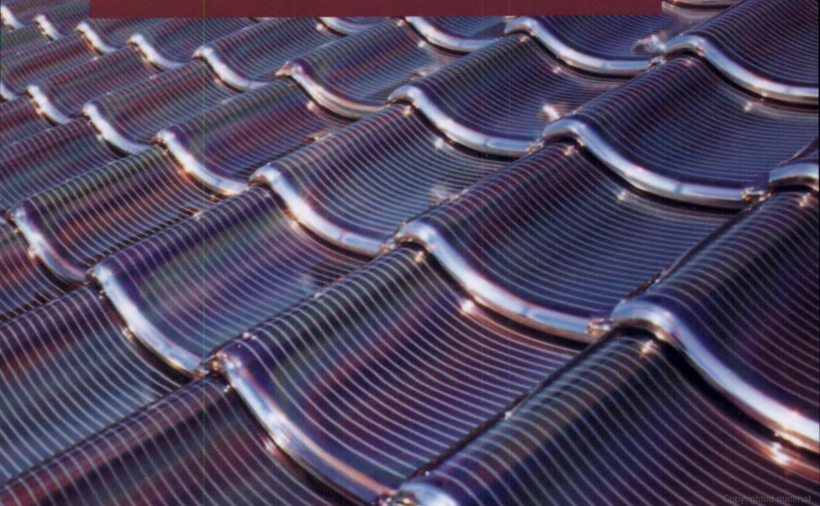
PHILIP BALL

MADE TO MEASURE

NEW MATERIALS FOR THE 21ST CENTURY



MADE TO MEASURE

NEW MATERIALS FOR THE 21ST CENTURY

Philip Ball

PRINCETON UNIVERSITY PRESS · PRINCETON, NEW JERSEY

Published by Princeton University Press, 41 William Street, Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press, Chichester, West Sussex

Fourth printing, and first paperback printing, 1999
Paperback ISBN 0-691-00975-9

The Library of Congress has cataloged the cloth edition of this book as follows

Ball, Philip
Made to measure : new materials for the 21st century / Philip Ball
p. cm.
Includes bibliographical references and index.
ISBN 0-691-02733-1 (cloth : alk. paper)
1. Materials—Technological innovations. I. Title.
TA403.B2247 1997
620.1′1—dc21 97-4027

This book has been composed in Times Roman

The paper used in this publication meets the minimum requirements
of ANSI/NISO Z39.48-1992 (R1997) (*Permanence of Paper*)

http://pup.princeton.edu

Printed in the United States of America

5 6 7 8 9 10

CHAPTER ONE

Light Talk

PHOTONIC MATERIALS

> Every day you play with the light of the universe.
>
> —*Pablo Neruda*

The next revolution in information technology will dispense with the transistor and use light, not electricity, to carry information. This change will rely on the development of photonic materials, which produce, guide, detect, and process light.

By a few years into the twenty-first century, the whole world will be "online." Just about every nation on Earth will be linked up to a communications network in which information can flow in the blink of an eye between computer terminals in Denver and Beijing, Mombasa and Copenhagen. This is the information superhighway, a web of information channels that knows no territorial, cultural, or political barriers. That it will coexist with the most appalling poverty in some parts of the world, with wars and ethnic conflicts, is a stark reminder that information alone solves no human problems. Yet however you regard it, a communications system of this sort will be like nothing we have seen before, and it will change our lives.

The flow of data that this system will have to support is immense. Many millions of individual messages will be routed along the superhighway's arteries, simultaneously and without interfering with one another. Their transmission must take place over long distances without deterioration of the signal. Computer networks like the Internet create an ever-expanding demand for efficient communications systems, and already threaten to strain existing systems to overload. The nascent digital video technology will add to the pressure; sending digital video data "down the line" so that distant viewers can receive live pictures from a video camera requires around five hundred times more data-transmission capacity than a telephone call.

All this is simply the latest development in a succession that has led from the telegraph of the early nineteenth century to the telephone, the television, the communications satellite, and the fax machine. Until the early 1970s, the demand on long-distance communications could be met by the electronics industry. But it has become ever more clear that electronic transmission of information will be unable to accommodate the growth in data flow that the future promises to bring. A new technology is needed.

That technology is with us already, but only in a form comparable to that of the early days of electronic communications. It is called *photonics*, and it replaces electrical currents with light: instead of being conveyed by electrons in a copper wire, information is borne by photons, the particles of light. The first long-distance photonic transmission cable was laid down in 1988; today such cables are replacing copper telecommunications cables in just about all long-distance and most short-distance applications. These cables, made from glass optical fibers, can carry many thousands of times more information than electrical wires, and at lower power consumption.

At present, just about all of the data handling at each end of a fiber-optic transmission cable is still done by electronics. So it has been necessary to devise ways of converting an electronic signal into a series of light pulses that are fed into the optical cable, and to turn those pulses back into electricity at the other end. This integration of optical and electronic data processing is called *optoelectronics*.

Optoelectronic circuits are now an essential part of information technology. The practical difficulties of making optoelectronic devices that can be integrated with silicon-based circuits on a single microchip are far from trivial, however, and are still being tackled. Quite aside from this integration problem, the use of electronics will ultimately set a speed limit on the rate with which data can be handled—photonics alone could do it much faster. So engineers are now asking whether this cumbersome method of converting a signal first to one form and then to another is really the best way of going about the problem. Why not, they suggest, do it *all* with light? That is to say, why not dispense with electronics altogether and make chips that perform data processing purely by photonic means?

The scientific underpinnings of an all-photonic technology are already in place: we know how to make miniaturized components that guide beams of light and use them to perform logical operations—the central steps of computation. A photonic transistor, a device that is still in the early stages of development, would be switchable much more quickly than the electronic varieties, and this might allow a photonic computer to run a thousand times more speedily than modern electronic computers. Moreover, photonic devices permit engineers to contemplate entirely new types of circuit design and architecture. Optical circuit components should in principle contain fewer constituent parts than their electronic counterparts, making them cheaper and easier to package onto chips. All in all, photonics should be a cleaner, faster, more compact, and more versatile form of information processing.

None of these developments can happen without the right materials. For optical communications, the optical properties of glass fibers have been honed to an astonishing degree. Optoelectronics has been wholly dependent on the identification of suitable materials for making the solid-state lasers that act as light sources and photodetectors for converting light back to electricity. Performing information processing with light requires materials whose response to light is highly unusual and very different from that of our everyday experience. When the photonic era arrives, it will be materials scientists who will act as the midwife.

A Revolution Written in Silicon

Telecommunications—literally, long-distance discourse—became an instant affair only with the advent of the electronic age. First came the telegraph, tapped out in code in the manner beloved of movies of the Old West; then in the 1870s Alexander Graham Bell's telephone, regarded in its early days with almost superstitious awe; and in the 1890s Guglielmo Marconi's "wireless telegraph," which showed that words could be sent through the air rather than through copper wire. Electronic communications, then and now, use modulated electrical signals—a current that varies in time—to carry information down copper wires. By the 1970s, the U.S. telecommunications industry was consuming around 200,000 tons of copper per year in cabling.

As the traffic of information grew, the task of processing it—modulating the signal at the transmitting end, routing the data correctly, and interpreting it at the receiving end—became ever more challenging. The turning point in electronic data processing came in 1947 with the invention of the transistor by John Bardeen and Walter Brattain at Bell Telephone Laboratories. Previously, the modulation and amplification of electrical signals were performed by vacuum tubes, which were fragile, cumbersome, and consumed a lot of power. Transistors did away with all of these problems in a single swoop—they are compact and robust and consume a minuscule fraction of the power of vacuum tubes (even the very first transistor ran on a millionth of the power of a tube). What is more, they are much faster and more reliable switches. It is no coincidence that the invention of the transistor was soon followed by a rapid growth in the power and commercialization of computers—automated devices for handling and processing electronic information. The earliest computers, such as the ENIAC device developed by engineers at the University of Pennsylvania in the 1940s, were tube-driven analog machines that occupied entire rooms and were of questionable reliability. Today many thousands of transistors and other electronic devices can be carved into semiconducting materials on a single chip no more than a millimeter square (fig. 1.1), and computers can fit into a briefcase.

The transistor's central place in modern electronics has been gained only through diligent research on the materials from which it is made, of which the most important is silicon. It is hard to think of any other industry that has become more intimately associated with the material on which it depends. We hear talk of the silicon revolution and of silicon chips pouring out of America's heartland of

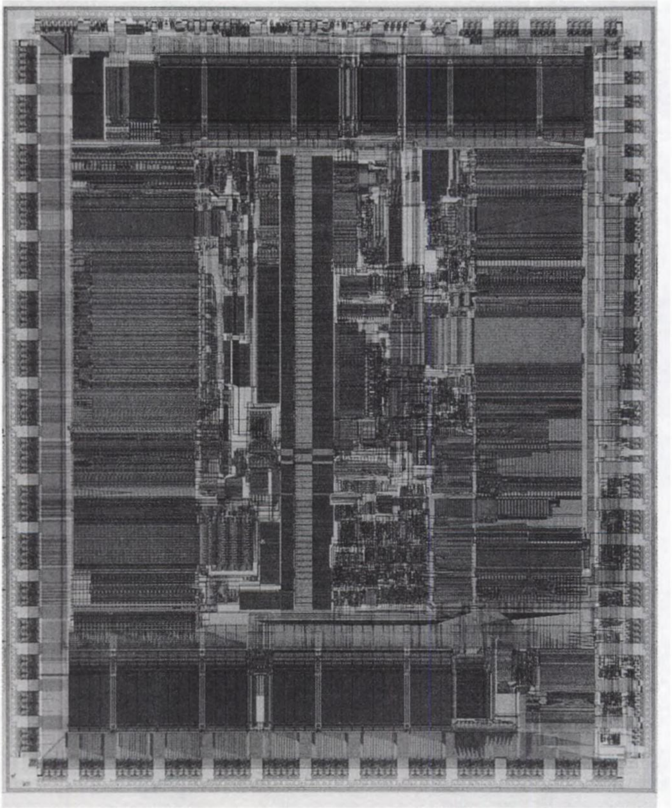


FIGURE 1.1 A silicon microchip manufactured by Digital Equipment Corporation. This chip, the Alpha 21164, is the world's fastest single-chip microprocessor, able to execute over one billion instructions per second. (Photograph courtesy of Digital Equipment Corporation.)

information technology, Silicon Valley in California. So closely has silicon become linked with "thinking" machines that it is the staple of science-fiction writers searching for plausible life forms not based on carbon.

The key to silicon's central role in microelectronics is the fact that it is a *semiconductor*—a material whose electrical properties can be influenced in a variety of subtle ways. A material's electrical conductivity is determined by its electronic structure, by which I mean the disposition of its electrons. The chemical bonds that hold materials together are formed by overlap of the veils of electrons (called orbitals) that surround atoms; these are called covalent bonds.[1] In solids these overlapping electron orbitals give rise to extended networks of "electron density" throughout the material; in general, different networks can be ascribed to the overlap of different sets of atomic orbitals. The energies of electrons in these extended states, or "bands," are restricted by quantum mechanics to a certain range of values, and so the electronic structure of solids can be depicted as electronic bands separated by gaps of forbidden energies, called *band gaps* (fig. 1.2*a*).

An electrical current corresponds to the flow of electrons (or sometimes of other charged particles). Although electronic bands are notionally extended throughout a solid, the mobility of the electrons that each contains depends on the extent to which the band is filled. Each band has only a limited electron capacity; once a band is filled, additional electrons in the material have to go into the band of next highest energy. Electrons in filled bands are relatively immobile, being constrained to stay more or less in the vicinity of individual atoms. Electrons in bands that are only partially filled, on the other hand, can move throughout the solid when a voltage is applied across it. So solids with only fully filled electronic bands cannot conduct—they are insulators—whereas those with partly filled bands (a category that includes most metals) are electrical conductors.

In all solids, the fully filled electronic band that has the highest energy is called the *valence band*. (Valence electrons are those that are available for forming chemical bonds; this naming of the uppermost filled band reflects the fact that it is these higher-energy electrons that are primarily responsible for the bonds between neighboring atoms.) The next band above the valence band is called the *conduction band*; in insulators this is empty, in metals it is partly filled (fig. 1.2*a*). A voltage applied across a material makes the electrons' energies vary in space; they are lower in energy close to the positive terminal and higher close to the negative terminal. So a voltage introduces a tilt to the band structure (fig. 1.2*b*), and electrons that are free to move (that is, those that are in a partially filled band) flow down the slope.

Semiconductors typically have an electrical conductivity somewhere between metallic conductors such as copper and insulators such as diamond. This suggests that they have some mobile charge carriers, but far fewer than metals. The electronic band structure of pure semiconductors like silicon is of the same type as that of insulators: the uppermost electronic band (the valence band) is completely filled, and a band gap separates this from a completely empty conduction band. But the crucial distinction between a semiconductor like silicon and an insulator like diamond is the size of this gap: in silicon it is small enough that a few electrons can pick up enough thermal energy to hop up into the conduction band, where they are free to move (fig. 1.2*a*). This hopping leaves behind an electron

[1] There are other types of chemical bond too. Many solids are held together by ionic bonds, in which charged atoms of opposite charge attract one another. The distinction between ionic and covalent bonds is not absolute—in fact, most bonds between different atoms have some ionic (electrostatic) and some covalent (electron-cloud overlap) character.

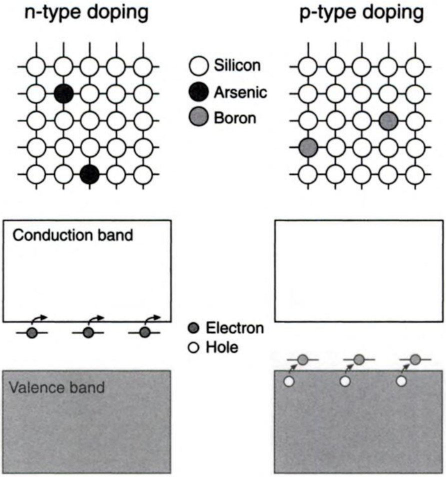


FIGURE 1.3 The conductivity of semiconducting materials can be enhanced by adding dopant atoms that inject more electrons into the conduction band. The dopant atoms have extra electrons, relative to the atoms of the bulk material. These sit in energy levels in the band gap, close to the bottom of the conduction band, and can be readily excited thermally into this band. This is called n-type doping. Alternatively, dopant atoms with a deficit of electrons provide energy levels into which electrons from the conduction band can jump, leaving behind mobile holes (which can be considered as positively charged pseudo-particles) in the conduction band. This is p-type doping.

vents further transfer of electrons into the valence band of the p-type region. Processes of this kind in semiconductor devices are often easier to envision in terms of a diagram of energies of the charge carriers. As charge migration across the interface sets up an electric field, the electronic energy bands across a p–n junction are tilted, effectively pulling the bands of the p-type and n-type regions out of alignment (fig. 1.4). This means that, in order to continue recombining with holes, free electrons have to first mount the slope up to the conduction band of the p-doped region, something for which they have insufficient energy. Recombination is therefore stopped.

