VERTEX DETECTORS: THE STATE OF THE ART AND FUTURE PROSPECTS

C. J. S. Damerell

Rutherford Appleton Laboratory

Chilton, Didcot, OX11 0QX, England

ABSTRACT

We review the current status of vertex detectors (tracking microscopes for the recognition of charm and bottom particle decays). The reasons why silicon has become the dominant detector medium are explained. Energy loss mechanisms are reviewed, as well as the physics and technology of semiconductor devices, emphasizing the areas of most relevance for detectors. The main design options (microstrips and pixel devices, both CCD's and APS's) are discussed, as well as the issue of radiation damage, which probably implies the need to change to detector media beyond silicon for some vertexing applications. Finally, the evolution of key performance parameters over the past 15 years is reviewed, and an attempt is made to extrapolate to the likely performance of detectors working at the energy frontier ten years from now.

© 1995 by C. J. S. Damerell.

TABLE OF CONTENTS

1 **Introduction**

2 **Energy Loss of High-Energy Charged Particles in Silicon**

- 2.1 Simplified Treatment
- 2.2 Improved Treatment
- 2.3 Implications for Tracking Detectors
- 2.4 Summary

3 **Physics and Properties of Semiconductors**

- 3.1 Conduction in Pure and Doped Silicon
- 3.2 The *pn* Junction
- 3.3 Charge Carrier Transport in Silicon Detectors

4 **Microstrip Detectors**

- 4.1 Introduction
- 4.2 The Generic Microstrip Detector
- 4.3 Microstrip Detectors; Detailed Issues
  - 4.3.1 Design Optimization
  - 4.3.2 Spatial Precision in Microstrip Detectors
  - 4.3.3 Electronics for Microstrip Detectors
- 4.4 Physics Performance and Future Trends

5 **Pixel-Based Detectors**

- 5.1 Introduction
- 5.2 Charge-Coupled Devices (CCD's)
  - 5.2.1 Structure and Basic Operation
  - 5.2.2 CCD Charge Transfer and Readout; Detailed Issues
    - 5.2.2.1 Charge Transfer Process
    - 5.2.2.2 Charge Detection
    - 5.2.2.3 Vertex Detector Readout Options
  - 5.2.3 Physics Performance and Future Trends
- 5.3 Active Pixel Sensors (APS's)
  - 5.3.1 Design Options
    - 5.3.1.1 Monolithic Detectors
    - 5.3.1.2 Hybrid Detectors
  - 5.3.2 Performance and Future Trends

6 **Radiation Damage in Silicon Detectors**

- 6.1 Introduction
- 6.2 Ionizing Radiation
- 6.3 Displacement Damage
- 6.4 Detector-Specific Effects
  - 6.4.1 Microstrip Detectors and APS Devices
  - 6.4.2 CCD's
  - 6.4.3 Local Electronics
- 6.5 Future Prospects

7 **Beyond Silicon**

- 7.1 Gallium Arsenide Detectors
- 7.2 CVD Diamond
- 7.3 Local Electronics

8 **Conclusions**

Acknowledgments

References

1 Introduction

There is for me a considerable sense of nostalgia in giving these lectures, since I previously gave such a series at the Summer Institute of 1984, which was especially noteworthy since it was coupled with the Pief-Fest to mark the retirement of Panofsky as Director of SLAC. Younger readers will be surprised to learn that the 1984 Institute, on the theme of the sixth quark, included evidence for the discovery of top with a mass of 40 ± 10 GeV.

In my 1984 lecture series, I suggested that these candidate top events really needed additional experimental evidence in order to be proved or disproved, and that this would best be provided by a precision vertex detector able to resolve the associated B decays. At the time, this suggestion was not taken particularly seriously. A lecture series relating to experimental methods of heavy quark detection at the same Institute made no mention of vertex detectors. Detectors with the required precision were only beginning to be used in the *fixed target* regime, and many of these were based on technologies such as bubble chambers that were manifestly not applicable to the collider environment. My own lectures made mention of techniques which have subsequently fallen into disuse for this reason. However, my main aim in those lectures was to establish a case for silicon vertex detectors in the collider environment. Our community was at that time in the early stages of planning the LEP and SLC detectors, and I focused particularly on Z^0 decays as the firm ground on which to base the case for these silicon vertex detectors. One was heavily dependent on Monte Carlo simulations of events with heavy flavor decays, where the possibilities for flavor tagging and some measure of topological vertexing could be demonstrated. Physicists at the time could be forgiven for not being wholly convinced by these simulations. Silicon detectors in those days were limited in size to a few square centimeters, were typically serviced by a huge amount of local electronics (easily accommodated in a fixed target experiment, but completely excluded in a collider), and detector reliability was a major problem. Here again, access for servicing which was easy in the fixed target environment would become much more difficult at the heart of a hermetic collider detector. In 1984, these Monte Carlo studies left on one side a host of technical problems which required many years of hard work to solve. Due to the loosely coupled R&D projects of many collaborations, the progress made since then has been immense. We now have a large variety of silicon vertex detectors in use in fixed target as well as collider experiments around the world. New designs are constantly being fabricated and tried out in test beams. The associated local electronics has shrunk spectacularly, and at the same time, become much faster and more powerful.

My task is thus made easier than 11 years ago; silicon vertex detectors have become well-established within the standard toolkit of high-energy experiments. I no longer need to rely on Monte Carlo studies to prove their usefulness; we can just look at the data. On the other hand, the array of detector types available has become somewhat bewildering, and I shall aim to provide some systematic guidance for nonexperts. Furthermore, despite the fact that the proponents of silicon detectors have been able to expand their horizons, even planning in some cases to displace gaseous tracking detectors with tens of square meters of strip detectors, they have begun to run into serious challenges in some vertex applications. In various hadron beam experiments, most spectacularly the LHC at its design luminosity, silicon detectors as we now know how to build them will fail after an unacceptably short time, when placed close to the interaction region. This has stimulated a major effort with other materials of greater radiation resistance, as we shall see towards the end of these lectures.

We are seeing the beginning of a technology division between e^+e^- colliders and hadron colliders, in regard to vertex detection at the energy frontier. Both are well-suited to the use of silicon at large radii, for general purpose tracking. But it is likely that at the luminosities needed for "discovery physics" at the TeV energy scale, silicon detectors will continue to be useful for high resolution vertex studies in the e^+e^- collider environment but not at LHC.

There are clearly great advantages in remaining with the silicon technology as far as possible. A major reason for its rapid growth as a material for tracking detectors is that the *planar process* for manufacturing silicon integrated circuits has been developed to an extremely fine art. These developments are continuing at a pace which reflects the billions of dollars annually invested, for purposes which have nothing to do with scientific research, let alone particle physics.

Before plunging into our rather specialized topic in fine detail, it is useful to take a brief look at the overall scene of silicon devices, particularly regarding their utility as radiation detectors. For, unlike some detection materials which are not widely used outside of our field (e.g., liquid argon), silicon finds applications in a vast range of scientific sensors. We in particle physics are concerned with its use for tracking microscopes that allow us to probe the smallest and shortest lived particles in nature. Silicon devices also provide the means to see the largest and oldest structures in the universe. Between these extremes, these sensors find a vast number of diverse applications, some of great importance to mankind (e.g., in medical imaging). Technically, all these areas are closely linked, so progress in one field may be significant to many others. All these scientific applications are dwarfed by the use of silicon sensors in the mass consumer markets, notably in video cameras but with applications now extending into other areas. What makes this field particularly dynamic is the flow of ideas from people with very different aims and agendas. The next major advance for HEP detectors may come from an astronomer concerned about radiation damage to his space-based telescope, or from the designer of an output circuit able to function at HDTV readout rates. Similarly, those designing devices for HEP use may dream up an advance that happens to be much more significant for some other field.

Why is silicon the preferred material for high-precision tracking detectors, as well as for such a wide range of radiation detectors?

Firstly, a *condensed medium* is essential if point measurement precision below about 10 μm is required. Gaseous tracking detectors are limited by diffusive spreading of the liberated electron cloud to precision of typically some tens of microns. Such detectors are entirely adequate for a host of particle tracking applications, but not for precision vertex detectors. Having established the need for a condensed medium, one should in principle consider liquids. There was some work done on high precision liquid xenon tracking detectors in the '70s [1] but there were many problems, not least of which was maintaining purity in conditions where the high mobilities of many contaminants rendered them particularly potent. In contrast, silicon wafers refined to phenomenal purity levels can then be sawn, exposed to the atmosphere, and assembled in complex geometries, with no degradation of their bulk electron lifetime characteristics. For these reasons, silicon and other solids are generally to be preferred, as opposed to liquids, for high-precision tracking purposes. There are, however, many possible solid state detection media, so why pick silicon?

Silicon has a band gap of 1.1 eV, *low* enough to ensure prolific production of liberated charge from a minimum-ionizing particle, hereafter referred to as a MIP (about 80 electron-hole pairs per micron of track length), but *high* enough to avoid very large dark current generation at room temperature (kT at room temperature = 0.026 eV). Being a low Z element of excellent mechanical properties (high modulus of elasticity) makes silicon an ideal material for use in tracking detectors where multiple scattering is of concern. This is nearly always the case in vertex detectors where tracks need to be extrapolated to the interaction region, and the dynamics of the fragmentation process ensures that even at the highest CM energies, many of the particles produced are of relatively low energy.

Besides these detector-related reasons, one has the vast IC technology developed specifically for this material. Silicon is currently unique in the combination of assets it brings with it; the growth of huge crystals of phenomenal purity, the possibility of *n*- and *p*-type doping, the possibility of selective growth of highly insulating layers (SiO_2 and Si_2N_3), and the possibility of doing all these using microlithographic techniques, allowing feature sizes of around one micron (and falling with time). A very readable account of the remarkable human stories associated with these amazing developments is to be found in George Gilder's book on the subject [2]. Very small feature sizes are, of course, precisely what one requires in order to construct detectors of precision below ten microns. Overall, the art of producing integrated circuits is probably by far the most sophisticated, fastest developing area of technological growth in the history of mankind. Without these developments, silicon as a detector of nuclear radiation would have remained a minor player, subject to arcane production procedures, of limited use for the spectroscopy of low-energy gamma rays, and wholly inappropriate for particle tracking purposes.

Though the scientific applications are of great importance, they are dwarfed by the use of silicon detectors for mass market consumer products and commercial interests.

Accurate figures are not readily available, but it seems that approximately $100M per year is spent on R&D of CCD's for domestic video and still cameras. These are interline transfer devices of no direct use for most scientific imaging applications. About $10M is spent on CCD development for medical and other scientific imaging applications (mostly X-rays). Silicon devices specifically aimed at particle tracking (microstrip detectors, CCD's, and active pixel sensors, hereafter referred to as APS devices) probably attract only $1M (order of magnitude) in R&D per year.

Even the consumer market for silicon sensors is dwarfed by the really hot commercial areas. For example, it was recently reported that NEC demonstrated a 1 Gbit DRAM. Production devices are expected to follow in three year's time, after the expenditure of *a further* $1.5B of R&D funding. Much of this will go in the development of submicron manufacturing capability, which ultimately will benefit the particle physics instrumentation community. We can eventually look forward to *submicron* tracking precision with *subnanosecond* timing information. However, the pace of such developments will be determined by the major players outside our own field, and there will inevitably be a time lag of several years between a technology being available for mass produced IC's and it being affordable for our purposes.

While the silicon processing infrastructure and R&D for a specific device can be enormously expensive, once production begins the costs can be modest. The ingredients of integrated circuits (sand, air, aluminum) are ridiculously cheap, and this benefit can be seen dramatically in large production runs. For example, SONY produces approximately five million CCD's per year for the domestic video camera market, at a production cost of only around $10, including the microlens and color filter system. This is a truly amazing achievement, as you can convince yourself by just looking through a microscope at one of these devices.

In summary, the match between silicon (and its attendant technologies) to the aspirations of the experimentalist wishing to construct tracking detectors of the highest possible precision, is evident. Were it not for the problems of radiation damage (which are most serious in the context of hadron colliders), there is little doubt that our field would by now have standardized completely on this material for vertex detection. Some time ago, test devices even surpassed photographic nuclear emulsions in precision, and with all the advantages of electronic readout. The challenge of hadron machines has stimulated some brave souls to undertake the monumental task of achieving similar technical performance using more radiation-resistant materials than silicon. They have, of course, to solve the problems not only of the detectors but also of the local electronics. We shall take a brief look at what they are doing in Sec. 7 of this paper. Other than that section, we shall devote ourselves exclusively to a discussion of silicon detectors and electronics.

2 Energy Loss of High-Energy Charged Particles in Silicon

High-energy charged particles traversing crystalline silicon can lose energy in two ways. Firstly, by ionization of the atomic electrons. This simple picture becomes rather more complex in regard to the valence electrons, as we shall see. The second energy loss mechanism (the so-called non-ionizing energy loss or NIEL) consists of displacement of silicon atoms from the crystal lattice, mostly by the process of Coulomb nuclear scattering. Only if the energy transfer to the nucleus exceeds approximately 25 eV can the atom be displaced from its lattice site. Below that, the energy is dissipated by harmless lattice vibrations. This implies an effective threshold energy for displacement damage with incident electrons (for example) of around 250 keV. Displacement of silicon atoms to *interstitial* positions (creating a *vacancy* in the lattice where the atom had previously been located) is one of the main radiation damage mechanisms. For a high-energy particle, the fraction of energy loss going into the NIEL mechanism is relatively small, but the cumulative effects on the detector performance can be severe.

A detector placed in a neutron flux experiences no signal from primary ionization, but the interactions can cause a high level of NIEL in view of the large neutron-silicon scattering cross section. For both charged hadrons and neutrons, other mechanisms of energy loss and radiation damage exist, notably neutron capture followed by nuclear decay, and inelastic nuclear scattering. The effects of non-ionizing energy loss on silicon detectors are considered in Sec. 6. In this section, we focus on the ionization energy loss only.

2.1 Simplified Treatment

Let us first imagine all the atomic electrons to be free, as if the crystal consisted of the silicon nuclei neutralized electrically by a homogeneous electron plasma. As a charged particle traverses the material, it loses energy by collisions (Coulomb scattering) with the electrons. Close collisions, while rare, will result in large energy transfers, while the much more probable distant collisions give small energy transfers. The process can be thought of classically in terms of the impulse generated by the attractive or repulsive Coulomb interaction between the projectile and the electron. The net impulse will be a kick transverse to the direction of travel of the projectile (see Fig. 1). The greater probability of remote collisions arises simply

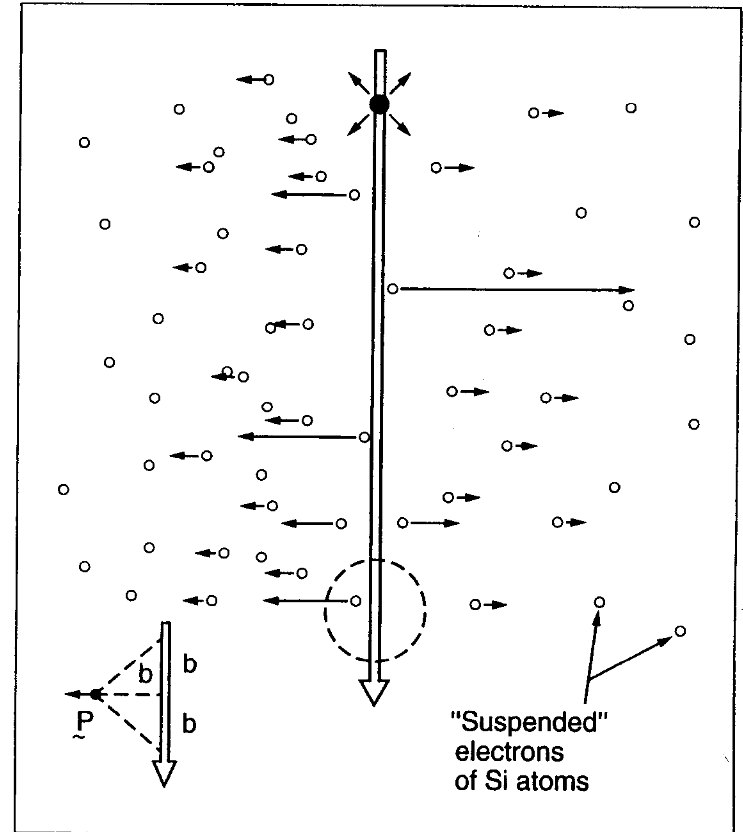


Fig. 1. Passage of charged particle through matter. Close collisions (electrons with small impact parameter b, shown by the inset) receive a powerful transverse impulse. Distant electrons receive a weak impulse.

from the greater volume of material available for collisions with a given impact parameter range, as the corresponding cylinder (of radius equal to the impact parameter) expands. In this simple case, the probability for a collision imparting energy E to an atomic electron is given by the Rutherford cross section

$$\frac{d\sigma_R}{dE} = \frac{2\pi q_e^4}{m_e c^2 \beta^2} \times \frac{1}{E^2}, \tag{2.1}$$

where q_e and m_e are the charge and mass of the electron.

Note the mass of the struck particle in the denominator. This explains why scattering off the silicon nuclei, which are much more massive, causes very little energy loss, though these collisions do make the major contribution to the deviation in angle of the incident particle trajectory, via the process of multiple nuclear Coulomb scattering. Also, for sufficiently large momentum transfers, these collisions contribute to the NIEL referred to above.

