidence for supersymmetry, many models exists which predict new physics in a regime mostly invisible to the LHC. Together ILC and LHC would explore essentially the complete phase space in the kinematic regime accessible at the energy of the ILC.

Should a new state of matter be found at either the LHC or the ILC, electron-positron collisions would allow to study this sign of new physics with great precision.

In addition to direct signs of new physics, as represented by new particles, the ILC would allow to indirectly explore the physics at the Terascale through precision measurements, up to energy scales which in many cases are equivalent if not higher than those at the LHC. It might well be, if no new physics is found at the LHC, that these precision measurements at comparatively low energies are our only way to learn more about the high-energy behaviour of the standard model, and to point at the right energy regime where new physics will manifest itself.

Even though the ILC has been at the focus of the discussions in this chapter, all other electron-positron collider options will have a very similar physics reach—for those energies which are reachable at each facility.

22.5 Experiments at a Lepton Collider

As discussed in the previous section, high energy lepton collisions offer access to a broad range of scientific questions. A hallmark of this type of colliding beam experiments is the high precision accessible for many measurements. A detector at such a facility therefore has to be a multi-purpose detector, which is capable to look at many different final states, at many different signatures and topologies. In this respect the requirements are similar to the ones for a detector at a hadron collider. The direction in which an lepton collider detector is optimised however is very different. Lepton collider detectors are precision detectors—something which is possible because the lepton collider events are comparatively clean, backgrounds are low, and rates are small compared to the LHC. The collision energy at the lepton collider is precisely known for every event, making it possible to measure missing mass signatures with excellent precision. This will make it possible to measure masses of supersymmetric particles with precision, or, in fact, masses of

any new particle within reach of the collider. The final states are clean and nearly background-free, making it possible to determining absolute branching ratios of essentially every state visible at the lepton collider. The reconstruction also of hadronic final states is possible with high precision, opening a whole range of states and decay modes which are invisible at a hadron machine due to overwhelming backgrounds.

This results in a unique list of requirements, and in particular on very high demands on the interplay between different detector components. Only the optimal combination of different parts of the detector can eventually deliver the required performance.

Many of the interesting physics processes at an LC appear in multi-jet final states, often accompanied by charged leptons or missing energy. The reconstruction of the invariant mass of two or more jets will provide an essential tool for identifying and distinguishing W's, Z's, H's, and top, and discovering new states or decay modes. To quantify these requirements the di-jet mass is ofter used. Many decay chains of new states pass through W or Z bosons, which then decay predominantly into two jets. To be able to fully reconstruct these decay chains, the di-jet mass resolution should be comparable or better than the natural decay width of the parent particles, that is, around 2 GeV for the W or Z:

$$\frac{\Delta E_{di-jet}}{E_{di-jet}} = \frac{\sigma_m}{M} = \frac{\alpha}{\sqrt{E\,(\mathrm{GeV})}}, \tag{22.1}$$

where E denotes the energy of the di-jet system. With typical di-jet energies of 200 GeV at a collision energy of 500 GeV, $\alpha = 0.3$ is a typical goal. Compared to the best existing detectors this implies an improved performance of around a factor of two. It appears possible to reach such a resolution by optimally combining the information from a high resolution, high efficiency tracking system with the ones from an excellent calorimeter. This so called particle flow ansatz [28, 29] is driving a large part of the requirements of the LC detectors.

Table 22.3 summarises several selected benchmark physics processes and fundamental measurements that make particular demands on one subsystem or another, and set the requirements for detector performance.

22.5.1 Particle Flow as a Way to Reconstruct Events at a Lepton Collider

Particle flow is the name for a procedure to optimally combine information from the tracking system and the calorimeter system of a detector, i.e. to fully reconstruct events. Particle flow has been one of the driving forces in the optimisation of the detectors at a Lepton Collider.

Typical events at the LC are hadronic final states with Z and W particles in the decay chain. In the resulting hadronic jets, typically around 60% of all stable

Table 22.3 Sub-Detector performance needed for key LC physics measurements (from [30])

Physics process	Measured quantity	Critical system	Critical detector characteristic	Required performance
$ZHH\ HZ \to q\bar{q}b\bar{b}$ $ZH \to ZWW^*$ $\nu\bar{\nu}W^+W^-$	Triple Higgs coupling Higgs mass B($H \to WW^*$) $\sigma(e^+e^- \to \nu\bar{\nu}W^+W^-)$	Tracker and calorimeter	Jet energy resolution, $\Delta E/E$	$3 \sim 4$ % or $30\%/\sqrt{E}$
$ZH \to \ell^+\ell^- X$ $\mu^+\mu^-(\gamma)$ $ZH + H\nu\nu \to \mu^+\mu^- X$	Higgs recoil mass luminosity weighted E_{cm} B($H \to \mu^+\mu^-$)	Tracker	Charged particle momentum resolution, $\Delta p_t/p_t^2$	5×10^{-5}
$HZ, H \to b\bar{b}, c\bar{c}, gg$ $b\bar{b}$	Higgs branching fractions b quark charge asymmetry	Vertex detector	Impact parameter, δ_b	$5\,\mu\text{m} \oplus$ $10\,\mu\text{m}/p(\text{GeV/c})\sin^{3/2}\theta$

particles are charged, slightly less than 30% are photons, only around 10% are neutral long lived hadrons, and less than 2% are neutrinos. At these energies charged particles are best re-constructed in the tracking system. Momentum resolutions which are reached in detectors are $\delta p/p^2 \approx 5 \times 10^{-5}\,\mathrm{GeV}^{-1}$, much better than any calorimeter system at these energies. Electromagnetic energy resolutions are around $\delta E_{em}/E = 0.15/\sqrt{E}(\mathrm{GeV})$, typical resolutions achieved with a good hadronic calorimeter are around $\delta E_{had}/E = 0.45/\sqrt{E}(\mathrm{GeV})$. Combining these with the proper relative weights, the ultimate energy resolution achievable by this algorithm is given by

$$\sigma^2(E_{jet}) = w_{tr}\sigma_{tr}^2 + w_\gamma\sigma_\gamma^2 + w_{h^0}\sigma_{h^0}^2, \tag{22.2}$$

where w_i are the relative weights of charged particles, photons, and neutral hadrons, and σ_i the corresponding resolution. Using the above mentioned numbers an optimal jet mass resolution of $\delta E/E = 0.16/\sqrt{E}(GeV)$ can be reached. This error is dominated by the contribution from the energy resolution of neutral hadrons, assumed to be $0.45/\sqrt{E}(GeV)$. This formula assumes that all different types of particles in the event can be individually measured in the detector. This implies that excellent spatial resolution in addition to the energy resolution is needed. Thus fine-grained sampling calorimeters are the only option currently available which can deliver both spatial and energy resolution at the same time. This assumption is reflected in the resolution numbers used above, which are quoted for modern sampling type calorimeters. Even though an absorption-type calorimeter—for example a crystal calorimeter as used in the CMS experiment—can deliver better energy resolution, it falls significantly behind in the spatial resolution, thus introducing a large confusion term in the above equation.

Formula 22.2 describes a perfect detector, with perfect efficiency, no acceptance holes, and perfect reconstruction in particular of neutral and charged particles in the calorimeter. In reality a number of effects result in a significant deterioration of the achievable resolution. If effects like a final acceptance of the detector, missing energy e.g. from neutrinos etc. is included, this number easily increases to $25\%/\sqrt{E}$ [31]. All this assumes that no errors are made in the assignment of energy to photons and neutral hadrons. The optimisation of the detector and the calorimeter in particular has to be done in a way that these wrong associations are minimised.

From the discussion above it is clear that three effects are of extreme importance for a detector based on particle flow: as good hadronic energy resolution as possible, excellent separation of close-by neutral and charged particles, and excellent hermeticity. It should also be clear that the ability to separate close-by showers is more important than ultimate energy resolution: it is for this reason that total absorption calorimeters, as used e.g. in the CMS experiment, are not well suited for the particle flow approach, as they do not lend themselves to high segmentation.

Existing particle flow algorithms start with the reconstruction of charged tracks in the tracking system. Found tracks are extrapolated into the calorimeter, and linked with energy deposits in there. If possible, a unique assignment is made between a track and an energy deposit in the calorimeter. Hits in the calorimeter belonging

to this energy deposit are identified, and are removed from further considerations. The only place where the calorimeter information is used in the charged particle identification is in determining the type of particle: calorimeter information can help to distinguish electrons and muons from hadrons. A major problem for particle flow algorithms are unassigned clusters, and mis-assignments between neutral and charged deposits in the calorimeter. The currently most advanced particle flow algorithm, PandoraPFA, tries to minimise these effects by a complex iterative procedure, which optimises the assignments, goes through several clean-up steps, and tries to also take the shower sub-structure into account [31].

What is left in the calorimeter after this procedure is assumed to have come from neutral particles. Clusters in the calorimeter are searched for and reconstructed. With a sufficiently high segmentation both transversely and longitudinally, the calorimeter will be able to separate photons from neutral hadrons by analysing the shower shape in three dimensions. A significant part of the reconstruction will be then the reconstruction of the neutral hadrons, which leave rather broad and poorly defined clusters in the hadronic calorimeter system.

Particle flow relies on a few assumptions about the event reconstruction. For it to work it is important that the event is reconstructed on the basis of individual particles. It is very important that all charged tracks are found in the tracker, and that the merging between energy deposits in the calorimeter and tracks in the tracker is working as efficiently as possible. Errors in this will quickly produce errors for the total energy, and in particular for the fluctuations of the total energy measured. Not assigning all hits in the calorimeter to a track will also result in the creating of additional neutral clusters, the so called double counting of energy. Reconstructing all particles implies that the number of cracks and the holes in the acceptance should be minimised. This is of particular importance in the very forward direction, where the reconstruction of event properties is complicated by the presence of backgrounds. However, small errors in this region will quickly introduce large errors in the total energy of the event, since many processes are highly peaked in the forward direction.