But it can be switched on again by applying a voltage across the junction with the negative terminal attached to the n-doped side and the positive terminal to the p-doped side. This counteracts the field at the interface and pulls the bands back into alignment. Migration and recombination of electrons and holes across the interface can then take place. But if the direction of the voltage is reversed, the two charge carriers are both drawn away from the interface, so no current passes.

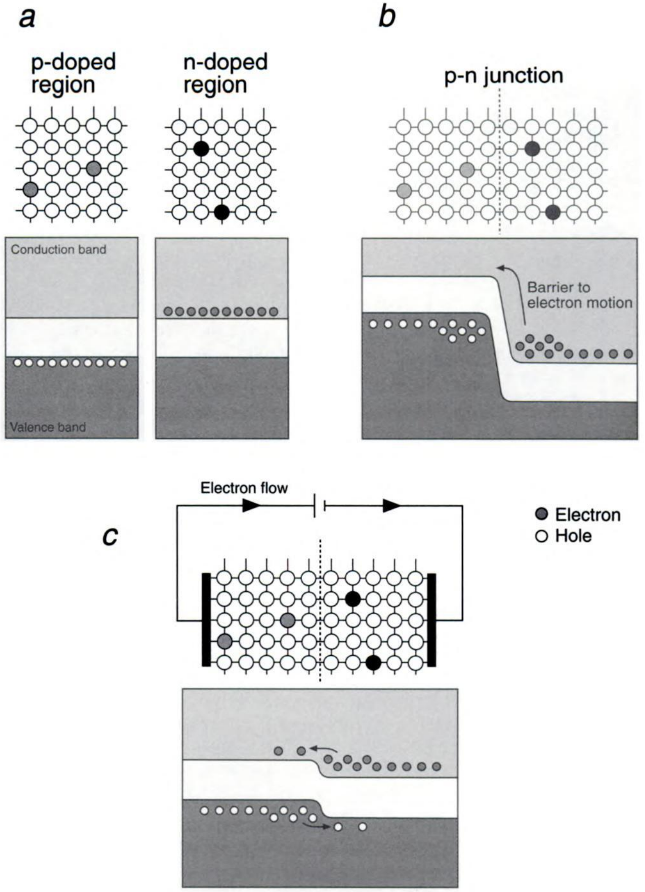


FIGURE 1.4 At a p–n junction, a p-type and n-type semiconductor (*a*) are placed back to back. Electrons in the n-type material and holes in the p-type material can meet at the interface and annihilate each other—the electrons fall into the holes, a process called recombination. The electrons lose energy in doing so, and this can be carried off as heat or light. Because there is a passage of electrons from the n-type side to the p-type side, there is a net current flow, in one direction only, across the junction until an internal electric field is set up that opposes this flow (*b*). By applying a voltage across the junction to counteract the internal field, the flow of charge is resumed (*c*). This allows a p–n junction to behave as a diode.

The p–n junction is therefore a kind of gate which lets current through in one direction but not in the other. This kind of behavior is called rectification, and is characteristic of a device called a diode.

Wired for Light

The transmission of information via pulses of light is a technology far older than electronic communication: it was used by the ancient Greeks, whose winking heliographs turned the Sun's rays into a coded photonic signal. Nor did this mode of communication cease at sunset; beacons burning on hilltops would also broadcast a message far and wide. But this approach needed an efficient system of relays to get over the horizon. Today we can channel light signals right around the world by using optical fibers, wires that carry light rather than electricity.

One advantage of transmitting information in this way is that optical fibres can potentially carry much more information than copper wires. Imagine all of the telephone conversations taking place across the United States at any one instant passing between your fingertips. That's one busy wire! If you have in mind a copper telecommunications cable, you can forget it—you'd be unable to get both arms around the cable needed to carry that much information. But in theory, a single optical fiber can do the job with room to spare—it can carry up to twenty-five trillion bits per second, one of those numbers too large to be meaningful (unless we can accommodate the awesome thought of all those chattering voices). In practice, however, the capacity of optical fibers falls considerably short of this theoretical maximum, although it still exceeds that of current-carrying wires. The very first transatlantic optical telephone cable, which was installed by the AT&T Bell Corporation and began operating in 1988, straightaway boosted the number of phone conversations that a single cable could carry by a factor of four, relative to its electronic counterparts. Fibers for carrying optical signals are now rapidly replacing electrical cables for all long-distance communications.

I should say a few words about light itself at this point. It is an electromagnetic wave, in the form of oscillating electric and magnetic fields perpendicular to one another. The frequency of these undulations is related to the wavelength: the higher the frequency, the shorter the wavelength. Within the visible spectrum, red light has the longest wavelength (around 700 nanometers) and violet the shortest (around 400 nanometers). At still longer wavelengths are infrared waves (0.8 nanometers to hundreds of micrometers), then microwaves (millimeters to centimeters) and radio waves (up to many kilometers). And beyond the short-wavelength (violet) end of the visible spectrum are ultraviolet light, X rays and gamma rays. The wavelengths of X rays are typically of the same order as the distance between two atoms in a crystal.

But within the quantum-mechanical description of light that was developed in the early part of this century, it has an alternative character: a collection of "light particles" called *photons*. These are discrete packets of electromagnetic energy, each with a characteristic wavelength and frequency. The energy of a photon increases in proportion to its frequency. In talking about photonic technology, I will sometimes need to use the wave picture of light and sometimes the particle picture. They are simply two ways of describing the same thing.

Optical fibers work by trapping light within a transparent central core (generally of silica glass, in essence the same material as is used for windowpanes) which is surrounded by a cladding of a material with different optical properties—specifically, with a lower refractive index than the core. The refractive index of a material is the ratio of the speed of light in a vacuum to the speed of light in the material. When light passes from one material into another with a different refractive index, the differing speeds cause light rays to bend at the interface: this is the phenomenon of refraction, which causes the distortion of objects seen through water. If the angle at which a light ray impinges on an interface between two substances of different refractive index is oblique enough, it can be completely reflected; this angle of total reflection depends on the difference in refractive index. An optical fiber confines light by total reflection—rays in the core are reflected back completely when they encounter the interface with the cladding, and the rays therefore bounce back and forth along the axis of the fiber (fig. 1.5).

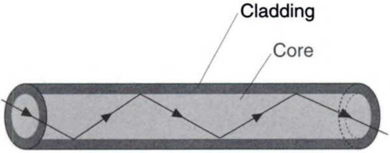


FIGURE 1.5 Optical fibers confine light through total reflection at the walls, a consequence of the difference in refractive index between the fiber core and its cladding.

Silica-glass optical fibers are made by inserting a glass rod into a glass tube with a slightly lower refractive index. This composite, called a preform, is then heated and drawn out into very thin fibers, as little as a tenth of a millimeter in diameter. Preforms are now routinely made from high-purity glass by burning a volatile silicon compound such as silicon tetrachloride to deposit a cladding of silica (silicon dioxide) on the surface of an inner glass core. A difference in refractive index between the core and the cladding is created by doping the former with germanium or phosphorus oxides, or the latter with fluorine. Because of the abrupt change in refractive index between core and cladding, these are called stepped-index fibers.

Although a single optical fiber can replace several copper cables, they are more expensive to produce, and the economics depend rather critically on the performance—the transmission efficiency—of the fibers. One of the key issues is the extent to which the intensity of a light signal becomes reduced over long transmission distances; this is equivalent to electrical losses suffered in electrical cables due to resistive heating. Light is lost from optical fibers primarily by being scattered by impurities and defects within the core material. Some scattering is inevitable in glass fibers, because the disordered atomic structure sets up small, random variations in composition and density which act as scatterers. But because the extent of scattering increases as the wavelength of the light decreases, this is not too great a limitation for optical communications that use infrared signals, which have a longer wavelength than visible light. And although impurities such as metals, water, and air bubbles are common in window glass, the more advanced procedures used to make glass fibers reduce or even eliminate these.

Some light is also lost by absorption by the glass. All materials absorb radiation at frequencies that match those with which the bonds between the constituent atoms vibrate, since light at these frequencies excites the bonds into resonant vibration. For silica glass these resonant absorption frequencies correspond to wavelengths of about 8 to 15 micrometers, in the near-infrared part of the spectrum; but absorption already starts to become significant at rather shorter wavelengths, so silica glass fibers can be used only for infrared signals with wavelengths no longer than about 2.5 micrometers. To make optical fibers that can carry longer-wavelength signals, the only option is to use another material with markedly different bond-vibration frequencies. Such materials are not common, since many interatomic bonds vibrate in about the same frequency range. Some metal fluorides, however, have vibration frequencies at considerably longer infrared wavelengths: zirconium tetrafluoride, for instance, exhibits resonant absorption only at wavelengths of 17 to 25 micrometers. Glasses made from mixtures of metal fluorides can provide optical fibers capable of conveying near-infrared signals, in the wavelength range from 0.3 to 8 micrometers. Because these materials have a lower melting temperature than silica, a different fabrication route becomes possible in which the core and cladding are melted in separate crucibles and drawn out together into strands from concentric nozzles. Metal fluoride glass optical fibers should theoretically be able to attain a transparency around twenty times smaller than that of the best silica glass fibers currently available. You would be able to see through window panes 200 kilometers thick if they were made with this degree of transparency. But at present we are far from this theoretical limit—the best fluoride glass fibers are only half as transparent as the best silica fibers.

Zirconium fluoride has also been used in *crystalline* optical fibers, which, because they have an ordered atomic structure, contain none of the random compositional fluctuations that scatter light in glass fibers. Crystalline fibers have also been fabricated from arsenic triselenide and potassium iodide, which possess large atoms connected by bonds with sluggish, low-frequency vibrations. Arsenic triselenide remains transparent to wavelengths up to about 10 micrometers. But making crystalline fibers that are free from defects is a slow and expensive process; the fibers have to be grown as single crystals, for which the growth rates can be as small as a few centimeters per minute. So this approach faces formidable obstacles before it can become practical.

Another problem that crops up in fiber-optic communications is that light pulses have a tendency to be smeared out as they travel through a stepped-index marily a triumph for physicists; but its practical application, especially in information technology, quickly became a challenge that required the expertise of materials scientists. The earliest lasers were bulky devices of bench-top size, but to integrate lasers into microelectronics they have to be much more compact—small enough, indeed, to fit on a silicon chip.

The light that lasers produce is not the same as that given out by an electric bulb or by the Sun—you could call it "designer" light. The photons in a sunbeam oscillate out of step with one another: this is called *incoherent light*. But in laser light, the oscillations are all synchronized: the light is *coherent*. This synchrony means that the photons do not disperse as they do in a beam of normal light; to put it crudely, they do not tread on each other's toes because they are all in step. The consequence is that a laser beam does not spread out in the same way that normal light does; it remains pencil-thin over long distances.

The physical principles that give rise to the emission of coherent light were elucidated in the first half of this century, by Albert Einstein among others. But not until 1954 were these principles put into practice in a working device that emitted coherent radiation. This device, developed by Charles Townes of the University of California at Berkeley, was not really a laser at all but a *maser*, which generated coherent microwave radiation rather than light. Townes and others strove in the ensuing years to create devices that would emit coherent light at visible wavelengths, and the first genuine laser was demonstrated in 1960 by H. Maiman, who obtained red laser light from a ruby laser.

A laser produces coherent light by *stimulated* emission of photons from excited atoms or molecules. An excited atom or molecule (which has an energy greater than its equilibrium thermal energy) can cast off this excess energy packaged in a photon—this is called radiative decay. The rules of quantum mechanics dictate that the energies of molecules cannot take arbitrary values, but are restricted to a set of discrete *energy levels* between which there are gaps of forbidden energies (these same rules give rise to the discrete energy bands and band gaps in solids). The energy levels define the rungs of an energy ladder (fig. 1.7). Emission from an excited species will inevitably take place spontaneously sooner or later—an excited species cannot hold on to extra energy indefinitely, just as a red-hot poker will gradually lose its heat. This radiative decay is governed by chance, like the decay of a radioactive nucleus. But emission can also be *induced* by exposing the excited species to light of the same wavelength as the photon that will be emitted. This is called stimulated emission (fig. 1.7). Its origin lies in quantum mechanics, which says that the probability of emission is enhanced in the presence of an electromagnetic field oscillating at the emission frequency. The field stimulates the excited species into resonant oscillation, much as a sung note will excite an undamped piano string into vibration. The resonant oscillation is synchronized with the stimulating oscillation—that is, the electromagnetic undulations of a photon of stimulated emission are in step with those that stimulated it. The two photons are, in other words, coherent.

