This opening chapter will concentrate on the changes in the world of semiconducting materials and devices over the latter half of the twentieth century. Within this field we have chosen to concentrate on a few developments and cannot claim to cover all of the major areas. What we plan to do is give a sense of perspective of how the science and technology of these materials has come to its current state and to present a brief overview of why certain materials are chosen for particular device applications.

We start by identifying some of the earliest developments in our understanding of electronic materials; follow the development of silicon technology from the first demonstration of the transistor through to today’s integrated circuit; track some of the key electronic and optoelectronic uses of the conventional III–V semiconductors; and end with a review of the last decade’s explosion of interest in the III–nitride materials. The band gaps of the semiconductors encountered in this chapter are shown in Fig. 1.1 – a figure which will be frequently referred to in explaining the choice of materials for specific applications.

We sometimes forget how remarkable electronic and optoelectronic materials are. Take the light emitting diode (LED) as an example: an electric current is passed through nothing more than a tiny stack of layers of slightly different compounds and brilliant coloured light is emitted. Of course the stack in question is carefully designed, and there are theories to explain the behaviour, but that should not detract from the initial moment of wonderment that it works at all! Devices from LEDs through to microprocessors containing tens of millions of transistors influence life at the start of this twenty first century to an incalculable extent. These complex devices are only available today because an array of electronic materials have been developed using increasingly sophisticated research methods over the past six decades.

**Fig. 1.1** Parameter perspective: band gaps and lattice parameters of selected semiconductors discussed in the text. The important wavelengths for optical storage (CD, DVD and Blu–Ray) and the 1.55 µm used for efficient data transmission through optical fibres are labelled.
Most semiconductor-based appliances which affect us every day are made using silicon, but many key devices depend on a number of different compound semiconductors. For example GaP-based LEDs in digital displays; GaAs-based HEMTs which operate in our satellite television receivers; AlGaAs and AlInGaP lasers in our CD and DVD players; and the InP-based lasers delivering the internet and telecommunications along optic fibres.

None of these devices could be manufactured without a basic understanding (either fundamental or empirical) of the materials science of the components. At the same time the overwhelming reason for scientific study of the materials is to elicit some improvement in the performance of the devices based on them. The rest of this book concentrates on the materials more than their devices but to give some idea of how a world largely untouched by electronic materials in the 1940s has become so changed by them we will consider the developments in the two fields in parallel in this chapter. Often they are inseparable anyway.

1.1 The Early Years

The exploitation of electronic materials in solid-state devices principally occurred in the second half of the twentieth century but the first serendipitous observations of semiconducting behaviour took place somewhat earlier than this. In 1833, Faraday found that silver sulphide exhibited a negative temperature coefficient [1.1]. This property of a decrease in electrical resistivity with increasing temperature was to be deployed in thermistor components a century later. In the 1870s scientists discovered and experimented with the photoconductivity (decreased resistivity of a material under incident light) of selenium [1.2, 3]. Amorphous selenium was to be used for this very property in the first Xerox copying machines of the 1950s.

While these discoveries had limited immediate impact on scientific understanding, more critical progress was made such as Hall’s 1879 discovery of what was to become known as the Hall Effect. The discovery of quantum mechanics was of fundamental importance for our understanding of semiconductors. Based on the advances in quantum theory in the early 1900s a successful theory to explain semiconductor behaviour was formulated in 1931 [1.4, 5]. However, the semiconductors of the 1930s were too impure to allow the theory to be compared with experiment. For example it was believed at the time that silicon, which was to become the archetypal semiconducting material, was a substance belonging to a group of materials which were “good metallic conductors in the pure state and... therefore to be classed as metals” [1.5]!

However, a solid theoretical understanding of semiconductors was in place by the 1940s. Hence when the device development focus of the second world war-time research was replaced by peace-time research into the fundamental understanding of real semiconductors, the foundations had been laid for working devices based on elemental semiconductors to be realised.

1.2 The Silicon Age

1.2.1 The Transistor and Early Semiconductor Materials Development

As its name suggests, electronics is about the control of electrons to produce useful properties; electronic materials are the media in which this manipulation takes place. Exactly fifty years after J. J. Thompson had discovered the electron in 1897, mankind’s ability to control them underwent a revolution due to the discovery of the transistor effect. It could be said that the world began to change in the final couple of weeks of 1947 when John Bardeen and Walter Brattain used germanium to build and demonstrate the first “semi-conductor triode” (a device later to be named the point contact transistor to reflect its transresistive properties). This success at Bell Laboratories was obtained within just a few years of the post-war establishment of a research group led by William Shockley focussing on the understanding of semiconducting materials. It was to earn Brattain, Bardeen and Shockley the 1956 Nobel Prize for Physics.

The first point-contact transistor was based around three contacts onto an n-doped germanium block: when a small current passed between the “base” and “emitter”, an amplified current would flow between the “collector” and “emitter” [1.6]. The emitter and collector
contacts needed to be located very close to one another (50–250 µm) and this was achieved by evaporating gold onto the corner of a plastic triangle, cutting the film with a razor blade and touching this onto the germanium – the two isolated strips of gold serving as the two contacts [1.7]. At about one centimetre in height, based on relatively impure polycrystalline germanium and adopting a different principle of operation, the device bears barely any resemblance to today’s integrated circuit electronics components. Nonetheless it was the first implementation of a solid-state device capable of modulating (necessary for signal amplification in communications) and switching (needed for logic operations in computing) an electric current. In a world whose electronics were delivered by the thermionic vacuum tube, the transistor was immediately identified as a component which could be “employed as an amplifier, oscillator, and for other purposes for which vacuum tubes are ordinarily used” [1.6].