We are interested in evaluating the *mean* energy loss and also the *fluctuations*, for traversal of a given thickness detector. An apparently simple approach would be to perform the integration over all E to obtain the mean energy loss, and to run a Monte Carlo calculation with multiple traversals to determine the energy loss distribution (straggling formula). However, we see that the integral diverges like $1/E$. The stopping power of this free-electron plasma would indeed be infinite, due to the long-range Coulomb interaction. In practice, the electrons are *bound*, and this prevents very low energy transfers to the vast number of electrons which are distant from the particle trajectory. This divergence is conventionally avoided by introducing a semi-empirical cutoff (binding energy) $E_{\min}$ which depends on the atomic number Z of the material. This is necessarily an approximate approach, since (for example) it ignores the fact that the outer electrons are bound differently in gaseous media than they are in solids. We shall need a more refined treatment to handle the cutoff in collisions with small energy transfer.

However, the Rutherford formula (with one small correction) is extremely useful as regards the close collisions, which are most important in defining the fluctuations in energy loss in "thick" samples (greater than approximately 50 μm of silicon, for example). The required correction is the upper cutoff $E_{\max}$ in energy transfer imposed by the relativistic kinematics of the collision process. If the projectile mass is much greater than m_e, we have $E_{\max} = 2m_e c^2 \beta^2 \gamma^2$. Due to the $1/E^2$ term in the Rutherford formula, we find that there is for each sample thickness, an energy transfer range in which the integrated probability of such transfers through the sample falls from almost unity to nearly zero. The Poisson statistics on energy transfers in this range gives rise to fluctuations on the overall energy loss for each traversal. Thus, the overall energy loss distribution consists of an approximately Gaussian core plus a high tail, populated by traversals for which a few close collisions occurred, each generating several times the mean energy loss. While the energy transfer region in which the probability function falls almost to zero is dependent on the sample thickness, this merely introduces an overall scale factor, so the *form* of the overall energy loss distribution is constant (the famous Landau distribution) over a wide range of detector thicknesses.

The rare close collisions with energy transfer greater than approximately 10 keV generate δ-electrons of significant range, which may be important in tracking detectors due to their potential for degrading the precision. For these close collisions, all atomic electrons behave as if they are free and the Rutherford formula may be used with confidence.

For thin samples, the energy loss fluctuations are not adequately handled by the Rutherford formula with cutoffs $E_{\min}$ and $E_{\max}$ In this case, the bulk of the energy loss arises from low-energy transfer collisions for which the binding of the atomic electrons must be handled in detail. We shall now consider the improved treatment of this case, specifically for crystalline silicon, though the same principles apply in general.

2.2 Improved Treatment

We note that energy loss is a discrete quantum mechanical process. We shall see that for very thin samples, a particle has even a finite probability of traversing the detector with no energy deposition at all.

For the low-probability close collisions, as noted above, it is valid to consider all atomic electrons as free, and the Rutherford formula applies. Ejected electrons of energies greater than approximately 10 keV will release further atomic electrons along their path. See Refs. [3] and [4] for a detailed treatment. For our purposes, it

is sufficient to note that the ultimate products that concern us are electrons, promoted into the conduction band of the material and holes (vacancies in the valence band), and that the generation of each electron-hole pair requires a mean creation energy W of approximately 3.6 eV. The precise value depends weakly on the temperature, see Fig. 2, and reflects the temperature dependence of the silicon band gap. Since this is around 1.1 eV, we note that electron-hole generation is a somewhat inefficient process; approximately 2/3 of the energy transferred from the primary (hot) electrons gives rise to phonon generation, eventually appearing as heat in the detector.

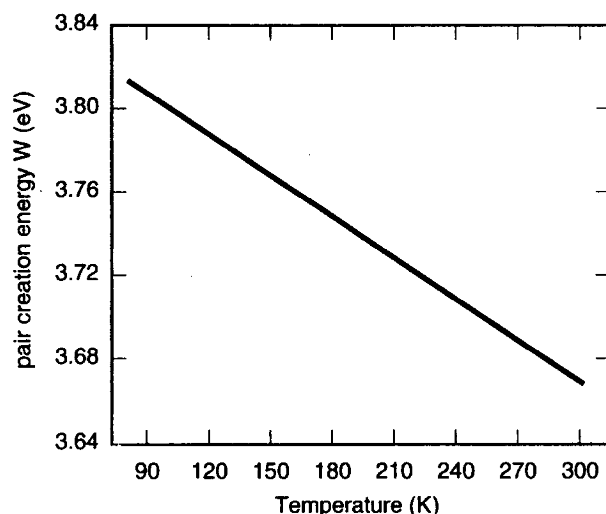


Fig. 2. Temperature dependence of the pair-creation energy W in silicon.

Beware, this has nothing to do with the non-ionizing energy loss (NIEL) referred to in the introduction to this section! Phonon generation (in contrast to NIEL) is a benign process which does not disrupt the crystal lattice and is usually ignored other than by enthusiasts for bolometric detectors. For our purposes, the δ-electrons ejected in close collisions can be considered to generate further electron-hole pairs at a mean rate of one per 3.6 eV of energy loss, *locally* on the track of the projectile, or *distributed* in the case that the δ-electron range is significant.

Qualitatively, the effect of the binding of the atomic electrons is to generate resonance-like enhancements in the energy loss cross section, above the values expected from the Rutherford formula. The K-shell electrons produce an enhancement in the 2 to 10 keV range, the L-shell in the 100 eV to 1 keV range, and the M-shell a resonance at around 20 eV. Below this resonance, the cross section rapidly falls to zero, in the region around 15 eV where the Rutherford formula would be cut off by the empirical ionization threshold energy.

The most satisfactory modern treatment proceeds from the energy-dependent photo-absorption cross section (a clean *point-like* process in the terminology of solid-state physics). This is, of course, closely linked to the energy loss process for charged particles, which fundamentally proceeds via the exchange of virtual photons. Combining photo-absorption and EELS (electron energy loss spectroscopy) data, Bichsel [5] has made a precise determination of the MIP energy loss cross section for silicon. The most subtle effects are connected with the valence (M-shell) electrons.

These valence electrons behave as a nearly homogeneous dense gas (plasma) embedded in a fixed positive-charge distribution. The real or virtual photons couple to this by generating longitudinal density oscillations, the quantum of which is called a *plasmon* and has a mean energy of 17 eV. The plasmons de-excite almost entirely by electron-hole pair creation. These somewhat energetic charge carriers are referred to as "hot carriers." Like the δ-electrons produced in the close collisions, they lose energy by thermal scattering, optical phonon scattering, and ionization. The topic of hot carriers is a major area of research, but for our purposes (as with the δ-electrons), we can ignore the details, since the end product that concerns us is again electron-hole pair creation at a rate of one per 3.6 eV of primary energy deposition. Figure 3 shows the photo-absorption cross section for silicon. The plasmon excitation is responsible for the extremely large cross section in the ultraviolet. It is by virtue of the low energy tail of this cross section in the visible that silicon has its optical sensing applications. The material becomes almost perfectly transparent once the photon energy falls below the 1.1 eV band gap energy.

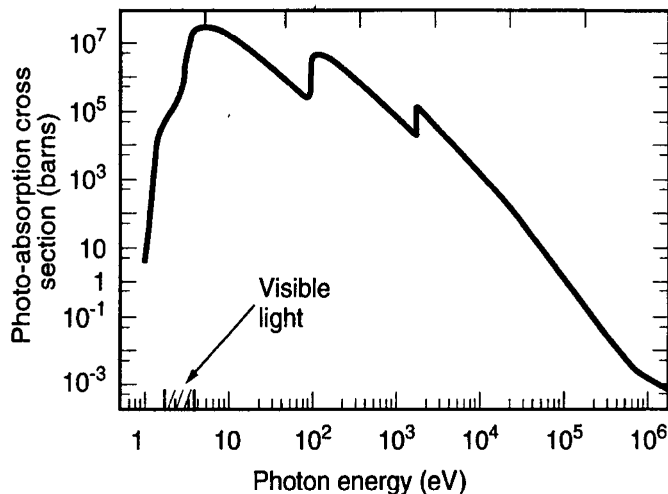


Fig. 3. Photo-absorption cross section of silicon versus photon energy.

The closely related energy loss cross section for a MIP is show in Fig. 4. Note that on average, it exceeds the Rutherford cross section by a factor of several in the energy range 10 eV to 5 keV. Above 10 keV, it is very close to the Rutherford value. By integrating the different components of this cross section, we can deduce the total mean collision rates associated with the different processes. These are as follows:

Electrons	Collision probability per micron
K (2)	8.8 x 10^{-3}
L (8)	0.63
M (4)	3.2

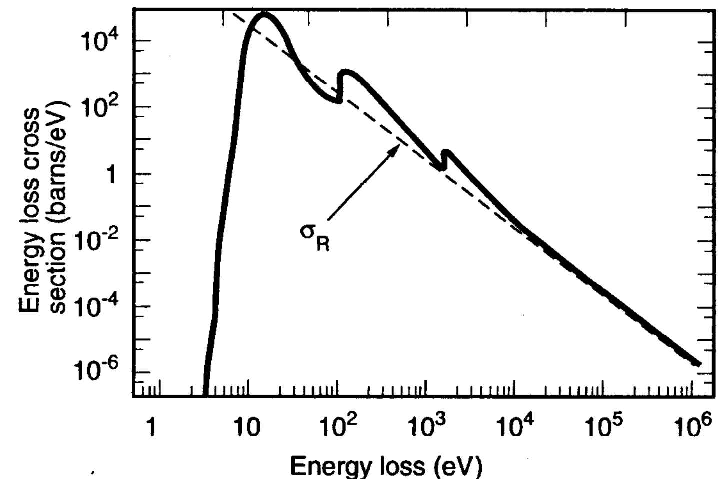


Fig. 4. Energy loss cross section for minimum-ionizing particles in silicon vs energy loss in primary collisions. The Rutherford cross section σ_R is also plotted.

Thus, despite the fact that on average a slice of silicon 1 μm in thickness will yield 80 electron-hole pairs, the Poisson statistics on the *primary* process (on average 3.8 collisions per micron) clearly implies a very broad distribution, with even a non-negligible probability of zero collisions, i.e., absolutely no signal. For thin samples, a correct statistical treatment of the primary process is essential if realistic energy loss (straggling) distributions are to be calculated. Their shapes are a strong function of the sample thickness, quite unlike the thickness-independent Landau distribution. The situation is depicted graphically in Fig. 5.

The area of each circle represents energy loss in a primary collision process. Those of smallest size correspond to plasmon excitation, while the larger ones represent the ionization of L-shell electrons. For these ten randomly selected tracks, the total energy deposition in the sample ranges from 37 eV to 390 eV.

2.3 Implications for Tracking Detectors

For high-precision tracking, there are clear advantages in keeping the silicon detector as thin as possible. A physically thin detector is optimal as regards multiple scat-

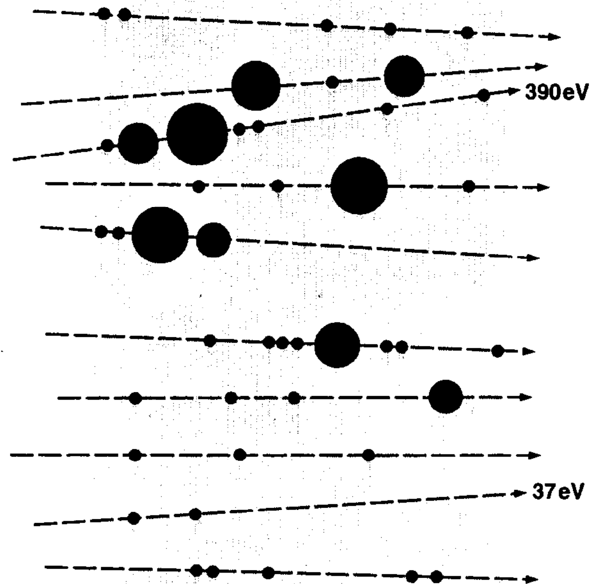


Fig. 5. Monte Carlo calculation of energy deposition in a 1 μm thick silicon detector. Area of a blob represents the energy deposited in each primary collision process.

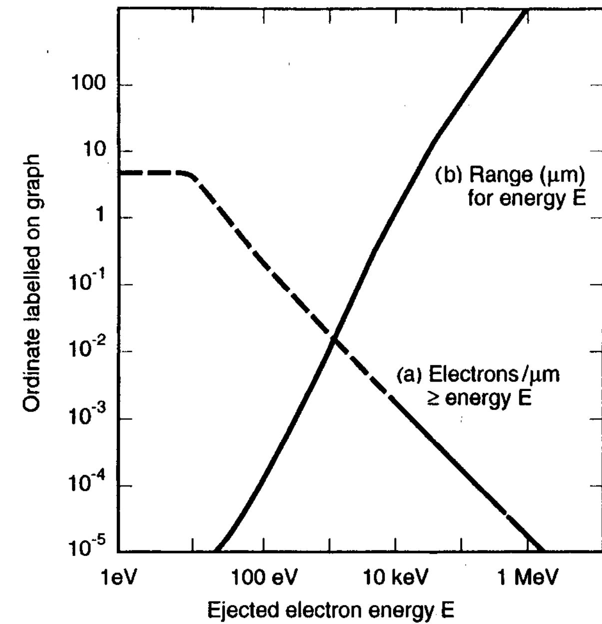


Fig. 6. (a) shows the number of electrons per micron of MIP track above a given energy, and (b) shows the range in silicon corresponding to that energy.

tering. A detector with the thinnest possible active region (which may be less than the physical thickness, as we shall see) is optimal as regards point measurement precision, for *two distinct reasons.*

For normal incidence tracks, the concern arises from δ-electrons of sufficient range to pull the centroid of the charge deposition significantly off the track. Figure 6(a) is an integral distribution of the number of primary electrons per micron of energy greater than a given value, and Fig. 6(b) shows the range of electrons of that energy in silicon. The range becomes significant for high-precision trackers for E greater than approximately 10 keV, for which the generation probability is less than 0.1% per micron. Thus, a detector of thickness 10 μm is much less likely to yield a "bad" co-ordinate than one of thickness 100 μm.

If the magnitude of the energy deposition in the detector is measured (by no means always possible), some of the bad co-ordinates will be apparent by the abnormally large associated energy. They could then be eliminated by a cut on the energy deposit, but this usually leads to unacceptable inefficiency and is rarely implemented. The situation is summarized in Fig. 7, which indicates the probabilities of the centroid for a track being pulled by more than a certain value (1 μm and 5 μm) as a function of detector thickness. The advantage of a thin active medium is apparent.

The second reason for preferring detectors to be as thin as possible applies to the case of angled tracks. In principle (and occasionally in practice), it may be possible to infer the position of such a track by measuring the entry and exit points in the detector, but more usually, the best one can do is to measure the centroid of the elongated charge distribution and take this to represent the track position as it traversed the detector mid-plane. In this case, large fluctuations in the energy loss (due to ejection of K- and L-shell electrons and δ-electrons) may be sufficient to cause serious track pulls for thick detectors. This is illustrated in Fig. 8. In the thin detector, there is a 10% probability of producing a δ-electron which, if it occurs near one end of the track, pulls the co-ordinate from its true position by 4 μm. In the thick detector, there is the same probability of producing a δ-electron which can pull the co-ordinate by 87 μm.

However, our enthusiasm for thin, active detector layers must be moderated by the primary requirement of any tracking system, namely a high efficiency per layer. Figure 9 (based on Ref. [5]) illustrates the problem we could already anticipate from Fig. 5. For very thin detectors (e.g., 1 μm Si), we see a very broad energy loss distribution with peaks corresponding to 0, 1, 2, ... plasmons excited, followed by a long tail extending to very large energy losses. An efficient tracking detector could never be built with such an active layer. Even at 10 μm silicon thickness, the true distribution is much broader than Landau and has a dangerous low tail. By 300 μm, the Landau distribution gives an adequate representation. Thus, while very thin detectors are ideal from the viewpoint of tracking precision, great care must be taken to assure that *system noise* allows a sufficiently *low threshold* to achieve the desired detector efficiency.

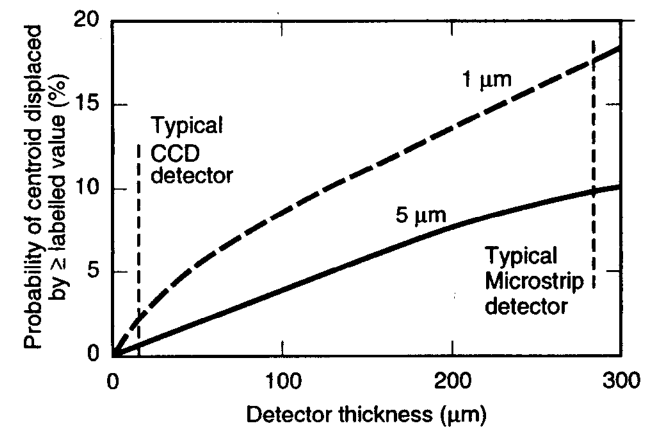


Fig. 7. Detector precision limitations from δ-electrons for tracks of normal incidence, as a function of detector thickness.