The phenomenon of stimulated emission leads to the possibility of a "feedback" process for accelerating the radiative decay of a whole ensemble of excited

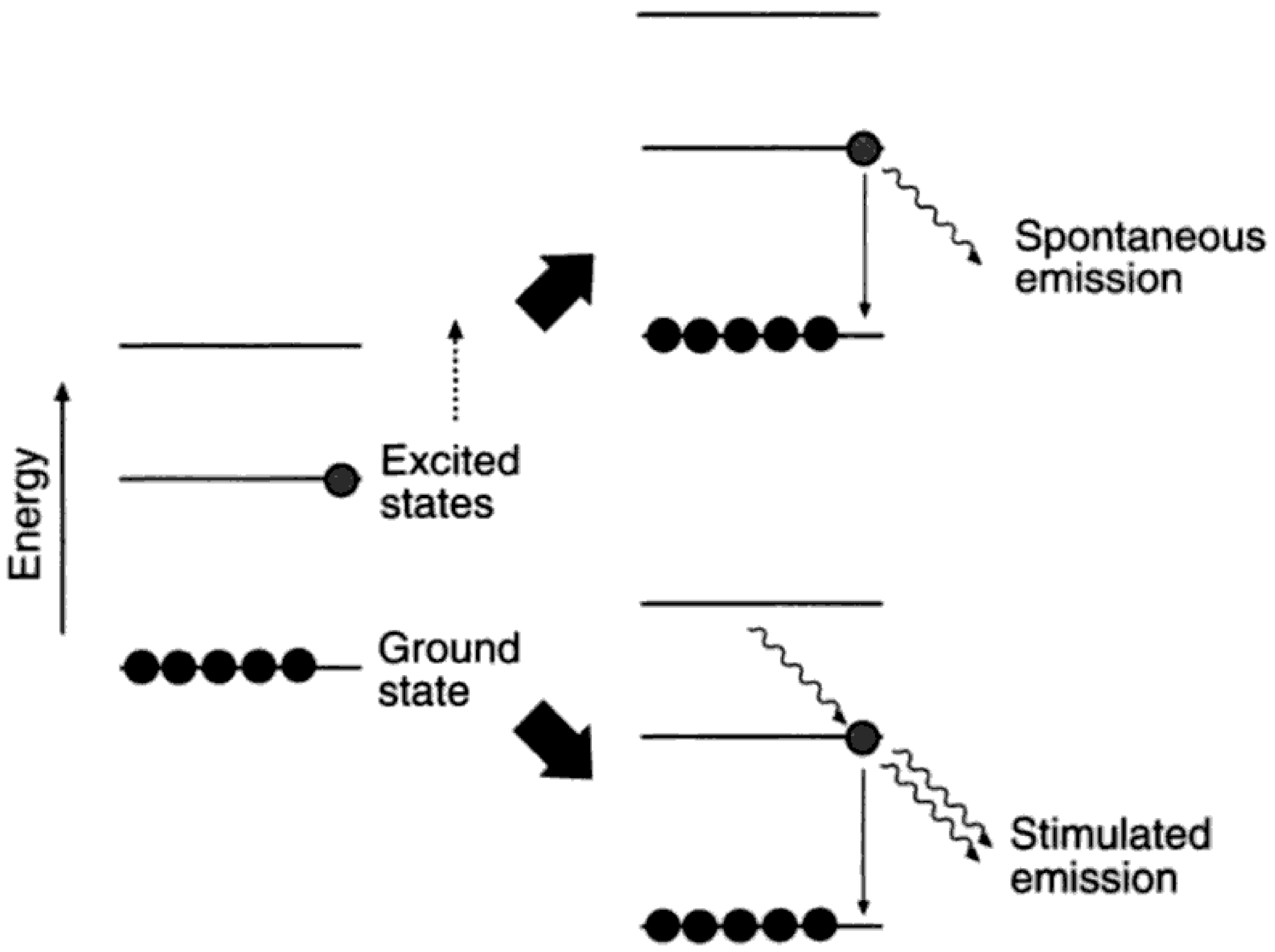


FIGURE 1.7 The energies of atoms and molecules are quantized—they can take only certain discrete values. An atom or molecule that is excited—for example, by absorbing a photon or by collisions—to a high-energy state will ultimately decay back to its original (ground) state by throwing off its excess energy. The energy can be dissipated as heat, carried off by collisions, or radiated in the form of a photon. The latter is called spontaneous emission (top). A photon can stimulate the decay of an excited atom or molecule if the decay generates a second photon of the same energy. This process, called stimulated emission (bottom), is the basis of laser action. The stimulated photon has its electromagnetic oscillations in step with the stimulating photon—they are coherent.

species. If a single excited entity undergoes spontaneous emission, the photon that it emits has the potential to *stimulate* emission from another excited molecule. This then creates two (coherent) photons, which can stimulate emission of two more and so forth. In a laser, the emitting ("active") material, which contains energetically excited atoms or molecules, is enclosed in a cavity between two mirrors that reflect emitted photons back into the cavity. This ensures that, once emission occurs, the light stays within the active material to stimulate more emission rather than escaping out into space. As more and more emission is stimulated, more photons bounce back and forth in the cavity until very rapidly all of the excited species undergo radiative decay in an avalanche that produces a burst of coherent light (fig. 1.8). Of course, if we are to make use of this light it has to be able to get out of the cavity somehow, so one of the mirrors is only partially reflective: enough to stimulate the avalanche, while being transparent enough to let the burst of laser light escape.

In order to create this avalanche process, one needs to start with a large number of excited species. This, however, is not a stable situation. An ensemble of molecules at thermal equilibrium with its surroundings contains a distribution of energies; the number of molecules with energies greater than average falls off very rapidly with increasing energy. So at equilibrium there are ever fewer molecules on each successive rung of the energy ladder. But for the avalanche of stimulated

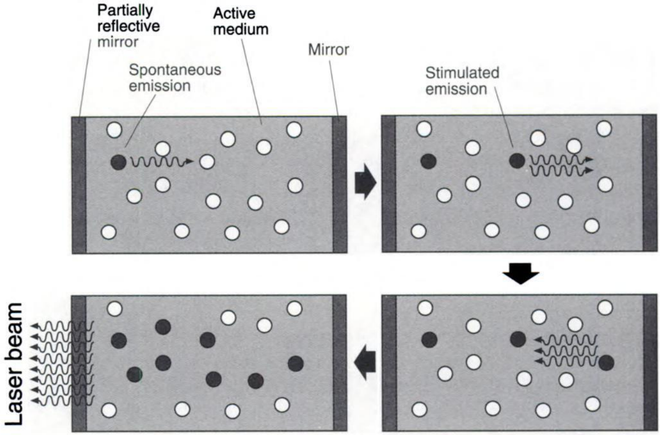


FIGURE 1.8 Stimulated emission is turned into laser action by confining the photons emitted from an excited ("active") material within a cavity. This leads to an avalanche-like amplification of the stimulated emission process—each photon stimulates the emission of others as they bounce back and forth in the cavity, and all of them are coherent. A burst of light is produced, which escapes through a partially mirrored cavity wall.

emission that leads to laser light, we need a disproportionate number of excited species perched on an upper energy level. In other words, rather than the energy ladder being populated in decreasing numbers the higher we go, it must have a region in which the population increases with increasing energy. This highly out-of-equilibrium situation is called a population inversion. It can be brought about by applying some kind of energetic pulse to kick a large number of species up onto a higher rung. In many lasers the pulse is an optical one: a pulse of light is applied through the active medium, and the absorption of this light creates many excited species in a population inversion. Sometimes laser pulses are themselves used to drive other lasers by supplying this optical kick. Population inversions can also be created electronically, while in lasers in which the active medium is a gas, atomic collisions may do the job.

The earliest lasers used a variety of active materials: Maiman's laser used a ruby crystal, which can be excited to emit red light, while other lasers used gases such as carbon dioxide or argon, or liquids such as solutions of dyes. The ruby laser was the forerunner of one of the mainstay lasers of modern technology, the neodymium-YAG laser. In this device, the active material is closely related to ruby, being a garnet mineral containing the elements yttrium and aluminum, to which small quantities of the "dopant" neodymium are added. The neodymium-YAG laser provides high-power emission at infrared wavelengths.

Drawing Light from a Well

For photonic technology, the attractive features of laser light are that it has a very narrow spread of photon frequencies (it is single-colored, or monochromatic), that it is coherent and can be emitted as a parallel-sided beam, and most of all that it can be modulated (switched on and off) extremely rapidly. But the high-power lasers used for bench-top science are of little value for a technology in which one wants devices little bigger than a bacterium. To make miniature lasers for photonic information processing, it was necessary to find a material that could be stimulated by electronic means to emit laser light efficiently from cavities of microscopic dimensions.

Lasing action in any medium is achieved by setting up a population inversion. Semiconducting materials have an electronic structure that allows one to do this electronically, by simultaneously feeding a disproportionate number of electrons into the conduction band and holes into the valence band. Provided that these charge carriers are mobile enough to migrate through the material until they encounter one another, the excess electrons can then fall back into the excess holes, releasing a photon as they do so: this is the process of electron-hole recombination. It happens spontaneously, but can also be stimulated by a photon of an energy equal to that which is lost in recombination.

To make lasers that are compatible with existing microelectronic circuitry, one would ideally like to be able to use silicon as the active material. But it is very hard to make silicon emit light efficiently, for reasons that I shall discuss later. So photonics has relied so far on different kinds of semiconducting materials, and in particular on alloys of elements from columns III and V of the periodic table, called III–V semiconductors. These materials are good light emitters. The most widely used of III–V semiconductors are the binary alloy gallium arsenide (GaAs) and the ternary (three-component) alloy gallium aluminum arsenide (GaAlAs).

One can think of a III–V semiconductor as an extreme form of both n- and p-type doped silicon, in which the dopants have essentially replaced all the silicon. As elements from group III have three valence electrons and those from group V have five, an alloy of an equal amount of the two elements has, on average, four valence electrons per atom, just like silicon. Such an alloy has an electronic structure that is entirely analogous to silicon—a filled valence band separated from an empty conduction band by a small band gap—but with the difference that the material can absorb and emit light efficiently in the near-infrared region of the spectrum (that is, between wavelengths of around 800 to 1,500 nanometers). This is because photons with this wavelength have energies corresponding to the size of the band gap, so an electron can be shunted between the valence and conduction bands by absorption and emission of one such photon. The precise width of the band gap, and thus the color of the emission, can be adjusted by varying the composition of the material, either by changing the relative proportions of the two elements or by adding additional elements from the same groups of the periodic table. Alternatively, composite "designer" materials with a specified band gap can be engineered by sequentially depositing very thin layers of different semiconductor alloys in sandwich structures, an approach called band-gap engineering. The ability to tune the color of emission more or less continuously by altering the nature of the emitting material is one of the great advantages of III–V semiconductors as photonic materials.

To make a laser based on recombination processes within one of these semiconductors, one has to be able to control the recombination electrically (in other words, to switch the laser emission on and off) and to find a way of triggering stimulated emission (giving coherent laser light) from the spontaneous emission generated by random recombination events. What the first requirement means in practice is that one has to be able to keep the electrons and holes separated until one wishes them to begin recombining. As we saw earlier, a semiconductor can be given excess electrons in the conduction band by n-type doping, and can be enriched in holes in the valence band by p-type doping. Semiconductor lasers generally contain electron and hole reservoirs, consisting of thin films of n-doped and p-doped semiconductor alloys, between which is sandwiched a layer of another semiconductor with a smaller band gap. These structures are called quantum wells, because their well-like band structure (fig. 1.9) has its origin in quantum

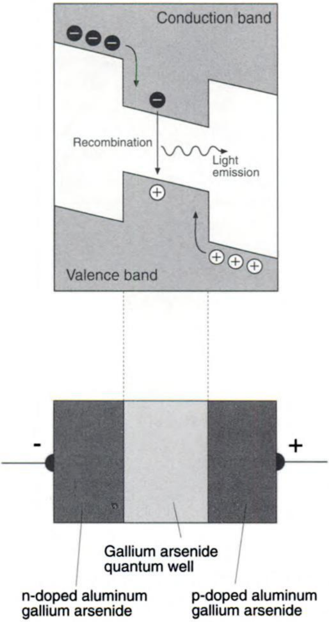

Conduction band
Recombination
Light emission
Valence band
-
+
Gallium arsenide quantum well
n-doped aluminum gallium arsenide
p-doped aluminum gallium arsenide

FIGURE 1.9 In semiconductor diode lasers, a "quantum well" of a photonic semiconductor such as gallium arsenide is sandwiched between two semiconductors with a larger band gap. One side of the sandwich is n-doped, and injects electrons into the quantum well when a voltage is applied; the other is p-doped, and injects holes from the other direction. Within the well, these charge carriers are trapped by the step in band gaps, so they can recombine efficiently with consequent light emission.

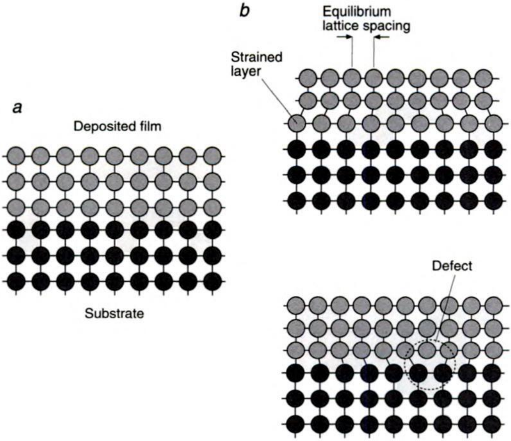


FIGURE 1.11 One crystalline material can be deposited on another without strain or defect proliferation if the spacing between atoms (the lattice spacing) is the same in both materials. This is called epitaxial growth (*a*). If, as is usually the case, the lattice spacings of the two materials do not match exactly, the deposited overlayer is strained—either compressed or expanded (*b*). This strain can be relieved by the formation of defects, where the regular crystal structure breaks down. Defects are detrimental to a semiconductor's electrical conductivity.

"Epitaxy" refers to the relationship between the atomic structure of the deposited film and that of the substrate on which it grows. To obtain good electronic properties in the semiconductor films, they must have highly ordered, crystalline structures. Defects such as misalignments of atoms degrade the conductivity of the material, for example by trapping charge carriers in their vicinity. If the lattice spacing of atoms in the deposited film matches that in the substrate, the film is said to be epitaxial (fig. 1.11*a*). This situation encourages the growth of highly ordered films, because it means that atoms in the deposited film can form bonds with atoms on the surface of the substrate without having to distort the normal crystal structure. But when the film and the substrate are comprised of different materials, their equilibrium lattice spacings will generally differ, and atoms of the two materials at the interface can then bond to one another only at the cost of either altering the lattice spacing in the first few layers of the overlying film (which imposes a strain on the material) or putting up with a few defects every so often to absorb the mismatch (fig. 1.11*b*).