In Fig. 22.5 the performance of one particular particle flow algorithm, PandoraPFA [31] is shown, as a function of the dip angle of the jet direction, $\cos\theta$. The performance for low energies of the jets, 45 GeV is close to the optimally possible resolution if the finite acceptance of the detector is taken into account. At higher energies particles start to overlap, and the reconstruction starts to pick up errors in the assignment between tracks and clusters. This effect, called confusion, will deteriorate the resolution, and will increase at higher energies. Jets at higher energies are boosted more strongly, resulting in smaller average distances between particles in the jet. This results in a worse separation of particle inside the jet, and thus a worse resolution. Figure 22.6 shows an event display of a simulated hadronic jet in the ILD detector concept for the ILC with particle flow objects reconstructed by PandoraPFA. The benefit of a highly granular detector system is clearly visible.

Over the last 10 years, the Pandora algorithm has matured into a robust and stable algorithm. It is now used not only in the linear collider community, but also in long baseline neutrino experiments, and is under study at the LHC experiments.

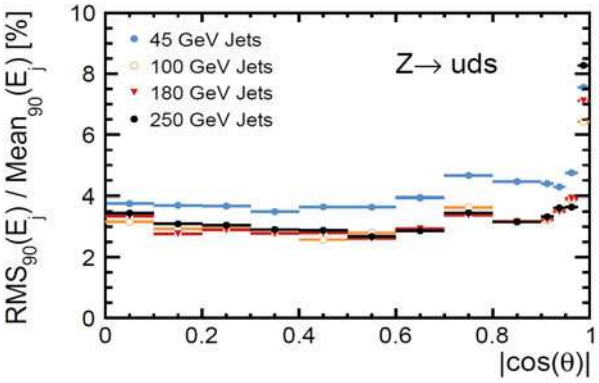

Fig. 22.5 The jet energy resolution, α, as a function of the dip angle $|\cos\theta_q|$ for jets of energies from 45 GeV to 250 GeV

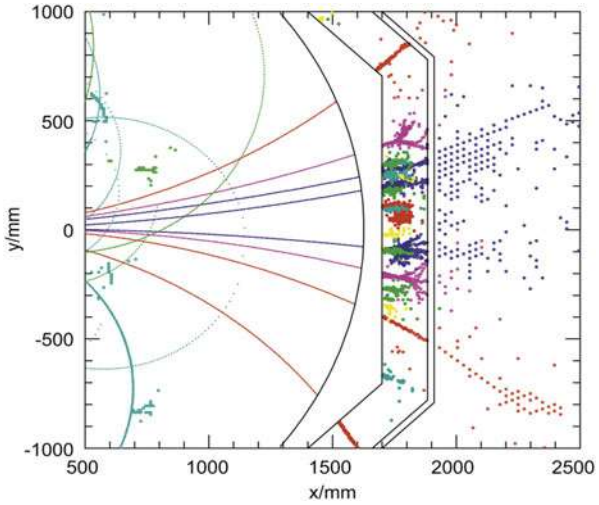

Fig. 22.6 Simulated jet in the ILD detector, with particle flow objects reconstructed by the Pandora algorithm shown in different colors

22.5.2 A Detector Concept for a Lepton Collider

Over the years a number of concepts for integrated detectors have been developed for use at a lepton collider [32–36]. Broadly speaking two different models exist: one based on the assumption, that particle flow is the optimal reconstruction technique, the other not based on this assumption. Common to all proposals is that both the tracking system and the calorimeter systems are placed insides a large superconducting coil which produces a large magnetic field, of typically 3–5 T. Both concepts use high precision tracking and vertexing systems, inside solenoidal

fields, which are based on state of the art technologies, and which really push the precision in the reconstruction of the track momenta and secondary vertices. Differences exist in detail in the choice of technology for the tracking devices, some rely heavily on silicon sensors, like the LHC detectors, others propose a mixture of silicon and gaseous tracking. The calorimeters are where these detectors are most different from current large detectors. The detectors based on the particle flow paradigm propose calorimeters which are more like very large tracking systems, with material intentionally introduced between the different layers. Systems of very high granularity are proposed, which promise to deliver unprecedented pictures of showering particles. Another approach is based on a more traditional pure calorimetric approach, but on a novel technology which promises to eventually allow the operation of an effectively compensated calorimeter [34].

At the ILC, detectors optimised for particle flow have been chosen as the baseline. The two proposed detector concepts ILD [32] and SiD [33] differ in the choice of technology for the tracking detectors, and on the overall emphasis based on particle flow performance at higher energies. However, both detectors have been optimised for collision energies of less than 1 TeV, while within the CLIC study the detector concepts have been further evolved to be optimised for operation at energies up to 3 TeV [35].

A conceptual picture of the ILD detector, as proposed for the ILC, is shown in Fig. 22.7. Visible are the inner tracking system, the calorimeter system inside the coil, the large coil itself, and the iron return yoke instrumented to serve as a muon identification system. A cut view of a quadrant with the sub-systems of ILD is shown in Fig. 22.8.

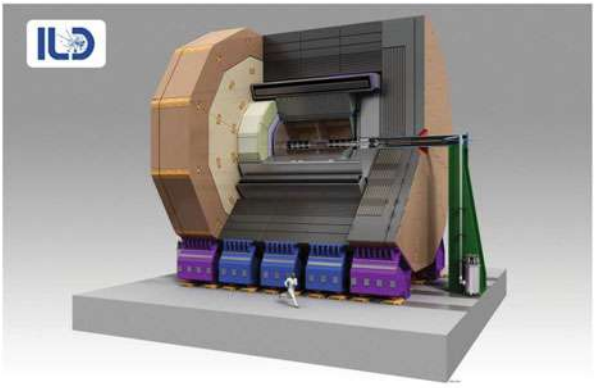

Fig. 22.7 Three-dimensional view of a proposed detector concept for the ILC, the ILD detector [32] (credit Ray Hori, KEK)

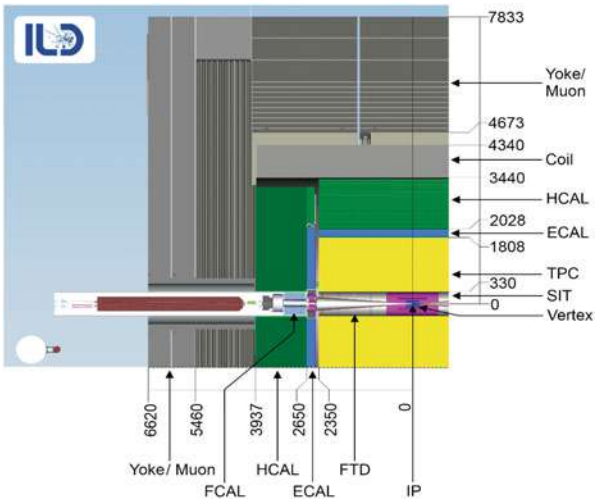

Fig. 22.8 Cut through the ILD detector in the beam plane, showing one quarter of the detector [37]

22.6 Detector Subsystems

A collider detector has a number of distinct sub-systems, which serve specific needs. In the following the main systems are reviewed, with brief descriptions of both the technological possibilities, and the performance of the system.

22.6.1 Trends in Detector Developments

Detector technologies are rapidly evolving, partially driven by industrial trends, partially itself driving technological developments. New technologies come into use, and disappear again, or become accepted and well-used tools in the community. A challenge for the whole community is that technological trends change faster than ever, while the design, construction and operation cycles of experiments become longer. Choosing a technology for a detector therefore implies not only using the very best available technology, but also one which promises to live on during the expected lifetime and operational period of the experiment. An example of this are Silicon technologies, which are very much driven by the demands of the modern consumer electronics industry. By the times Si detectors are operational inside an experiment, the technology used to built them is often already outdated, and replacements or extensions in the same technology are difficult to get. Even more so than to the sensors this applies to readout and data acquisition systems.

Because of the rapid progress in semiconductor technology, feature sizes in all kind of detector are getting ever smaller. Highly integrated circuits allow the integration of a great deal of functionality into small pixels, allowing the pixellation of previously unthinkable volumes. This has several consequences: the information about an event, a particle, a track, becomes ever larger, with more and more details at least potentially available and recorded. More and more the detection of particles, of properties of particles, rely no longer on averaging its behaviour over a volume large compared to the typical distances involved in the process used to measure the particle, but allows the experimenter to directly observe the processes which eventually lead to a signal in the detector. Examples of this are e.g. the Si-TPC (Silicon readout Time Projection Chamber, described in more detail below) where details of the ionisation process of a charged particle traversing a gas volume can be observed, or the calorimeter readout with Si-based pixellated detectors, given unprecedented insights into the development of particle showers. Once the volume read out becomes small compared to the typical distances involved in the process which is being observed, a digital readout of the information can be contemplated. Here, only the density of pixels is recorded, that is, per pixel only the information whether or not a hit has occurred, is saved. This results in potentially a much simpler readout electronics, and in more stable and simpler systems. These digital approaches are being pursued by detectors as different as a TPC and a calorimeter.

Increasing readout speed is another major direction of developments. It is coupled but not identical to the previously discussed issue of smaller and smaller feature sizes of detectors. Because of the large number of channels, faster readout systems need to be developed. An even more stringent demand however comes from the accelerators proposed, and the luminosities needed to make the intended experiments. They can only be used if data are readout very quickly, and stored for future use. To give a specific example: the detector with the largest numbers of pixels ever built so far (until the Phase-II upgrades of the LHC detectors) has been the SLD detector at SLAC which operated during the 1990. The vertex detector, realised from charged coupled sensors with some 400 Million channels, was readout with a rate of around 1 MHz. For the ILC readout speeds of at least 50 MHz, maybe even more, are considered, to cope with larger data rates and smaller inter-bunch spacings.

Technological advances in recent years have made it feasible to consider the possibility to do precision timing measurements with semi-conductor detectors. Timing resolutions in the range of 100 ps or better are becoming feasible, something completely unthinkable only a few years ago. This capability—somewhat orthogonal to the readout speed discussed above—can significantly extend the capabilities of semiconductor tracker, into the direction of so-called 4D tracking or calorimeter systems. Timing information at this level of precision can be used to measure the mass of particles through time-of-flight, and can help to separate out-of-time background from collision related events.