In spite of this, after the public announcement of the invention at the end of June 1948 the response of both the popular and technical press was somewhat muted. It was after all still “little more than a laboratory curiosity” [1.8] and ultimately point-contact transistors were never suited to mass production. The individual devices differed significantly in characteristics, the noise levels in amplification were high and they were rapidly to be superseded by improved transistor types.

A huge range of transistor designs have been introduced from the late 1940s through to today. These successive generations either drew upon, or served as a catalyst for, a range of innovations in semiconductor materials processing and understanding. There are many fascinating differences in device design but from a materials science point of view the three most striking differences between the first point contact transistor and the majority of electronics in use today are the choice of semiconductor, the purity of this material and its crystalline quality. Many of the key electronic materials technologies of today derive from the developments in these fields in the very early years of the post-war semiconductor industry.

Both germanium and silicon had been produced with increasing purity throughout the 1940s [1.9]. Principally because of germanium’s lower melting temperature (937 °C compared with 1415 °C) and lower chemical reactivity its preparation had always proved easier and was therefore favoured for the early device manufacture such as the first transistor. However, the properties of silicon make it a much more attractive choice for solid state devices. While germanium is expensive and rare, silicon is, after oxygen, the second most abundant element. Silicon has a higher breakdown field and a greater power handling ability; its semiconductor band gap (1.1 eV at 300 K; Fig. 1.1) is substantially higher then germanium’s (0.7 eV) so silicon devices are able to operate over a greater range of temperatures without intrinsic conductivity interfering with performance.

The two materials competed with one another in device applications until the introduction of novel doping techniques in the mid-1950s. Previously p- and n-doping had been achieved by the addition of dopant impurities to the semiconductor melt during solidification. A far more flexible technique involved the diffusion of dopants from the vapour phase into the solid semiconductor surface [1.10]. It became possible to dope with a degree of two-dimensional precision when it was discovered that silicon’s oxide served as an effective mask to dopant atoms and that a photoresist could be used to control the etching away of the oxide [1.11, 12]. Successful diffusion masks could not be found for germanium and it was soon abandoned for mainstream device manufacture. Dopant diffusion of this sort has since been superseded by the implantation of high-energy ions which affords greater control and versatility.

Shockley was always aware that the material of the late 1940s was nothing like pure enough to make reliable high performance commercial devices. Quantum mechanics suggested that to make a high quality transistor out of the materials it was necessary to reduce the impurity level to about one part in 10¹⁰. This was a far higher degree of purity than existed in any known material. However, William Pfann, who worked at Bell Laboratories, came up with the solution. He invented a technique called zone-refining to solve this problem, and showed that repeated zone refining of germanium and silicon reduced the impurities to the level required. The work of Pfann is not widely known but was a critical piece of materials science which enabled the practical development of the transistor [1.13, 14].

At a similar time great progress was being made in reducing the crystalline defect density of semiconducting materials. Following initial hostility by some of the major researchers in the field it was rapidly accepted that transistor devices should adopt single crystalline material [1.15]. Extended single crystals of germanium several centimetres long and up to two centimetres in diameter [1.15, 16] and later similar silicon crystals [1.17] were produced using the Czochralski technique of pulling a seed crystal from a high purity melt [1.18]. The majority of material in use today is derived from this route. To produce silicon with
the very lowest impurity concentration, an alternative method called float zoning was developed where a polycrystalline rod was converted to a single crystal by the passage of a surface tension confined molten zone along its length [1.19–21]. No crucible is required in the process so there are fewer sources of impurity contamination. Float zoning is used to manufacture some of the purest material in current use [1.22]. The early Czochralski material contained dislocation densities of $10^6 – 10^7 \text{cm}^{-2}$ but by the start of the 1960s dislocation free material was obtained [1.23–26]. Initially most wafers were on the silicon (111) plane, which was easiest to grow, cut and polish [1.27]. For field-effect devices, which are discussed below, use of the (100) plane was found to offer preferable properties so this was introduced in the same decade. The impurity concentration in dislocation-free silicon has been continually reduced up to the present day and wafer diameters have increased almost linearly (though accelerating somewhat in recent years) from about 10 mm in the early 1960s to the “dinner plate” 300 mm today [1.22]. These improvements represent one of the major achievements in semiconductor materials growth and processing.

A series of generations of transistors followed in rapid succession after Brattain and Bardeen’s first triumph. Here we only mention a few of the major designs whose production have traits in common with technology today. Early in 1948 Shockley developed a detailed formulation of the theory of p–n junctions that concluded with the conception of the junction transistor [1.28, 29]. This involved a thin n-doped base layer sandwiched between p-doped emitter and collector layers (or vice versa). This p–n–p (n–p–n) structure is the simplest form of the bipolar transistor (so-called because of its use of both positive and negative charge carriers), a technology which remains important in analogue and high-speed digital integrated circuits today. In April 1950, by successively adding arsenic and gallium (n-type) and phosphorus (p-type) impurities to the n-type substrate wafers were manufactured using the so-called “mesa” process [1.31]. This process (which is still at the heart of device production today) was subsequently developed, in which the p–n junctions were all formed inside the substrate using oxide masking and diffusion from the surface. This resulted in a flat surface to which contacts could be made using a patterned evaporated film [1.33]. This processing technique was combined with some exciting thoughts at the end of the 1950s and led to the application of transistor devices and other components in a way which was to transform the world: the integrated circuit.