If epitaxial growth is possible, then the substrate can act as a template which ensures that the deposited film grows in an ordered manner. Atoms will not necessarily impinge on the surface in exactly the right place for forming an epitaxial layer, however, and so the substrate is generally kept hot so that the deposited

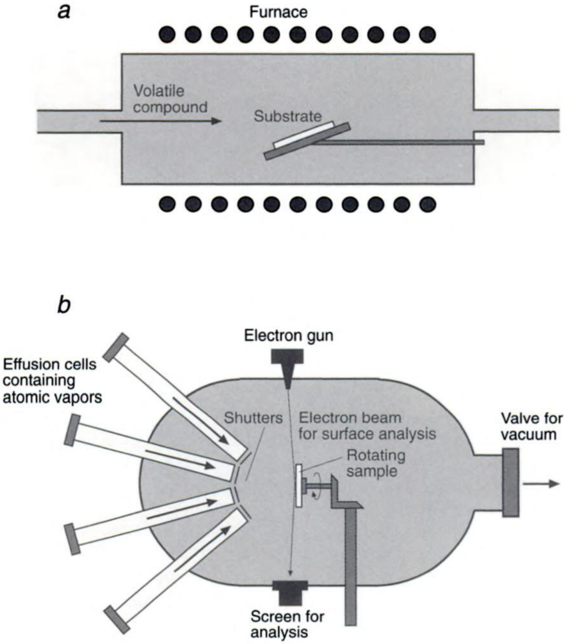


FIGURE 1.12 Microscopic semiconductor devices are generally prepared by depositing thin films from vapors of the relevant elements onto a substrate. In chemical vapor deposition (CVD; *a*) each element is introduced to a heated vacuum chamber in the form of a volatile compound, which is broken down into its constituent atoms or small molecular units by the heat of the furnace. In molecular beam epitaxy (MBE; *b*) the elements are introduced in the form of a pure elemental vapor, generated by heating a lump of the pure element in a separate chamber. The composition of the deposited film is controlled by varying the mix of vapors in the deposition chamber. These techniques allow the thickness of the deposited films to be specified with a precision of virtually a single atomic layer.

atoms have enough thermal energy to shuffle around into the right positions. In vapor-phase epitaxy (VPE, a form of the more general technique called chemical vapor deposition) the substrate is placed in a chamber under high vacuum, to eliminate gaseous contaminants that might otherwise stick to the surface and degrade the purity of a deposited film. Volatile compounds of the elements to be deposited are then injected as vapors into the chamber (fig. 1.12*a*). For instance, to deposit films of an arsenic-containing alloy, the arsenic atoms might be provided in the form of the volatile compound arsine, AsH_3. The deposition chamber is placed inside a furnace, which raises the temperature to a point where the volatile compounds decompose into their constituent parts. If the substrate is kept cooler than the rest of the furnace, the atoms in the vapor will stick to it to form a film. The composition of the film can be controlled by injecting different amounts of the various volatile compounds.

Early forms of vapor-phase epitaxy had the drawback that the volatile compounds used were highly toxic, necessitating careful and expensive safety measures. A more recent variant, called metalorganic chemical vapor deposition (MOCVD), makes use of somewhat less dangerous compounds in which the metallic or "semimetallic" elements from group III of the periodic table are introduced in the form of organometallic compounds, with organic groups bound to the metals. Aluminum, for instance, is supplied in the form of trimethylaluminum, $(CH_3)_3Al$.

Molecular beam epitaxy (MBE) differs from these approaches in that the material to be deposited is supplied in the form of beams of the pure elements, made by heating lumps of each material until they start to vaporize. Each element is held within a separate cell, called an effusion cell, and the atomic vapor passes down the cells and exits from a port at the far end into a chamber where the substrate sits under a high vacuum. The flux of material out of the effusion cells is monitored very accurately, and can be controlled by varying the amount, temperature, and position of the source material within the cell and by shutters at the output ports. The composition of the films deposited on the substrate is determined by the relative sizes of the fluxes of different elemental beams (fig. 1.12*b*).

Both of these approaches have their strong points and their drawbacks. In vapor-phase epitaxy the fact that there is a relatively high pressure of gases in the deposition chamber creates problems. First, it prevents one from being able to monitor the structure and composition of the growing film by electron diffraction (see chapter 10), as is done in the MBE technique, because the vapor scatters the electron beam. Second, the deposition process cannot be controlled too precisely, in part because of turbulence in the vapor. On the other hand, the technique allows large-area films to be grown rapidly. Films grown by MBE are deposited much more slowly, but the high-vacuum conditions afford thinner films with very smooth interfaces between layers. Recently there have been attempts to combine the two techniques so as to share their advantages while minimizing the disadvantages, for example by depositing the kinds of compounds used in VPE using the effusion cells and high-vacuum environment of MBE.

The control over the composition and thickness of epitaxial layers offered by these techniques opens up tremendous possibilities for band-gap engineering. The band gap of very thin quantum-well structures is determined not only by the composition of the well layer but its thickness too: when charge carriers are confined within narrow wells, quantum-mechanical effects shift their energy bands. So the band gap, and hence the color of light emission, of III–V semiconductor photonic heterostructures can be varied simply by changing their dimensions. Moreover, films can be deposited with band gaps that vary smoothly in the lateral direction by gradually changing the composition of the mix of elements deposited.

In the early days of epitaxial growth of heterostructures it was thought that perfect epitaxy would be possible only when the lattice constants—the distance between neighboring atoms in the crystal structure—were the same for the substrate and the overlying layer. This lattice matching is possible for GaAs/AlGaAs heterostructures, but the lattice constant of silicon is rather different from that of the III–V alloys, suggesting that films of the latter would have to incorporate defects to make up for the mismatch, when grown on silicon. This would pose a serious problem for the integration of III–V semiconductor devices with silicon technology. But advances in epitaxial growth technology have now made it possible to marry layers with mismatched lattice constants in a way that enables them to take up the strain of the mismatch without suffering badly from defects. Gallium arsenide, for example, can now be grown on silicon without too great a proliferation of defects, although some problems remain to be overcome before such composite sandwiches become technologically useful.

Once layers of semiconductors have been deposited on a substrate, they need to be cut up, sandwich-style, into discrete devices. For both photonic and electronic devices, the cutting is done by various kinds of etching agent, and is restricted to selected areas by coating the topmost layer of the flat multilayer films with a patterned "resist" that prevents etching of the regions below. This resist is generally a photosensitive polymer containing chemical groups that form cross-linking bonds when excited by light. When a uniform polymer film spread on the top surface is irradiated with light through a mask patterned as a "negative" of the circuit structure that one wants to etch, those parts of the polymer film that are irradiated become cross-linked, making them insoluble; the rest of the film is washed away with a solvent. This leaves a patterned polymer film on the sandwich surface, which acts as a protective coat against the etching agent (fig. 1.13).

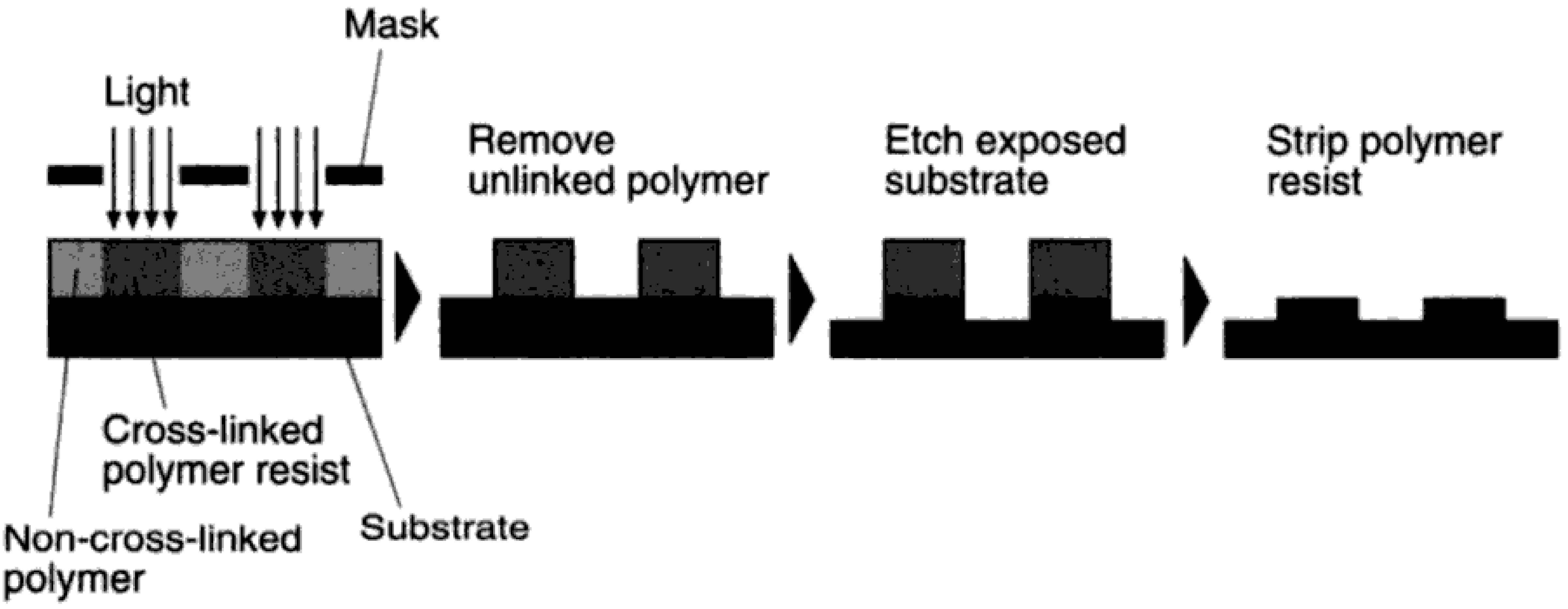


FIGURE 1.13 Device structures are carved into semiconductors by means of photolithography. A polymer film is deposited on the surface of the semiconductor, and illumination through a patterned mask fixes the pattern into the film—either by cross-linking of the irradiated parts of the polymer to form a robust "negative resist" or by breaking bonds in the irradiated parts, leaving behind a "positive resist." The non-cross-linked parts of the polymer film are then washed away, and the patterned resist confers protection to the semiconductor below from an etching process, which commonly involves exposure to a reactive plasma. After etching is completed, the resist is broken down chemically.

An agent such as a plasma of oxygen ions is then used to scour away the exposed parts of the material, leaving behind the patterned heterostructure capped with the polymer photoresist, which is subsequently removed by chemical means. This approach is said to involve a negative resist, because the pattern of the resist is the negative of that cut into the mask. Positive resists are also used, in which exposure to light induces a chemical reaction that makes the film *more* soluble, leaving behind a replica of the mask. These procedures, collectively called photolithography, can be used to inscribe extremely narrow surface patterns: the smallest features presently achieved on commercial chips are just 350 nanometers across. The lower limit to this resolution is set by the wavelength of the light used to irradiate the resist: by using ultraviolet light, researchers at the Massachusetts Institute of Technology have developed a photolithographic process that can inscribe features 200 nanometers across. But as miniaturization of integrated electronic circuits continues apace, even this is not enough: the Intel Corporation hopes by the year 2001 to be able to make silicon chips whose smallest features measure just 180 nanometers, while the dream for the further future is to push this limit to 100 nanometers. To do so, X rays rather than UV light will have to be used to irradiate the films, and that requires components for X-ray optics, such as lenses and lasers, that are still in the early stages of development.

Small Is Beautiful

Despite their proven usefulness in optoelectronic communications systems, laser diodes have a drawback—they are big beasts. To fully integrate photonics with electronics, it will be necessary to fit semiconductor lasers on a chip. In very recent years, new kinds of ultrasmall semiconductor lasers called microlasers have been constructed. The crucial aspect of most of these devices is that they emit laser light perpendicular to the layers of the sandwich structure, rather than parallel to the layers like conventional laser diodes. This difference means that the laser structure occupies less surface area, so that many more can be packed onto a chip. The development of surface-emitting structures was pioneered by Kenichi Iga and coworkers at the Tokyo Institute of Technology in the late 1970s, and led to efficient surface-emitting microlasers in the 1980s.

Typically, a surface-emitting microlaser is a cylindrical structure in which many disk-shaped layers of III–V semiconductors are stacked like coins (fig. 1.14). The principles of the laser's operation are much the same as those of a laser diode: the active region is a quantum well sandwiched between layers that inject electrons and holes into the well. But in the surface-emitting microlaser there are commonly several such wells stacked on top of each other in the active region. And most importantly, the wells are very thin—perhaps just 10 nanometers across—to keep the laser's power requirements as small as possible. This is important if they are to be used in their millions as vast emitting arrays. It means, however, that for lasing action to be achieved, the photons emitted in the active layers must bounce back and forth within them many more times than in conventional laser diodes to induce stimulated emission. And that in turn means that the semiconductor lasers is ultimately set by the optical properties of the materials from which it is made.

But a new kind of laser devised in 1994 by Federico Capasso and colleagues at AT&T Bell Laboratories has done away with this restriction. In this laser, called a quantum-cascade laser (QCL), the color of the laser light is controlled by the physical dimensions of the component parts, which can be varied at will across a wide range. The device achieves this flexibility by virtue of a lasing mechanism that is fundamentally different from that of all preceding semiconductor lasers. We saw earlier that light emission in these conventional devices arises from recombination of electrons and holes injected into the lasing cavity. In the QCL, in contrast, the emission of light relies on the injection of electrons alone.

The device consists of a sequence of about five hundred layers of doped III–V semiconductors, laid down by molecular beam epitaxy. Many lasing "cells" repeat throughout this sequence, each consisting of three narrow quantum wells of indium gallium arsenide sandwiched between indium aluminum arsenide barriers (fig. 1.16). The energy levels in the quantum wells are determined by their thickness, and they form a three-step staircase in each lasing cell, down which electrons cascade (from left to right in the figure) with the consequent emission of photons. The photon is emitted during the first step, when the electrons tunnel through the barrier from the left-most well to the middle well, which has a lower energy level. To turn this photon emission into laser emission, it must be able to stimulate the decay of other electrons down this step, and this in turn relies on maintaining a population inversion between the first and second wells, so that there are more electrons in the former than in the latter. This is the function of the third, right-most well, which siphons off the electrons from the middle well because it has an energy level just a little lower.

The electrons are supplied by an electron-injecting region to the left of the first well, and can tunnel from the third well to an electron-collecting region to the right. Both of these regions are made of a "graded alloy," comprising a multilayer sequence (a superlattice) of indium aluminum arsenide and indium gallium arsenide layers doped with silicon, whose band gaps are graded such that they provide a gentle "sloping" electric field down which the electrons pass to the next cell.

The wavelength of the laser light emitted by the QCL is determined by the difference between the energy levels in the first and second wells of each cell, which varies according to their thicknesses. Capasso and colleagues have demonstrated laser emission at a wavelength of 4.2 micrometers, well into the infrared; but there is no obvious reason why, with suitable engineering of the band gaps, it should not be possible to make visible-light lasers this way. At present, the QCL needs to be cooled to around minus 170 degrees Celsius to operate, and even then it produces only pulsed light rather than the continuous laser light needed for many applications. But the crucial point is that it shows that new laser colors can be achieved not by chemical tinkering (by altering the composition, and thus the band gap, of the active medium) but by physical engineering (altering the quantum-well thickness). As Capasso has put it, we are no longer "slaves of the band gap."