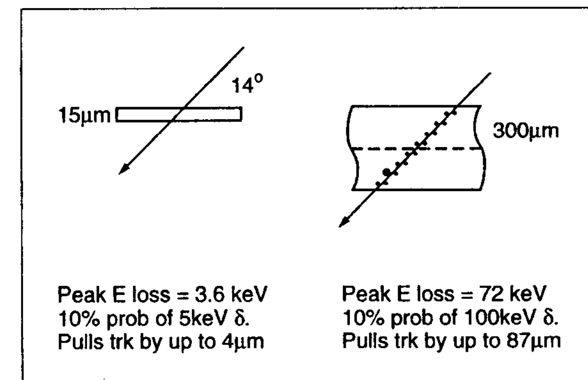


Fig. 8. Effect of energy loss fluctuations on detector precision for angled tracks.

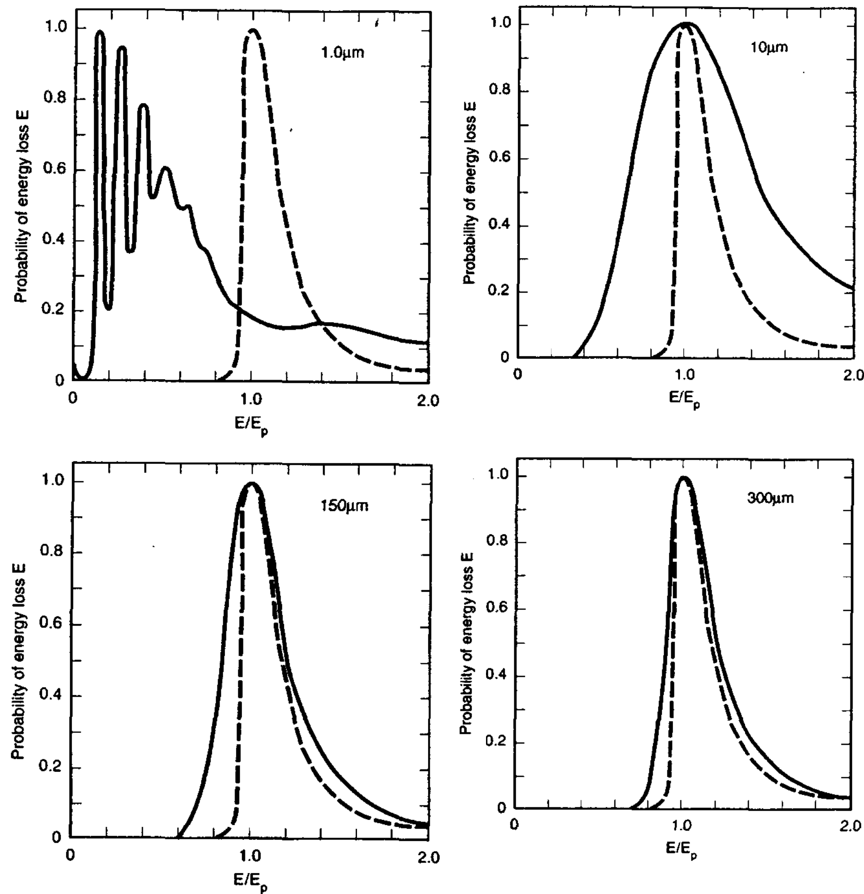


Fig. 9. Energy loss distributions for various silicon detector thicknesses, with (in each case) a Landau distribution for comparison. The separate peaks corresponding to 0, 1, 2 ... plasmon excitation are already merged by a thickness of 10 μm.

2.4 Summary

The valence electrons of silicon are very easily excited into plasmon oscillations from which they dislodge a small number (typically five) of electrons into the conduction band. A MIP thus creates a fine trail of electron-hole pairs along its track. The quantity W (energy needed to create an electron-hole pair) is approximately 3.6 eV, but depends on the band gap and hence (weakly) on the temperature. This energy loss process allows, in principle, unprecedented precision (much better than 1 μm) even compared to a nuclear emulsion (which needs typically a 400 eV δ-electron to blacken a grain). One does need to be prepared to exclude the measurements associated with large energy deposition, but these are rare in thin detectors.

How can this potential performance be achieved in practice? Standard IC processing (the planar technology) provides us with a host of suitable tools. This is after all one of the few areas of engineering in which submicron tolerances are now standard practice. In Secs. 4 and 5, we shall explore some types of detectors currently available. But first, we consider some of the basic properties of silicon which allow us in principle to collect and sense the signal charges we have been discussing in this section.

3 Physics and Properties of Semiconductors

Gaseous silicon has a typical structure of atomic energy levels (see Fig. 10). It has an ionization potential of 8.1 eV, i.e., it requires this much energy to release a valence electron, compared with 15.7 eV for argon. As silicon condenses to the crystalline form, the discrete energy levels of the individual atoms merge into a series of energy bands in which the individual states are so closely spaced as to be essentially continuous. The levels previously occupied by the valence electrons develop into the *valence band*, and those previously unoccupied become the *conduction band*. Due to the original energy level structure in gaseous silicon, it turns out that there is a gap between these two bands. In conductors, there is no such gap; in semiconductors, there is a small gap (1.1 eV in silicon, 0.7 eV in germanium), and in insulators, there is a large band gap. In particular, the band gap in silicon dioxide is 9 eV. This makes it an excellent insulator, and coupled with the ease with which the surface of silicon can be oxidized in a controlled manner, accounts partly for the pre-eminence of silicon in producing electronic devices.

We shall denote as E_v and E_c the energy levels of the top of the valence band and the bottom of the conduction band (relative to whatever zero we like to define). The energy needed to raise an electron from E_c to the vacuum E_0 is called the electron affinity. For crystalline silicon, this is 4.15 eV.

3.1 Conduction in Pure and Doped Silicon

To understand the conduction properties of pure silicon, the *liquid analogy* is helpful. This is illustrated in Fig. 11: (a) shows the energy levels in silicon under no applied voltage with the material at absolute zero temperature. All electrons are in the valence band, and under an applied voltage, (b) there is no change in the population of occupied states, and so no flow of current; the material acts like an insulator. At a high temperature, (c) a small fraction of the electrons are excited into the conduction band, leaving an equal number of vacant states in the valence band. Under an applied voltage, (d) the electrons in the conduction band can flow to the right and

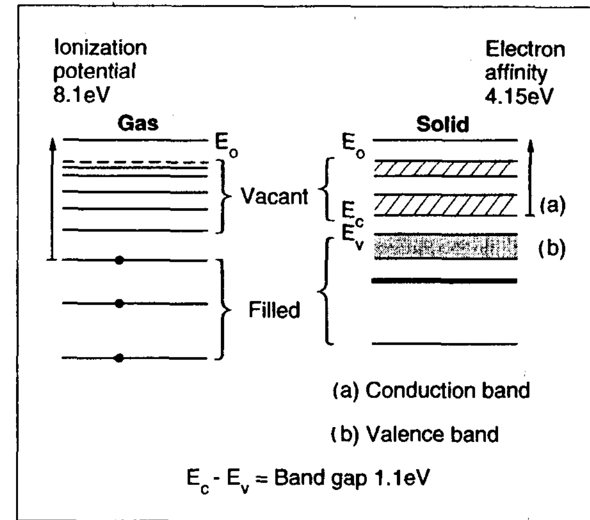


Fig. 10. Sketch of allowed energy levels in gaseous silicon which become energy bands in the solid material.

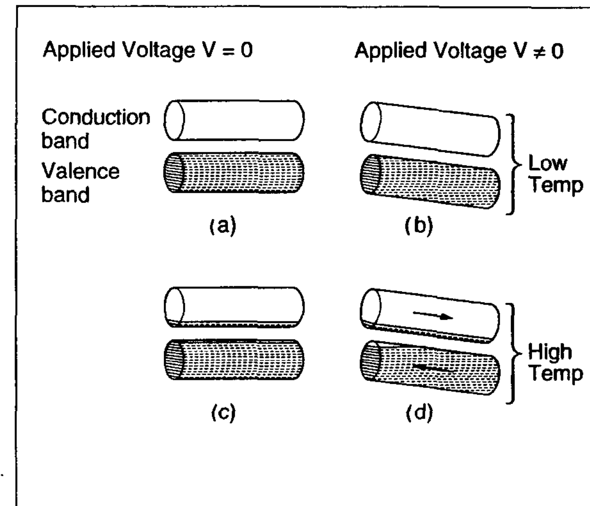


Fig. 11. Liquid analogy for a semiconductor.

there is a repopulation of states in the valance band which can be visualized as the leftward movement of a bubble (holes) in response to the applied voltage.

Now kT at room temperature is approximately 0.026 eV. This is small compared with the band gap of 1.1 eV, so the conductivity of pure silicon at room temperature is very low. To make a quantitative evaluation, we need to introduce the Fermi-Dirac distribution function $f_D(E)$ which expresses the probability that a state of energy E is filled by an electron. Figure 12(a) shows the form of this function

$$f_D(E) = \frac{1}{1+\exp\left(\frac{E-E_f}{kT}\right)}. \tag{3.1}$$

Note that E_f, the Fermi level, is the energy level for which the occupation probability is 50%. Figure 12(b) sketches the density of states $g(E)$ in silicon. The concentration of electrons in the conduction band is given by the product $f_D g$, and the density of holes in the valence band by $(1-f_D)g$, as shown in Fig. 12(c). In pure silicon, the Fermi level is approximately at the midband gap, and the concentrations of electrons and holes are, of course, equal. These concentrations, due to the form of f_D, are much higher in a narrow band gap semiconductor, Fig. 12(d), than in a wide gap material, Fig. 12(e).

So far, we have been discussing pure (so-called *intrinsic*) semiconductors. Next, we have to consider the *doped* or extrinsic semiconductors. These allow us to achieve high concentrations of free electrons [n-type, Fig. 12(f)], or of holes [p-type, Fig. 12(g)], by moving the Fermi level very close to the conduction or valence band edge. The procedure for doing this is to replace a tiny proportion of the silicon atoms in the crystal lattice by dopant atoms with a different number of valence electrons.

Figure 13 shows the lattice structure characteristics of diamond, germanium, and silicon crystals. Silicon, with four valence electrons, forms a very stable crystal with covalent bonds at equal angles in space. It is possible (e.g., by ion implantation) to introduce a low level of (for example) pentavalent impurities such as phosphorus. By heating (*thermal activation* as it is called), the phosphorus atoms can be induced to take up lattice sites in the crystal. For each dopant atom, four of its electrons share in the covalent bonding with neighboring silicon atoms, but its fifth electron is extremely loosely bound. At room temperature, this electron would be free, and hence available for conduction in a sea of fixed positive charge (the phosphorus ions, present at precisely the same average density as the liberated electrons). At absolute zero, all valence electrons would be bound and the phosphorus-doped (n-type) silicon effectively an insulator. The mathematical description of the effect of doping in silicon is to retain the Fermi-Dirac distribution function, but to raise the Fermi level (50% occupation probability) close to the binding energy of the fifth electron, i.e., close to the conduction band edge. The population of those electrons within the conduction band is again given by the overlap of the Fermi-Dirac distribution function (now shifted in energy) and the density of states in the conduction band. Except at very low temperatures (where the Fermi-Dirac function is extremely sharp), the result is a high density of electrons (*majority carriers*) and a negligible density of holes (*minority carriers*) in the n-type material in equilibrium, as shown in Fig. 12(f).

Alternatively, silicon may be doped with trivalent impurities such as boron. In this case, three strong covalent bonds are formed, but the fourth bond is incomplete. This vacancy (hole) is easily filled by an adjacent electron. Thus, as in the intrinsic material, holes behave as reasonably mobile, positively charged carriers in a sea of fixed negative charge (the boron atoms with an additional electron embedded in the fourth covalent bond). The carrier concentrations (now with holes as majority carriers) are given by shifting the Fermi-Dirac distribution to within the hole binding energy, i.e., close to the valence band edge as shown in Fig. 12(g).

The general situation regarding doped silicon is sketched in Fig. 14, which indicates the energy levels corresponding to various commonly used dopant atoms. Pentavalent atoms are referred to as *donors* and trivalent atoms as *acceptors*. Note that the carriers are bound by only approximately 0.045 eV in the common n- and p-type dopants phosphorus and boron, compared to kT at a room temperature of 0.026 eV.

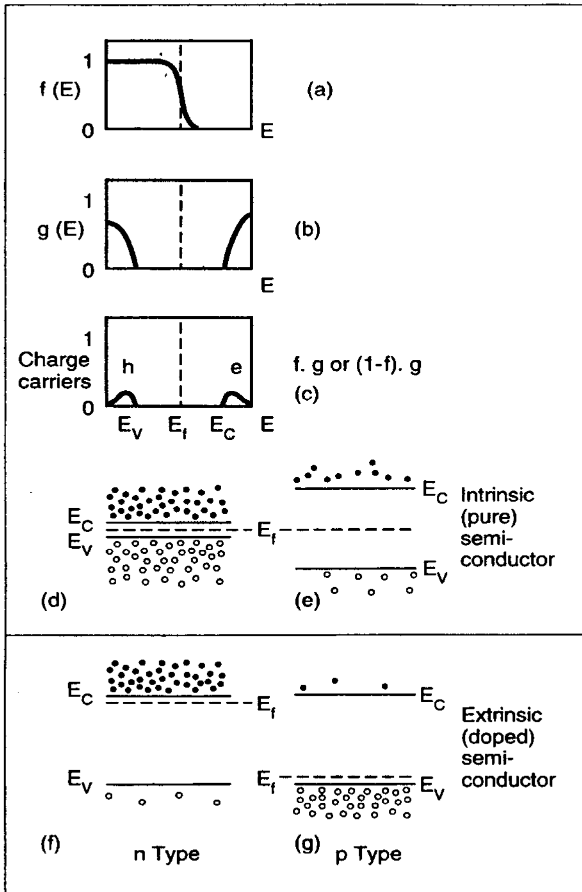


Fig. 12. (a) Fermi-Dirac distribution function. The slope increases as the temperature is reduced. (b) Density of states below and above forbidden band gap. (c) Concentration of charge carriers (electrons and holes) available for conduction. (d) and (e) Charge carrier distributions in narrow and wide band gap semiconductors. (f) and (g) Charge carrier distributions in *n*- and *p*-type semiconductors.

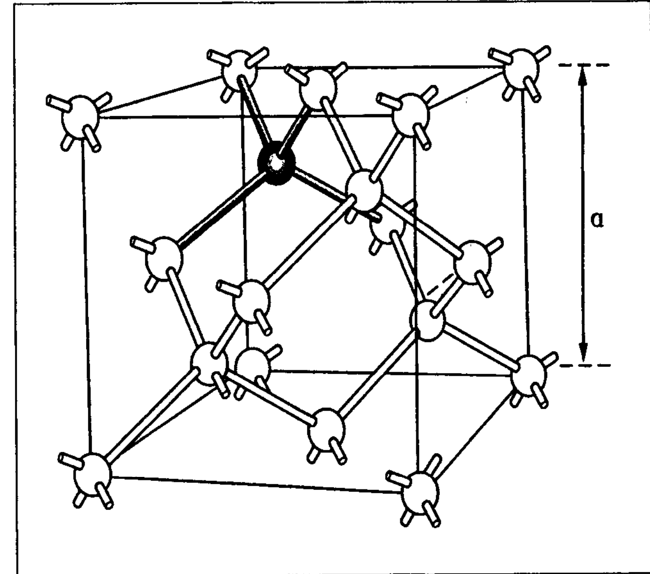


Fig. 13. Lattice structure of diamond, germanium, silicon, etc. where *a* is the lattice constant.

Figure 15 shows the concentration of electrons in *n*-type silicon (1.15 x 10^{16} arsenic dopant atoms per cm^3) as a function of temperature. Below about 100 K, one sees the phenomenon of *carrier freeze-out*, loss of conductivity due to the binding of the donor electrons. This is followed by a wide temperature range over which the electron concentration is constant, followed above 600 K by a further rise as the thermal energy becomes sufficient to add a substantial number of intrinsic electrons to those already provided by the dopant atoms. These will, of course, be accompanied by an equal concentration of mobile holes. The general behavior shown in Fig. 15 is typical of all doped semiconductors.

The *resistivity* ρ of the material depends not only on the concentration of free holes and electrons but also on their *mobilities*. As one would intuitively expect, the hole mobility is lower than that for electrons. Both depend on temperature and on the impurity concentration. At room temperature, in lightly doped silicon, we have

electron mobility $\mu_n = 1350\ cm^2\ (V\ s)^{-1}$,

hole mobility $\mu_p = 480\ cm^2\ (V\ s)^{-1}$,

and the resistivity is given by

$$\rho = \frac{1}{q_e(\mu_n \times n + \mu_p \times p)} \tag{3.2}$$

(n and p are the electron and hole concentrations).

For pure silicon at room temperature, $n_i = p_i = 1.45 \times 10^{10}\ cm^{-3}$ which gives $\rho_i = 235\ K\ \Omega\ cm$.