For many applications, particularly at the LHC, radiation hardness is at a premium. Major progress has been made in recent years in understanding damage mechanisms, an understanding, which can help to design better and more radiation hard detectors. For extreme conditions novel materials are under investigation.

22.6.2 Vertex Detectors: Advanced Pixel Detectors

Many signals for interesting physics events include a long lived particle, like e.g. a B or charmed hadron, with typical flight distances in the detector from a few 10 μm to a few mm. The reconstruction of the decay vertices of these particles is important to identify them and to distinguish their decay products from particles coming from the primary vertex, or to reconstruct other event quantities like vertex charge.

To optimally perform these functions the vertex detector has to provide high precision space points as close as possible to the interaction point, has to provide enough space points, so that an efficient vertex reconstruction is possible over the most relevant range of decay distances, of up to a few cm in radius, and present a minimal amount of material to the particle so as to not disturb their flight path. Ideally, the vertex detector also offers enough information that stand-alone tracking is possible based only on vertex detector hits.

At the same time a vertex detector has to operate stably in the beam environment. At a hadron collider it has to stand huge background rates, and cope with multiple interactions. At a lepton collider, very close to the interaction point a significant number of beam background particles may traverse the detector, mostly originating from the beam–beam interaction. These background particles are bent forward by the magnetic field in the detector. The energy carried away by this beamstrahlung may be several 10 TeV, which, if absorbed by the detector, would immediately destroy the device. The exact design of the vertex detector therefore has to take into account these potential backgrounds. At a hadron collider, the largest challenge will be to design the detector such that it can survive the radiation dose and is fast enough and has small enough pixels to cope with the large particle multiplicity. Here pixel size, readout speed, and radius of the detector are the main parameters which need to be optimised. At a lepton collider, both size and magnetic field can be used to make sure that the detector stays clear of the majority of the background particles. The occupancy at any conceivable luminosity is not driven by the physics rate, but only by the background events. Since they are much softer than the physics events, a strong magnetic field can be used to reduce the background rates, and allow small inner radii of the system. Nevertheless, the remaining hits from beam background particles dominate the occupancies, especially at the innermost layers of a vertex detector, and therefore require fast read-out speeds.

The particular time structure of the collider has an important impact on the design and the choice of the technology. At the ILC collisions will happen about every 300 ns to 500 ns, in a train of about 1 ms length, followed by a pause of around 200 ms. About 1300 bunches are expected to be in one train. A fast readout of the vertex detector is essential to ensure that only a small number of bunches are superimposed within the occupancy of the vertex detector. At CLIC the inter-bunch spacing is much smaller, putting a premium on readout speed. At LHC the typical time between collisions in a bunch crossing is order 100 ps decreasing to about 10 ps at the high luminosity LHC-HL.

A Si-pixel based technology is considered the only currently available technology which can meet all these requirements. A small pixel size ($< 20 \times 20\,\mu m^2$) combined with a fast read out will ensure that the occupancy due to backgrounds and from expected signals together remain small enough to not present serious reconstruction problems. It also allows for a high space point resolution, and a true three dimensional reconstruction of tracks and vertices essentially without ambiguities. Several silicon technologies are available to meet the demands. Increasingly, sensors based on the CMOS process are considered. Most recently devices with intrinsic gain larger than one are studied intensely, as they promise excellent performance combined with very good timing properties.

Quite a number of different technologies are currently under study. Broadly they can be grouped into at least two categories: those which try to read the information content as quickly as possible, and those which try to store information on the chip, and which are readout during the much longer inter-bunch time window. Another option under study is a detector with very small pixels, increasing the number of pixels to a point where even after integration over one full bunch train the overall occupancy is still small enough to allow efficient tracking and vertexing.

A fairly mature technology is the CCD technology [38, 39], which for the first time was very successfully used at the SLD experiment at the SLC collider at SLAC, Stanford. Over the past decade a number of systems based on this concept have been developed.

Newer approaches use the industrial CMOS process to develop monolithic active pixel sensors (MAPS) that are at the same time thin, fast, and radiation hard enough for particle physics experiments [40]. A smaller scale application of this technology is a series of test-beam telescopes, based on the Mimosa families of chips [41], built under the EUDET and AIDA European programs [42, 43] and operated a CERN, DESY and SLAC. The Phase-II upgrade of the ALICE experiment at the LHC contains a new inner tracking system that will be completely based on the CMOS-MAP sensor ALPIDE [44]. With a pixel size of $24.9\,\mu m \times 29.3\,\mu m$, a spatial resolution of $\approx 5\,\mu m$ and a time resolution of 5–10 µs is envisaged for hit rates of about $10^6/cm^2/s$. The CBM experiment, planned for the FAIR heavy-ion facility in Darmstadt, foresees to use MAPS for the microvertex detector. It will be based on the MIMOSIS chip, that is an advancement of the ALPIDE chip with similar pixel size and spatial resolution, but that has to cope with a much higher event rate of about $10^8/cm^2/s$ (and the associated radiation load) at the cost of a higher power consumption. The MIMOSIS chip already aims for a higher readout speed of about 5 µs.

In Fig. 22.9 a measured point resolution achieved with the CMOS-MAPS technology in a test beam experiment is shown [49]. Other technologies are at a similar level of testing and verifying individual sensors for basic performance.

Studies are underway to push the CMOS-MAPS towards even higher readout speeds [45]. The two parameters that currently govern the process are the time required for the pixel address encoding and the signal shaping during the pre-amplification. Changing the algorithm of the pixel address encoding and increasing the internal clock, could lead to an improvement from 50 ns to 25 ns for this step.

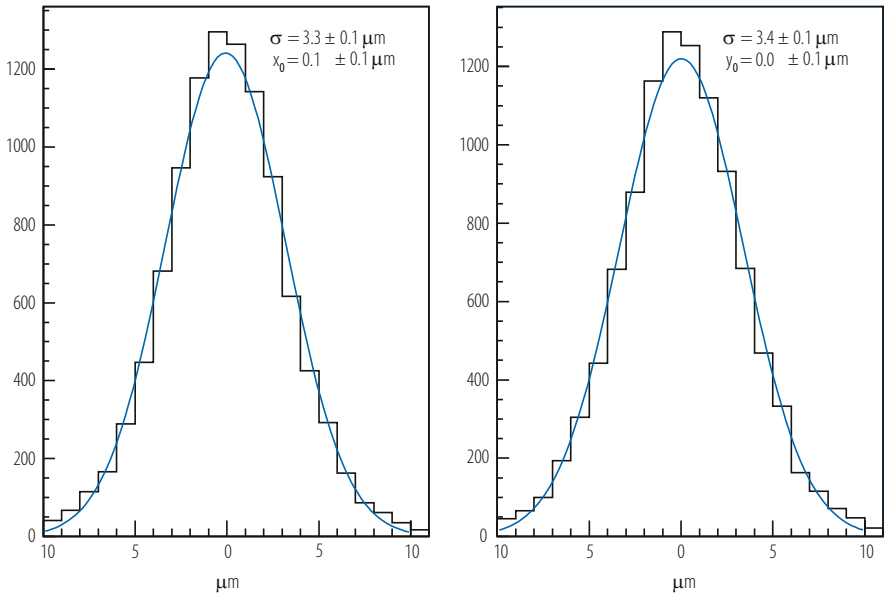

Fig. 22.9 (Left): Biased residual distribution measured in a CMOS pixel detector with 6 GeV electrons. (Right) Measured residual width in a 6 layer setup with a layer spacing of 20 mm [49]

The signal shaping currently takes about 2 μs and could be shortened to about 500 ns at the price of increasing the pixel current and therefore also increasing the power consumption. However, as the detectors at a linear collider would be operated in power-pulsing mode, the impact on the cooling requirements would be minor. Such an optimised CMOS detector for the ILC would have a readout speed of about 1 μs, i.e. it could be read out every two to three bunch crossings. Other groups explore the possibility to store charge locally on the pixel, by including storage capacitors on the pixel. Up to 20 timestamped charges are foreseen to be stored, which will then be readout in between bunch trains.

The most recent example of a pixel detector at a lepton collider has been the pixel detector for the Belle-II experiment. This system is based on the DEPFET technology [46]. Charge generated by the passage of a charged particle through the fully depleted sensitive layer is collected on the gate of a DEPFET transistor, implemented into each pixel. DEPFET sensors can be thinned to remove all silicon not needed for charge collection, to something like 50 μm, or 0.1% of a radiation length. This makes this technology well suited for lepton collider applications, where minimal material is of paramount importance [48].

A problem common to all technologies considered is the amount of material present in the detector. A large international R&D program is under way to reduce significantly the material needed to build a self-supporting detector. The goal, driven by numerous physics studies, and the desire for ultimate vertex reconstruction, is a single layer of the detector which in total presents 0.1% of a radiation length,

including sensor, readout and support. This can only be achieved by making the sensors thin, and by building state-of-the-art thin and light weight support structures. To compare, at the LHC the total amount of material present in the silicon based trackers is close to 2 radiation length, implying that per layers, close to 10% of a radiation length is present.

Very thin sensor layers are possible with technologies based on fully depleted sensors. Since here only a thin layer of the silicon is actually needed for the charge collection the rest of the wafer can be removed, and the sensor can be thinned from typically 300 µm, used e.g. in the LHC experiments, to something like 50 µm or less. Several options are under study how such thin Si-ladders can then be supported. One designs foresees that the ladders be stretched between the two endcaps of the detectors, being essentially in the active area without additional support. Another approach is to study the use of foam material to build up a mechanically stiff support structure. Carbon foam is a prime candidate for such a design, and first prototype ladders have come close to the goal of a few 0.1% X_0 [47]. Another group is investigating whether Si itself could be used to provide the needed stability to the ladder. By a sophisticated etching procedure stiffening ribs are built into the detector, in the process of removing the material from the backside, which will then stabilise the assembly. This approach has been successfully implemented for the vertex detector at the Belle-II experiment [48].