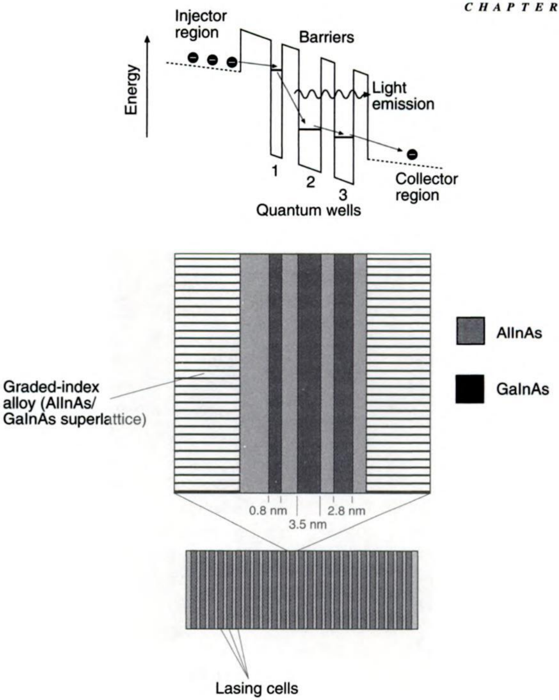


FIGURE 1.16 The quantum-cascade laser devised by Federico Capasso, Jerome Faist and colleagues at AT&T Bell Laboratories is the first solid-state laser whose emission wavelength does not depend on the chemical composition of the active lasing medium. Instead, the color of the laser light is determined by the physical dimensions of the device. The laser is a highly complex heterostructure, comprising twenty-five separate lasing cells, each containing three quantum wells of gallium indium arsenide sandwiched between barriers of aluminum indium arsenide. The whole structure has no fewer than five hundred different layers, each deposited by molecular-beam epitaxy with a precisely controlled thickness.

The quantum wells are so thin that their energy levels are influenced by the quantum-mechanical effects of confinement on charge carriers in the well. The first (left-most) well in each cell is the thinnest, and therefore has the highest energy level. Electrons injected into this well tunnel through to the lower energy level of the middle well, emitting a photon. They are then removed from the middle well by tunneling to the right-most well, which has a very slightly lower energy level—this maintains the population inversion between the first and second wells that is necessary for laser action. The electrons tunnel from the third well into a "graded alloy" superlattice of AlInAs and GaInAs to the right, through which they travel to the next cell; the band gap of this graded alloy varies smoothly so that the electrons reach the next cell with the right energy to tumble into the first (narrow) well. Thus the electrons cascade down a "staircase" in energy as they pass from left to right through the device.

Out of Step

Not all optical telecommunications systems rely on laser light, in which all of the photons are synchronized. The spontaneous emission that occurs in gallium arsenide/aluminum gallium arsenide heterostructures can be used "as is," without enclosing the device in a lasing cavity to bring about stimulated emission. Rather than a laser diode, the device is then simply a light-emitting diode (LED), and the light that it emits is incoherent—the photons are out of step.

Semiconductor LEDs are generally cheap to make and easy to switch on and off (to modulate), but they have neither the high light intensity nor the well-defined emission wavelengths of laser diodes. LEDs that emit in the infrared region, based on AlGaAs/GaAs and indium phosphide/indium gallium arsenide phosphide heterostructures, are used for short-distance optical-fiber communications links, while visible-light LEDs made from gallium arsenide phosphide find applications in display devices and as illumination sources. LEDs based on these materials are now commercially available and cover virtually the entire wavelength range from the near infrared (around 1.6 micrometers) to the green part of the spectrum. In their simplest form they consist of an active layer (say, GaAs) sandwiched between an n-doped and a p-doped layer, which inject electrons and holes respectively when a voltage is applied—just as was shown in figure 1.9. Like laser diodes, these LEDs come in several shapes or forms: some emit their light from the surface (perpendicular to the semiconductor layers), while others are edge emitters (fig. 1.17).

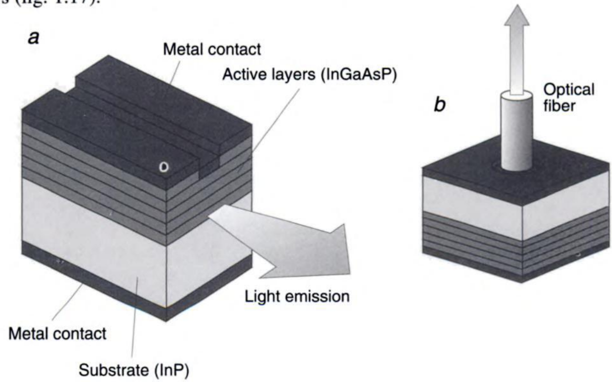


FIGURE 1.17 Photon emission due to recombination of charge carriers within a photonic semiconductor is exploited in light-emitting diodes (LEDs). Despite their low brightness and the incoherent nature of the light they emit, semiconductor LEDs are widely used in communications and display devices. The former, which emit in the infrared, commonly use indium gallium arsenide phosphide as the emitting layer. Both edge- and surface-emitting LEDs (*a* and *b*, respectively) have been fabricated; in the latter, the light can be fed efficiently into an optical fiber glued to a cavity in the surface.

Seeing the Light

At present it is necessary to convert the light-encoded information borne by optical-fiber transmission lines back into electronic form so that microelectronic circuitry at the other end can process it. This is a job performed by photodetectors. The simplest of these use photoconductive materials, in which absorbed photons excite charge carriers into mobile states so that they can carry an electrical current when a voltage is applied across the material. In so-called intrinsic photoconductors this excitation involves kicking an electron up from the material's valence band to the conduction band; in extrinsic photoconductors the electrons are excited to or from energy states of dopants. Although there are several semiconductor materials that show good photoconductivity at infrared wavelengths, photoconductive detectors are not used very much in optical telecommunications because their light sensitivity is rather low and their response is slow.

Instead, most photodetectors in information technology are photodiodes, which consist of layers of p-doped and n-doped materials. In its simplest manifestation, a photodiode is a plain old p–n junction across which a "reverse" voltage is applied: the p-doped region is given a negative voltage and the n-doped region a positive one. The result is that holes in the former and electrons in the latter are attracted away from the p–n interface, preventing recombination and creating a region depleted in mobile charge carriers on either side of the interface—this is called a "space charge" region, since the immobile countercharges that are left behind set up an electric field (fig. 1.18*a*). Provided that their energy is at least as

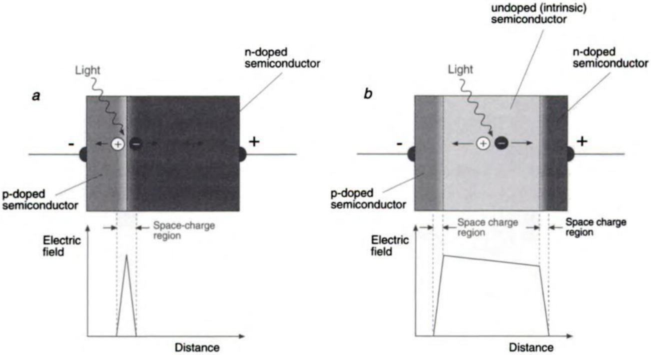


FIGURE 1.18 (*a*), Photodetectors can be constructed from p–n junctions. When a reverse bias is applied (so that the negative terminal is attached to the p-type layer and vice versa), charge carriers are drawn away from the p–n junction and a depleted region called the space-charge region is set up. When electron–hole pairs are generated in this region by absorption of a photon, they are rapidly swept outward toward the two terminals by the electric field created in the

great as the band gap, photons falling onto the material create electron–hole pairs by exciting electrons to the conduction band, and the electric field in the space-charge region will rapidly sweep the holes in one direction and the electrons in the other. So an electrical current—a photocurrent—flows in response to the light.

To maximize this photocurrent, all of the light should be absorbed in the space-charge region, which means making this region as large as possible. One way of doing so is simply to apply a large reverse voltage, but a better way is to leave a wide *un*doped ("intrinsic") region between the p and n materials: this creates a p–i–n structure (fig. 1.18*b*). Photoexcited charge carriers in the undoped region pass rapidly down the electric-field gradient set up by the space-charge regions at either end. These p–i–n photodiodes form the basis of just about all detector systems in optical telecommunications: for wavelengths from 800 to 900 nanometers (the near-infrared), doped silicon is used, while indium gallium arsenide, sometimes sandwiched between p- and n-doped indium phosphide in a heterostructure, provides photodetectors for longer wavelengths of around 1 to 1.6 micrometers.

Putting It All Together

With semiconductor-based light sources, optical fibers and photodetectors, we have all that we need for transmitting information with light. Laser diodes and LEDs provide light sources for sending the data, optical fibers are the wires that carry it, and photodetectors read out the data at the other end. At present, these are the only operations carried out with light in information technology—the difficult job of interpretation and processing the signals is left to electronic devices. This marriage of optical and electronic technology is called optoelectronics, in which light effects the communication between "thinking" electronics. In the present generation of optoelectronic telecommunications systems, the optical components are linked up to microelectronic information-processing devices in *hybrid* circuits, in which the electronic devices are mounted on silicon chips but the optical devices remain separate. This separation is not ideal—not only does it make the circuits less compact and harder to package, but it also degrades their performance (the interconnections cause signal losses) and their durability. The next challenge is to combine electronic and optical devices on a single chip.

space-charge region. (*b*), The standard photodetector in fiber-optic telecommunications technology is the p–i–n photodiode, consisting of a thick layer of an undoped semiconductor sandwiched between p- and n-doped materials. The device is again reverse biased. When light is absorbed in the undoped region, electron–hole pairs are created, and the two charge carriers are pulled in opposite directions by the applied electric field, and are again rapidly swept up when they enter the depletion regions, where the field varies sharply. Silicon is used for photodetectors sensitive to infrared wavelengths of 0.8 to 0.9 micrometers—these devices have relatively low light-to-electricity conversion efficiencies, but they are cheap. For longer wavelengths, indium gallium arsenide is used—these devices are more expensive but also more efficient than silicon-based ones.

If optical logic devices are going to talk to each other on a chip, we need to be able to wire them up. What that means is that there is a need for a way of guiding light signals around on a chip—a microscopic equivalent of the optical-fiber technology used for long-distance communications. One possibility that exists for light but not for electronics is simply to send the signals through free space—to shine a laser beam from one component to the next through air. This solution has the advantage that there is extremely little absorption of the light by the medium through which it travels; but it requires very precise alignment of components and can take the signal from place to place only in straight lines. Free-space optical signaling is very attractive for getting one integrated chip to communicate with another, because it means that there need be no physical connection between the chips.

For directing the flow of light on a chip, however, it is more usual to contain it within a "waveguide," a kind of light wire that is really nothing more than a miniaturized optical fiber. In general, waveguides consist of a transparent material with a refractive index higher than that of the surrounding medium. Just as in optical fibers, this difference in refractive index allows for total reflection at the walls of the waveguide, confining the light within. Such waveguides are commonly thin strips of transparent material deposited on a substrate; the surrounding medium is then the substrate on one face and air on the others (fig. 1.20). Waveguides can also be made from strips buried beneath layers of lower-refractive-index material. They are made by depositing an entire flat layer of waveguide material and then using photolithography to etch away part of the layer, leaving only the strips in the desired pattern. This is just the same process as is used to make copper wiring on microelectronic circuit boards. The favored waveguide materials are silica and silicate glasses, the III–V semiconductors gallium arsenide and indium phosphide, and crystalline lithium niobate doped with titanium or hydrogen ions. There is now increasing interest in developing organic polymers for optical waveguides, partly because these can conceivably be laid down using much simpler technologies such as injection molding and embossing.

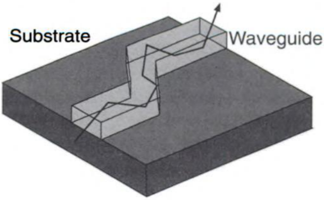


FIGURE 1.20 Waveguides on a chip are transparent channels whose refractive index differs from that of the materials on all sides. They confine light in the same way as optical fibers.

Nonlinear Thinking

To make photonic circuits, one needs devices that can perform all of the tasks of which electronic devices are capable. In general, devices that manipulate and process data are controlled or driven by a signal separate from that which carries the data; these are therefore active devices, and their need for a driving signal means that they consume power.

Just about every kind of active photonic device makes use of a material property called *optical nonlinearity*. "Nonlinear" can generally be regarded as a kind of technical term for "not obvious," in the sense that a nonlinear response of any sort is simply one that does not vary in direct proportion to the stimulus. Linear phenomena are easier to understand intuitively: the bigger the cause, the bigger the effect. If you increase the power supplied to a light bulb by turning up a dimmer switch, it gets proportionally brighter. Nonlinear responses embrace just about any other alternative to this kind of simple behavior: for example, if, as you turned up the dimmer switch, the light flashed alternately on and off, or if it got abruptly brighter and then stayed that way, or if it blew, those would be examples of a nonlinear response. All physical systems exhibit nonlinear behavior when driven hard enough: commonly the nonlinear response takes the form of saturation (where further increases in the driving force make no more difference to the output) or breakdown (where the system simply fails above a certain threshold).

Nonlinear behavior is often something to be avoided—for example, in the case of an audio amplifier, where a signal starts to distort when amplified too far. But just as rock guitarists have found this effect useful for their own purposes, so nonlinear effects in other situations can be put to (some might say more desirable) use. A diode is an example of a nonlinear electronic component: its output current remains negligibly low until the driving voltage reaches a threshold value, whereupon the output current increases sharply. This makes the diode a kind of switch, which is "off" when driven by voltages lower than the threshold and "on" for above-threshold voltages. A transistor can act as a more advanced version of a diode: a switch that also amplifies the signal through it.

A nonlinear *optical* response means that the amount of light transmitted through a substance is not proportional to the amount of incident light: doubling the latter will not necessarily double the former. Instead, the incident light actually changes the way that the material responds to light—for example, by altering the material's transparency or refractive index. Not only does this mean that the light that comes out might be rather different, in frequency for example, from that which went in; it also means that one light beam can exert a strong influence on a second as they pass through the material. In this way, nonlinear optical (NLO) materials may act as switches, in which one light beam can be used to control and direct the path of another.

At root, this nonlinear optical behavior is a consequence of the ease with which charges in the material can be shifted around (polarized) by an electric field.

These charges—electrons or ions in the NLO material—can be displaced by the oscillating electric field of a photon. Because of this rearrangement, the electric fields of subsequent photons interact with the material in a different way.

One example of such behavior is called second-harmonic generation (SHG). When light of a sufficiently high intensity is shined on a material that exhibits SHG, the transmitted beam takes on a different color from the incident beam. This may not sound so surprising at first: after all, stained-glass windows produce a kaleidoscope of colors even though illuminated with plain white light. But all that is happening here is that the glass is extracting some of the colors from the white light by absorbing them: those that are not absorbed pass through the glass and give it its color. Thus, the glass is not actually generating any colors that were not present already in the incident light. In nonlinear optical materials that exhibit SHG, on the other hand, light of one pure color can be shined onto the material, and light of a different pure color can emerge at the other side. Infrared light, for instance, can be converted to red or blue. The incident photons are converted into photons with a frequency twice as high (and thus a wavelength half as long). This frequency is the "second harmonic" of the oscillating electromagnetic field set up in the material by the incident light, just as higher harmonics can be heard in the note emitted by a guitar string or an organ pipe. In some materials a strong third harmonic can be excited instead, and the frequency of the light is then tripled.