The carrier drift velocity (v_p for holes and v_n for electrons) is related to the mobility by $v_{p,n} = \mu_{p,n}\mathcal{E}$, where $\mathcal{E}$ is the electric field strength. This relationship applies only up to a maximum field, beyond which saturation effects come into play and one enters the realm of "hot carriers" which lose energy by impact ionization (creation of additional electron-hole pairs). Figure 16 shows the situation for silicon, as well as

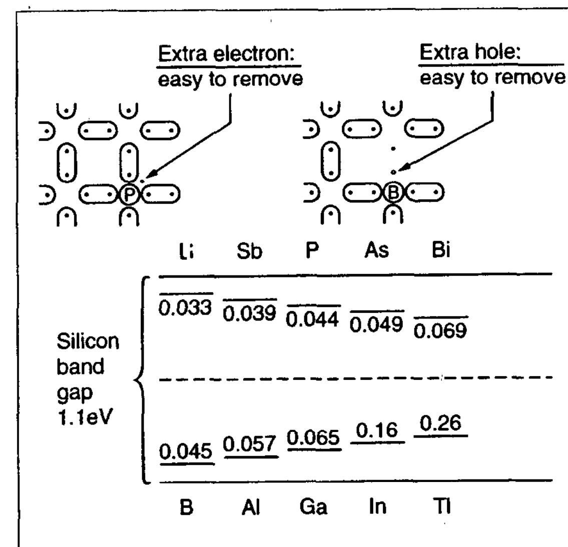


Fig. 14. Sketch of band occupation in doped silicon (upper) and energy levels within the band gap corresponding to various *n*- and *p*-type dopants (lower). Levels of acceptor atoms are conventionally measured from the top of the valence band, and levels of donor atoms are measured from the bottom up the conduction band.

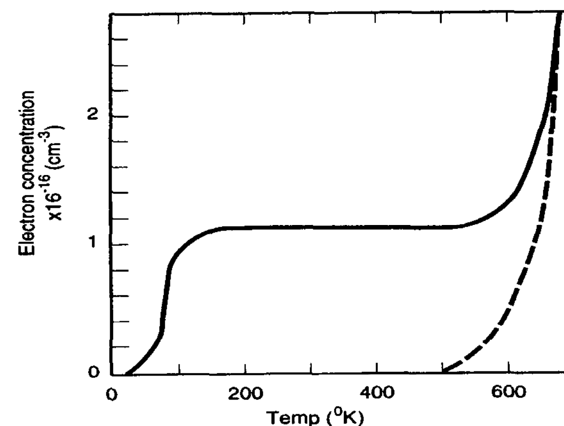
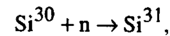


Fig. 15. Electron concentration versus temperature for *n*-type (arsenic doped) silicon. The dashed curve shows the concentration for intrinsic material.

the fact that nearly ten times higher electron drift velocities are achievable in gallium arsenide, which therefore has the potential for much faster electronic circuits.

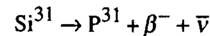
The ionization rate is defined as the number of electron-hole pairs created per unit of distance travelled by an electron or hole. It depends primarily on the ratio $q_e\mathcal{E} / E_i$ where E_i is the effective ionization threshold energy, damped by terms reflecting the energy loss of carriers by thermal and optical phonon scattering, see Ref. [6]. For silicon, E_i is approximately equal to W (3.6 eV) for electrons and 5.0 eV for holes. The ionization rate becomes significant for electric fields in the range 10^5 to 10^6 V/cm in silicon, leading to the saturation of carrier drift velocity shown in Fig. 16.

The resistivity as a function of dopant concentration is shown in Fig. 17. For silicon detectors, as will be explained in the next section, we are often concerned with unusually high resistivity material, some tens of KΩ cm. From Fig. 17, one sees, for example, that 20 KΩ cm *p*-type material requires a dopant concentration of 5 x 10^{11} cm^{-3}. Remembering that crystalline silicon has 5 x 10^{22} atoms per cm^3, this implies an impurity level for the *predominant* impurities of 1 in 10^{11}. Even in the highly developed art of silicon crystal growing, this presents a major challenge. The resistivity noted above in connection with pure silicon (over 200 KΩ cm) is entirely unattainable in practice. Very high resistivity *n*-type silicon can be produced in the form of *compensated* material. The most uniformly doped material which can be grown is (for technical reasons) *p*-type, and this (with a resistivity of about 10 KΩ cm) is used to start with. It is then turned into *n*-type material by the procedure known as neutron doping. The crystal is irradiated with slow neutrons and by means of the reaction



$$Si^{30} + n \rightarrow Si^{31},$$

followed by



$$Si^{31} \rightarrow P^{31} + \beta^- + \bar{\nu}$$

is turned into *n*-type material. The resistivity is monitored and the irradiation ceases when this, having passed through a maximum, falls to the required value. In this way, material of resistivity as high as 100 KΩ cm can be made. Achieving

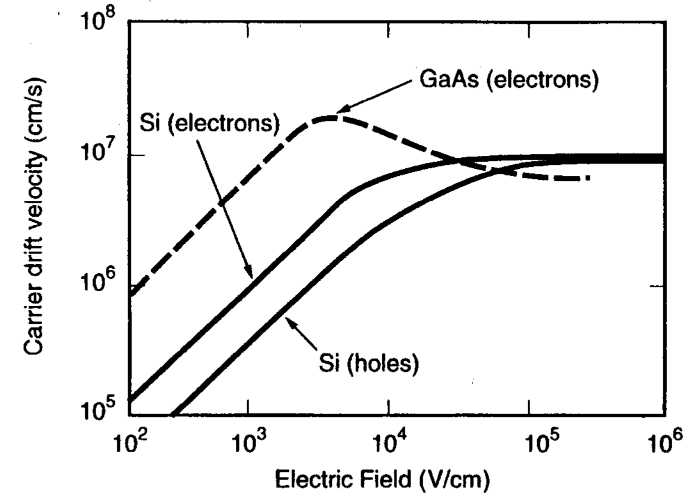


Fig. 16. Carrier drift velocity (electrons and holes) for silicon, and electron velocity for gallium arsenide as a function of electric field in the material.

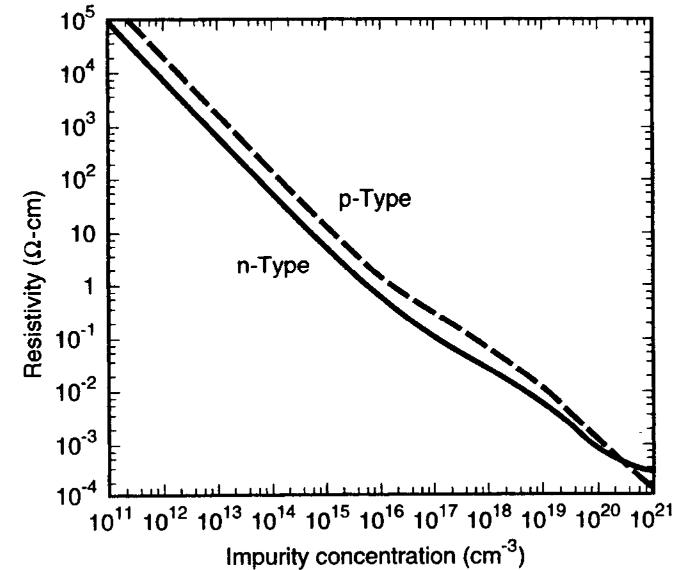


Fig. 17. Resistivity of silicon at room temperature as a function of acceptor or donor impurity concentration.

reasonable uniformity through the wafer of such a high resistivity is obviously extremely difficult.

We now consider more quantitatively the relationship between the carrier concentration and the Fermi level. The number of conduction band states occupied by electrons is given by

$$n = \int_{E_c}^{E_t} N(E) f_D(E) dE.$$

E_c and E_t are the energy at the bottom and top of the conduction band; $f_D(E)$ is the function (3.1); $N(E)$, the density of states, is given by the band theory of solids and is proportional to $(E-E_c)^{1/2}$. For the commonly encountered situation where Boltzmann statistics applies, for which the Fermi level is at least several times kT below E_c, the above integral can be approximately evaluated to yield

$$n = N_c \exp\left(-\frac{E_c - E_f}{kT}\right). \quad (3.3)$$

N_c is called the effective density of states. Its meaning is not as intuitively clear as the simple density of states $N(E)$; unlike $N(E)$, it is temperature dependent, being proportional to $T^{3/2}$.

The equivalent approximation for the hole concentration is

$$p = N_v \exp\left(-\frac{E_f - E_v}{kT}\right). \quad (3.4)$$

For *intrinsic* semiconductors, thermal agitation excites electrons from the valence band to the conduction band, leaving an equal number of holes in the valence band. In this case, $n = p = n_i$, where n_i is the intrinsic carrier density. There is a dynamic equilibrium between thermal generation on the one hand, and recombination of electrons in the conduction band with holes in the valence band, on the other. The neutrality condition obtained by equating Eqs. (3.3) and (3.4) leads to

$$E_f = E_i = \frac{E_c + E_v}{2} + \frac{kT}{2} \ln\left(\frac{N_v}{N_c}\right). \quad (3.5)$$

Thus, the Fermi level of an intrinsic semiconductor lies very close to the middle of the band gap. The intrinsic carrier density is given from Eqs. (3.3) and (3.4) also:

$$pn = n_i^2 = N_c N_v \exp(-E_g / kT) \quad (3.6)$$

where $E_g = E_c - E_v$.

Note that

$$n_i = \sqrt{N_c N_v} \exp(-E_g / 2kT) \quad (3.7)$$
$$\propto T^{3/2} \exp(-E_g / 2kT).$$

Thus, n_i has a rapid temperature dependence, doubling for every 12°C rise for silicon around room temperature.

For doped silicon, e.g., *n*-type, the neutrality condition is between the ionized donors and the conduction band electrons created by the ionization process. For a dopant energy level E_d, the number of ionized donors is related to the Fermi level by the relation

$$N_d^+ = \frac{N_d}{1 + 2\exp\left(\frac{E_f - E_d}{kT}\right)}. \quad (3.8)$$

See Ref. [7]. From Eqs. (3.3) and (3.8), we have the neutrality condition

$$N_c \exp\left(-\frac{E_c - E_f}{kT}\right) = \frac{N_d}{1 + 2\exp\left(\frac{E_f - E_d}{kT}\right)}. \quad (3.9)$$

Figure 18 shows graphically the solution of Eq. (3.9) for two temperature values. At room temperature, the donor atoms are completely ionized and the carrier

concentration is essentially equal to N_d, with $E_f = E_{f1}$ a little below E_d. At the reduced temperature, $E_f = E_{f2}$ falls in the small energy range between E_d and E_c, and the carrier concentration plummets. Conversely, at very high temperatures, thermal excitation of valence band electrons would become dominant, causing the carrier concentration to rise rapidly, and the Fermi level to stabilize near the middle of the band gap, off-scale to the left in the figure. For *p*-type material, the number of ionized acceptors is given by

$$N_a^- = \frac{N_a}{1 + 4\exp\left(\dfrac{E_a - E_f}{kT}\right)}. \tag{3.10}$$

The difference in the factors in the denominator arises from the difference between the ground-state degeneracy for donor and acceptor levels.

In general, for doped material, we have

$$\left.\begin{aligned} n &= n_i \exp\left(\frac{E_f - E_i}{kT}\right) \\ p &= n_i \exp\left(\frac{E_i - E_f}{kT}\right) \end{aligned}\right\} \tag{3.11}$$

and $pn = n_i^2 = N_c N_v \exp(-E_g / kT)$ *just as for intrinsic material.* Thus, the deviation of a doped semiconductor from the intrinsic condition can be simply represented by a shift in the Fermi energy level with respect to the intrinsic level. The constancy of the *pn* product for different doping conditions is a particular example of the very important *law of mass action* which applies as much in semiconductor theory as it does in chemistry. In thermal equilibrium, the increase in electron concentration by donor doping causes a decrease in the concentration of mobile holes (by recombination) such that the *pn* product is constant. The ionized donors in this sense are passive bystanders, serving to preserve charge neutrality. It is generally valid to think of *n*-type material in equilibrium as containing only mobile electrons, and *p*-type material as containing only mobile holes, the majority carriers in each case.

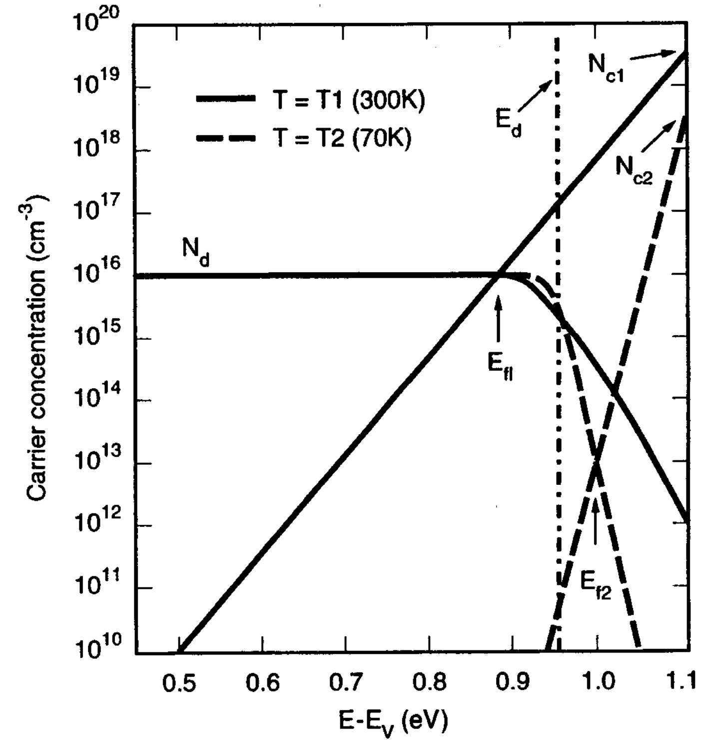


Fig. 18. Number of ionized acceptors and number of conduction band electrons versus the Fermi energy level E_f.

3.2 The *pn* Junction

We now need to introduce a most important fact related to conducting materials which are electrically in contact with one another and in thermal equilibrium; *they all must establish the same Fermi energy.* This applies to

metal/semiconductor systems
n-type/*p*-type systems, etc.

Charge flows from the high- to low-energy region for that carrier type until this condition is established. For example, at a *pn* junction, there develops a fixed space charge of ionized donors and acceptors, creating a field which opposes further drift of electrons and holes across the junction. The *depletion approximation* says that the semiconductor in this condition changes abruptly from being neutral to being fully depleted. This is far from obvious, and in fact, there is a finite length (the *Debye length*, typically 0.1 μm) over which the transition takes place. But the depletion approximation will be adequate for all the examples we need to consider. Let us look in some detail at the important case of the *pn* junction. Before contact [Fig. 19 (a)], the surface energy E_0 is equal in both samples; the *p*-type Fermi level is close to E_v and the sample is densely populated by holes; the *n*-type Fermi level is close to E_c and the sample is densely populated by electrons.

On contact, the electrons diffuse into the electron-free material to the left, and the holes diffuse to the right. In so doing, the electrons leave exposed donor ions (positively charged) over a thickness x_n in the *n*-type material, and the holes leave exposed acceptor ions (negatively charged) over a thickness x_p in the *p*-type material. This builds up an electric field which eventually just balances the tendency for current to flow by diffusion. Once this condition is reached [Fig. 19 (b)], the Fermi levels in the materials have become equal. The electrical potentials in the two samples (for example, the potential energy at the surface E_0 or at the conduction band edge E_c) are now unequal.

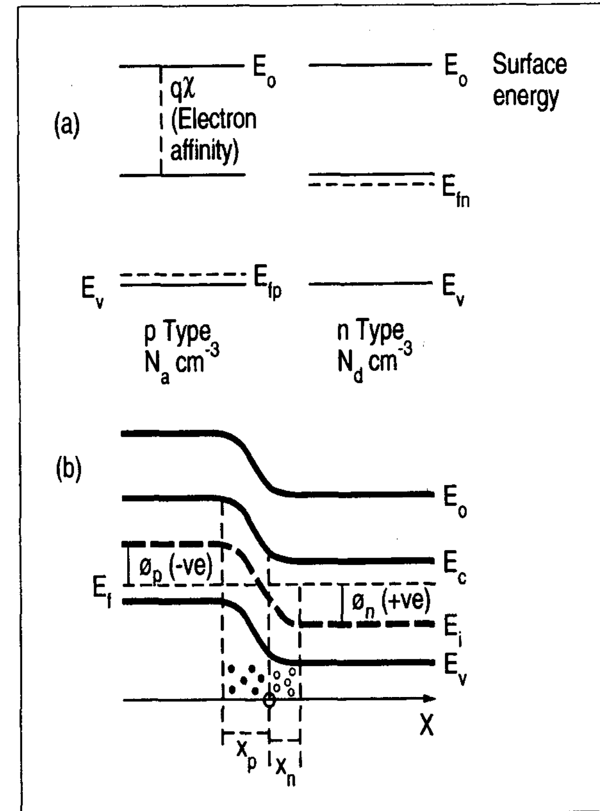


Fig. 19. (a) Energy levels in two silicon samples (of *p* and *n* type) when electrically isolated from one another. (b) When brought into contact, the Fermi level is constant throughout the material. The band edges bend in accordance with the space charge generated.

Intuitively, this can be understood as follows. Initially, the electrons at a particular level in the conduction band of the n-type material see energy levels in the p-type material at equal or lower energy which are unpopulated, so they diffuse into them. The developing space charge bends the energy bands up, so that these levels become inaccessible. Eventually, only very high-energy electrons in the n-type material see anything other than the absence of states of the band gap in the p-type material, and conversely for the holes in the p-type material.

Let us develop this quantitatively, adopting a coordinate system in which the pn junction of Fig. 19(b) is at position $x = 0$. E_0, E_c, E_i, and E_v all follow the same x dependence. The zero of the electric potential ϕ is arbitrary, so we define

$$\phi = -\frac{(E_i - E_f)}{q_e}. \tag{3.12}$$

Thus, ϕ is 0 for intrinsic material

positive	for n-type
negative	for p-type.