Material reduction is an area where close connections exist between developments done for the ILC and developments done for the LHC and its upgrade. In both cases minimum material is desired, and technologies developed in the last few years for the ultra-low material ILC detector are of large interest to possible upgrade detectors for LHC and LHC Phase-II.

The readout of these large pixel detectors present in itself a significant challenge. On-chip zero-suppression is essential, but also well established. Low power is another important requirement, consistent with the low mass requirement discussed above. Only a low power detector can be operated without liquid cooling, low mass can only be achieved without liquid cooling. It has been estimated that the complete vertex detector of an ILC detector should not consume on average more than 100 W, if it is to be cooled only through a gas cooling system. Currently this is only achievable if the readout electronics located on the detector is switched off for a good part of the time, possible with the planned bunch structure of the ILC. However such a large system with pulsed power has never been built, and will require significant development work. It should not be forgotten that the system needs to be able to operate in a large magnetic field, of typically 4 T. Each switching process therefore, which is connected with large current flows in the system, will result in large Lorentz forces on the current leads and the detectors, which will significantly complicate the mechanical design of the system. Nevertheless, with current technologies power pulsing is the only realistic option to achieve the desired low power operation, and thus a central requirement for the low mass design of the detector. In Fig. 22.10 the conceptual layout of a high precision vertex detector is shown.

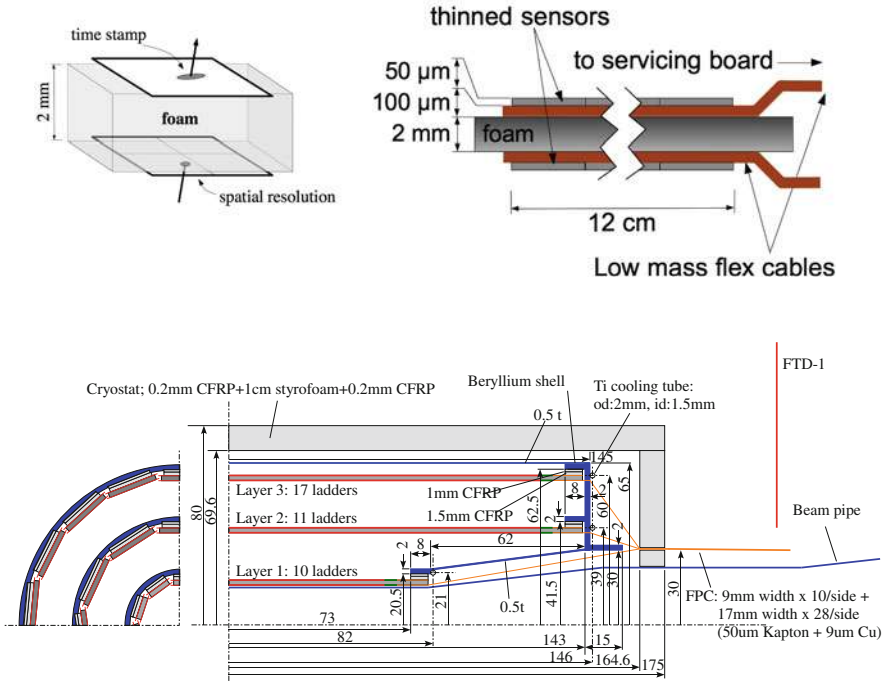

Fig. 22.10 Top: Concept of a double-layer vertex detector system developed within the PLUME project. Bottom: Vertex detector for the ILD concept, based on a layout with three double layers [37]

One of the key performance figures of a vertex detector is its capability to tag heavy flavours. At the ILC b-quarks are an important signature in many final states, but more challenging are charm quarks as the are e.g. expected in decays of the Higgs boson. Obtaining a clean sample of charm hadrons in the presence of background from bottom and light flavour is particularly difficult. Already at the SLC and the LEP collider, the ZVTOP [50] algorithm has been developed and used successfully. It is based on a topological approach to find displaced vertices. Most tracks originating from heavy flavour decays have relatively low momenta, so excellent impact parameter resolution down to small ($\approx$1GeV) energies is essential. On the other hand, due to the large initial boost of the heavy hadrons, the vertices can be displaced by large distances, up to a few cm away from the primary vertex, indicating that the detector must be able to reconstruct decay vertices also at large distances from the interaction point. The algorithms have been further developed and adapted to the expected conditions at a linear collider [51]. The performance of a typical implementation of such a topological vertex finder is shown in Fig. 22.11.

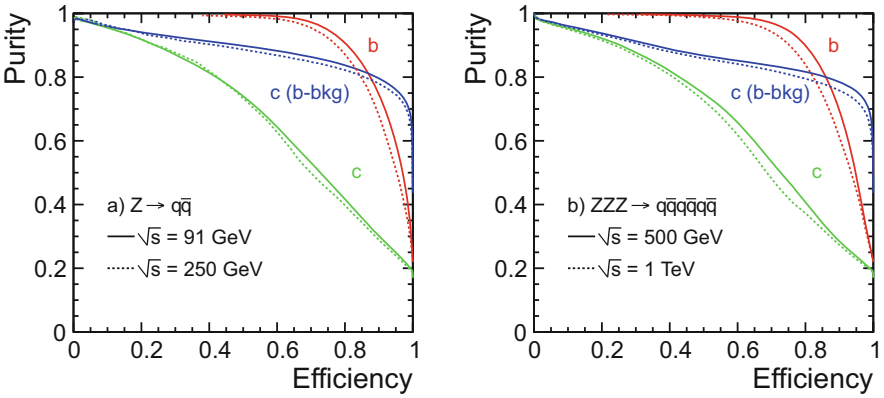

Fig. 22.11 Purity versus efficiency curve for tagging b-quarks (red points) and c-quarks (green points) and c-quarks with only b-quark background (blue points) obtained in a simulation study for Z-decays into two (left) and six (right) jets, as simulated in the ILD detector [37]

22.6.3 Solid State Tracking Detectors: Strip Detectors

To determine the momentum of a charged particle with sufficient accuracy, a large volume of the detector needs to be instrumented with high precision tracking devices, so that tracks can be reliably found and their curvature in the magnetic field can be well measured. Cost and complexity considerations make a pixel detector for such large scale tracking applications at present not feasible. Instead strip detectors are under development, which will provide excellent precision in a plane perpendicular to the electron-positron beam.

Silicon microstrip detectors are extremely well understood devices, which have been used in large quantities in many experiments, most recently on an unprecedented scale by the LHC experiments. A typical detector fabricated with currently established technology might consist of a 300 µm thick layer of high resistivity silicon, with strips typically every 50 µm running the length of the detector. Charge is collected by the strips. These detectors measure one coordinate very well, with a precision of <10 µm. The second coordinate can be measured e.g. by arranging a second layer of strip detectors at a small stereo angle. Double sided detectors, with two readout structures on either side, with strips running also at an angle to each other, have in the past proved to be a costly and not very reliable alternative to the combination of two single sided detectors back-to-back.

Strip detector have received a major boost through the upgrade program for the LHC experiments. The large area tracking systems for both ATLAS and CMS will need to be replaced in time for the start of the high luminosity phase of the LHC, scheduled to start around 2026. Several hundred square meters of Silicon detectors need to be produced, to build up these large detector systems. Compared to the previous ones, the radiation hardness of these devices had to be improved by at least an order of magnitude, and the total amount of material in the system will be reduced

significantly. This requires novel approaches to the structures, and to powering and cooling of these detectors, which will be discussed in a separate section.

A major R&D goal needed for the application of these devices to the ILC detector is the significant reduction of material per layer. As for the vertex pixel detector, thinning the detectors is under investigation, as is the combination of thinned detectors with light weight support structures and power-pulsed readout electronics. New schemes to deliver power to the detectors—like serial powering—are being studied.

22.6.4 Gaseous Tracking

Even though solid state tracking devices have advanced enormously over the last 20 years, gaseous tracking is still an attractive option for a high precision detector like an ILC detector. Earlier in this section the concept of particle flow has been discussed. Particle flow requires not the very best in precision from a tracking detector, but ultimate efficiency and pattern recognition ability. Only if charged tracks are found with excellent efficiency can the concept of particle flow really work. A large volume gaseous tracker can assist in this greatly by providing a large number of well measured points along a track, over a large volume. In addition a gaseous detector can assist in the identification of the particle by measuring the specific energy loss, dE/dx, of the particle, which for moderate momenta up to 10–20 GeV is correlated to the particle type.

A particularly well suited technology for this is the time projection chamber, TPC [52]. It has been used in the past very successfully in a number of colliding beam experiments, most recently in the ALICE experiment at the LHC [53]. A time projection chamber (see Chapt. C1 ii) essentially consists of a volume of gas, onto which a uniform electric and magnetic field is superimposed. If a charged particle crosses the volume, the produced ionisation is drifted under the influence of the field to the anode and the cathode of the volume. Since the electrons drift typically about 1000 times faster than the ions, they are usually used in the detection. A gas amplification system at the anode side is used to increase the available charge which is then detected on a segmented anode plane, together with the time of arrival. Combining both, a three dimensional reconstruction of the original crossing point is possible.

Traditionally time projection chambers are read out at the anode with multi-wire proportional chambers. They operate reliably, have a good and well controllable gas gain, and give large and stable signals. However wires are intrinsically one dimensional, which means, that a true three-dimensional reconstruction of the space point is difficult. Wires need to be mechanically stretched, which restricts the distance between them to something larger than typically 1 mm. More importantly though, the fact that all electrons produced in the drift volume are eventually collected by these wires, and that this collection happens in a strong magnetic field, limits the achievable resolution. Very close to the wire the electric field lines and the

magnetic field lines are no longer parallel, and the particle will start to deviate from the ideal straight track toward the anode. It will start to see a strong Lorentz force, which will tend to distort the drift path. This distortions will be different whether the electron approaches the wire from below or from above, and will introduce biases in the reconstruction of the space coordinate which might be similar in size to the spacing between the wires. Corrections might be applied, and can correct in part this effect, but typical uncertainties around 1/10 of the inter-wire distance might remain. This does limit the ultimately achievable resolution in a wire-equipped TPC.