Materials that show second-harmonic generation are very useful for expanding the range of available colors from laser light, and in particular for obtaining green and blue laser light. For many years attempts to make blue-light diode lasers were impaired by the lack of suitable materials. Although, as I mentioned earlier, II–VI alloys (particularly zinc selenide) and more recently nitride III–V materials are now filling this gap, second-harmonic generation provides a shortcut to blue laser light. Pass the infrared light (of wavelength around 860 nanometers) of a standard diode laser through a frequency-doubling material and out comes coherent blue light of twice the frequency, and thus half the wavelength. In the first demonstration of frequency doubling by SHG in 1961, a quartz crystal was used to turn red light from a ruby laser into ultraviolet. Subsequent studies of SHG at first used the natural mineral potassium dihydrogen phosphate; but today inorganic niobates, particularly lithium niobate, are the most commonly used frequency doublers, and can be grown in perfect crystals several centimeters across.

In 1995, the first commercial blue laser appeared on the market. Developed by Coherent Inc. in Santa Clara, California, in collaboration with IBM's Almaden research center in San Jose, it uses crystalline potassium niobate to double the frequency of an infrared laser. By employing a special trick called resonant doubling, the Coherent device achieves the high-intensity output needed for practical applications such as reading optical disks. Its high cost is sure to curtail widespread use, however, and most researchers now see these frequency-doubling measures as a stopgap before intrinsic blue-light lasers, perhaps based on gallium nitride, become commercially available.

Switching On

Probably the most important active component in any information-processing circuit is the switch, one of the roles played by the transistor in electronics. To achieve electronic switching, electrons can be rerouted along a conducting pathway by an applied voltage—in other words, one electronic signal is used to control another. In a truly photonic switch, one light beam will control another. But a kind of halfway house is a switch in which an electric field directs the path of a light beam. This kind of switch is an optoelectronic device, and can be fabricated by making use of the nonlinear optical effect known as the electro-optic effect, whereby the refractive index of a material is changed by an applied electric field.

A typical electro-optic switch comprises two parallel waveguides of a nonlinear optical material such as crystalline lithium niobate, which come together at a bottleneck (fig. 1.21). The waveguides are defined in the crystalline film by doping the channels (commonly with titanium) to change their refractive index relative to the surrounding material. The light in one channel can be switched back and forth between the other channel by applying a voltage across the bottleneck—this electric field polarizes the NLO material and alters the refractive indices of the two waveguides. The application of a voltage does not, as you might think, cause the light to switch channels; rather, this happens in the *absence* of an applied field, because of the nonlinear way in which the electric field of the light in one channel interacts with the charges in the other, "empty" channel at the bottleneck. The applied voltage, meanwhile, keeps the light in the same channel

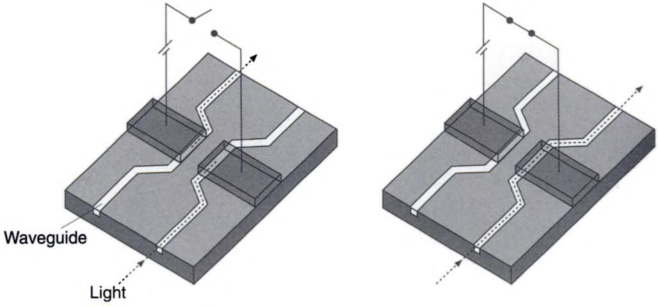


FIGURE 1.21 An electro-optic switch controls the path of a light signal through a waveguide. Light traveling through the lower waveguide interacts with the upper waveguide in a nonlinear way at the bottleneck, causing the light to switch channels. But if an electric field is applied across the bottleneck, the nonlinear interaction is modified and the light stays in the lower channel.

as it traverses the neck. Electro-optic devices of this sort are now produced commercially for integrated optoelectronic circuits.

The electro-optic effect is also exploited in devices for modulating light signals—for chopping them up into discrete, "digital" pulses or for altering the frequency of a pulsed signal. These modulators again take advantage of the effect of an electric field on the refractive index of an NLO material. In the modulators known as Mach–Zehnder interferometers, an incoming laser beam is split into two beams, each of which is directed along a separate waveguide in the NLO material before they are merged again. When a voltage is applied across one of the branches, its refractive index is altered and the light beam along that branch is slowed down slightly. As a result, when the two beams merge again their peaks may no longer coincide and so they interfere destructively. If the refractive index of one channel is changed just enough to slow the beam it carries until the peaks coincide with the troughs of the other, unperturbed beam, there will be complete destructive interference when the two beams are reunited, and the light signal will be switched "off." So by modulating the applied voltage, the light beam can be modulated with the same frequency.

These electro-optic modulators, which are again optoelectronic devices, can achieve extremely high modulation frequencies—they can switch the light beam on and off very rapidly. This means that the modulated light beam can ferry a lot of data very quickly (each pulse can be regarded as a "bit" of information). This illustrates one of the fundamental attractions of photonic computing—the speed of data transmission and processing. Commercial electro-optic modulators, which use lithium niobate as the NLO material, can reach switching rates of twenty billion times per second, while laboratory prototypes have been made that are almost four times faster than this.

But these switches are not cheap (their cost can run to tens of thousands of dollars), primarily because the crystalline NLO materials are difficult to grow, and hence expensive. This is one of the main reasons why there is so much interest in developing organic polymers that carry out the same function as the NLO crystals. Some of the polymers that have been successfully used for electro-optic switching and other photonic functions derived from nonlinear optical behavior have their NLO properties "built in" to the polymer chains, which have rather sloppy (highly polarizable) electron clouds. Several polyimides fall into this category (fig. 1.22*a* shows one such). A polyimide called Pyralin 2611D, developed by Du Pont, has been used in an electro-optic modulator that provides modulation frequencies of up to 20 billion times per second—equal to the commercial lithium niobate devices. Other NLO polymers have "optically inert" backbones (such as polyacrylates and polyurethanes) to which optically responsive organic side groups are attached (fig. 1.22*b*). Pilot-scale commercial production of modulators made from these materials is underway. But the advantages of low cost and easy processing of polymers are somewhat offset by question marks over their long-term stability—these materials may start to break down after prolonged exposure to intense light.

The nearest thing to a commercial all-optical switch at present is a device that sion down a fiber-optic network. At the other end, the signal must be demultiplexed—separated back into the original streams of data—and a NOLM can perform this separation too. For instance, the demultiplexing NOLM might pick out every fourth data bit from a stream of optical pulses carrying forty billion bits per second, thus converting this single signal into one that carries ten billion bits per second and one that carries thirty billion. This process can be regarded as the *selective* optical switching of every fourth bit in the input data stream: the control beam consists of pulses fired off at a rate of ten billion per second, so that a control pulse accompanies every fourth input bit and reroutes it.

But fiber-based all-optical switches and multiplexers have not yet left the laboratory. One key reason for this is their size: it is no easy matter to pack kilometers of fiber into a shoebox, let alone reduce them to the proportions that would fit on a chip. Researchers are now striving to reproduce the same kind of effect in microscopic semiconductor devices in which the coiled fiber is replaced by waveguides etched into photonic materials such as indium gallium arsenide phosphide. Demultiplexing in these miniaturized semiconductor devices has already been demonstrated in the laboratory.

Toward the All-Optical Computer

The battle between electronics and photonics is being waged in two major arenas. One is in information transmission, and here the photons are already starting to taste victory. For the long-distance cables that carry the data, there is no longer any doubt that optical fibers have outmatched copper wires; but if optical data transmission is to achieve its full potential, these transmission networks will have to become *all*-optical, dispensing with any form of electronic mediation to modulate, direct, or amplify the signals between the sending and receiving hardware. Only then will the kinds of transmission speeds that photonics makes possible be realized. We've seen how progress is being made on all of these fronts. The question of multiplexing of optical signals—sending several independent signals simultaneously down the same fiber—is particularly important, because it is here, as much as in the issue of modulation speed, that light-based transmission holds such promise.

The most well-developed kind of multiplexing at present is that in which each signal is assigned a different wavelength band—this is equivalent to the way in which radio broadcasting is divided into bands, although in that case the signal wavelengths are in the radio-wave part of the spectrum rather than the optical or infrared. By tuning the receiver appropriately, one particular band can be picked up without interference from the others, provided that their wavelength ranges do not overlap. A small-scale all-optical network that carries twenty wavelength-multiplexed signals, each at a different wavelength and each bearing up to ten billion bits of data per second, has been constructed in eastern Massachusetts by a collaboration involving AT&T Bell Laboratories, the Massachusetts Institute of Technology, and the Digital Equipment Corporation, and was demonstrated successfully in 1995. Other all-optical wavelength-multiplexing projects have been launched in the United States, Europe, and Japan. These projects serve to demonstrate that the technologies needed for all-optical networks are already at hand.

The second arena in which electronics and photonics compete lies at either end of the transmission network: the signal-processing systems, and in particular the computer. Here electronics still has the upper hand; indeed, the challenge posed by photonics remains rather feeble at present. The all-optical computer, in which logical data processing is carried out on photonic integrated circuits which need electricity for little more than just driving the semiconductor microlasers—this is something that remains a distant dream. For as we have seen, the photonic integrated circuit, which processes light on a chip, is by far the most immature of the new optical technologies. But when it arrives—maybe in the next five years, maybe in the next fifty—we will see computers change *qualitatively*. Not only will they be faster, but entirely new kinds of computer architecture should become possible. In others words, we will discover new ways to make machines think.

The Silicon Glow

While there is no lack of photonic materials for making light-emitting devices that can be miniaturized and coupled to silicon circuitry in hybrid structures, integration of photonic devices onto silicon chips has been hampered by the difficulties of growing the photonic materials, such as gallium arsenide, on silicon surfaces. Photodetectors, waveguides, and light modulators can all be made from silicon-based materials instead, but the light sources—semiconductor lasers and light-emitting diodes (LEDs)—cannot, because silicon does not emit light efficiently. Overcoming this obstacle would revolutionize optoelectronics by making it an all-silicon technology.

Like gallium arsenide, silicon can be made to produce photons—to luminesce—by creating electron–hole pairs which then recombine. In a simple light-emitting diode, the electrons and holes are injected by applying a voltage across the material, and the resulting light emission is known as electroluminescence. A good guide to the electroluminescent potential of a material is its efficiency of photoluminescence, in which the pair of charge carriers is excited by light absorption. In general, the energy given out by recombination is slightly less than the energy taken in by initial light absorption, so the wavelength of photoluminescence is somewhat longer (the photon is less energetic) than that of the light used to stimulate it.

Measurements of the photoluminescence of silicon in the 1950s appeared to write off the possibility of silicon-based photonics: if irradiated with visible light, silicon emitted radiation in the infrared part of the spectrum, but only very weakly. For every photon of light emitted, about a million photons had to be absorbed. This is because there are several alternative ways in which the electron–hole pair can recombine *without* emitting light. Since in silicon these processes are much more rapid than radiative recombination, most of the pairs get squandered. Radiative recombination is so slow in silicon because a peculiarity of its electronic band structure means that the net momentum of the recombined pair is different from that of the excited pair (its band gap is "indirect"). So the principle of conservation of momentum requires that some of the energy of recombination be carried away by lattice vibrations, which make up the difference in momentum. In a sense, electrons and holes have to wait for a suitable lattice vibration to pass by before they can recombine radiatively.

But this conservation rule is relaxed if the material is not perfectly crystalline, which has led to the exploration of disordered forms of silicon as photonic materials. Disordered silicon-based materials called polysilanes, which contain polymers of silicon and hydrogen, do indeed show good photoluminescence, but at the expense of the good electrical conductivity that is also central to silicon's use for microelectronics. Another way to introduce disorder without sacrificing crystallinity altogether is to introduce defects into the silicon crystal. But these must be chosen with care, since most defects actually enhance the ability of the electron–hole pair to recombine without emitting a photon. Certain defects, however, make the light-emitting transition easier.

Among them are defect atoms that have the same electronic makeup as silicon itself: carbon, germanium, and tin, which, like silicon, all have four electrons available for chemical bonding. When introduced into the silicon crystal lattice in place of a silicon atom, these are known as isoelectronic ("same-electron") defects. Efficient electroluminescence can be achieved in carbon-rich silicon, but only at the temperature of liquid nitrogen (minus 196 degrees Celsius); good electroluminescence at room temperature by isoelectronic defect doping of silicon has yet to be achieved. Another approach is to introduce defect atoms that themselves have strong luminescent properties, the idea being that the electron–hole pair will find its way to the defect and there recombine with efficient photon emission. Rare-earth metals such as erbium have been shown to enhance the luminescence properties of silicon in this way, but again the practical value of the approach remains limited.

These approaches show much promise but remain very far from providing a silicon-based device that will emit light efficiently at room temperature. This is why the discovery made by Leigh Canham of the Defence Research Agency in Malvern, U.K., has excited such interest. Canham announced in 1990 that silicon can be turned into a good emitter of light by cutting most of it away—specifically, by using an electrochemical method to etch channels throughout the material, leaving a ramified array of thin silicon filaments, like a sponge. When the husband-and-wife team of A. and I. Ulhir, working at Bell Laboratories in New Jersey, first found in 1956 that electrochemical etching of silicon produced this result, they were far from pleased, because it undermined their attempts to make smooth, polished silicon surfaces for microelectronics applications. But four decades later, this porous form of silicon may prove to be the optical material that photonic engineers have been dreaming of. A disk of porous silicon illuminated by ultraviolet light will emit an orange-red glow, without any need for careful surface treatments or sophisticated defect doping (plate 2). Indeed, the material is produced by a technology so simple that Canham can do it in a bucket.

The etching process used first by the Ulhirs and later by Canham involves immersing a silicon wafer in a bath of hydrofluoric acid and passing an electrical current through the solution, with the silicon acting as the positive electrode. Under these conditions the acid eats away at the silicon in such a way as to carve out tiny channels. As more and more of the material is eroded away, what is left is a delicate filigree of silicon consisting mostly of empty space. The remaining material forms an interconnected web of fine silicon wires, no more than a few nanometers thick (fig. 1.24). About 20,000 of these wires side by side would be no thicker than a human hair.