1.2.2 The Integrated Circuit

With the benefit of hindsight, the integrated circuit concept is quite simple. The problem faced by the electronics industry in the 1950s was the increasing difficulty of physically fitting into a small device all of the discrete electronic components (transistors, diodes, resistors and capacitors), and then connecting them together. It was clear that this problem would eventually limit the complexity, reliability and speed of circuits which could be created. Transistors and diodes were manufactured from semiconductors but resistors and capacitors were best formed from alternative materials. Even though they would not deliver the levels of performance achievable from the traditional materials, functioning capacitors and resistors could be manufactured from semiconductors so, in principle, all of the components of a circuit could be prepared on a single block of semiconducting material. This reasoning had been proposed by Englishman G.W.A. Dummer at...
a conference in 1952 [1.34] but small-scale attempts to realise circuits had failed, largely because they were based on connecting together layers in grown-junction transistors [1.35]. In 1958, however, Jack Kilby successfully built a simple oscillator and “flip-flop” logic circuits from components formed in situ on a germanium block and interconnected to produce circuits. He received the Nobel Prize in 2000 for “his part in the invention of the integrated circuit”.

Kilby’s circuits were the first built on a single semiconductor block, but by far the majority of the circuit’s size was taken up by the wires connecting together the components. Robert Noyce developed a truly integrated circuit (IC) in the form that it was later to be the components. Robert Noyce applied the planar technique to form transistors on silicon and photolithographically defined gold or aluminium interconnects. This was more suited to batch processing in production and was necessary for circuits with large numbers of components.

Most integrated circuits manufactured today are based around a transistor technology distinct from the bipolar device used in the first chips but one still dating from the 1960s. In 1960 the first metal oxide semiconductor field effect transistor (MOSFET) was demonstrated [1.36]. In this device a “gate” was deposited onto a thin insulating oxide layer on the silicon. The application of a voltage to the gate resulted in an inversion layer in the silicon below the oxide thereby modifying the conducting channel between “source” and “drain” contacts. This structure was a p-MOS device (current transfer between the collector and emitter was by hole conduction) grown on (111) silicon using an aluminium gate. Earlier attempts at such a device had failed because of trapped impurities and charges in the gate oxide – this new structure had reduced the density of these to below tolerable levels but the device still could not compete with the bipolar transistors of its time [1.27]. By 1967, however, (100) silicon (which offered lower densities of states at the Si/SiO₂ interface) was used together with a polycrystalline silicon gate to construct a more effective and more easily processed device with advantages over the bipolar transistor. In the early 1970s the n-MOS device, which was even less tolerant to the positive gate oxide charges, was realised thanks to much improved cleanliness in the production environment. With conduction occurring by the transfer of electrons rather than holes these were capable of faster operation than similar p-MOS structures (the mobility of electrons in silicon is about three times that of holes). By the 1980s, these two devices were combined in the complementary MOS (CMOS) device which afforded much lower power consumption and simplified circuit design [1.37]. This remains the principal structure used in microelectronics today. Of course now it is much smaller and significantly faster thanks to a range of further advances. These include improved control of the doping and oxidation of silicon and developments in optical lithographic techniques [1.37]. The minimum dimension of components which can be lithographically patterned on an integrated circuit is ultimately limited by the wavelength of radiation used in the process and this has continually been decreased over the past few decades. In the late 1980s wavelengths of 365 nm were employed; by the late 1990s 248 nm were common and today 193 nm is being used. Research into extreme ultraviolet lithography at 13.5 nm may see this being adopted within the next decade enabling feature sizes perhaps as low as 25 nm.

Though ICs implementing CMOS devices are the foundations for computing, silicon-based bipolar transistors maintain a strong market position today in radio frequency applications. In particular germanium is making something of a comeback as a constituent of the latest generation of SiGe bipolar devices [1.38]. The combination of the two forms of technology on a single chip (BiCMOS) offers the potential for computing and communications to be integrated together in the wireless devices of the coming decade.

The development in complexity and performance of silicon devices, largely due to materials science progress, is unparalleled in the history of technology. Never before could improvements be measured in terms of a logarithmic scale for such a sustained period. This is often seen as the embodiment of “Moore’s Law”. Noting a doubling of the number of components fitted onto integrated circuits each year between 1959 and 1965, Moore predicted that this rate of progress would continue until at least ten years later [1.39]. From the early 1970s, a modified prediction of doubling the number of components every couple of years has been sustained to the current day. Since the goals for innovation have often been defined assuming the continuation of the trend, it should perhaps be viewed more as a self-fulfilling prophecy. A huge variety of statistics relating to the silicon microelectronics industry follow a logarithmically scaled improvement from the late 1960s to the current day: the number of transistors shipped per year (increasing); average transistor price (decreasing); and number of transistors on a single chip (increasing) are examples [1.40]. The final member of this list is
Fig. 1.2 The realisation of Moore’s law by commercial Intel processors: The logarithmic increase in the number of transistors in each processor chip plotted in Fig. 1.2. A graphic example of the continuous scaling down of IC component dimensions is the reduction in the thickness of the MOSFET gate insulator. Figure 1.3 shows high-resolution transmission electron microscopy images of typical commercial devices since the mid-1990s: improved device performance has predominantly been achieved by reducing all dimensions of the IC components. With state-of-the-art gates now approaching the sub-nanometre thicknesses at which SiO2 breaks down as an insulator, it is clear that achieving device performance by scaling alone cannot be sustained for much longer. Overcoming this limit, perhaps through the substitution of SiO2 for a material with a higher dielectric constant, is one of the major challenges in silicon IC research today.