From Eq. (3.11), in the case of fully ionized donors and acceptors,

$$\phi_n = \frac{kT}{q_e}\ln\left(\frac{N_d}{n_i}\right)$$

$$\phi_p = -\frac{kT}{q_e}\ln\left(\frac{N_a}{n_i}\right).$$

The potential barrier

$$\phi_i = \phi_n - \phi_p = \frac{kT}{q_e}\ln\left(\frac{N_d N_a}{n_i^2}\right). \tag{3.13}$$

Notice that the potential barrier falls linearly with temperature since it is sustained by the thermal energy in the system. We may deduce the electric field strengths $\mathcal{E}(x)$ near the junction by using Poisson's equation

$$\frac{d^2\phi}{dx^2} = -\frac{d\mathcal{E}}{dx} = -\frac{q_e}{\varepsilon_s}\rho(x).$$

ε_s is the permittivity of silicon $= \varepsilon_r \varepsilon_0$.
ε_0 is the permittivity of space $= 8.85 \times 10^{-14}\,\mathrm{F\ cm^{-1}}$
$= 55.4\ e^-/\mathrm{V}\ \mu\mathrm{m}$.
ε_r is the dielectric constant or
relative permittivity of silicon $= 11.7$.

For $x_n \geq x \geq 0$,

$$\frac{d\mathcal{E}}{dx} = +\frac{q_e N_d}{\varepsilon_s} \qquad \therefore \mathcal{E}(x) = -\frac{q_e N_d}{\varepsilon_s}(x_n - x).$$

For $-x_p \leq x \leq 0$,

$$\frac{d\mathcal{E}}{dx} = -\frac{q_e N_a}{\varepsilon_s} \qquad \therefore \mathcal{E}(x) = -\frac{q_e N_a}{\varepsilon_s}(x + x_p). \tag{3.14}$$

The *undepleted* silicon on either side of the junction is *field free*. The depleted silicon close to the junction experiences an electric field whose strength is maximum at the junction and is directed always to the left, i.e., opposing the flow of holes to the right and opposing the flow of electrons to the left.

Requiring continuity of the field strength at $x = 0$ implies

$$N_a x_p = N_d x_n. \tag{3.15}$$

Thus, if one wants to make a deep depletion region on one side of the junction (important, as we shall see, for many detectors), we need to have a very low dopant concentration, i.e., very high resistivity material.

The electric field strength varies linearly with x; the electric potential, by integration of Eq. (3.14), varies quadratically.

For $x_n \geq x \geq 0$,

$$\left.\begin{aligned} \phi(x) &= \phi_n - \frac{q_e N_d}{2\varepsilon_s}(x_n - x)^2 \\ \text{For } -x_p \leq x \leq 0, \quad & \\ \phi(x) &= \phi_p + \frac{q_e N_a}{2\varepsilon_s}(x + x_p)^2. \end{aligned}\right\} \tag{3.16}$$

Requiring continuity of the potential at $x = 0$ implies

$$x_n + x_p = \left[\frac{2\varepsilon_s}{q_e}\phi_i\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{1/2}. \tag{3.17}$$

From Eq. (3.13), ϕ_i depends only weakly on N_a and N_d.

If, for example, $N_a >> N_d$, we have $x_p \approx 0$ and Eq. (3.17) gives $x_n \propto N_d^{-1/2}$.

So a factor of two increase in resistivity leads to a factor of only $\sqrt{2}$ increase in depletion depth.

Figure 20 summarizes these results on the characteristics of an unbiased *pn* junction, with the inclusion of some typical numerical values based on $N_a = 10^{14}\,\text{cm}^{-3}$ and $N_d = 2 \times 10^{14}\,\text{cm}^{-3}$. The peak field in this case is about 3 kV/cm.

We now consider the effect of applying a voltage across the junction. Under equilibrium conditions, electron-hole pairs are continually generated by thermal excitation throughout the semiconductor. In the case of zero bias [Fig. 21(a)], the electrons and holes generated within the bulk of the semiconductor recombine. Those generated in the depletion region are swept into the undepleted silicon, holes to the left, electrons to the right. This effect would act to reduce the potential barrier and so is compensated by a small flow of *majority carriers* which find themselves with just sufficient energy to diffuse across the barrier in the opposite directions at just the rate needed to cancel the charge generation in the depleted material. The overall effect is of zero current flow, i.e., equilibrium.

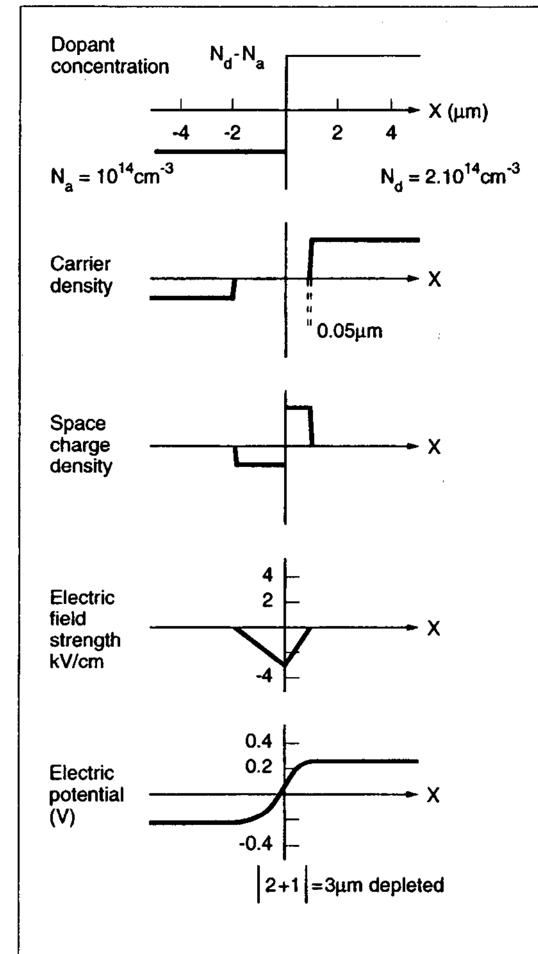


Fig. 20. Summary of various quantities across an unbiased *pn* junction.

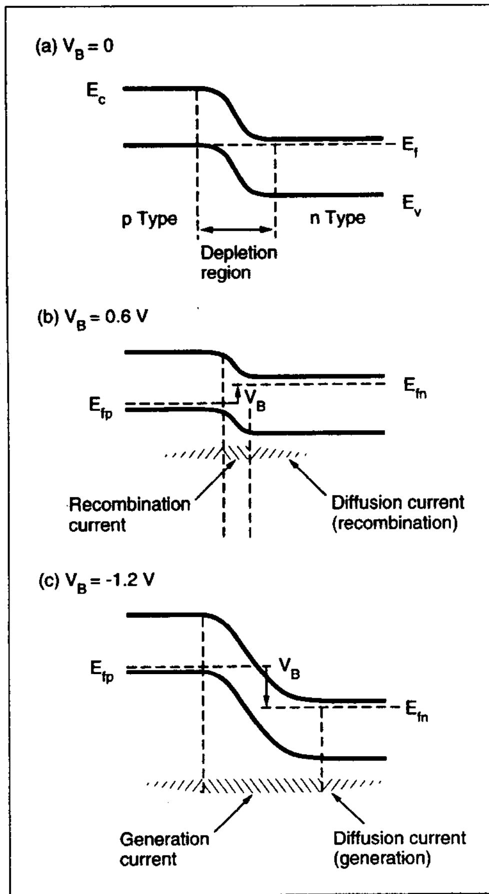


Fig. 21. Effect of an applied voltage across the semiconductor junction.

By applying a forward bias [Fig. 21(b)], we separate the previously equal Fermi levels by an amount equal to the bias voltage; the system is no longer in thermal equilibrium or this condition could not be maintained. Although there is still an electric field in the depletion region which is directed against the current flow, the depletion region is narrowed and the potential barrier is now inadequate to prevent majority carriers from flooding across it, holes from the left and electrons from the right. Many of these will recombine within the depletion region giving rise to the *recombination current.* Those which survive are absorbed within one or two diffusion lengths by recombination with the majority carriers on that side of the junction, giving rise to the *diffusion current.* Beyond these regions, there is just a steady flow of majority carriers supplied from the voltage source to keep the current flowing. Notice that in a forward-biased junction, the current flow results entirely in electron-hole *recombination.*

With a reverse bias, we have the situation shown in Fig. 21(c). The depletion region is now much wider and electron-hole pairs generated within it are efficiently swept into the undepleted silicon, electrons to the right and holes to the left, giving rise to the *generation current.*

Unlike the case of the unbiased junction, there is now no supply of majority carriers able to overcome the increased potential barrier across the junction. On the contrary, the thermal generation of *minority carriers* within one or two diffusion lengths of the depletion region leads to some holes generated in the *n*-region reaching this depletion region and then being briskly transported across it, and conversely for electrons generated in the *p*-region. This leads to the so-called *diffusion current.* In the case of the reverse-biased junction, the current flow is thus caused entirely by electron-hole *generation.* The current flow across reverse-biased junctions is of great importance in determining the noise limits in silicon detectors. An immediate observation is that since this current arises from *thermal* generation of electron-hole pairs, the operating temperature will be an important parameter.

Fig. 22. Operating principles of a simple microstrip detector. It consists essentially of a reverse biased diode, operated at a voltage sufficient to fully deplete the thick, high-resistivity n layer.

Before continuing to discuss this point, it is worth noting that we have finally collected up enough information to calculate the characteristics of a typical particle detector, and it is instructive to do so. Referring to Fig. 22, we have a silicon detector of thickness ℓ made of good-quality, high-resistivity n-type silicon (ρ = 10 KΩ cm). On the front surface, we make shallow implants of acceptor atoms (the p strips), and on the back surface, we make a highly doped n-type implant to provide a good low-resistance ohmic contact. The terms n^+ and p^+ are conventionally used to represent high doping levels, n and p represent moderate levels, n^-, p^-, or π and ν represent low levels, and i is used for intrinsic or compensated material of the highest possible resistivity.

Now we apply a positive voltage V_B to the n-type surface with the aim of completely depleting the detector. In this way, we shall ensure complete collection of the electrons and holes generated by the passage of a charged particle; with incomplete depletion, we would lose signal by recombination. Equation (3.17) applies, with the difference that we replace ϕ_i by $V_B + \phi_i$ since the junction is biased in the direction which assists the previously existing depletion voltage.

We have

$$x_n + x_p \approx x_n = \left[\frac{2\varepsilon_s}{q_e}(V_B + \phi_i)\left(\frac{1}{N_a} + \frac{1}{N_d}\right)\right]^{1/2}$$

$$\approx \left[\frac{2\varepsilon_s}{q_e} \times \frac{V_B}{N_d}\right]^{1/2}.$$

From Fig. 17, we see that $N_d \approx \dfrac{6\times10^{15}}{\rho}$, and we require $x_p = \ell$

$$\therefore V_B = \frac{q_e}{2\varepsilon_s} \times \frac{6\times10^{15}}{\rho} \times \ell^2$$

$$= \frac{10^{-4}}{2\times55.4\times11.7} \times \frac{6\times10^{15}}{\rho}\ell^2 \times 10^{-8},$$

where ℓ is in μm and ρ in Ω cm

$$\therefore V_B = \frac{4.63\ell^2}{\rho}.$$

For the above example, $V_B = 42$ V is the potential needed to fully deplete the detector. We also find $\mathscr{E}_{max} = 2.8$ kV/mm. This looks comfortable in terms of the breakdown field in silicon. However, in a real detector, it is important to pay

attention to the regions near the edges of the p strips, where the fields can be very much higher.

Returning to the general properties of the reverse biased junction, the most important parameter influencing the leakage current is the operating temperature. At high temperatures, above 100°C typically, the leakage current is dominated by thermal electron-hole generation within approximately one diffusion length of the depletion edge. The diffusion length for minority carriers is

$$L_D = \sqrt{D\tau_m}, \tag{3.18}$$

where D is the diffusion constant and is related to the mobility μ by the Einstein relation

$$D = \frac{kT}{q_e}\mu. \tag{3.19}$$

$$\left.\begin{array}{ll}\text{For electrons} & D_n = 34.6\ \text{cm}^2\text{s}^{-1} \\ \text{For holes} & D_p = 12.3\ \text{cm}^2\text{s}^{-1}\end{array}\right\} \text{at room temperature.}$$

τ_m is the minority carrier lifetime, and it can vary from about 100 ns to more than 1 ms depending on the quality of the silicon. This point will be discussed further. This leakage current (termed the diffusion current, as previously noted) depends only weakly on the reverse bias voltage but is highly temperature dependent due to its origin in the thermal generation of minority carriers.

At lower temperatures (less than about 100°C), the diffusion current becomes negligible and the generation current dominates. This continues to show a similarly fast temperature dependence, but is now also quite voltage dependent, since the depletion width is proportional to $V_B^{1/2}$.

The diffusion and generation currents depend on the rate of generation of electron-hole pairs, and the diffusion current depends also on the minority carrier lifetime. These quantities are, in fact, closely related. Direct thermal generation of an electron-hole pair is quite rare in silicon for reasons which depend on the details of the crystal structure. Most generation occurs by means of intermediate generation-recombination centers (impurities and lattice defects) near the band gap center. Thus, an electron-hole pair may be thermally created in a process where the hole is released into the valence band and the electron is captured by the trapping center in one step, to be subsequently emitted into the conduction band. These *bulk trapping states* vary enormously in their density and can be held down to a low level by suitable processing. It is precisely these states which determine the minority carrier lifetime already mentioned. Reducing the density of bulk trapping states does two things. It cuts down the thermal generation of charge carrier pairs in the material, so reducing the concentration of minority carriers available for the generation of current across a reverse-biased junction. It also increases the minority carrier lifetime and so the diffusion length (but only at $\tau^{1/2}$). The first effect vastly outweighs the second, so that a low density of bulk trapping states is highly advantageous in ensuring low leakage current. As we shall see later, even originally high-grade silicon can deteriorate due to the production of bulk trapping states by radiation damage. Mid-band gap impurities such as gold are a particularly serious source of bulk trapping centers. Even in low concentrations, gold atoms strongly reduce the carrier lifetimes and lead to greatly increased leakage current.

These effects are less serious in cases where one is collecting large signals promptly. But in cases of small signals and/or long storage times (such as in a silicon drift chamber, or CCD), particular care is needed. One important design criterion is to keep the stored charges well away from the surface of the silicon, since the silicon/silicon dioxide interface always has a high level of lattice defects. This criterion has led to the development of various forms of *buried channel* radiation detectors.

3.3 Charge Carrier Transport in Silicon Detectors

While the charge generated by an ionizing particle is being transported by the internal field in the detector, the process of diffusion spreads out the original very fine column of charge. In the case of very highly ionizing particles (such as alphas), the original density of electrons and holes can be so high that space-charge effects are important. In the case of MIP's, however, such effects are negligible and the time

development of the electron and hole charge distributions may be treated by simple diffusion theory.

Consider a local region of electron charge, for example, a short section of the particle track length within the silicon. Under the influence of the internal field, this will be drifted through the material, and at the same time, will diffuse radially as indicated in Fig. 23.

The RMS radius of the charge distribution increases as the square root of drift time t_d, as in Eq. (3.18), with standard deviation $\sigma = \sqrt{2Dt_d}$. Thus, 50% of the charge is contained within a radius of 0.95 $\sqrt{Dt_d}$. Assuming a "typical" drift field in depleted silicon of 1 kV/cm and using the fact that the drift velocity $V_d = \mu_n \mathcal{E}$, we obtain the following indication of the growth of a charge packet with time:

Drift Time	Charge Radius	Drift Distance
10 ns	6 μm	135 μm
1 μs	60 μm	14 mm.

Diffusive charge spreading is an attractive option for improving spatial precision beyond the limits of the detector granularity. For example, one might hope to achieve precision of one or two microns from a strip detector with 25 μm pitch, by centroid finding on the basis of measured charge collection in adjacent strips. This depends on achieving a charge radius of $\geq$ 30 μm which (from the above table) implies large drift distances and/or gentle drift fields. Ideas for improved precision by centroid finding may be limited by the available resistivity of silicon.

There is, however, an alternative approach that has so far been applied only to CCD detectors but which could be of more general interest. A wafer cut from a silicon crystal will normally have a rather uniform dopant concentration. It is possible subsequently to grow relatively thick (up to around 100 μm) *epitaxial layers* on the substrate wafer, of excellent crystalline quality and quite different (but also uniform) dopant concentration. For detector applications, a low-resistivity substrate with a high-resistivity epi layer is of particular interest. In the CCD case, as we shall see, the epi layer would be implanted with an *n* layer and biased so as to deplete only approximately 3 μm depth. The charge carrier transport associated with (for example) a charged particle track traversing such a structure is depicted in Fig. 24. Electrons within the thin depletion region are promptly collected into the buried channel, with no time for lateral diffusion. Electrons from the highly doped p^+ bulk are completely disposed of by recombination (very short minority carrier diffusion length in this material). However, electrons generated in the undepleted epitaxial layer find themselves able to diffuse homogeneously in all directions. Those which approach the p/p^+ junction experience a potential barrier as we have already discussed in the case of the unbiased *pn* junction, of magnitude

$$\phi_B = \frac{kT}{q_e} \ln\left(\frac{N_{p^+}}{N_p}\right).$$

For a 20 Ω cm epi layer on a highly doped 0.1 Ω cm substrate, we find

$$\phi_B = 180 \text{ mV compared with } \frac{kT}{q_e} = 26 \text{ mV}$$

at 300 K. The p/p^+ interface therefore acts as a *perfect mirror*, and the electrons continue diffusing until they happen to approach the *pn* depletion edge, at which point they are stored. Thus, a MIP leaves an electron charge cluster which is transversely spread by an amount related to the epi layer thickness. Such a detector made with partially undepleted thick epi material is in principle better for precision tracking by centroid finding than a fully depleted detector. To fully exploit this concept, one has to pay attention to the detector granularity, epi layer thickness, readout noise, etc. The most spectacular results in precision centroid finding in CCD's have been obtained not as yet with MIP's but with defocused star images in a satellite guidance system, where precision below 0.1 μm has been achieved using 20 μm pixels. This constitutes a very important demonstration of the inherent pixel-to-pixel homogeneity possible with high-quality silicon processing.