An alternative which is being studied intensely is the use of micro-pattern gas detectors as readout systems in a TPC [54]. Gas electron multipliers (GEM) [55, 56] or Micro Mesh Chambers Micromegas (MM) [57, 58] are two recent technologies under investigation.

A GEM foil consists of a Polyamide foil of a typical thickness of 50–100 μm, copper clad on both sides. A regular grid of holes of 50 μm diameter spaced typically 150 μm apart connects the top and the bottom side. With a potential of a few hundred volts applied across the foil a very high field develops inside the hole, large enough for gas amplification. Gains in excess of 10^3 have been achieved with such setups. In Fig. 22.12 the cross section of a hole in a GEM is shown, together with field lines, showing clearly the high field region in the center of the hole. A challenge for the GEM based system is the development of a mechanically stable readout system. A system based on ceramic spacer structures has been developed and successfully tested [61].

A MM is constructed by stretching a metal mesh with a very fine mesh size across a readout plane, at a distance of typically less than 1 mm from the readout plane. A potential is applied between the mesh and the readout plane. The resulting field is large enough for gas amplification. Spacers at regular intervals ensure that the system is mechanically stable, and withstands the electrostatic forces.

Both systems have feature sizes which are one order of magnitude smaller than the ones in conventional wire-based readout systems, thus reducing the potential errors introduced through the gas amplification system. The smaller feature sizes in addition reduce the spatial and temporal spread of the signals arriving at the readout structure, thus promising a better two particle separation. The spatial resolution obtained in a prototype TPC equipped with a Micromegas readout is shown in Fig. 22.13.

The positive ions which are produced both in the initial ionisation along the track, and in the amplification process at the anode, will drift slowly to the cathode. Thus, the drift volume of the TPC will slowly fill with positive charge, if nothing is done, which will tend to change the space-to-time relation central to the TPC principle. Both GEM and MM suppress the drift of positive ions to the cathode, by catching a large percentage (over 98%) on the GEM foil or on the mesh [60]. To reduce the amount of positive ions even further a gating electrode can be considered. This is an electrode mounted on top of the last amplification stage, facing towards the drift volume. The potential across the gate can be changed to change the transparency of the gate for ions. At the ILC the gate can be opened for one complete bunch train, and then be closed for the inter-bunch time. This would reduce the volume affected by significant ion densities to only the first few cm in drift, above the readout plane.

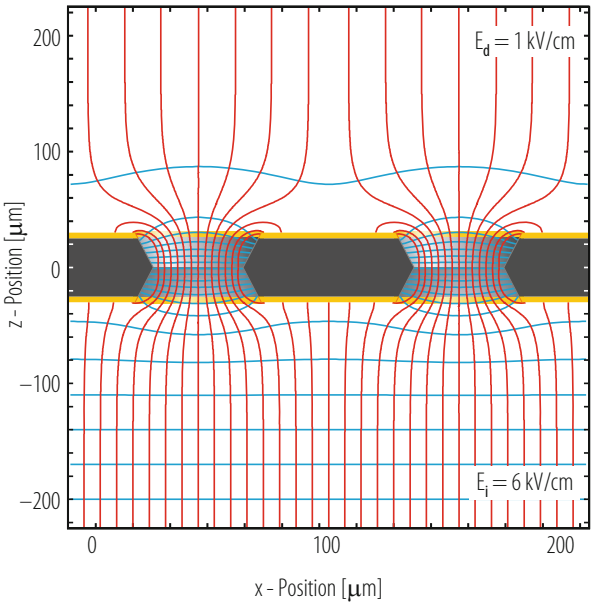

Fig. 22.12 Cross section of a hole in a GEM foil, with simulated field lines (picture credit Oliver Schäfer, DESY)

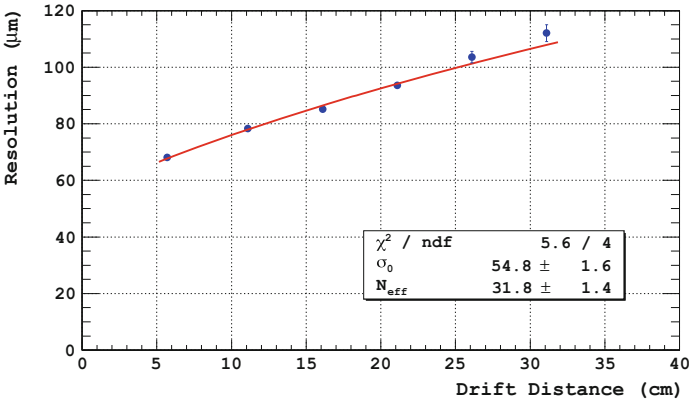

Fig. 22.13 Preliminary result of the spatial resolution of Micromegas readout as a function of drift length. A resistive pad plane was used to spread the charge [59]

Recently specialised GEM foils have been developed, which show a very large optical transparency. Experimentally is has been shown that such devices allow a large change in electron transparency, from close to 90% to 0%, by changing the potential across the GEM by some 50 V. This is expected to translate into a very similar change in ion-transparency, but the final experimental proof for this is still missing.

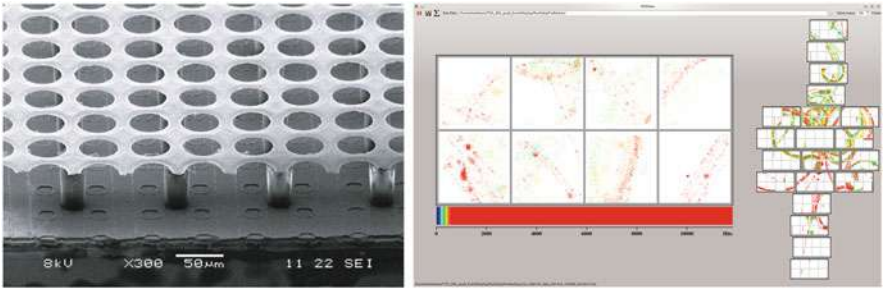

Fig. 22.14 Left: Microscopic picture of an Ingrid: a micromegas detector implemented on top of the read out chip by post-processing; Right: Event display of test beam electrons in a Pixel-TPC setup with Ingrids and Micromegas readout [62] (Credit Michael Lupberger, Bonn)

A recent development tries to combine the advantages of a micro-pattern gas detector with the extreme segmentation possible from silicon detectors. A Si pixel detector is placed at the position of the readout pad plane, and is used to collect the charge behind the gas amplification system. Each pixel of the readout detector has a charge sensitive amplifier integrated, and measures the time of arrival of the signal. Such a chip was originally developed for medical applications (MediPIX [63]), without timing capability, and has since been further developed to also include the possibility to record the time (Timepix [64]). This technology, which is still in its infancy, promises exciting further developments of the TPC concept. The close integration of readout pad and readout electronics into one pixel allows for much more compact readout systems, and also for much smaller readout pads. Pad sizes as small as $50 \times 50\,\mu$ m have been realised already. This allows a detailed reconstruction of the microscopic track a particle leaves in the TPC, down to the level of individual ionisation clusters. First studies indicate that a significantly improved spatial resolution can be obtained through silicon pixel readout of the TPC. In Fig. 22.14 a picture of a track segment recorded in a small test setup equipped with a Micromegas and the Medipix chip is shown.

The size of charge clouds in a typical TPC is of the order of a few hundred µm to mm, depending on the choice of gas, on environmental parameters like pressure and magnetic field, and on the drift distance. The feature size of the proposed silicon based readout is significantly smaller than this, which may allow the operation of the TPC in a different mode, the so called digital TPC mode. In this case no analogue information about the size of the charge collected at the anode is recorded, but instead only the number and the distribution of pixels which have fired are saved. The distribution of the hits is used to reconstruct the position of the original particle, much as it is done in the case of a conventional TPC. It can be shown that as long as the pixel size is small compared to the size of the electron cloud the number of pixels is a good measure for both the position of the cluster and the total charge in the cluster. One advantage of recording only the number of hits is that the sensitivity to delta rays is reduced. Delta rays are energetic electrons which are kicked out of a

gas molecule by the interaction with the incoming particle, and which then rapidly loose energy in the gas. Delta rays produce large charge clusters along the track, which are not correlated any more with the original particle. They also produce charge some distance away from the original track, and thus limit the intrinsic spatial resolution. Altogether delta rays are responsible for the tails in the charge distribution along a particle track, and for a deterioration of the possible spatial resolution. In digital readout mode these effects are less pronounced. The tails in the charge distribution are reduced, and the excellent spatial resolution through small pads allows the removal of at least some delta rays on a topological basis. Recent studies indicate that the spatial resolution of a Si based TPC readout might be better by about 30%, while the capability to measure the specific energy loss, dE/dx, might increase by 10–20% [65].

22.6.5 Electromagnetic Calorimeters

The concept of particle flow discussed above requires an excellent granularity in the calorimeters to separate charged from neutral particles in the calorimeter. Some hypothetical New Physics scenarios are associated with event topologies where high energetic photons do not originate at the interaction region, so that the device should in addition be able to also reconstruct the direction of a photon shower with reasonable accuracy.

Electromagnetic calorimeters (ECAL) are designed as compact and fine-grained sandwich calorimeters optimised for the reconstruction of photons and electrons and for separating them from depositions of hadrons. Sandwich calorimeters are the devices of choice, since they give information on the development of the cluster both along and transverse to the direction of the shower development. This capability is very difficult to realise with other technologies, and is essential to obtain an excellent spatial reconstruction of the shower. To keep the Moliére radius small, tungsten or lead are used as absorber. Sensor planes are made of silicon pad diodes, Monolithic Active Pixel sensors (MAPS) or of scintillator strips or tiles.