1.3 The Compound Semiconductors

It has been said that silicon is to electronics what steel is to mechanical engineering [1.42]. Steel is very effectively used for most of the world’s construction but there are some tasks which it is incapable of performing and others for which alternative structural materials are better suited. In the same way there are some crucial applications – such as optoelectronics and very high speed electronics – that silicon cannot usually deliver but which a wide range of compound semiconductors are better equipped to perform.

Silicon’s band gap is indirect (an electron-hole recombination across the band gap must be accompanied by an interaction with a phonon in the lattice) which severely limits the potential efficiency of light emission from the material. Many of the important compound semiconductors, such as the alloys AlxGa1−xAs, InxGa1−xN, AlxGa1−xN, and AlxInyGa1−x−yP exhibit direct band gaps (no phonon interaction is required) so can efficiently emit brilliant light in light emitting diodes (LEDs) and laser diodes. Furthermore, in these alloy systems, where the band gap can be adjusted by changing the composition, there is a means of selecting the energy released when an electron and hole recombine across the gap and therefore controlling the wavelength of the photons emitted. From the AlxInyGa1−x−yAs and AlxInyGa1−x−yN alloy systems there is, in principle, a continuous range of direct band gaps from deep in the infra red (InAs; \( \lambda = 3.5 \mu m \)) to far into the ultraviolet (AlN; \( \lambda = 200 \text{ nm} \)). The semiconductor band gaps of these materials and the corresponding photon wavelengths are put into context with the visible spectrum in Fig. 1.1.

Compounds are also very useful in high speed electronics applications. One of the determining factors in the speed of a transistor is the velocity of the charge carriers in the semiconductor. In GaAs the electron drift velocity is much higher than in silicon so its transistors are able to operate at significantly higher frequencies. The electron velocity in InAs is higher still. Furthermore, in the same way that silicon was preferred to...
germanium, devices manufactured using semiconductors such as GaN, which have much wider band gaps than silicon (3.4 eV compared with 1.1 eV), are capable of operating in much higher temperature environments.

Aside from these advantageous properties of compound semiconductors, the use of different alloy compositions, or totally different semiconductors, in a single device introduces entirely new possibilities. In silicon, most device action is achieved by little more than careful control of dopant impurity concentrations. In structures containing thin layers of semiconductors with different band gaps (heterostructures) there is the potential to control more fundamental parameters such as the band gap width, mobilities and effective masses of the carriers [1.42]. In these structures, important new features become available which can be used by the device designer to tailor specific desired properties. Hebert Kroemer and Zhores Alferov shared the Nobel prize in 2000 “for developing semiconductor heterostructures used in high-speed- and opto-electronics”.

We will mainly consider the compounds formed between elements in Group III of the periodic table and those in Group V (the III–V semiconductors); principally those based around GaAs and InP which were developed over much of the last forty years, and GaN and its related alloys which have been most heavily studied only during the last decade. Other families are given less attention here though they also have important applications (for example the II–VI materials in optoelectronic applications). It can be hazardous to try and consider the “compound semiconductors” as a single subject. Though lessons can be learnt from the materials science of one of the compounds and transferred to another, each material is unique and must be considered on its own (that is, of course, the purpose of the specialised chapters which follow in this handbook!).

It is worth repeating that the power of the compound semiconductors lies in their use as the constituent layers in heterostructures. The principal contribution from chemistry and materials science to enable successful devices has been in the manufacture of high-quality bulk single crystal substrates and the creation of techniques to reliably and accurately produce real layered structures on these substrates from the plans drawn up by a device theorist. In contrast to silicon, the compound semiconductors include volatile components so encapsulation has been required for the synthesis of low-defect InP and GaAs substrates such as in the liquid encapsulated Czochralski technique [1.43, 44]. The size and crystalline quality of these substrates lag some way behind those available in silicon. Crucial to the commercialisation of electronic and optoelectronic heterostructures were the improvements over the last few decades in the control of epitaxial growth available to the crystal grower. The first successful heterostructures were manufactured using deposition onto a substrate from the liquid phase (liquid phase epitaxy; LPE) – “a beautifully simple technology but with severe limitations” [1.42]. However, the real heterostructure revolution had to wait for the 1970s and the introduction of molecular beam epitaxy (MBE) and metalorganic chemical vapour deposition (MOCVD) – also known as metalorganic vapour phase epitaxy (MOVPE) provided that the deposition is epitaxial.

MBE growth occurs in an ultra-high vacuum with the atoms emitted from effusion cells forming “beams” which impinge upon, and form compounds at the substrate surface. It derives from pioneering work at the start of the 1970s [1.45]. MOCVD relies on chemical reactions occurring on the substrate involving metalorganic vapour phase precursors and also stems from initial work at this time [1.46]. In contrast to LPE, these two techniques permit the combination of a wide range of different semiconductors in a single structure and offer a high degree of control over the local composition, in some cases on an atomic layer scale. The successful heterostructure devices of the late 1970s and 1980s would not have been achievable without these two tools and they still dominate III–V device production and research today.