Porous silicon shows both photoluminescence—visible-light emission in response to irradiation—and electroluminescence, where the emission is stimulated by an applied voltage. The reason for this behavior was hotly debated when the discovery was first announced. Some researchers believed that the dissolution process was forming silicon polymers, called siloxenes, on the surfaces of the material; these compounds were already known to have luminescent properties. Others suggested that silicon hydrides at the surface were responsible. But Canham believed that the key to the effect lay with the narrowness of the silicon wires themselves. He suggested that the emission was the result of the same electron–hole recombination processes that in bulk silicon cause inefficient emission in the infrared region of the spectrum. The reason that in porous silicon the emission is both more efficient and shifted to shorter (visible) wavelengths might, Canham said, be connected with the confinement of the electrons and holes to very narrow channels.

The effects of confinement are quantum-mechanical in origin: the discrete energy levels of a quantum particle (like an electron) held in a box depend on the size of the box. Although that might seem a little strange, it can be compared with the effect on a flowing stream of water of confining it within an ever-narrower channel: as the channel gets tighter, the stream flows ever faster, and the kinetic energy of the water increases.

For a semiconductor like silicon, quantum-confinement effects bring about a change in the energies of the electronic bands: in a narrow silicon wire the conduction and valence bands are shifted farther apart, and so the band gap is increased. Consequently there is a greater energy loss in the recombination process, leading to the emission of a photon of greater energy than from the bulk material. So the luminescence shifts to shorter wavelengths as the wire gets thinner. In addition, when confined to a narrow space, electrons and holes attract each other more strongly, and so radiative recombination is able to compete more effectively with the nonradiative processes that suppress luminescence. A considerable body of experimental results now supports this interpretation.

Perhaps the most notable benefit of the quantum-confinement effect is that it allows the color of the emitted light to be tuned simply by varying the width of the silicon wires, which can be done by carrying out the etching process for different lengths of time: the longer one etches, the more silicon is eaten away and so the finer the wires are. As finer wires have a larger band gap, the color of the

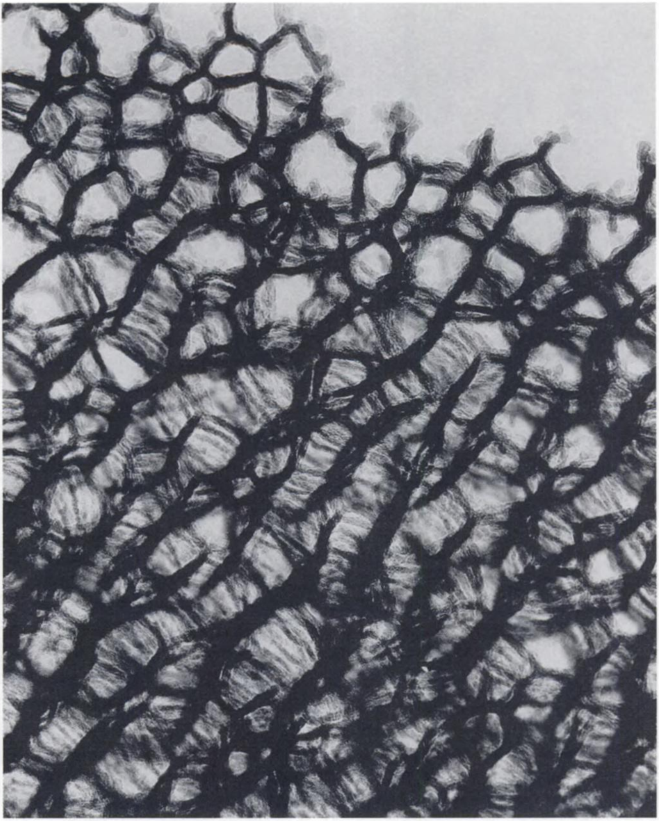


FIGURE 1.24 When etched into tiny wires no more than a few nanometers across (dark regions in this electron micrograph), silicon will emit visible light efficiently. (Photograph courtesy of Leigh Canham, Defence Research Agency, Malvern, U.K.)

luminescence moves from red to green and toward blue for more highly etched silicon.

Porous silicon that emits red, orange, yellow, and green light has now been prepared. But once the wires get too thin, they start to become so fragile that they break during the drying process, making it hard to extend the range of colors even to green. One answer to this problem is to use a process called supercritical drying, which is described in chapter 7. Canham and colleagues have used this approach to make highly porous silicon that emits uniform green photo-

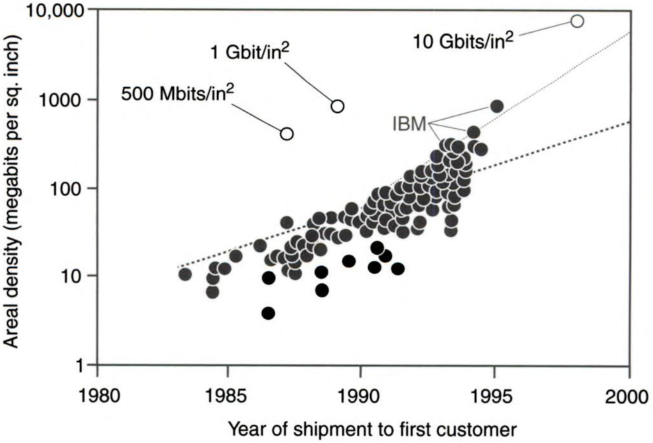


FIGURE 2.1 Over the past decade, the areal (two-dimensional) data storage density of commercial disk drives for computers has roughly doubled every three years. The bold dashed line represents a conservative estimate of the upward trend; over the last five years or so there are signs that this trend could begin to rise even more steeply (thin dashed line). The points in white represent demonstration models rather than commercial devices.

Today's memory banks, whether they be on cassette tape, on disk, on videotape, or on hard drives, are almost all magnetic: the data is recorded as a pattern in a magnetic medium. And magnetic recording technology is taking giant strides forward: to ever higher storage densities (the amount of information in a given volume of material—or, for media in which the information is stored two-dimensionally, with items of data side-by-side like words on a page, the amount in a given area), and to ever greater fidelity, better durability, and reliability. One reads, from time to time, the suggestion that the magnetic recording industry has reached the limit of its potential, that it has no further scope for improvement. This is, of course, an enticing line if one is in the business of singing the praises of a new technology for data storage (examples of which we shall encounter in this chapter); but it is scarcely accurate.

A glance at figure 2.1 should persuade you that the magnetic storage industry is anything but moribund. This figure shows the way in which the areal storage density of commercial disk drives has been changing over the past decade. Analysts argue about how steeply the curve is rising, but there is no question that the trend is upwards—to ever more amounts of data packed into an ever smaller area of disk. The ballooning of the home computer business has been powered by this trend, which allows magnetic hard drives to hold immensely sophisticated software without the need for battalions of add-on devices to expand the memory.

This is the sort of graph that industrialists like to present to show just how rosy the future looks. But its upward slope will persist only so long as fundamental research continues to push at the envelope. Magnetic recording has existed as a burgeoning commercial enterprise since the 1950s, but the advances have not been a matter of refining old ideas; instead, new ways of imprinting data have been devised, and new materials developed to contain the information revolution.

Are there *fundamental* limits to this growth? Is there anything in the physics of magnetic recording, or in the properties of the magnetic media used for storage, that will ultimately limit the rise, in the same way that some say physical laws will ultimately curtail the miniaturization and increase in speed of computers? So far, there seems to be little prospect of that. On the contrary, the potential for growth looks ever brighter, as recent years have produced entirely new approaches to data storage that could accelerate the trend even further if their technological potential can be realized.

But magnetic storage, unchallenged for decades, now has a rival. Just as photons threaten to usurp electrons for data processing, so too does light promise to become the memory medium of the future. Optical memories claim to offer not only potentially higher storage densities, but also entirely new kinds of memory, some of which operate more like our own minds than like the linear data banks of magnetic tape. They are, at present, the David to magnetism's Goliath, but the story may yet have the biblical conclusion.

Getting It Taped

Like many technologies that today we take for granted—the fax machine is another example—magnetic recording was an idea that had to wait a long time for applied science to catch up with the theory. It was mooted in the nineteenth century, and was first put into practice by the Danish engineer Valdemar Poulsen, who in 1898 invented the Telegraphone, a sound-recording device that employed steel wires as the magnetic medium. Magnetic recording became commercial only in the 1930s, however, when the German company AEG Telefunken created the Magnetophon, a device that recorded sound signals on magnetic tape developed by the German company BASF. Versions of this device found their way into German broadcasting during the Second World War, and after the war similar sound-recording devices were developed outside of Germany, notably by the Ampex company in California. The sound quality improved steadily, and domestic tape recording was given a new impetus by the invention in 1964 of the compact cassette, a tape reel enclosed in a plastic box for ease of handling.

Pointing the Way

Although all of these early magnetic-tape recorders were analog devices (in which the input signal is modulated continuously), the principles of magnetic recording are more easily understood for digital data—increasingly the prevalent means of encoding data in today's information age. Digital information—whether it represents a sound, a picture, or a stream of numbers or words—is recorded as a sequence of pulses (bits), each of which takes one of two values. This sequence of data pulses can be written as a series of 1's and 0's; in practical terms these bits are stored in a memory that can be thought of as comprising an array of switches, which are flicked one way ("on," say) to represent a "1" and the other ("off") to represent a "0."

A string of data—the words on this page, for instance—can be encoded using just 1's and 0's by creating a code in which a string of several bits represents a single character (the letter "a," for example, or a space or a comma). As there are thirty-two ways of arranging a sequence of five 1's and 0's, and sixty-four ways of arranging six, it is clear that one quickly acquires enough scope to include all of the characters on a keyboard in a binary code of just a few bits per character. To store a more complicated kind of data set—say, the information in a photograph—requires more ingenuity, but only a little. One can break the image down into tiny squares, like the pixels of a television screen, and for each pixel one can use a binary code to record the relative intensity of the primary colors red, blue, and green in that pixel.

An information storage device working on this basis needs three basic components: a medium for storing the binary data (this will comprise, in effect, an array of switches that can be flipped one way or the other, or a substance that can locally exist in one of two states), a means of writing the information into the storage medium (recording), and a means of getting the information back out (readout or playback). In many storage systems it is also desirable to erase information (to reset the switches) and to overwrite old information with new (which is essentially the same operation as erasing). Magnetic audiotape can of course be erased and over-written; but CD-ROM storage media (which are optical memory devices, described later) cannot (ROM stands for "read-only memory," which means that once the information is written onto the disk, it can be read but not erased).

Magnetic materials are ideal candidates for storage media because magnets are rather like switches—they can be made to point in one direction or another if one places them in an external magnetic field. The extent to which this simple fact changed the world is not always sufficiently appreciated. After the invention in ancient China, nearly two millennia ago, of the floating compass—a magnetic needle that aligned itself with the Earth's magnetic field—the magnetic iron oxide mineral magnetite became known as the lodestone (the "leading stone"), and ensured that seafarers were no longer at the mercy of the stars for accurate navigation. Without this navigational device, European explorers, and later settlers, in the sixteenth century might never have found their way to the New World.

These discoveries of the properties of natural magnets were very much of an empirical nature; just as the earliest navigators had no real conception of *why* the stars showed them the way, so the mysterious rotation of the lodestone was unexplained for many centuries. It was only in 1600 that William Gilbert, a scientific genius of the order of Isaac Newton, suggested in his book *De Magnete* that perhaps the Earth itself is a giant magnet.

In the nineteenth century, Michael Faraday showed that magnetism and electricity are like opposite sides of a coin—the flow of electricity induces a magnetic field, and a changing magnetic field induces an electrical current. These two fields, electric and magnetic, can be self-supporting in the sense that a varying electric field can propagate out into empty space with a varying magnetic field propagating perpendicular to it. These propagating fields constitute electromagnetic radiation—radio waves, light, X rays, and all the rest.

The unified theory of electricity and magnetism, developed by James Clerk Maxwell in the 1880s, provides one of the central ideas in physical science. It is not an easy theory, but for the purposes of understanding magnetic recording we need know nothing more than Faraday's principle of electromagnetic induction: a (changing) electric field can induce a magnetic field, and vice versa. We will see why this is useful shortly.

A magnet, then, is a potential switch: place it in a magnetic field, and it will tend to align its poles with those of the field (fig. 2.2*a*). Reverse the direction of the field (so that the north pole becomes the south), and the magnet will switch to point in the other direction. Here we have a simple system for storing binary information: the magnet represents a "1" when pointing in one direction, and a "0" when pointing in the other. A magnetic memory can be constructed from a whole array of magnets, in which the binary data can be encoded by applying an external magnetic field to each magnet in turn, the direction of which determines whether we write a "1" or a "0" (fig. 2.2*b*).

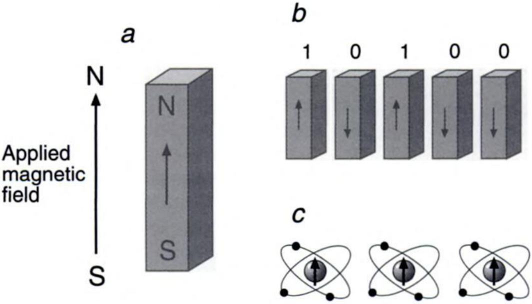


FIGURE 2.2 (*a*), A magnet that is free to rotate will align itself in an external magnetic field so that its poles point in the same direction as those of the field. (*b*), By using an applied field to orient magnets in one direction or the other, one can record binary data in a magnetic array. (*c*), Some atoms behave like tiny magnets, possessing a magnetic dipole (a "magnetic moment") which can be oriented by an external field.

Just a Moment

This is all very well, but if we want to have a realistic storage density, then our magnets will have to be very small. In principle that is simple enough, since certain individual *atoms* act as magnets. That is to say, these atoms (iron is the typical example) can be thought of as tiny magnets possessing a north and south pole, which can be flipped in one direction or another in a magnetic field.

This is not a property that all atoms possess. Rather, it is specific, in general, to atoms that have a certain arrangement of electrons. Electrons are negatively charged particles that surround the central, positively charged nucleus of an atom. They can be considered (in the crude so-called classical picture of an atom, which retains a great deal of value even though it has been superseded by quantum mechanics) to be circulating the nuclei in orbits. Now, a circulating, negatively charged electron is really nothing more than a circulating electrical current, which, because of electromagnetic induction, gives rise to a magnetic field. In many atoms electrons are paired up in each orbit; crudely speaking, they can be considered to be circulating in opposite directions, giving rise to opposing magnetic fields which therefore cancel out.