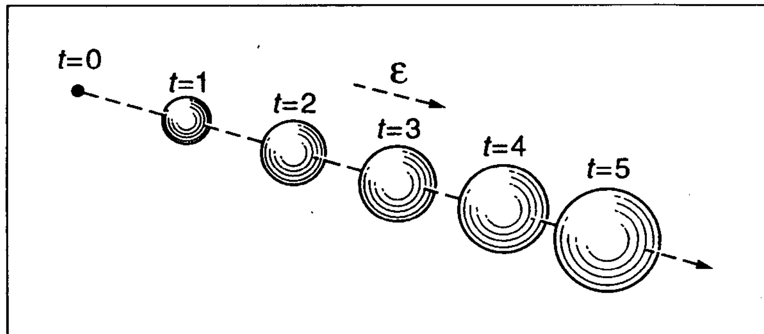


Fig. 23. Combined drift and diffusion of an initially compact charge cluster (electrons or holes) as a function of time over equal time intervals.

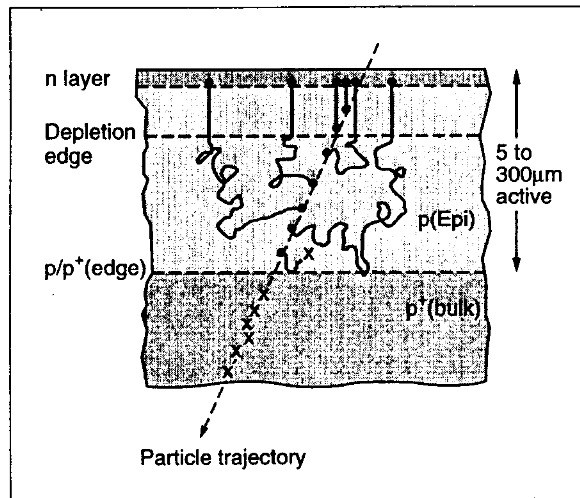


Fig. 24. Charge collection from a silicon structure as used in some pixel devices.

4 Microstrip Detectors

4.1 Introduction

Charged particles deposit a significant fraction of their energy by ionization in all types of materials, but only some are suitable as detector media. The conceptually most elementary detector types are insulators in which the signal is collected simply by applying a voltage to a pair of metal plates attached to the opposite faces of the detector layer, so creating an electric field within the material. The detection medium may be a gas (ionization chamber), a liquid (e.g., liquid argon calorimeter), or a solid (e.g., diamond detector). However, this principle cannot be applied to semiconductor detectors since even the highest purity material would generally have unacceptably low resistivity (i.e., excessive leakage current) except at extremely low temperature. As we have seen, it is possible to generate a region of internal electric field devoid of free charge carriers, and hence having greatly reduced leakage current, by creating a reverse-biased junction. Electron-hole pairs generated within the depletion region, for example, by thermal or optical excitation, or by the passage of a charged particle, are promptly swept to the surface for collection. This principle has been used for the detection of ionizing particles in silicon for over 40 years [8]. We have already noted some variations on this theme in connection with pixel devices (collection of minority carriers from undepleted material adjacent to depleted silicon), but the microstrip detector follows exactly this simple tradition.

The pioneering microstrip detectors of the early '80s (Ref. [9]) were based on the processes used for many years to manufacture nonsegmented semiconductor detectors for nuclear physics applications. The diodes were simply formed by the surface barrier between metal (aluminum) strips and the high-resistivity substrate. The strips were wire bonded to huge fanout boards which housed local pre-amplifiers connected to every Nth strip ($N \approx 5$). The principle of capacitive charge division was used to interpolate the track coordinates for signals collected on floating strips. The ratio of board area to detector area was almost 1000 to 1; this was tolerable in fixed-target experiments having unlimited space for local equipment outside the aperture of the forward spectrometer.

Closely following on these early developments, two revolutions took place which totally changed the technology of these detectors, opening up for them a much more powerful role in particle physics.

The first of these revolutions was to switch from surface barrier detectors to ion implantation, thus adopting the highly developed techniques used for processing integrated circuits. The microstrip detector becomes essentially a *p-i-n* diode structure, as we discussed in Sec. 3.2. The *p* strips (Fig. 22) were overlaid with metal (aluminum) to provide a low resistant path and connected to external electronics. This development [10] had been considered impossible by many semiconductor detector experts at the time. The high-resistivity material used almost uniquely by detector people was supposedly incompatible with the high-temperature processing required for the activation stage of ion implanted material. Kemmer showed that these experts were incorrect; it was problems of cleanliness in processing, rather than the high temperatures themselves, which led to the dreaded resistivity drops. The first result of this revolution was *more* robust detectors and hence the possibility of much larger areas. As important, the door was opened for the inclusion of a host of features already developed for IC's, such as techniques for isolating edge-related leakage currents (guard rings), for biasing with high dynamic resistance, and so on. Some of these will be discussed in Sec. 4.3.

The second revolution was the development of readout chips with high-density front-end amplifiers [11, 12]. Using integrated circuit technology, the front-end could be shrunk to a pitch of 50 μm, permitting the microstrip channels to be wire bonded directly to these compact IC's located along the edge of the detector. Furthermore, the readout chips embodied resettable storage of the analogue signals, and multiplexed readout. Thus, the number of cables needed for the detector readout was reduced by about a factor of 100. We shall in Sec. 4.3.3 record great ongoing progress in developing special readout IC's to suit a wide range of experimental conditions.

The combination of robust, sophisticated microstrip detectors and extremely compact electronics has led to their application in a host of experiments. With the SLC, Mark II, and LEP detectors, they crossed the barrier from fixed-target experiments into the collider environment, with excellent results in heavy flavor physics.

4.2 The Generic Microstrip Detector

Microstrip detectors come in a large variety of designs, each with its own strengths and weaknesses, each with a certain range of applications.

Due to the fact that high resistivity *n*-type material is more readily available, most detectors have used *n*-type wafers as starting material, though this may be changing in some application areas. The 111 crystal-orientation is conventionally used, but reasons why this too may be changing are discussed in the next section. As already mentioned, the pioneering detectors all used p^+ strips, collecting holes from the track of the ionizing particle. More recently, the back surface (n^+ implant) has also been subdivided into strips (which can as well be angled, perhaps at 90° to the *p* strips) giving us double-sided microstrip detectors.

Such a detector, and the associated internal electric field, is sketched in Fig. 25. The reverse bias is achieved by applying a positive voltage to the *n* strips, the *p* strips being grounded. In each case, series resistors (usually on-chip polysilicon) are used to create a high impedance path. The electric field (directed in the negative *Z* direction) would be uniform across the depleted n^- substrate, were it not for the finite resistivity and hence the presence of a low density of fixed positive charges. Due to this space charge, the magnitude of the field falls steadily from its peak value at the *pn* junction, towards the *n* side. The sketch shows an overdepleted detector. For the just-depleted case, the field would sink to zero at the surface of the *n* strips. Once we enter the heavily-doped *p*- or *n*-strip region, the field develops a large gradient, falling abruptly to zero.

The sketch indicates an AC coupled detector. The metal readout strips are isolated from the implanted strips by a thin layer of dielectric (silicon dioxide). Thus, the amplifier inputs sense the fast signal without also being obliged to sink the DC leakage current. Both AC and DC coupled microstrip detectors are common. In applications where radiation levels are low, and hence degradation in leakage current

is not a problem, the extra simplicity of DC coupled detectors may be advantageous. Early microstrip detectors were all DC coupled.

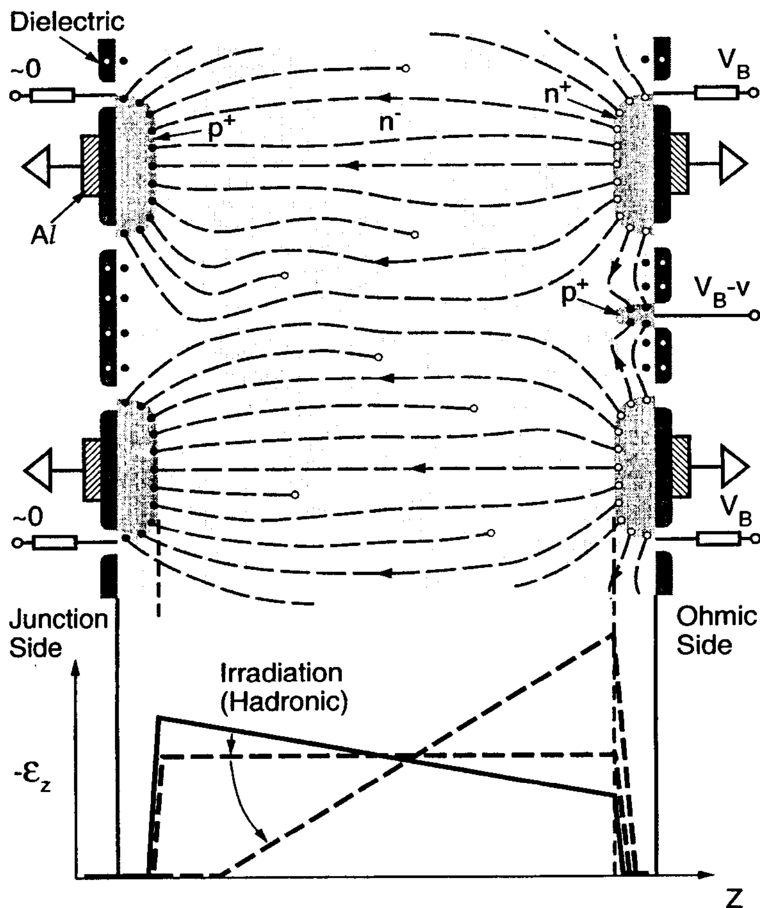


Fig. 25 Sketch of a cross section of a generic double-sided microstrip detector. Exposed fixed charges are shown by open circles (positive) and filled circles (negative). Also shown is the electric fields distribution in such a detector before and after radiation-induced displacement damage in the silicon.

Between neighboring charge collection strips on both sides is a passivation layer of silicon dioxide. Such oxide layers inevitably collect some positive charge (holes trapped as interface states) which is compensated by a very thin accumulation layer of mobile electrons in the bulk material. On the *p* side, these are repelled by the exposed negatively charged dopant atoms in the *p* strips. However, on the *n* side, they create a low-resistance interstrip leakage path. Signal electrons collected on one *n* strip will readily flow to neighboring strips; the strips are effectively shorted together. This problem can be overcome in a number of ways; Fig. 25 shows one of the cleanest solutions which is drawn straight from the textbooks of IC design. p^+ "channel stops" are implanted between the *n* strips. They are biased somewhat negatively relative to the strips, and hence, acquire a negatively charged depletion layer which repels the mobile electrons in the surface accumulation layer, so blocking the leakage path that would otherwise be present.

Before leaving this figure, there is one further point worthy of note, relating to the collection of signal charge. After the passage of an ionizing particle, holes begin to drift to the left, electrons to the right. Once the charges separate, the space-charge self-repulsion in principle leads to expansion of the charge cloud during the drift time. A localized charge distribution of N carriers (holes or electrons) will expand with time to a sphere of radius r_s, where

$$r_s = \left[4.5 \times 10^{-7} \frac{\mu N}{\varepsilon_s} t_d \right]^{1/3} \text{cm}.$$

ε_s is the permittivity of silicon, and t_d is the drift time in seconds. For collection of holes or electrons in a microstrip detector, r_s amounts to less than 1 μm and can be neglected (while the signal from an α particle can expand to $r_s \approx 10\ \mu$m; see Ref. [9]). As we saw in Sec. 3.3, diffusive charge spreading can, on the other hand, be considerable. This is sensitively dependent on the type of charge carrier collected, on the detector resistivity, and on the biasing conditions.

For the conditions shown in Fig. 25, a strongly overdepleted detector, the electric field is reasonably uniform. For a just-depleted detector, the *holes* would all pass through the high-field region close to the *pn* junction, and those generated in that half of the detector would be entirely drifted through a fairly high field. For the *electrons*, on the contrary, all would pass through the low field region before reaching the *n* strips. Hence (even without the effect of the relative mobilities), the electron cloud will experience greater diffusive charge spreading than the hole cloud. In principle, this would give us higher precision (by centroid fitting) on the *n* side than on the *p* side. This question is discussed in more detail in the next section.

There are, however, several reasons why such fine tuning of detector parameters may not yield the desired improvement in precision.

Firstly, in a radiation environment, the effective dopant concentration varies with time. As depicted in Fig. 25 and discussed in detail in Sec. 6, hadronic irradiation causes the depleted material to become steadily more *p* type. Having passed through the compensated condition (when it could be depleted with a few volts), the resistivity falls steadily. After a certain dose (for fixed operating voltage), the detector would fail to deplete fully and the hole signal would be lost (no longer collected on an individual *p* strip). The electron signal would still be collected, but from a steadily decreasing thickness of detector. Thus, any precision advantage gained by fine tuning the depletion conditions could not be preserved through the life of the detector.

Secondly, due to their thickness, microstrip detectors have a significant probability of loss of precision due to δ-electrons, as discussed in Sec. 2.3. Results published from test beams often limit the signal charge to less than approximately 1.7 times the MIP mean value, in order to restrict the tails on the coordinate residuals. In tracking detectors with a limited number of points per track, one would not normally have the luxury of such a filter. For binary readout detectors, one would not even know which were the large signal clusters.

Thirdly, detector precision is seriously degraded for angled tracks, as we shall see in detail in the next section.

Finally, most tracking detectors in experiments operate in a magnetic field which (because of the Lorentz angle) degrades the measurement precision. In a conventional collider geometry with a solenoid magnet, the Z measurements are unaffected but the precision of the $R\phi$ measurement is degraded. For details, see the next section.

4.3 Microstrip Detectors: Detailed Issues

4.3.1 Design Optimization

All silicon microstrip detectors are of approximately 300 μm thickness. For much thinner detectors, the loss of signal charge, exacerbated by the reduction in signal voltage due to the increased capacitance from strip to substrate, results in a poor signal-to-noise performance. Even thicker detectors might be required, for example, in cases of modules having several long strips linked together and to a single readout chip. The capacitance to substrate is a particularly important issue in cases where capacitive charge division is used for the readout of floating strips. To avoid serious signal loss, it is essential that the geometry be chosen so that the interstrip capacitance greatly exceeds the strip-to-substrate capacitance, or one would suffer from serious loss of signal from floating strips. In some large systems currently under design (e.g., the ATLAS Silicon Central Tracker or SCT), the individual modules are 12 cm in length, with strip capacitances of around 18 pF (1–2 pF/cm is typical). Such large capacitances represent a considerable challenge for readout electronics, as we shall see in Sec. 4.3.3.

As already mentioned, a high-resistivity *n*-type substrate is conventionally used. High-resistivity *p*-type material is now available (both bulk and epitaxial), providing an interesting option for detector fabrication. Such detectors would have the advantage that under irradiation, they simply become steadily more *p* type. Thus, one would avoid the complications (e.g., in guard-ring structures) associated with the junction shifting over from the *p* side to the *n* side during the life of the detector.

The 111 crystal orientation is conventionally used in microstrip detectors, since it provides the densest surface, and hence the lowest probability of "spiking" (growth of aluminum deeply into the crystal in local regions, possibly shorting out the diode

structure). For IC manufacture (and also for MOS detector types such as CCD's), the 100 crystal orientation is preferred due to the lower density of dangling bonds at the silicon/silicon dioxide surface, and hence lower trapped charge at the interface. This may be particularly important in some microstrip detector applications, and for this reason some groups are doing exploratory work with 100 material. For AC coupled detectors, the area of metal in contact with silicon is reduced by many orders of magnitude compared to the early DC coupled devices. Also, metallization equipment is now extremely refined compared to 10 years ago, so the problem of spiking should be largely in the past.

For biasing microstrip detectors, the most commonly used method (also the simplest) is via on-chip polysilicon resistors. A problem with this approach is that as one has to allow for higher leakage current (due to radiation damage and/or longer strips), the resistance value needs to be reduced in order not to disturb the bias voltage excessively. This in turn can lead to loss of signal and worsening signal-to-noise ratio. The ideal solution would be a low DC resistance and a high dynamic resistance. Two approaches have been adopted, the reach-through structure [13] and the FOXFET biasing scheme [14]. This Field OXide FET structure, which employs a thick gate oxide, is vulnerable to radiation damage effects [15, 16]. The present situation appears to be that polysilicon biasing is the only safe solution for detectors to be used in a high-radiation environment.

For the *n*-strip isolation in detectors (one- or two-dimensional) where the electron signal is collected, two methods have been adopted. The channel stop approach [13] has been illustrated in Fig. 25. An alternative "field plate" method uses an MOS gate structure, in the form of "wings" attached to the aluminum readout strips in AC coupled detectors [17]. This is illustrated in Fig. 26.