1.3.1 High Speed Electronics

The advantages of the III–V materials over silicon for use in transistors capable of operating at high frequencies were identified early in the semiconductor revolution [1.47]. Shockley’s first patent for p-n junction transistors had included the proposal to use a wide-gap emitter layer to improve performance and in the 1950s Kroemer presented a theoretical design for a heterostructure transistor [1.48]. Some years later the structure of a GaAs metal semiconductor FET (MESFET) was proposed and realised soon afterwards [1.49, 50]. In these devices a Schottky barrier surface potential was used to modulate the conductivity of the GaAs channel. One of the earliest applications of the III–V’s was as low noise amplifiers in microwave receivers which offered substantial improvements relative to the silicon bipolar transistors of the time. The devices were later used to demonstrate sub-nanosecond switching in monolithic digital ICs [1.51]. Today they form the core of the highest speed digital circuits and are used in
high speed electronics in microwave radar systems and wireless communications which incorporate monolithic integrated circuits.

For at least 30 years there have been repeated attempts to replicate the MOSFET, the dominant transistor form in silicon ICs, on GaAs material. These attempts have been frustrated by the difficulty of reproducibly forming a high quality stoichiometric oxide on GaAs. In direct analogy with the initial failure of constructing working n-MOSFETs on silicon, the GaAs devices have consistently been inoperative because of poor quality gate oxides with a high density of surface states at the GaAs-insulator interface [1.44]. One of the research efforts focussed on realising this device was, however, to be diverted and resulted in the discovery of probably the most important III–V electronic device: the high electron mobility transistor (HEMT).

The background to this invention lies in the beautiful concept of modulation doping of semiconductors which was first demonstrated in 1978 [1.52]. One of the tenets of undergraduate semiconductor courses is the demonstration that as the dopant density in a semiconductor increases, the mobility of the carriers is reduced because the carriers are scattered more by the ionised dopants. It was found that in a multilayer of repeating n-AlGaAs layers and undoped GaAs layers, the electrons supplied by donor atoms in the AlGaAs moved into the adjacent potential wells of the lower-band gap GaAs layers. In the GaAs these suffered from substantially less ionised impurity scattering and therefore demonstrated enhanced mobility.

While working in a group attempting to create GaAs MOSFETs (and seemingly despairing at the task! [1.53]), Mimura heard of these results and conceived of a field effect transistor where the conducting channel exploited the high mobility associated with a modulation doped structure. In essence, a doped AlGaAs layer was formed above the undoped GaAs channel of the transistor. Donated carriers gathered in the AlGaAs immediately below the interface where they did not suffer from as much ionised impurity scattering and so their mobility would approach that of an ultra-pure bulk semiconductor. The current was conducted from the source to the drain by these high mobility carriers and so the devices were able to operate in higher frequency applications [1.53]. Realisation of the structure required a very abrupt interface between the GaAs and AlGaAs and was considered beyond the capability of MOCVD of the time [1.53]. However, following the advances made in MBE procedures during the 1970s the structure was achieved by that technique within a few months of the original conception [1.53, 54]. The first operational HEMT chips were produced on 24th December 1980: by pleasing coincidence this was the anniversary of Brattain and Bardeen’s demonstration of their point contact transistor to the management of Bell labs in 1947! Structures based on the same principle as Mimura’s device were realised in France very shortly afterwards [1.55].

The commercialisation of the HEMT became significant in the late 1980s thanks to broadcasting satellite receivers. The improved performance of the devices compared with the existing technology allowed the satellite parabolic dish size to be reduced by at least a factor of two. Structures similar to these have since played a crucial role in the massive expansion in mobile telephones.

The evolution in HEMT structures since the early 1980s is a fine example of how fundamental compound semiconductor properties have been exploited as the technology has become available to realise new device designs. The electron mobility in InAs is much higher than in GaAs and rises as the indium content in In$_x$Ga$_{1-x}$As is increased [1.56]. The introduction of an InGaAs (as opposed to GaAs) channel to the HEMT structure resulted both in increased electron mobility and a higher density of carriers gathering from the doped AlGaAs layer (because of the larger difference in energy between the conduction band minima of InGaAs and AlGaAs than between GaAs and AlGaAs). This so-called pseudomorphic HEMT (pHEMT) demonstrates state of the art power performance at microwave and millimetre wave frequencies [1.43]. The indium content and thickness of the channel is limited by the lattice mismatch with the GaAs (Fig. 1.1). If either is increased too much then misfit dislocations are formed within the channel. The restriction is reduced by growing lattice matched structures on InP, rather than GaAs, substrates. Al$_{0.48}$In$_{0.52}$As and In$_{0.53}$Ga$_{0.47}$As are both lattice matched to the InP (Fig. 1.1) and their conduction band minimum energies are well separated so that in the InGaAs below the interface between the two compounds a high density of electrons with a very high mobility is formed. Compared with the pHEMTs these InP based HEMTs exhibit significant improvements, have been shown to exhibit gain at over 200 GHz and are established as the leading transistor for millimetre-wave low noise applications such as radar [1.43].