A major problem of fine-grained calorimeters is one of readout and data volume. For a typical electromagnetic calorimeter considered for the ILC, where cell sizes of $5 \times 5\,\mathrm{mm}^2$ are investigated, the number of channels quickly passes the million. With the progress in highly integrated electronics, more and more of the readout electronic is going to be integrated very close to the front-end. The design of the electromagnetic calorimeter by the CALICE collaboration [66] or by a North-American consortium [67, 68] has the silicon readout pads integrated into a readout board which sits in between the absorber plates. A special chip reads out a number of pads. A 12-bit ADC is included on the chip, and data are then sent on thin Kapton tape cables to the end of the module. There data from the different chips are concentrated, and sent on to the central data acquisition system. Such highly integrated detector designs have been successfully tested in large scale prototypes

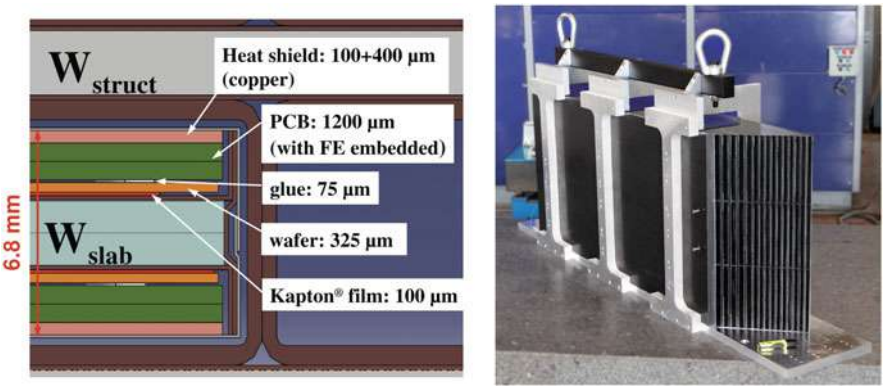

Fig. 22.15 Schematic figure of an integrated silicon-tungsten layer for an ILC ECAL (left) and tungsten absorber prototype (right) [37]

in test beams at CERN and Fermilab, although with a earlier version of the readout electronics, with a lesser degree of concentration (Fig. 22.15).

It is only with the progress in integration and in the resulting price reduction per channel that large scale Si-based calorimeter systems will become a possibility. Nevertheless the price for a large electromagnetic calorimeter of this type is still rather high, and will be one of the most expensive items in a detector for a linear collider. A cheaper alternative investigated is a more conventional sampling calorimeter readout by Scintillator strips. Two layers of strips at orthogonal orientation followed by a somewhat larger tile can be used to result in an effective granularity as small as $1 \times 1\,\mathrm{cm}^2$, nearly as good as in the case of the Si-W calorimeter. Light from the strips and tiles is detected from novel silicon based photo-detectors (for a more detailed description, see the section on hadronic calorimeters). The reconstruction of the spatial extent of a shower in such a system is more complicated, since ambiguities arise from the combination of the different layers. In addition the longitudinal information of the shower development is less detailed, but still superb compared to any existing device. This technology as well has been successfully tested in test beam experiments, and has shown its large potential.

Whether or not this technology or the more expensive Si-W technology is chosen for a particular detector depends on the anticipated physics case, and also the center-of-mass energy, at which the experiment will be performed. Simulation studies have shown that at moderate energies, below 250 GeV, both technologies perform nearly equally well, only at larger energies does the more granular solution gain an advantage. To some extent this advantage can be compensated by a larger detector in the case of the scintillator, though the price advantage then quickly disappears.

An extreme ansatz is a study trying to use vertex detector technology as readout planes in a calorimeter. The MAPS technology has been used to equip a tungsten absorber stack with sensors. This results in a extremely fine granular readout, where again only digital information is used—that is, only the number of pixels hit within

a certain volume is used, not any analogue information. This in turn means a much simpler readout electronics per channel, and a potentially more robust system against noise and other electrical problems. The amount of detail which can be reconstructed with such a system is staggering, and would open a whole new realm of shower reconstruction. However the cost at the moment is prohibitive, and many technical problems would need to be solved should such a system be used on a large scale [69].

22.6.6 Hadronic Calorimeters

In a particle flow based detector the distinction between an electromagnetic and a hadronic calorimeter conceptually disappears. Finely grained systems are needed to reconstruct the topology of the shower, both for electromagnetically and for hadronically interacting particles. Nevertheless, the optimization of the hadronic section of the calorimeter results in a coarser segmentation.

The traditional approach is based on a sampling calorimeter, typically with iron as absorber, maybe with lead, and with scintillators as active medium. New semiconductor photo detectors allow the individual readout of comparatively small scintillator tiles. These photo detectors are pixellated Si diodes, with of order 1000 diodes on an area of 1 mm^2. Each diode is operated in the limited Geiger mode, and the number of photons detected is read out by counting the number of pixels which have fired. This is another example of the previously discussed digital readout schemes. These so-called silicon photo multipliers (SiPM), also called Multi Pixel Photon Counters (MPPC), are small enough that they can be integrated into a calorimeter tile. To operate they only need to be provided with a potential of below 100 V, and the power lines are used to read out the signal from the counter. This makes for a rather simple system, which allows the instrumentation of a large number of tiles, and thus the construction of a highly granular scintillator based calorimeter. Complications which in the past severely limited the number of available channels—e.g., the routing of a large number of clear fibers from the tile to the photon counter, the operation of a larger number of bulky photo-multipliers of rather high voltage, etc all do not apply any more.

Light created through scintillation in the tile is collected by a Silicon photomultiplier, attached to each tile Earlier systems needed a wave-length shifting fibre, to adopt to the spectral sensitivity of the sensor (c.f. Fig. 22.16). A calibration of the energy response of the tile and SiPM system has two components: For small signals, the output signal shows contributions from one, two, three and more photons by clearly separate peaks in the amplitude spectrum. These can be used to establish the response of the system to single photons. At high signals, because of the limited number of pixels on the sensor, saturation leads to a non-linear response of the system. This needs to be measured and calibrated on the test bench, using a well calibrated photon source.

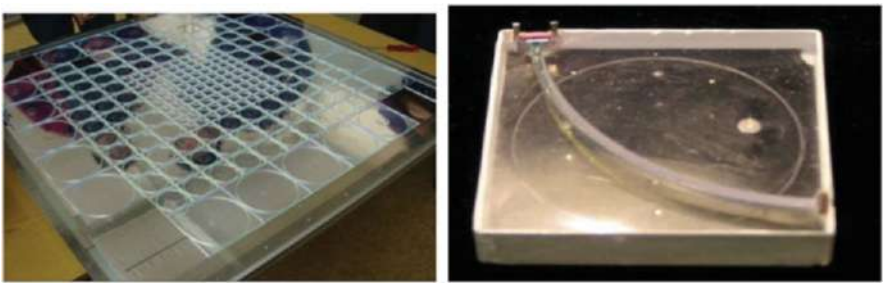

Fig. 22.16 Picture of a prototype readout plane for a highly segmented tile calorimeter (left) and one scintillator tile with wavelength shifting fibre and SiPM readout (right) [70]

The CALICE collaboration has designed a calorimeter based on this technology to be used in a detector at the linear collider. It is based on steel as absorber material, and uses $3 \times 3\,\mathrm{cm}^2$ scintillator tiles as sensitive elements. Each tile is readout by a silicon photomultiplier. A prototype readout plane is shown in Fig. 22.16. Groups of tiles are connected to a printed circuit board, which provides the voltage to the SiPM's, and routes the signals back to a central readout chip. This chip, which has been derived from the one developed for the Si-W calorimeter readout described in the previous section, digitises the signals, multiplexes them and sends them out to the data acquisition. Again, nearly all of the front-end electronics is integrated into this printed circuit board, and as such becomes part of the readout plane. This makes for a very compact design of the final calorimeter, with minimum dead space, and only a small number of external connections. This calorimeter has successfully passed a series of stringent beam tests in recent years, giving confidence that this technology is mature and can be used for a large scale detector application.

Recently the technology for SiPM advanced and pushed the sensitivity into the ultra-violet range, making a direct coupling between scintillator and silicon sensor possible (c.f. Fig. 22.17). The SiPM-on-tile technology has been proposed for the upgrade of the CMS endcap calorimeter. This system will use many of the developments done for an ILC detector, and be the first large-scale real-life application of this technology in an experiment. Through significantly smaller in size than the anticipated linear collider experiment, it will be a major asset for the LC community. Figure 22.17 shows a prototype readout HCAL plane using the SiPM-on-tile technology.

A potentially very interesting development in this area is again a digital version of such a calorimeter [72]. If the tile size can be made small enough - for hadronic showers this means a few $1 \times 1\,\mathrm{mm}^2$—a digital readout becomes possible. Counting the number of tiles belonging to a shower gives a good estimate of the showers energy. However scintillator tiles are difficult to built and read out for sizes this small—a major problems is the coupling between the light and the photo detector—so that a gaseous option is considered for this digital approach. Resistive plate chambers offer a cheap and well tested possibility to instrument large areas. They

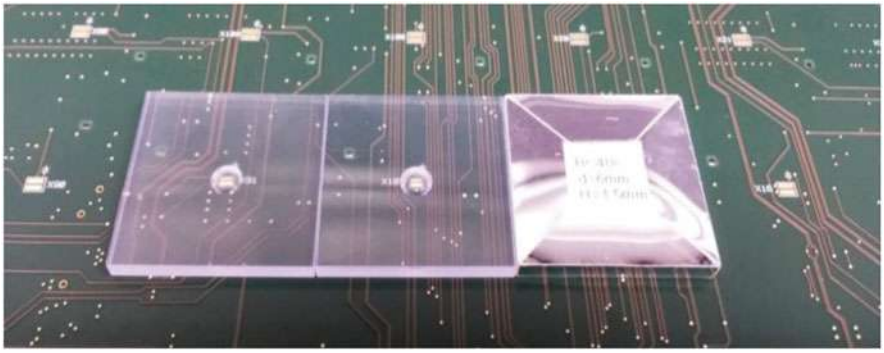

Fig. 22.17 Picture of HCAL scintillator tile with direct SiPM-on-tile readout [71]

are readout by segmented anode planes, which can be easily constructed with small pads of order of $1 \times 1\,\mathrm{cm}^2$. The principle of such a digital calorimeter has been established, and seems to meet specifications [72]. A major challenge however is to produce readout electronics for the very large number of channels which is about an order of magnitude cheaper per channel than the one for the analogue tile technology.