For all these various microstrip detector structures, careful attention should be paid (by two-dimensional simulation) to the peak electric fields induced near the strip edges. Poorly understood leakage current has characterized many of the designs which at first glance looked quite reasonable. In a pioneering paper [18], Ohsugi and coauthors demonstrated the sensitivity to geometrical details in the specific case of AC coupled *p*-strip sensors. Breakdown was demonstrated in structures where the

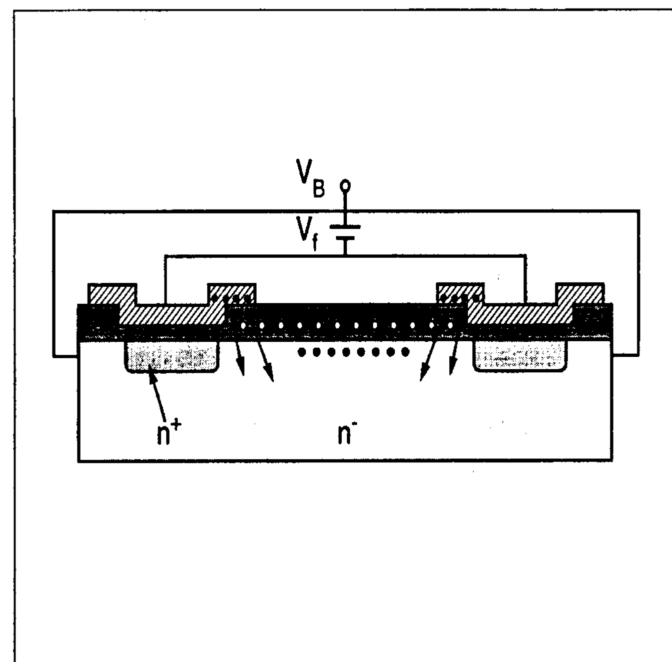


Fig. 26. The technique of *n*-stop isolation by field plate separation with extended AC coupled electrodes (one of several field plate approaches).

relative edges of the p^+ implant and the aluminium electrode led to peak fields at the edge of the implanted strips exceeding the breakdown field in silicon of 30 V/μm. While such problems can in principle be avoided by careful design, it is very easy to encounter some local variations, edge effects at the ends of the strips, etc., which can still cause problems. To this end, the diagnostic tool demonstrated in their paper is of enormous value. Using an infrared microscope equipped with a CCD camera, very small regions of avalanche breakdown can be seen clearly. This marvellous tool [19] is of value wherever anomalous leakage currents are encountered either due to design deficiencies or to process faults. One of the problems that has plagued manufacturers of large area microstrip detectors, particularly in the case of double-metal structures (see below), is that of pinholes in the dielectric, permitting unwanted leakage paths. An infrared microscope can be used to explore the positions of these defects, and possibly to suggest solutions (e.g., improved step coverage across gate edges). Similar problems have been encountered and solved in this way in the world of CCD detectors. For n-strip microstrip detectors, there is evidence (not surprisingly) that field plate devices are more susceptible to microdischarges than p-stop devices. However, much depends on the specific design details.

It is hardly surprising that another issue which still causes problems in microstrip detector design is that of uncontrolled oxide layers (e.g., interstrip, as depicted in sketch form in Fig. 25). In other detector types such as CCD's, care is taken to avoid even fine cracks between gate electrodes (by overlapping neighboring electrodes) since gate oxide inevitably contains trapped interface charge, the magnitude of which increases with irradiation. The electrical effects of such trapped charge can be minimized by the presence of a metal or polysilicon cover layer held at a well-defined potential. Microstrip detectors do not easily lend themselves to such design rules, but one may escape from trouble due to the accumulation layer of electrons already referred to. However, particularly if one is aiming for high efficiency for detection of (say) soft X-rays which deposit their signal near the surface, there are numerous examples of anomalous dead layers and other effects probably related to the uncontrolled oxide. This is an area for ongoing concern regarding the design of microstrip detectors.

The use of high resistivity silicon is driven by the desire to have a manageable operating voltage for full depletion; 150 V is commonly considered an upper limit. Under intense hadronic irradiation, this may set an uncomfortably short lifetime for the detector. It has been pointed out [20] that careful design of microstrip detectors (particularly as regards implant profiles, strip edges, guard rings, etc.) may enable operating voltages to be set even above 1 kV before microdischarges or breakdowns occur. Such a design would considerably extend the usefulness of microstrip detectors in high-radiation environments. Note that it is usually the breakdown voltage rather than the leakage current which shortens the lifetime of a detector in a radiation environment. The leakage current can always be reduced by cooling. There is long experience of this in the area of CCD detectors, and large systems of cooled microstrip detectors are now in the planning stages [21].

We have discussed briefly the availability of double-sided detectors, which are of interest in that they provide apparently two advantages over (for example) a pair of single-sided detectors: firstly, less material (of particular significance for vertex detectors), and secondly, some degree of resolution of the ambiguity problem for multihit events. Regarding the latter, the idea is that one can measure the signal charges in the p- and n-side clusters and use the correlation between them to rule out some of the associations (e.g., between a below-average cluster in one view and a multi-MIP cluster in the other view). In fact, this is not a very practicable idea, since the level of ambiguity is not greatly reduced.

Regarding extraneous material in the active volume, much depends on the angle between the strips on the two sides. If this is small (e.g., a few degrees), both sides of the module can be read out from the end without complications. If, however, one requires a large angle between the two strip planes (e.g., 90°), there are two options. Consider the case of a Z view as well as the conventional $R\phi$ view in a collider environment. The most obvious option, implemented in the pioneering double-sided ALEPH vertex detector [22], would be to place the Z readout chips along the long edge of the module. This results in a large amount of electronics in the active volume of the barrel detector system, which is not a good idea if precision vertexing is the goal. Later detectors have followed one of two different approaches. Both move the Z readout chips to the ends of the barrel, outside the active volume, in the same

general area as the $R\phi$ readout chips. The most ambitious approach is to integrate the linking traces onto the detector modules themselves, using a double-metal technology [23, 24]. A dielectric layer separates the Z-strips from the orthogonal readout strips, and metallized vias provide the connections between the two levels. Due to the larger number of Z-strips than readout strips in a typical module (a long rectangle), the Z-strips may be connected in a repeating pattern, resulting in some degree of ambiguity as to the spatial position (normally not a problem given the overall track-finding software). Alternatively, the Z-strip pitch may be made correspondingly coarser than the pitch of the $R\phi$ readout strips. There is one inevitable disadvantage to the double-metal approach, which is the increased capacitance of every strip; the detector strips and readout strips form a web of closely linked electrodes, separated only by the thickness of the dielectric layer. This, coupled with the fact that tracks at the ends of the polar angle range may deposit their charge over a number of Z-strips, can lead to a serious degradation in the signal-to-noise in the detector. The capacitance problem can be greatly reduced, with only a modest degradation in terms of material in the active volume, by routing the readout traces on separate thin substrates (e.g., copper traces on kapton) [25]. The Z-strips are wire bonded to diagonal readout strips at the edge of the detector, the signals being carried to the electronics in a zig-zag geometry, using additional Z-strips to link the diagonal readout strips. This idea is illustrated in Fig. 27. In this way, a low and acceptable ambiguity level as to which of a few widely separated strips was hit, is the price paid for accessing the data in an economical form with little additional material, and a generally acceptable overhead in capacitance.

There remains the choice between double-sided detectors and two back-to-back single-sided detectors, one for $R\phi$ and one for Z. As has been noted, the correlated cluster signal information is not often very useful, so the key issue is that of the additional material in the back-to-back approach. In vertexing applications, this is always important, though seldom decisive. There is necessarily additional material in the form of support structures, etc., so we are certainly not talking about a factor of two, and the multiple scattering is proportional to the square root of the thickness. If the double-sided option came free of additional costs, it would clearly be preferred. However, this is far from the case. Double-polished silicon wafers are available and

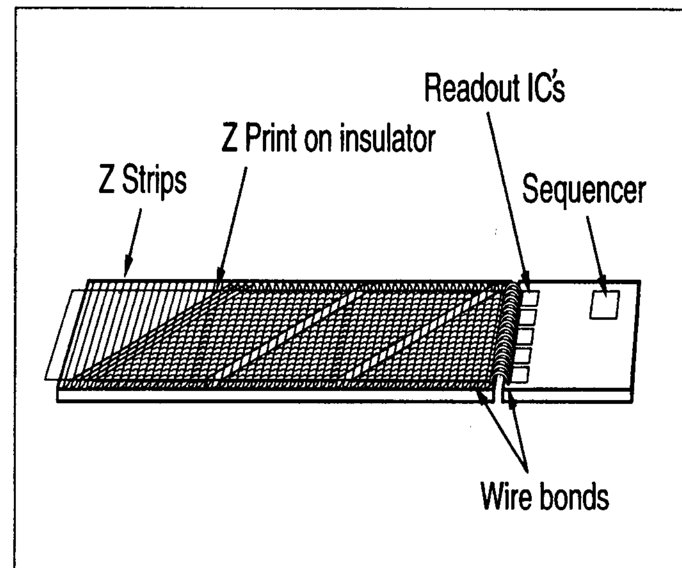


Fig. 27. A scheme for Z-strip readout using a separate metallized substrate (glass or kapton).

are not in themselves particularly expensive. However, for bulk production, it is desirable to use as far as possible the standard IC manufacturing equipment, which is all explicitly geared to single-sided processing. It has been claimed that the cost of double-sided relative to single-sided detectors is 3:1. This may be true for some small production runs, where it merely reflects the reduced yield of the double-sided devices. However, for large-volume production such as we are now seeing planned (e.g., for the LHC SCT's), it should be possible to greatly reduce the cost per unit area of detectors made with standard processing equipment. In this case, the cost ratio mentioned above is likely to become much more unfavorable. Time will tell.

4.3.2 Spatial Precision in Microstrip Detectors

Early microstrip detectors with very fine readout pitch (and huge fanout factors) had wonderful spatial precision but are now only of historical interest. We are at present effectively constrained by the readout pitch of all existing front-end electronics, namely 50 μm. This can be reduced by a factor of two by attaching readout IC's at each end of a module, and this has been done in environments of high track density. Also, one can include floating strips as has already been discussed. Spatial precision of approximately $\frac{25}{\sqrt{12}}\mu m = 7.2\mu m$ is thus in some ways natural for a silicon microstrip detector when read out with currently standard electronics. In large tracking systems, one has frequently to work very hard to achieve such levels of stability and systematic precision, for many reasons. Having said this, considerably better spatial precision has been achieved, mostly in test- beam situations.

Let us consider first the case of normal incidence tracks. As we saw in Sec. 4.2, the extra diffusive spreading would suggest that (for a given strip pitch) one might be able to achieve a higher precision in the charge collection on the n side (electrons) as opposed to the p side (holes). However, most experimental results to date have been obtained with detectors made with p strips on n-bulk silicon.

Using a single-sided detector with p strips on a 20 μm pitch and analogue readout on every strip, Belau *et al.* [26] were able to measure the *spatial distribution* of the hole charge collected. This varied from σ=2.5 μm to 1.9 μm as the operating voltage was raised from 120 V (just-depleted) to 200 V (overdepleted). From this, they *calculated* the precision achievable with a readout pitch of 20, 60, and 120 μm to be σ = 2.8, 3.6, and 5.9 μm, in the optimal case of the just-depleted detector. Measurements with 60 and 120 μm readout pitch [27] yielded precisions of 4.5 and 7.9 μm, a little worse than calculated, but better than $\frac{20}{\sqrt{12}}$ = 5.8 mm which would be the limit for a digital system with 20 μm readout pitch. Evidently, some degree of useful charge spreading is achieved with detectors having narrow strip pitch. For electron collection, the lower average electric field yields even better *calculated* precision, 0.8 μm to 3.6 μm, for the three cases mentioned above. In this case, they did not have data for comparison.

In all this, please remember the caveat about δ-electrons mentioned in Sec. 2.3. In these test beam studies, clusters with more than 1.7 times the mean MIP signal were discarded, with the consequential efficiency loss that could probably not be tolerated in a detector used for a physics experiment.

Results with a more typical arrangement of readout of every strip on a pitch of 50 μm have been reported for double-sided detectors [28]. For normal incidence, the precision achieved was 8.8 μm *on the p side*. This slightly worse figure is attributed to the higher electronic noise in that system. The signal-to-noise was 16 μm for the p side and degraded (for not completely clear reasons) to ten for the n side. The precision for the n-side signal was 11.6 μm, confirming the suggestion that the system noise played a large part in the measured spatial precision.

For normal incidence tracks, we may conclude that spatial precision in the region 5–10 μm is typical for strip pitch ≤ 50 μm, and with readout pitch ≤ 150 μm. The degradation in precision with increasing readout pitch is fairly modest. The usual reason for requiring a fine readout pitch (typically, equal to the strip pitch) is the need to preserve an optimal two-track resolution.

Once we permit angled tracks (which really only are of concern for the RZ view as opposed to the $R\phi$ view in colliders), the situation deteriorates fairly rapidly. The particle leaves a trail of charge carriers which are collected on a number of Z strips.

Taking the overall centroid is a bad approximation to the track position at the center plane of the detector, due to the energy-loss fluctuations along the track. The problem has been studied theoretically [29] and experimentally [30], as a result of which Hanai *et al.* have developed an algorithm ("convoluted Gaussian centroid") which leads to an experimental precision as a function of α, the track angle to the detector normal, varying from 8 μm ($\alpha = 0°$) to 40 μm ($\alpha = 75°$). These results were obtained using a single-sided *p*-strip detector with 25 μm strip pitch and 50 μm readout pitch.

A dangerous factor affecting spatial precision in microstrip detectors is the effect of magnetic fields. Empirical measurements have been reported in Ref. [26]; these agree well with calculations. For the *p*-strip signal in a just-depleted detector, a magnetic field of 1.7 T shifts the measured co-ordinate by about 10 μm and increases the width of the collected charge distribution from 5 to 12 μm. The relevant parameter determining these effects is the Hall mobility μ_n^H for electrons and μ_p^H for holes; see Shockley [7]. With the usual arrangement in collider barrel detectors (magnetic field $\mathcal{H}$ perpendicular to electric field), the charges drift at the Lorentz angle θ^L with respect to the electric field, where θ^L is almost independent of the magnitude of the electric field and is given by

$$\tan\theta^L_{n,p} = \mu^H_{n,p} \times \mathcal{H}.$$

Now

$$\mu_p^H \approx 310 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$$

and

$$\mu_n^H = 1650 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}.$$

For a typical case of a magnetic field of 1.5 T and a 300 μm thick detector, the charge distribution of the holes shifts by ≈ 7 μm, while that for the electrons shifts by ≈ 37 μm [31]. Thus, collection of the *electron* signal in future collider experiments is liable to serious systematic effects, unless the *n*-strips are oriented at least approximately along the direction of drift induced by the magnetic field (the $R\phi$ direction in a barrel detector).

Finally, a reminder that in any silicon detector of thickness approximately 300 μm, the production of δ-electrons of significant range is quite a common occurrence, so the residual distributions will inevitably have a significant non-Gaussian tail, unless steps are taken to exclude large-signal clusters, with the attendant loss of efficiency.

4.3.3 Electronics for Microstrip Detectors

We have seen that silicon microstrip detectors have developed and diversified to an extraordinary degree, due partly to the ingenuity of those involved, and partly to the tools and devices provided for them by the integrated circuit industry. As regards the readout electronics, the progress has been at least as spectacular, for the same two reasons. The current picture is in fact one of somewhat bewildering complexity, since the diversity of options is so great. Part of this diversity reflects the variable detector applications, but even for one single application (e.g., the ATLAS SCT), there is not yet unanimity among the experts as to the optimal approach. The issues are quite subtle and the boundary conditions keep shifting. In this section, we shall aim to take a general look at the principles leading to these various options and make some remarks about the relevant areas of application. What is clear, however, is that the ASIC designer now has enormous power and flexibility at his disposal, so that a new application area is likely to lead to the very rapid evolution of one or more new readout schemes full of wonderful ideas to handle the peculiarities of that particular application.

Even from the very beginning of the ASIC initiative which opened the door for silicon microstrip detectors to find a home in collider detectors, there was not a unanimous approach. At that time, there *was* unanimity at the level of the functional requirements (amplifier, sample-and-hold, multiplexed analogue output) but two technological solutions; nMOS [12] and CMOS [11] were pushed by different groups. In the event, the "low and slow" CMOS solution proved superior, largely due to the much lower power dissipation (around 2 mW per channel compared with ten times that for nMOS). This pioneering CMOS chip, the first of a family of CAMEX chips, was joined by others, of which the most commonly used are the MX

(3-7) (Ref. [32]), SVX (1-3) (Ref. [33]), and AMPLEX [34] families. More recently, a bipolar chip for the front-end electronics has made its appearance [35].