1.3.2 Light Emitting Devices

LEDs and laser diodes exploit the direct band gap semiconductors to efficiently convert an electric current into
photons of light. Work on light emission from semiconductor diodes was carried out in the early decades of the twentieth century [1.57] but the start of the modern era of semiconductor optoelectronics traces from the demonstration of LED behaviour and laser from P-n junctions in GaAs [1.58, 59] and GaAs$_{1-x}$P$_x$ [1.60]. The efficiency of these LEDs was low and the lasers had large threshold currents and only operated at low temperatures. A year later, in 1963, Kroemer and Alferov independently proposed the concept of the double heterostructure (DH) laser [1.61, 62]. In the DH device, a narrow band gap material was to be sandwiched between layers with a wider gap so that there would be some degree of confinement of carriers in the “active layer”. By the end of the decade DH devices had been constructed which exhibited continuous laser at room temperature [1.63, 64]. Alferov’s laser was grown by LPE on a GaAs substrate with a 0.5 µm GaAs active layer confined between 3 µm of Al$_{0.25}$Ga$_{0.75}$As on either side. The launch of the Compact Disc in 1982 saw this type of device, or at least its offspring, becoming taken for granted in the households of the world.

One of the major challenges in materials selection for heterostructure manufacture has always been avoiding the formation of misfit dislocations to relieve the strain associated with lattice parameter mismatch between the layers. Al$_x$Ga$_{1-x}$As exhibits a direct band gap for $x < 0.45$ and the early success and sustained dominance of the AlGAs/GaAs system derives significantly from the very close coincidence of the AlAs and GaAs lattice parameters (5.661 Å and 5.653 Å – see Fig. 1.1). This allows relatively thick layers of AlGAs with reasonably high aluminium content so to be grown lattice matched onto GaAs substrates with no misfit dislocation formation. The use of the quaternary alloy solid solution In$_x$Ga$_{1-x}$AsP$_{1-y}$ was also suggested in 1970 [1.63] to offer the independent control of lattice parameters and band gaps. Quaternaries based on three Group III elements have since proved very powerful tools for lattice matching within heterostructures. (Al$_x$Ga$_{1-x}$)$_{1-y}$In$_{y}$P$_z$ was found to be almost perfectly lattice matched to GaAs and additionally have a very similar thermal expansion coefficient (which is important to avoid strain evolution when cooling after growth of heteroepitaxial layers at high temperatures). By varying $x$ in this compound, direct band gaps corresponding to light between red and green could be created [1.65]. Lasers based on this alloy grown by MOCVD are a common choice for the red wavelengths (650 nm) used in DVD reading.

Obtaining lattice matching is not so crucial for layers thinner than the critical thickness for dislocation production and can be less of an issue these days because of probably the most important development in the history of optoelectronic devices: the introduction of the quantum well. In some ways a quantum well structure is an evolution of the double heterostructure but with a very much thinner active layer. It is the chosen design for most solid state light emitting devices today. With the accurate control available from MBE or MOCVD, and following from some early work on superlattices [1.66], very thin layers of carefully controlled composition could be deposited within heterostructure superlattice stacks. It became possible to grow GaAs layers much less than 10 nm thick within AlGaAs–GaAs heterostructures. The carriers in the GaAs were found to exhibit quantum mechanical confinement within the one dimensional potential well [1.67, 68]. Lasing from GaAs/Al$_{0.5}$Ga$_{0.5}$As quantum wells was reported the following year, in 1975, [1.69] but it was a few years before the performance matched that achievable from DH lasers of the time [1.70] and the quantum well laser was further advanced to significantly outperform the competition by researchers in the 1980s [1.71].

The introduction of heterostructures with layer thicknesses on the nanometre scale represents the final stage in scaling down of these devices. Similarly Brattain and Bardeen’s centimetre-sized transistor has evolved into today’s microprocessors with sub-micron FETs whose gate oxide thicknesses are measured in Angstroms. Throughout this evolution, materials characterisation techniques have contributed heavily to the progress in our understanding of electronic materials and deserve a brief detour here. As the dimensions have been reduced over the decades, the cross-sectional images of device structures published in the literature have changed from a period where optical microscopy techniques were sufficient [1.31] to a time when scanning electron microscopy (SEM) images were used [1.70] and to today’s high resolution transmission electron microscopy (TEM) analysis of ultra-thin layers (e.g. Fig. 1.3). For each new material family, understanding of defects and measurement of their densities (e.g. by TEM and X-ray topography) have contributed to improvements in quality. Huge improvements in X-ray optics have seen high-resolution X-ray diffraction techniques develop to become a cornerstone of heterostructure research and production quality control [1.72]. Scanning-probe techniques such as scanning tunnelling and atomic force microscopy have become crucial to the understanding of MBE and MOCVD growth. Chemically sensitive techniques such as secondary ion mass spectroscopy and Rutherford
Annular dark field-scanning transmission electron microscope (ADF-STEM) image of Sb-doped Si. The undoped region (right) shows atomic columns of uniform intensity. The brightest columns in the doped region (left) contain at least one Sb atom. The image is smoothed and background subtracted (After [1.76], with permission Elsevier Amsterdam). (Courtesy of Prof D.A. Muller)

Backscattering have improved to provide information on doping concentrations and compositions in layered structures with excellent depth resolution. Meanwhile recently developed techniques such as energy-filtered TEM [1.73] afford chemical information at extremely high spatial resolutions. The characterisation of doping properties is also coming of age with more quantitative measurement of dopant contrast in the SEM [1.74], analysis of biased junctions in situ in the TEM [1.75] and the recent exciting demonstration of imaging of single impurity atoms in a silicon sample using scanning TEM, Fig. 1.4. The materials characterisation process remains a very important component of electronic materials research.