An interesting compromise between digital and analogue readout calorimeters is the semi-digital approach. Here, moderately dimensioned cell sizes of $1 \times 1\,\mathrm{cm}^2$ are combined with a rather simple 2-bit electronics with three signal thresholds. This would allow for having a high enough granularity to study the fine details of the hadronic shower evolutions and at the same time use the semi-digital charge signal for the analysis. A prototype semi-digital calorimeter for the ILD concept has been built and tested in beams and shows promising results [73].

A gaseous readout system has another feature which might be of advantage for a particle flow calorimeter. In the development of hadronic showers many neutrons are produced. Because of their long mean free path the loose energy and get absorbed far away from the core of the shower. This makes it very hard to attach these hits to the correct shower, thus creating a deficit in the energy for the shower, and creating fake hits away from the shower which might be confused with other nearby showers. Because of the very low cross section for neutrons in typical counter gases hardly any hits due to neutrons are recorded in a RPC based system. In a scintillator system, because of the high hydrogen and carbon content of the scintillator, the opposite is the case, and significant numbers if hits from neutrons are observed. On the other hand, neutrons travel slowly, and hits from neutron are later in time than other particle. Timing information at the 10 ns level might be good enough to reject a large number of the neutron hits in a shower. Its impact on the shower reconstruction is a subject of intense study at the moment, for both technologies, and no final verdict can be given which technology in the end has more advantages.

Large prototypes of ECAL and HCAL calorimeter systems have been built and tested in testbeam experiments. Figure 22.18 shows an event display for a combined

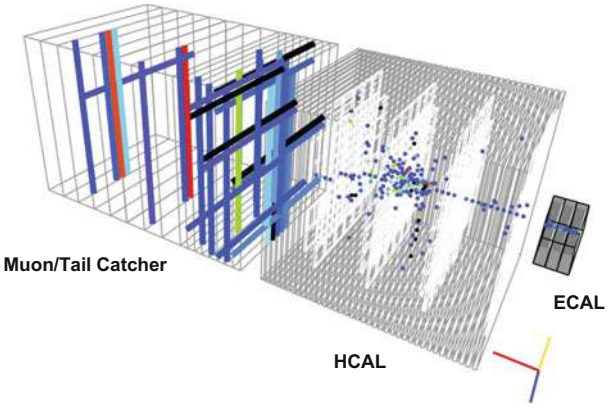

Fig. 22.18 Event in a combined testbeam where a 20 GeV pion (from the right) passes an ECAL prototype (small volume on the right), an analogue HCAL prototype with scintillator-tile readout (centre), and a muon system/tail catcher prototype with scintillator-strip readout (left) [71]

setup (from right to left) of a silicon-tungsten ECAL, an analogue scintillator-steel HCAL, and a muon/tailcatcher system with scintillator-strip readout (c.f. Sect. 22.6.7). A 20 GeV pion enters from the right, the details of the hadronic shower are clearly visible.

22.6.7 Muon Detectors

The flux return from the large field solenoids usually is realised as a thick iron return yoke. Often the iron is slit and detectors are integrated into the slots to serve as muon detectors. Many types of low-cost large-area charged particle detectors are possible and under investigation, e.g. resistive plate chambers, GEM chambers, or Scintillator based strip detectors. In a detector equipped with highly segmented calorimeters however a lot of the measurements traditionally done by such a muon system can be done in the calorimeters themselves. The identification of muons is greatly helped by the hadronic calorimeter, and its longitudinal sampling. Due to the high fields anticipated, muons below 3–4 GeV in fact never even reach the muon chambers, and need to be identified by the calorimeters together with the tracking system. The parameters of the muon are measured by the detector inside the coil, combining information from the tracker and the calorimeter. For these detector concepts the muon system in fact only plays a minor role, and can be used to backup and verify the performance of the calorimeter system.

An interesting approach is proposed by one of the ILC detector concepts [34]. Here the magnetic flux is returned not by an iron yoke, but by a second system of large coils. A smaller coil creates the high central field, of about 3 T, while a second larger coil creates a 1.5 T field in the opposite direction and serves as the flux return.

A system of planar coils in the endcap control the transition from the small to the large bore coil. In this concept muon chambers are mounted in between the two large solenoids. A similar approach is followed up in the studies for the very large detectors of potential very large hadron colliders.

22.6.8 Triggering at the ILC

The comparative cleanliness of events at the ILC allow for a radical change in philosophy compared to a detector at the LHC: the elimination of a traditional hardware based trigger. Triggering is a major concern at the LHC, and highly sophisticated and complex systems have been developed and built to reduce the very high event rate at the LHC to a manageable level [74, 75]. At the ILC with its clean events, without an underlying event, it is possible to operate the detector continuously and read out every bunch crossing. At a local level filtering is applied to the data to remove noise hits, and to eliminate as much as possible "bad hits", but overall no further data reduction is done. Events are written to the output stream unfiltered, and are only classified by software at a later stage. This allows the detector to be totally unbiased to any type of new physics, and to record events with the best possible efficiency. As a draw back the expected data rates are rather large. Great care has to be taken that the detector systems are robust and not dominated by noise, so that the data volume remains manageable, and the readout can keep up with the incoming data rate.

A slightly different approach has been suggested by the LHC experiments ALICE and LHCb, where upgrade plans foreseen to read out every event and to perform event selection and reconstruction in on-line processor farms.

22.7 Summary

Even though with the four LHC experiments, major experimental facilities recently built and commissioned, work on the next generation of experiments is proceeding. In particular the proposed linear collider poses very different and complementary challenges for a detector, with a strong emphasis on precision and details of the reconstruction. Significant work is happening worldwide on the preparation of technologies for this project. First results from test beam experiments show that many of performance goals are reachable or have already been reached. The move to ever increasing number of readout channels, with smaller and smaller feature sizes, has triggered a systematic investigation of "digital" detectors, where for a huge number of pixels only very little information per pixel is processed and stored. Whether or not such systems are really feasible in a large scale experiment is not proven yet. Tests over the next few years will answer many of these questions.

References

1. J. Gillies et. al (eds.), "Accelerating Science and Innovation", CERN-Brochure-2013-004-Eng. CERN 2013.
2. S. Ritz et. al, "Building for Discovery", Accessible at: http://www.usparticlephysics.org., 2014.
3. M. Nozaki et. al, "AsiaHEP/ACFA Statement on the ILC", Accessible at: http://www.acfa-forum.net/AsiaHEP/other_documents, 2014.
4. M. Nozaki et. al, "AsiaHEP/ACFA Statement on the ILC and CEPC/SPPC", Accessible at: http://www.acfa-forum.net/AsiaHEP/other_documents, 2016.
5. International Committee on Future Accelerators, "ICFA Statement on the ILC Operating at 250 GeV as a Higgs Factory", Accessible at: http://icfa.fnal.gov/statements, 2017.
6. T. Behnke *et al.*, "The International Linear Collider Technical Design Report - Volume 1: Executive Summary," arXiv:1306.6327 [physics.acc-ph].
7. M. Aicheler *et al.*, "A Multi-TeV Linear Collider Based on CLIC Technology : CLIC Conceptual Design Report," https://doi.org/10.5170/CERN-2012-007.
8. M. J. Boland *et al.* [CLIC and CLICdp Collaborations], "Updated baseline for a staged Compact Linear Collider," https://doi.org/10.5170/CERN-2016-004 arXiv:1608.07537 [physics.acc-ph].
9. M. Benedikt *et al.*, "Future Circular Collider : Vol. 2 The Lepton Collider (FCC-ee)," CERN-ACC-2018-0057.
10. M. Benedikt *et al.*, "Future Circular Collider : Vol. 3 The Hadron Collider (FCC-hh)," CERN-ACC-2018-0058.
11. [CEPC Study Group], "CEPC Conceptual Design Report: Volume 1 - Accelerator," arXiv:1809.00285 [physics.acc-ph].
12. CEPC-SPPC Study Group, "CEPC-SPPC Preliminary Conceptual Design Report." IHEP-CEPC-DR-2015-01, IHEP-TH-2015-01, IHEP-EP-2015-01.
13. M. Mangano, "Physics at the FCC-hh, a 100 TeV pp collider," CERN Yellow Report CERN 2017-003-M https://doi.org/10.23731/CYRM-2017-003 [arXiv:1710.06353 [hep-ph]].
14. CMS Collaboration, "Technical Proposal for a MIP Timing Detector in the CMS Experiment Phase 2 Upgrade," CERN-LHCC-2017-027.
15. I. Dawson [ATLAS and CMS Collaborations], "The SLHC prospects at ATLAS and CMS", J. Phys. Conf. Ser. **110** (2008) 092008.
16. J. Brau, Y. Okada, N. Walker (editors) [ILC Collaboration], "ILC Reference Design Report Volume 1 - Executive Summary", arXiv:0712.1950.
17. EUPRAXIA - European Plasma Research Accelerator with Excellence in Applications, see *http://www.eupraxia-project.eu*.
18. TESLA Technology Collaboration, see *http://tesla.desy.de*.
19. M. Altarelli *et al.*, "XFEL: The European X-Ray Free-Electron Laser. Technical design report," https://doi.org/10.3204/DESY_06-097
20. A. Grassellino *et al.*, "Nitrogen and argon doping of niobium for superconducting radio frequency cavities: a pathway to highly efficient accelerating structures," Supercond. Sci. Technol. **26** (2013) 102001 https://doi.org/10.1088/0953-2048/26/10/102001 [arXiv:1306.0288 [physics.acc-ph]].
21. C. Adolphsen *et al.*, "The International Linear Collider Technical Design Report - Volume 3.I: Accelerator R&D in the Technical Design Phase," arXiv:1306.6353 [physics.acc-ph].
22. C. Adolphsen *et al.*, "The International Linear Collider Technical Design Report - Volume 3.II: Accelerator Baseline Design," arXiv:1306.6328 [physics.acc-ph].
23. H. Baer *et al.*, "The International Linear Collider Technical Design Report - Volume 2: Physics," arXiv:1306.6352 [hep-ph].
24. L. Evans *et al.* [Linear Collider Collaboration], "The International Linear Collider Machine Staging Report 2017," arXiv:1711.00568 [physics.acc-ph].
25. K. Fujii *et al.*, "Physics Case for the International Linear Collider," arXiv:1506.05992 [hep-ex].