Why is the user of silicon microstrip detectors faced with such a large array of readout options? Some part of the reason is sociocultural. There never was a "standard" drift chamber preamplifier; different laboratories like to do their own thing, and this competition is extremely healthy in encouraging new ideas. But mostly, these various approaches have been driven by the need to equip detectors to work in increasingly varied and hostile conditions. Beam-crossing intervals at SLC (8 ms) and LEP Phase 1 (22 μs) allowed very relaxed shaping times of 1 or 2 μs. The detector modules were small (strip lengths ≤ 6 cm) and the radiation environment almost nonexistent. Under these benign conditions, the ASIC designers were able to achieve spectacularly good signal-to-noise from a variety of single- and double-sided detectors. Moving to the Tevatron (originally 3.5 μs, upgrading to 396 ns and eventually 132 ns), HERA (96 ns), and in the future, the SLAC and KEK *B* factories (4 ns), and LHC (25 ns) represents a phenomenal challenge. Compounded with the escalating beam-crossing rate is the need to increase the module sizes (strip lengths of 12 cm will be used in the large ATLAS SCT, for example), plus the fact that the detectors at all hadron machines will encounter significant, if not fatal, radiation damage. Some relief is provided by cooling the detectors to reduce leakage current, but for the most part, it has been up to the chip designers to get the physicists out of a very uncomfortable situation. This is a rapidly evolving story, and it is far from clear where we shall end up. In the case of the LHC detectors, several critical decisions have to be taken over the next year, and these will be based on the results of much hard work going on in design labs and in test beams. Let us review in very general terms the main approaches, all of which are certainly appropriate, for some applications.

Firstly, the generic analogue chip comprises typically 128 channels, one of which is shown in its essentials in Fig. 28. The amplifier/shaper may include a CR-RC filter. It has been shown [36] that more sophisticated filtering schemes do not lead to a major improvement in noise performance. On receipt of a trigger, the signals are sampled and stored on capacitors C_S, which are read out (sequentially for each channel on the chip) via the analogue output, for remote digitization. Such a readout chip minimizes the logic local to the detector (and hence, is optimal from the viewpoint of power dissipation, which is usually an important issue), but it cannot be used in high-rate environments where even the first-level trigger appears after several beam crossings. The most obvious response to this situation is firstly to reduce the shaping time so as to retain an analogue signal which is unambiguously associated with its beam crossing. However, this causes inevitably a penalty in noise performance and may not be necessary. Given the sparsity of the tracks in the detector, each strip has a low probability of being hit on successive beam crossings. Then one may retain a longer shaping time and use a filtering approach [37] to recover the fast timing information by deconvoluting the sampled voltages of a shaped pulse, to retrieve the original impulse signal with high precision. This ingenious approach may extend the range of applicability of CMOS front-end electronics into the realm of LHC operating conditions, and has been adopted by the CMS Collaboration. Their analogue signal (50 ns shaping time) is sampled at the beam crossing rate of 40 MHz. The samples are stored in an analogue pipeline of 128 cells, and if a positive level-1 trigger signal is received, are deconvoluted by the analogue signal processor. All this happens in parallel for each channel.

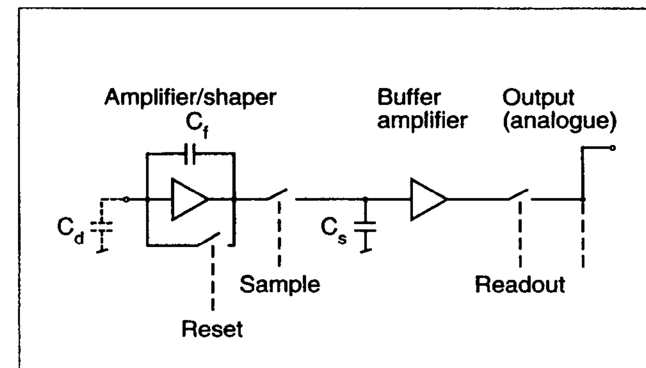


Fig. 28. Block diagram of one channel of a typical analogue readout chip.

The stored signals are read out at leisure via a multiplexer, connected off-chip to an electro-optical modulator. This consists of a multiquantum-well device which amounts essentially to a mirror of voltage-controlled reflectivity. Consuming almost no power, this device permits a change of reflectivity from 30% to 60% by changing the voltage across an InP/InGaAs sandwich [38]. The device is connected to an optical fiber, at the remote end of which is the drive laser, receiver module, flash ADC, and event builder memory. The beauty of such links is that they permit very high-speed data transmission with almost no power dissipation at the detector end. Used (as here) in analogue mode, they permit seven-bit resolution which is entirely adequate for microstrip detector applications.

The SVX family of readout chips has pioneered the digital approach. An example is shown in Fig. 29. Analogue signals are again put into a pipeline (one per channel). On receipt of a level-1 trigger, the relevant signal is transferred to a storage capacitor which serves as one input to a comparator used as a Wilkinson ADC circuit. The other comparator input is ramped at a fixed rate, and the time to reach equality of input is stored digitally as a measure of the signal amplitude. The digital data are then read out via a multiplexer.

Finally, we consider the bipolar option. Bipolar IC technology has been making great strides in recent years, and it has become possible to shrink amplifiers down to a pitch of 50 μm, as has been true for some time with CMOS systems. As a result, stray capacitances have been greatly reduced, and furthermore, very small transistors can be made with high bandwidth and low current. In short, the power dissipation has dropped to an extremely competitive level. At hadron colliders, even with cooled detectors, the problem of leakage current in long-strip modules after a few years of radiation damage will be considerable. The shot noise associated with the leakage current tends to favor short shaping times as opposed to the longer shaping time with deconvolution mentioned previously. The lower limit on the useful amplifier shaping time is given by the charge collection time of typically 20 ns. Below that, one encounters increasingly severe signal loss (the *ballistic deficit* effect). The superior transconductance of the bipolar transistor compared with CMOS (even if run in the weak inversion mode) suggests that to achieve adequate signal-to-noise performance

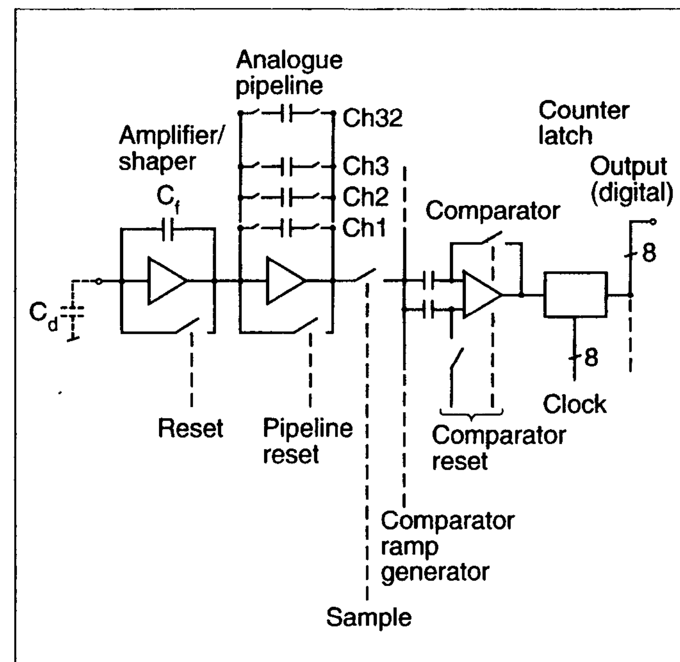


Fig. 29. Block diagram of one channel of a typical digital readout chip, of the SVX type.

for long microstrip modules in fast readout conditions, the bipolar option may be superior.

A disadvantage (possibly minor) of the bipolar approach is that (due to the near non-availability of rad-hard bi-CMOS) one necessarily has an analogue chip followed by a digital CMOS readout chip. Doubling the number of wire bonds in the system is not a major overhead, and there are advantages. For LHC applications, the size of the digital processing chip is such that yield is a significant consideration. Having the analogue front-end as a separate chip may be more economical overall.

This bipolar/CMOS combination has been used with excellent performance in the demanding environment of the ZEUS Leading Proton Spectrometer (LPS) at HERA [39, 40]. The basic system (Fig. 30) consists of a bipolar amplifier/comparator chip with 20 ns risetime, followed by a low-power digital pipeline. Not only does the front-end break with tradition in microstrip readout systems, but so does the digital system. The designers have adopted the simple "binary" approach of recording only the addresses of above-threshold strips, not the pulse heights. In fact, their system (which has been carefully designed to minimize common-mode noise) operates extremely stably with a constant threshold of 0.78 fC set for all channels.

Lack of pulse height information, of course, limits the spatial precision to $\frac{p}{\sqrt{12}}$, where p is the strip pitch, but as we have seen, this precision is in any case close to the limit achieved in nearly all systems. Furthermore, it is only in small radius vertex detectors that there are major physics advantages in pushing the point measurement precision to the highest achievable value.

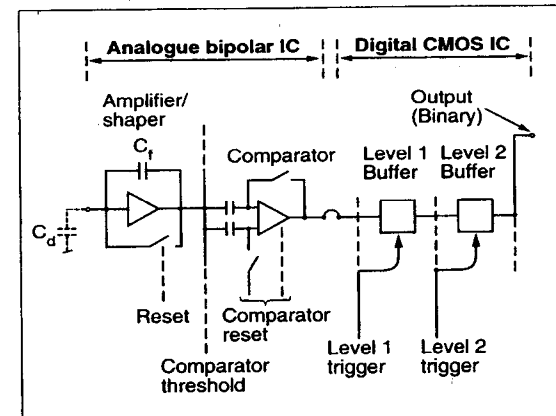


Fig. 30. Block diagram of an FEE system based on a bipolar analogue chip followed by a digital readout chip (binary readout).

The readout system takes advantage of the hierarchical trigger structure of ZEUS, which will also be followed in LHC. In the ZEUS application, they use a synchronous level-1 buffer of about 6 μs followed by an asynchronous level-2 buffer. Data are thus stored on-chip until a valid level-2 trigger arrives after about 1 ms.

All these considerations of readout options are complicated by another question, that of radiation damage. The move to hadron colliders has stimulated a major effort to develop rad-hard versions of the local detector electronics.

In the case of CMOS, a number of companies (Harris, UTMC, Honeywell, and DMILL) are involved with the chip designers already mentioned. For example, a 100 Mrad-hard version of the MX7 chip in 1.2 μm CMOS exists. These chips tend to somewhat exceed the 50 μm channel width, but for applications such as the LHC SCT's, this is acceptable. One cloud on the horizon is that, with the downturn in military spending, there is less funding for development of rad-hard electronics. However, as the industry moves into submicron processing, the devices have improved radiation resistance as a by-product (thinner oxide, etc.), so the trend may be to add a few steps to achieve adequate performance from a process not specifically developed for optimal radiation hardness.

For the bipolar IC's, the radiation damage situation is more favorable, due to the lack of sensitivity to oxide charge. The main cause of deterioration is bulk damage, which results in a reduction of the current gain β at high doses. Typically, an *npn* transistor suffers a β degradation of approximately a factor of two after 5 Mrads. The circuit designer can allow for such degradation, which makes these IC's usable at all but the smallest radii needed for vertex detectors at LHC. This region (as we shall see) is territory almost certainly out of bounds for silicon strip detectors due to the radiation damage in the detectors themselves.

Very recently, one company, DMILL (LETI), has produced some bi-CMOS chips using a rad-hard process. Whether they will find a sufficient market to sustain this initiative, and if so, whether these will offer a way to the future for HEP detectors, remains to be seen. At least for the time being, the combination of bipolar chips with rad-hard CMOS digital chips appears to be the safest means to satisfy our requirements.

Thus, in conclusion, both the CMOS and bipolar IC's we have discussed can, it appears, be designed to tolerate the worst radiation conditions likely to be encountered by silicon microstrip detectors. The inevitable noise degradation due to growth of leakage current in the detectors, plus other detector-related issues to be discussed in Sec. 6, are what finally limit the scope for these detectors. There is no possible cure in the electronics for these deficiencies, once they reach an unacceptable level in the detectors.

4.4 Physics Performance and Future Trends

Silicon microstrip detectors were originally developed as vertex detectors for charm physics at fixed-target experiments. Here, with the benefit of the high track momenta, they were able to achieve excellent impact parameter precision, and hence, clean reconstruction of a wide range of charm particle decays.

The move to e^+e^- colliders (initially SLC and LEP) called for much larger detector areas (and here the detector manufacturers were well able to oblige) and much more compact electronics (and, as we have seen, the ASIC designers solved these problems for us). Nevertheless, the physics capabilities of the detectors took a step backwards. Due to the lower particle momenta and the large radius beampipe (initially 10 cm at LEP, eventually reduced to 5.5 cm), the impact-parameter precision for tracks in hadronic jets was relatively poor. Nonspecialists were at first understandably ignorant of the situation, because all groups were proudly demonstrating beautiful miss-distance plots based on muon pairs. The situation for tracks in jets was, of course, much worse. The Holy Grail for vertex detectors is to present to the experimentalist a clear topology of the event, with high efficiency for associating all tracks uniquely with their true parent vertices. Fortunately for the detector builders, there is a host of lesser objectives which are still extremely useful for physics. The long lifetime of beauty particles means that b tagging is relatively straightforward. The cleanliness of the $\tau^+\tau^-$ signal at the Z^0 means that lifetime measurements (though imprecise at the individual event level) can be made accurately, given high event samples. Areas such as charm tagging and the separation between charged and neutral B states are much more challenging.

The one- and two-dimensional microstrip vertex detector systems at LEP have covered the range of radii typically 60 to 110 mm, and (in their finally upgraded forms) delivered precision in impact parameters as a function of momentum p GeV/c of:

$$\sigma^b_{XY} \approx 20 \oplus \frac{80}{p\sin^{3/2}\theta}\ \mu\text{m}$$

and

$$\sigma^b_{RZ} \approx 30 \oplus \frac{80}{p\sin^{3/2}\theta}\ \mu\text{m}.$$

With an average track momentum of about 5 GeV/c, this implies a mean impact parameter precision for normal incidence ($\theta = 90^\circ$) of around 30 μm. For reasonable topological vertexing (including charm), one would like to do five to ten times better than this. Another problem for the extraction of physics with microstrip detectors is that of poorly understood tails on residual distributions. These are presumably due to a combination of factors such as energy loss fluctuations, δ-electrons, cluster merging (by no means negligible in the core of jets), and so on. The general approach has been to artificially broaden the Monte Carlo residual

distributions to match the data. This is only a partial solution since it ignores the correlations that are surely present (e.g., a pair of tracks closely spaced in one view, giving poor coordinates on *all* planes due to partial merging of clusters).

The overall picture is one of impressively high efficiency and purity for b tagging, with flagging performance in the more challenging areas. For LEP2, the b-tagging requirement is considered to be of paramount importance (Higgs and SUSY searches). To do better as regards topological vertexing at the Z^0 would have required a smaller beampipe, giving a shorter extrapolation length to the interaction point (IP), and hence better impact parameter precision. But then, the track merging problem on the inner barrel would have been more severe. In any case, the time for such discussions is past.

The pioneering silicon microstrip vertex detector at hadron colliders has been the SVX family (same name as their readout chips) at the Tevatron. SVX1, the original detector, played a crucial role in the discovery of the top quark, again by performing the relatively simple task of b tagging. This is the first major discovery in particle physics in which a silicon vertex detector has been essential in achieving a convincing signal, and I am sure it will not be the last. After all the technical complications we have been discussing, it is somewhat comforting to note that the detector used for the top discovery was a single-sided, DC coupled, low signal-to-noise, radiation-soft detector. Such a detector would never survive the conditions after the Tevatron upgrades, and this vertex detector has already been upgraded at least once.

New microstrip vertex detectors are on the way. CLEO II has one (on a small-radius beam pipe, necessarily tackling the more challenging requirements of charm decay), and the SLAC and KEK B factories are building them, primarily to measure the longitudinal position of the B and $\overline{B}$ decay points with respect to the IP.

The ZEUS LPS set the trend for microstrip detectors to be used as momentum spectrometers in high-intensity conditions in which wire chambers could not survive. This trend continues with the D0 silicon tracker ($\approx 5\ m^2$) and the CMS and ATLAS SCT's ($40\ m^2$ for ATLAS). What has happened is that the high-energy, high-luminosity hadron collider environment has become too unfriendly for wire chambers on almost any radius. Therefore, silicon microstrips are taking over as detectors with tracking precision $< 100\ \mu m$ and are able to handle the hit rates and the integrated radiation doses. For such large detectors, spatial precision is less of an issue (it will, in fact, be a challenge to build them with few micron stability, so the intrinsic detector precision may not be the driving factor). This is one reason for the interest in (for example) binary readout.

However, these detectors clearly have their limitations. There is a nasty hole of radius ≈ 30 cm in ATLAS and CMS within which microstrip detectors dare not venture, due (as we shall see in Sec. 6) to problems of radiation damage. With the huge event multiplicities, track merging would also be very serious. In this region, silicon pixel detectors may find a home and (at the smallest radii) other detector options, as we shall discuss in Sec. 7. The overall result is that the main emphasis in the world of silicon microstrip detectors has shifted from aiming to achieve the ultimate in spatial precision with the minimal detector thickness (including pushing for double-sided detectors) to aiming to cover very large areas as economically as possible, with electronics having an extremely high rate capability. The pressure for the most economical solution may argue against double-sided detectors, particularly since the material associated with the additional silicon layers is not seriously detrimental to the momentum resolution of the tracks that are important for physics. Fortunately, the size of the collaborations has grown at least as fast as the areas to be covered, so there is every reason to believe that they will succeed in these challenging tasks.

To describe any advanced technology as mature is usually misleading. Silicon microstrip detectors and particularly the associated electronics will continue to evolve for many years. However, as the OPAL Collaboration demonstrated when they decided they needed a silicon microstrip vertex detector to retain LEP competitiveness, it is possible starting from scratch to build a sophisticated detector with this technology within a year, provided one does not try to invent a lot of new features.