Two commercial-product oriented aims dominate semiconductor laser research: the production of more effective emitters of infra-red wavelengths for transmission of data along optic fibres; and the realisation of shorter wavelength devices for reading optical storage media. In the first of these fields devices based on InP have proven to be extremely effective because of its fortuitous lattice parameter match with other III–V alloys which have band gaps corresponding to the low-absorption “windows” in optic fibres. While remaining lattice matched to InP, the In$_x$Ga$_{1-x}$As$_y$P$_{1-y}$ quaternary can exhibit band gaps corresponding to infra-red wavelengths of 1.3 µm and 1.55 µm at which conventional optic fibres absorb the least of the radiation (the absolute minimum is for 1.55 µm). Room temperature continuous lasing of 1.1 µm radiation was demonstrated from the material in 1976 [1.77] and InP based lasers and photodiodes have played a key role in the optical communications industry since the 1980s [1.43].

We have already mentioned the AlGaAs infra-red ($\lambda = 780$ nm) emitters used to read compact discs and the AlGaInP red ($\lambda = 650$ nm) devices in DVD readers (see Fig. 1.1). As shorter wavelength lasers have become available the optical disc’s surface pits (through which bits of data are stored) could be made smaller and the storage density increased. Though wide band gap II–VI compounds, principally ZnSe, have been researched for many decades for their potential in green and blue wavelengths, laser operation in this part of the visible spectrum proved difficult to realise [1.78]. In the early 1990s, following improvements in the p-doping of ZnSe, a blue-green laser was demonstrated [1.79] but such devices remain prone to rapid deterioration during operation and tend to have lifetimes measured in, at most, minutes. However, also in the early years of the 1990s, a revolution began in wide band gap semiconductors which is ongoing today: the exploitation of GaN and its related alloys In$_x$Ga$_{1-x}$N and Al$_x$Ga$_{1-x}$N. These materials represent the future for optoelectronics over a wide range of previously inaccessible wavelengths and the next generation of optical storage, the “Blu-ray” disc, will be read using an InGaN blue-violet laser ($\lambda = 405$ nm).

1.3.3 The III-Nitrides

The relevance of the In$_x$Ga$_{1-x}$N alloy for light emitting devices is clear from Fig. 1.1. The InN and GaN direct band gaps correspond to wavelengths straddling the visible spectrum and the alloy potentially offers access to all points in-between. The early commercially successful blue light emitters were marketed by Nichia Chemical Industries following the research work of Nakamura who demonstrated the first InGaN DH LEDs [1.80] and blue InGaN quantum well LEDs and laser diodes soon after [1.81]. Since this time the global research interest in the GaN material family has expanded rapidly and the competing technology (SiC and ZnSe for blue LEDs and lasers respectively) has largely been replaced.

The development of the III-nitride materials has much in common with the early research of other III–V systems. For example MOCVD and MBE technology
could be adapted for the nitride systems (the former has to date been more suited for creating optical devices) and one of the obstacles limiting early device development was achieving sufficiently high p-type doping. However, in some ways they are rather different from the other compound semiconductors. It is important to realise that while all III–Vs mentioned previously share the same cubic crystallographic structure, the nitrides most readily form in a hexagonal allotrope. Most significant in terms of device development over the past decade has been the difficulty in obtaining bulk GaN substrates. Due to the very high pressures necessary to synthesise the compound only very small pieces of bulk GaN have been produced and though they have been used to form functioning lasers [1.82] they remain unsuited as yet to commercial device production. There has consequently been a reliance on heavily lattice mismatched heteroepitaxial growth.

Many materials have been used as substrates for GaN growth. These include SiC, which has one of the lowest lattice mismatches with the nitride material and would be more widely used if it was less expensive, and silicon, which has considerable potential as a substrate if problems associated with cracking during cooling from the growth temperature can be overcome (there is a large difference between the thermal expansion coefficients of GaN and silicon). The dominant choice, however, remains sapphire (α-Al2O3) which itself is by no means ideal: it is electrically insulating (so electrical contacts cannot be made to the device through the substrate material) and, most significantly, has a lattice mismatch of ≈16% with the GaN [1.83]. This mismatch is relieved by the formation of misfit dislocations which give rise to dislocations threading through the GaN into the active layers (e.g. InGaN quantum wells) of the devices. The key discovery for reducing the defect densities to tolerable levels was the use of nucleation layers at the interface with the sapphire [1.83] but densities of ≈10^6 cm^-2 remain typical. More recently epitaxial lateral overgrowth (ELOG) techniques have allowed the dislocation densities to be reduced to ≈10^5 cm^-2 in local regions [1.84]. However, perhaps the most interesting aspect of GaN-based optoelectronic devices is that they emit light so efficiently in spite of dislocation densities orders of magnitude greater than those tolerated in conventional semiconductors.

Even though InGaN based light emitters have been commercially available for several years, the precise mechanism of luminescence from the alloy is still not fully understood. Having so far discussed the evolution of semiconductors with the benefit of hindsight we can now for a moment consider an unresolved issue which, no doubt, will be solved in the coming years. It is widely believed that the tolerance of InGaN optoelectronic devices to high densities of defects is caused by the presence of low-energy sites within the layers at which electrons and holes are localised. They are thus prevented from interacting with the dislocations at which they would recombine in a non-radiative manner. The origin of localisation remains a matter of debate. One popular explanation is that the InGaN alloy has a tendency for phase segregation [1.85] and indium-rich “clusters” form and cause the localisation. However, there is now evidence [1.86] that the results of some of the measurements used to detect the indium rich regions could be misleading so the clustering explanation is being re-assessed. InGaN remains a fascinating and mysterious alloy.