26. K. Fujii *et al.*, "Physics Case for the 250 GeV Stage of the International Linear Collider," arXiv:1710.07621 [hep-ex].
27. K. Seidel, F. Simon, M. Tesar and S. Poss, "Top quark mass measurements at and above threshold at CLIC," Eur. Phys. J. C **73** (2013) no.8, 2530 https://doi.org/10.1140/epjc/s10052-013-2530-7 [arXiv:1303.3758 [hep-ex]].
28. J. C. Brient and H. Videau, "The Calorimetry at the future e+ e- linear collider," eConf C **010630** (2001) E3047 [hep-ex/0202004].
29. V. L. Morgunov, "Energy flow method for multi - jet effective mass reconstruction in the highly granulated TESLA calorimeter," eConf C **010630** (2001) E3041.
30. J. Brau *et al.*, "International Linear Collider reference design report. 1: Executive summary. 2: Physics at the ILC. 3: Accelerator. 4: Detectors," https://doi.org/10.2172/929487
31. M. A. Thomson, "Particle Flow Calorimetry and the PandoraPFA Algorithm," Nucl. Instrum. Meth. A **611** (2009) 25 [arXiv:0907.3577 [physics.ins-det]].
32. The ILD concept group, see *http://www.ilcild.org*.
33. The SiD concept group, see *http://silicondetector.org*.
34. The 4th detector concept at the ILC, see *http://www.4thconcept.org*.
35. CLIC Detector and Physics Study, see *http://clicdp.web.cern.ch*.
36. CEPC Physics and Detector Working Group, see *http://cepc.ihep.ac.cn*.
37. T. Behnke *et al.*, "The International Linear Collider Technical Design Report - Volume 4: Detectors," arXiv:1306.6329 [physics.ins-det].
38. Y. Banda *et al.*, "Design and performance of improved column parallel CCD, CPC2," Nucl. Instrum. Meth. A **621** (2010) 192. https://doi.org/10.1016/j.nima.2010.05.055
39. Y. Sugimoto *et al.*, "CCD-based vertex detector for GLC," Nucl. Instrum. Meth. A **549** (2005) 87. https://doi.org/10.1016/j.nima.2005.04.032
40. G. Deptuch *et al.*, "Monolithic Active Pixel Sensors adapted to future vertex detector requirements," Nucl. Instrum. Meth. A **535** (2004) 366. https://doi.org/10.1016/j.nima.2004.07.152
41. C. Hu-Guo *et al.*, "First reticule size MAPS with digital output and integrated zero suppression for the EUDET-JRA1 beam telescope," Nucl. Instrum. Meth. A **623** (2010) 480. https://doi.org/10.1016/j.nima.2010.03.043
42. EUDET - Detector R&D Towards the International Linear Collider, see *https://www.eudet.org*.
43. AIDA2020 - Advanced European Infrastructure for Detectors at Accelerators, see *https://aida2020.web.cern.ch*.
44. G. Aglieri Rinella [ALICE Collaboration], "The ALPIDE pixel sensor chip for the upgrade of the ALICE Inner Tracking System," Nucl. Instrum. Meth. A **845** (2017) 583. https://doi.org/10.1016/j.nima.2016.05.016
45. M. Winter, "CMOS Pixel Sensors for ILC related Vertexing and Tracking Devices", presented at American Linear Collider Workshop 2017, SLAC, 2017, https://portal.slac.stanford.edu/sites/conf_public/AWLC17/Pages/default.aspx
46. DEPFET Collaboration for Vertex Detectors at ILC and Belle-II, see *https://www.depfet.org*.
47. A. Nomerotski *et al.* [PLUME Collaboration], "PLUME collaboration: Ultra-light ladders for linear collider vertex detector," Nucl. Instrum. Meth. A **650** (2011) 208. https://doi.org/10.1016/j.nima.2010.12.083
48. F. Luetticke [DEPFET Collaboration], "The ultralight DEPFET pixel detector of the Belle II experiment," Nucl. Instrum. Meth. A **845** (2017) 118. https://doi.org/10.1016/j.nima.2016.06.114
49. I. M. Gregor, "Summary of One Year Operation of the EUDET CMOS Pixel Telescope," arXiv:0901.0616 [physics.ins-det].
50. C. J. S. Damerell and D. J. Jackson, "Design of a vertex detector and topological vertex reconstruction at the future linear collider," *Prepared for 3rd Workshop on Physics and Experiments with e+ e- Linear Colliders (LCWS 95)*, Iwate, Japan, 1995.
51. D. Bailey *et al.* [LCFI Collaboration], Nucl. Instrum. Meth. A **610** (2009) 573 https://doi.org/10.1016/j.nima.2009.08.059 [arXiv:0908.3019 [physics.ins-det]].
52. LCTPC Collabortation, see *https://www.lctpc.org*.

53. K. Aamodt *et al.* [ALICE Collaboration], "The ALICE experiment at the CERN LHC," JINST **3** (2008) S08002. https://doi.org/10.1088/1748-0221/3/08/S08002
54. P. Schade *et al.* [LCTPC Collaboration], Nucl. Instrum. Meth. A **628** (2011) 128. https://doi.org/10.1016/j.nima.2010.06.300
55. R. Bouclier *et al.*, "The Gas electron multiplier (GEM)," IEEE Trans. Nucl. Sci. **44** (1997) 646 [ICFA Instrum. Bull. **1996** (1996) F53]. https://doi.org/10.1109/23.603726
56. M. Killenberg, S. Lotze, A. Munnich, S. Roth, M. Weber and J. Mnich, "Development of a GEM-based high resolution TPC for the International Linear Collider," Nucl. Instrum. Meth. A **573** (2007) 183. https://doi.org/10.1142/9789812773678_0183,10.1016/j.nima.2006.10.396
57. Y. Giomataris, P. Rebourgeard, J. P. Robert and G. Charpak, Nucl. Instrum. Meth. A **376** (1996) 29. https://doi.org/10.1016/0168-9002(96)00175-1
58. D. S. Bhattacharya *et al.*, "A Micromegas-based TPC for the International Linear Collider," DAE Symp. Nucl. Phys. **61** (2016) 962.
59. D. Attie [LC-TPC Collaboration], "Beam tests of Micromegas LC-TPC large prototype," JINST **6** (2011) C01007. https://doi.org/10.1088/1748-0221/6/01/C01007
60. M. Killenberg *et al.*, "Modelling and measurement of charge transfer in multiple GEM structures," Nucl. Instrum. Meth. A **498**, 369 (2003) [arXiv:physics/0212005].
61. L. Hallermann, "Analysis of GEM properties and development of a GEM support structure for the ILD Time Projection Chamber," https://doi.org/10.3204/DESY-THESIS-2010-015
62. M. Lupberger, "The Pixel-TPC: A feasibility study," Thesis, Bonn, 2015.
63. X. Llopart et el., "MediPix2, a 64k Pixel read-out with 55 μm square elements working in single photon counting mode", IEEE Trans. Nucl. Sci. NS-49 (2002) 2279
64. X. Llopart, R. Ballabriga, M. Campbell, L. Tlustos and W. Wong, Nucl. Instrum. Meth. A **581** (2007) 485 Erratum: [Nucl. Instrum. Meth. A **585** (2008) 106]. https://doi.org/10.1016/j.nima.2007.08.079, https://doi.org/10.1016/j.nima.2007.11.003
65. M. Hauschild, "Particle ID with dE/dx at the TESLA-TPC," *Prepared for 5th International Linear Collider Workshop (LCWS 2000), Fermilab, Batavia, Illinois, 24–28 Oct 2000*
66. CALICE Collaboration, see *https://twiki.cern.ch/twiki/bin/view/CALICE/WebHome*
67. J. E. Brau *et al.*, "An electromagnetic calorimeter for the silicon detector concept," Pramana **69** (2007) 1025. https://doi.org/10.1007/s12043-007-0222-2
68. J. Brau *et al.*, "A silicon-tungsten electromagnetic calorimeter with integrated electronics for the International Linear Collider," J. Phys. Conf. Ser. **404** (2012) 012067. https://doi.org/10.1088/1742-6596/404/1/012067
69. G. Nooren *et al.*, "The FoCal prototype - an extremely fine-grained electromagnetic calorimeter using CMOS pixel sensors," JINST **13** (2018) no.01, P01014 https://doi.org/10.1088/1748-0221/13/01/P01014 [arXiv:1708.05164 [physics.ins-det]].
70. F. Sefkow, A. White, K. Kawagoe, R. Pöschl and J. Repond, "Experimental Tests of Particle Flow Calorimetry," Rev. Mod. Phys. **88** (2016) 015003 https://doi.org/10.1103/RevModPhys.88.015003 [arXiv:1507.05893 [physics.ins-det]].
71. F. Sefkow, "The new scintillator-SiPM based analogue HCAL prototype", presented at International Workshop on Future Linear Colliders LCWS2017, Strasbourg, 2017. https://agenda.linearcollider.org/event/7645/overview.
72. C. Adloff, J. Blaha, J. J. Blaising, M. Chefdeville, A. Espargiliere and Y. Karyotakis, "Monte carlo study of the physics performance of a digital hadronic calorimeter," JINST **4** (2009) P11009 [arXiv:0910.2636 [physics.ins-det]].
73. V. Buridon *et al.* [CALICE Collaboration], "First results of the CALICE SDHCAL technological prototype," JINST **11** (2016) no.04, P04001 https://doi.org/10.1088/1748-0221/11/04/P04001 [arXiv:1602.02276 [physics.ins-det]].
74. S. George [ATLAS Collaboration], "Design and expected performance of the ATLAS trigger and event selection," Eur. Phys. J. direct **4** (2002) no.S1, 06. https://doi.org/10.1007/s1010502cs106
75. S. Dasu [CMS Collaboration], "CMS trigger and event selection," Eur. Phys. J. direct **4** (2002) no.S1, 09. https://doi.org/10.1007/s1010502cs109