Solid-state lighting will be a huge market for III-nitride materials in the coming decades. LEDs are perfectly suited to coloured light applications: their monochromatic emission is very much more energy efficient than the doubly wasteful process of colour filtering power hungry filament white light bulbs. InGaN LEDs are now the device of choice for green traffic signals worldwide and offer significant environmental benefits in the process. In principle there is also the opportunity to create white light sources for the home which are more efficient than the tungsten filament light bulbs used today and a variety of promising schemes have been developed for converting the coloured output of III-nitride LEDs to white sources. These include the use of three colour (red, green and blue) structures and blue InGaN or ultraviolet AlGaN based LEDs coated with a range of phosphor materials to generate a useful white spectrum [1.65]. In particular ultraviolet LEDs coated with a three-way phosphor (red, green and blue) can produce high quality white light that mimics sunlight in its visible spectrum. The main disadvantages preventing the widespread use of LEDs in white lighting are their high cost and the relatively low output powers from single devices but these obstacles are rapidly being overcome.

Many other applications for the III-nitrides are being investigated including the use of (Al,Ga,In)N solar cells which could offer more efficient conversion of light into electric current than silicon based devices [1.87]; the possibility of lasers and optical switches operating at the crucial 1.55 µm wavelength based on intersubband transitions in AlGaN/GaN quantum well structures; and the use of the compact InGaN LEDs to fluoresce labelled cancerous cells and aid detection of...
affected areas. The wide band gap is also very attractive for many electronic device applications—particularly in high-temperature, high-power applications. Exploiting its high thermal conductivity and insensitivity to high operating temperatures, GaN-based HEMTs may extend the power of mobile phone base stations and it has even been suggested that GaN devices could be used as an alternative source of ignition in car engines. There is also, of course, the possibility of monolithically integrating electronic and optoelectronic action onto a single chip. GaN-related materials should prove to have a huge impact on the technology of the coming decades.

1.4 From Faraday to Today

So, we have come 170 years from Faraday’s nineteenth century observation of semiconductor to a world dominated by electronic materials and devices. The balance of power between the different semiconductor families is an unstable and unpredictable one. For example if inexpensive, high quality, low defect density GaN substrates can be produced this will revolutionise the applications of GaN-based materials in both optoelectronics and electronics. The only inevitable fact is that the electronics revolution will continue to be crucially dependent on electronic materials understanding and improvement. And while reading the more focussed chapters in this book and concentrating on the very important minutiae of a particular field, it can be a good idea to remember the bigger picture and the fact that electronic materials are remarkable!

Further References

In particular we recommend the transcripts of the Nobel Lectures given by Brattain, Bardeen, Shockley, Kilby, Kroemer and Alferov. Available in printed form as set out below and, for the latter three in video, at www.nobel.se.

References

1.2 W. Smith: J. Soc. Telegraph Eng. 2, 31 (1873)
1.10 C. S. Fuller: Phys. Rev. (Ser 2) 86, 126 (1952)
1.12 J. Andrus, W. L. Bond: US Patent 3 122 817 (1952)
1.16 J. J. Little: Phys. Rev. (Ser 2) 78, 647 (1950)
1.17 G. K. Teal, E. Buehler: Phys. Rev. 87, 190 (1952)
1.18 J. Czochralski: Z. Phys. Chem. 92, 219 (1917)
1.20 H. C. Theurer: US Patent 3 060 123 (1952)
1.22 K. A. Jackson (Ed.): Silicon Devices (Wiley, Weinheim 1998)
1.23 W. C. Dash: J. Appl. Phys. 29, 736 (1958)
1.27 M. Grayson (Ed.): Encyclopaedia of Semiconductor Technology (Wiley, New York 1984) p. 734
1.29 W. Schockley: Bell Syst. Tech. J. 28(4), 435 (1949)
1.30 W. Schockley, M. Sparks, G. K. Teal: Phys. Rev. 83, 151 (1951)
<table>
<thead>
<tr>
<th>Reference Number</th>
<th>Author(s)</th>
<th>Title</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.36</td>
<td>D. Kahng, M. M. Atalla</td>
<td><em>Silicon–Silicon Dioxide Field-Induced Surface Devices</em> (Solid State Research Conference, Pittsburgh, Pennsylvania 1960)</td>
<td></td>
</tr>
<tr>
<td>1.39</td>
<td>G. E. Moore</td>
<td>Electronics 38(8) (1965)</td>
<td></td>
</tr>
<tr>
<td>1.43</td>
<td>D. Kahng, M. M. Atalla</td>
<td><em>Silicon–Silicon Dioxide Field-Induced Surface Devices</em> (Solid State Research Conference, Pittsburgh, Pennsylvania 1960)</td>
<td></td>
</tr>
<tr>
<td>1.46</td>
<td>H. M. Manasevit</td>
<td>Appl. Phys. Lett. 12, 156 (1968)</td>
<td></td>
</tr>
<tr>
<td>1.48</td>
<td>H. Kroemer</td>
<td>RCA Rev. 18, 332 (1957)</td>
<td></td>
</tr>
<tr>
<td>1.57</td>
<td>R. N. Hall</td>
<td>Phys. Rev. Lett. 9, 366 (1962)</td>
<td></td>
</tr>
<tr>
<td>1.59</td>
<td>H. Kroemer</td>
<td>Proc. IEEE 51, 1782 (1963)</td>
<td></td>
</tr>
<tr>
<td>1.60</td>
<td>I. Ho</td>
<td>G. B. Stringfellow</td>
<td></td>
</tr>
</tbody>
</table>