But when electrons are not paired up in this way, they can generate a net magnetic field. This contributes to the magnetic field of the atom, but that is not the whole story. Electrons also possess a property called *spin*. This is a quantum-mechanical property; it is not really so simple as the spinning of the electron about an axis (in the same way as the Earth rotates on an axis as it orbits the Sun), but that again is not a totally inept analogy. Electron spin also generates a magnetic field, and the overall contribution of an unpaired electron to the magnetism of an atom results from the combination of its orbital and spin components. An atom that possesses net magnetism is said to have a magnetic moment—a kind of magnetic dipole that points in a certain direction (fig. 2.2*c*). Metal atoms commonly have unpaired electrons, both in the pure metal and in some compounds (such as natural oxide minerals). These substances are therefore potential magnets.

Only "potential" magnets, mind you, because the existence of magnetic atoms in a substance does not in itself guarantee that the substance will be magnetic in the sense that we generally recognize it—picking up iron filings and so forth. Scientific techniques for revealing such things will show that an iron nail is made up of atoms possessing magnetic moments, yet the nail will not rotate to point north when hung on a string. Why not?

In iron, the magnetic moments on each atom have a tendency to line up so that they all point in the same direction. At first glance this seems intuitively natural, but it is in fact less obvious. If you place two bar magnets side by side, with the north poles of each facing the same direction, they will certainly *not* be happy to maintain this arrangement; they will rotate so that the like poles are not adjacent. If they can rotate freely, they will tend to align in opposite directions, and indeed may pull themselves together with the north pole of one facing the south pole of the other and vice versa. As in the case of electric charge, one can say that opposites attract.

One might therefore expect the magnetic moments in iron to line up in an alternating fashion, pointing first north and then south as one passes down a row. That this does not happen is because the way in which the moments interact is subtle, involving quantum-mechanical effects that are determined by the nature of stroying this alignment, and at room temperature a magnetized nail will gradually find its perfect alignment of atomic moments disrupted by thermal jostling. This disruption increases as the temperature increases; at 770 degrees Celsius thermal effects are so great that a piece of iron loses its ability to align the moments of neighboring atoms entirely in the absence of a magnetic field. Every ferromagnet has such a point, called the Curie temperature after the French physicist Pierre Curie who first explained this property.

Writing with Magnetism

The ability to influence the alignment of magnetic moments by applying a magnetic field forms the basis of magnetic recording. By applying a strong, localized magnetic field to one region of a thin layer of a ferromagnetic material such as iron or its magnetic oxide magnetite, we can create a domain in which all of the magnetic moments of the iron atoms are aligned (provided that the temperature is below the magnetic medium's Curie temperature). Applying the field elsewhere in the opposite direction, we can form a domain of opposite alignment. If we take one alignment direction to represent the binary digit "1," and the opposite alignment to represent "0," we can write into the magnetic layer a pattern of 1's and 0's—binary data, in other words. For practical reasons that will become clear shortly, in current magnetic storage media based on thin magnetic films the two different alignment directions are both parallel to the plane of the film. Alignment perpendicular to the plane is also possible in principle, and has potential advantages in terms of storage density, as we shall see; but it is difficult to achieve for technical reasons.

The earliest thin-film magnetic recording media, developed in the 1930s, used films of iron particles stuck to paper tape. But iron oxide soon became the preferred material—not only does pure iron go rusty, but in the form of a fine powder it is highly flammable. The early tapes developed by BASF used the plastic cellulose acetate in place of paper; today other plastics are used as the support for the magnetic particles. The information is written onto the tape by a recording head, which generates a magnetic field that varies the direction of alignment of the magnetic moments within the particles (the magnetized domains are bigger than the size of the particles themselves).

This writing process uses the phenomenon of electromagnetic induction. The information is fed into the recording head in electronic form—it might be, for example, audio information from a microphone, or visual (video) information from a television receiver. In most recording devices currently in use, the recording head contains an induction coil: an electromagnet, in which a coil of wire surrounds a core of magnetic material such as iron. The electric current circulating around the coil induces a magnetic field in the core; reversing the direction in which the current flows reverses the orientation of the induced magnetic field.

The core consists of a ring with a small gap. When the core is magnetized, a magnetic field bridges the gap. This field can be thought of as a series of *field lines*, or lines of *magnetic flux*, which can be made visible around a bar magnet by scattering iron filings around it—the particles become arranged into lines running from one pole to the other. The flux lines spread out somewhat as they pass from one side of the gap to the other, and so they can pass through a magnetic tape or disk lying just below the gap (fig. 2.5). The direction of the magnetic moments in the recording medium will align themselves with the direction of the field across the gap in the recording head. You can now see from figure 2.5 why the direction of alignment is parallel to the plane of the recording medium.

Thus, as the recording medium is pulled past the recording head, a series of current pulses in one direction or another will write into the medium a series of magnetized domains of differing orientation. To read this information back out, the same process is used in reverse. The magnetized regions of the recording medium also have associated magnetic flux lines which pass from one "pole" to the other (fig. 2.6). (Notice that the direction of the field lines of this so-called demagnetizing field is opposite to the direction of the magnetization.) These flux lines can induce a magnetization in the core of a readout head more or less identical to the recording head. Every time the direction of this induced magnetization changes (that is, every time the boundary between one magnetized region of the

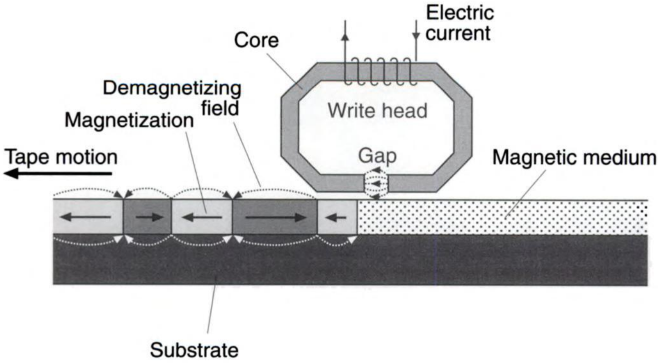


FIGURE 2.5 In conventional magnetic recording, regions within a magnetic film deposited on a substrate (a rigid disk for magnetic hard disks, a flexible plastic tape for magnetic tape) are magnetized in different directions by an electromagnetic head. The direction of magnetization between the poles of a gap in the head is determined by the direction of the current that flows through the electromagnetic induction coil wound around the head's core. This in turn determines the direction of magnetization in the storage medium, because the magnetic flux lines across the gap penetrate this medium. Data is thus encoded in binary form as a series of domains of differently oriented magnetization. A "demagnetizing field" that opposes the magnetization of a given domain is set up between the oppositely oriented domains on either side (see text); unless the resistance to reorientation (the coercivity) of the magnetic medium is large enough, this field will flip the orientation of a domain so that it is aligned with its neighbors.

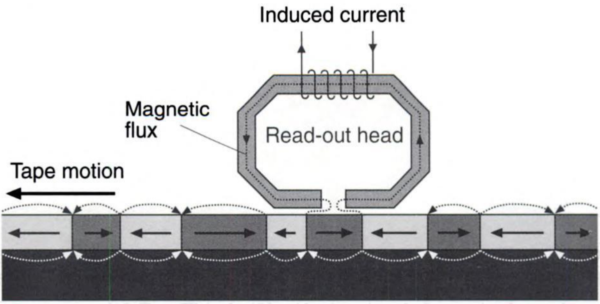


FIGURE 2.6 The readout process is essentially the reverse of the write process. As the tape passes below the readout head, the core experiences a changing magnetic flux each time a boundary between domains passes beneath the gap, owing to the demagnetizing fields that extend beyond the magnetic medium itself. This changing flux induces an electric current in the induction coil of the head.

recording medium and another of opposite orientation passes by the readout head), a current is induced in a coil of wire wound around the core of the head. Note that this induction of electric current requires a *change* in the magnetic field—the mere presence of a constant magnetic field in the core will not have this effect.

For magnetic tape devices, the tape is pulled past the recording head. But storing information in the essentially one-dimensional format of a magnetic tape has the drawback that one has to trawl through long stretches of tape in order to find a specific piece of information. The two-dimensional format of a disk allows for much more efficient writing and retrieval—the write or readout head can be moved rapidly to the relevant part of the disk. In modern computer technology, the magnetic disk has all but entirely replaced the magnetic tape. In disk drives, information is written and located both by spinning the disk and by moving the recording head across its surface. But the principles of the recording process are the same.

These principles involve physics that was well known at the end of the nineteenth century. But the practical requirements of today's top-range recording devices are awesome and are realized only thanks to the phenomenal capacity of modern electronic and mechanical engineering. Commercially available hard-disk storage devices can hold several billion bits of information in a square inch of magnetic recording medium (equivalent to a stack of typed pages about 100 meters tall). This density of information is useful only if it can be accessed on a realistic timescale, and modern readout devices can retrieve several million characters per second. This requires that the head skim over the surface of the disk at speeds of around 100 miles per hour, with jarring halts and reversals in direction. And all the while, the head sits poised only ten millionths of an inch above the recording medium. If this were a fairground ride, you wouldn't want to be on board!

Denser and Denser

My computer came supplied with an encyclopedia on CD-ROM that staggers me. It seems to have a limitless capacity to supply me with more than I need to know about every topic that I ask of it, and provides pictures and sometimes sound and even movies to boot. (The CD-ROM drive is in fact an optical, not a magnetic system, but the computer does not balk as I load this mountain of information into its magnetic hard drive.)

And yet I know that this is nothing compared with what will be available in just a few years, as storage densities and memory capacities become ever greater. What is it that is bringing about this explosion in data storage?

We saw above that data is stored in magnetic media as a series of regions magnetized in different directions. In principle, the storage density can be increased simply by making these regions smaller. But there is a limit to how far this can be taken, which is determined in large part by the detailed nature of the boundary between one region and the next. It is these boundaries that play the crucial part in data storage because, as explained above, the readout head produces a pulse of electricity only when the magnetization of the recording medium beneath it *changes*: in other words, when we pass from one magnetized region to the next.

The boundaries are not abrupt; they have a finite width, called the transition region. One can imagine increasing the storage density of the medium by reducing the width of the transition region, so that more magnetized regions can be fitted within a given area of material.

We can think of the transition regions as a kind of buffer between the magnetic poles of the magnetized regions. Because the poles of adjacent regions are oriented in different directions, poles that are alike abut one another on either side of the transition region (fig. 2.7). But there lies the rub: for, as is well known from experience with bar magnets, like poles repel one another. There is consequently a driving force for any given magnetized region to reverse its direction, so that its poles are aligned with those of its neighbors and opposite poles meet at the boundaries. In effect, one can think of the two neighbors on either side of a magnetized region in a track as setting up a magnetic field that opposes that of the central region and which thus has a tendency to reverse it. This is the demagnetizing field mentioned earlier.

If the transition region is made narrower, the poles of neighboring magnetized regions become closer together and the driving force for reversing the direction of magnetization becomes stronger. Whether or not this reversal actually occurs (thereby erasing data) depends on a property of the magnetic medium called its *coercivity*, which is a measure of the strength of the magnetic field required to flip the direction of magnetization. A material with a high coercivity will be more able

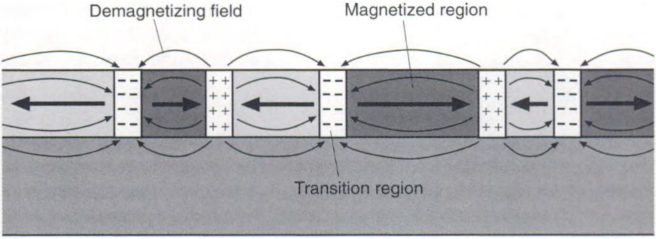


FIGURE 2.7 Between each magnetized domain in a magnetic storage medium there is a transition region in which the flux lines of the demagnetizing fields start and end. This transition region is a kind of buffer between the like poles of adjacent domains, which repel each other. The storage density of the medium is limited by the width of this transition region: if it is made too narrow, the repulsion between like poles can become so great that the demagnetizing field flips the orientation of a magnetized region to align it with those regions on either side.

to resist flipping, and so will be able to support narrower transition regions and higher storage densities.

The search for materials for high-density magnetic storage is therefore in part a search for materials with high coercivity. The coercivity of iron oxide (specifically, the form denoted γ-Fe_2O_3) is sufficient to make it still the mainstay of magnetic recording, but other, superior materials are now available. Chromium dioxide is used for improved sound fidelity on commercial audio cassettes, because it has a higher coercivity than iron oxide and therefore captures a more accurate record of the audio signal. If you really want exceptional sound quality, however, you can resort to metal tapes, which use fine particles of pure iron as the recording medium. The coercivity of pure iron is considerably greater than that of its oxide. The manufacture of metal tapes is a delicate process because, as indicated earlier, the metal particles are highly susceptible to oxidation (basically, to rusting, although this can manifest itself as inflammability for very small particles) and are hard to disperse evenly on the tape.

While these are the materials most commonly used today in magnetic disk and tape technology, more advanced media are required for significant improvements in storage density. The most promising new materials are continuous thin films of metal alloys, particularly cobalt–nickel alloys. Whereas recording media based on metal oxide particles are produced by dispersing the small particles within a binding matrix, thin-film media are made by depositing continuous films of the alloy on a rigid substrate—the resulting films are a mosaic of small crystals of the alloy, packed intimately together. These films are laid down either by a chemical plating process, in which the substrate is immersed in a bath containing the film medium, or by so-called sputtering, where high-energy ions accelerated by an electric field are used to knock atoms or clusters of atoms from a lump of the magnetic medium; these vaporized particles are then deposited on the substrate.

Stand-Up Recording

The storage density of a magnetic medium is set by the point at which its coercivity is no longer large enough to prevent reorientation of recording domains by the demagnetizing fields that are set up between adjacent domains. As the domains get smaller, the transition region between two domains gets thinner and the reorienting effect of the demagnetizing field increases. Although materials with higher coercivities therefore support smaller recording domains and greater storage densities, they also require stronger fields in the write head to encode the domains in the first place. An obvious solution is to do away altogether with the uncomfortable end-to-end orientation of neighboring domains, and instead to orient the magnetic moments in head-to-toe fashion, perpendicular to the plane of the medium. Then, adjacent domains of opposite magnetization have their opposite poles, not their like poles, next to one another (fig. 2.9). This means that the demagnetizing fields actually get weaker as the data density increases.

Obvious, in theory. But if you take another look at figures 2.5 and 2.6, you'll see one reason why this idea has not more readily been put into practice: any usual design for an inductive write or readout head involves magnetization of the recording medium *parallel* to the plane of the film (called longitudinal recording). Perpendicular recording requires a considerable amount of rethinking of both the device design and choice of recording medium.

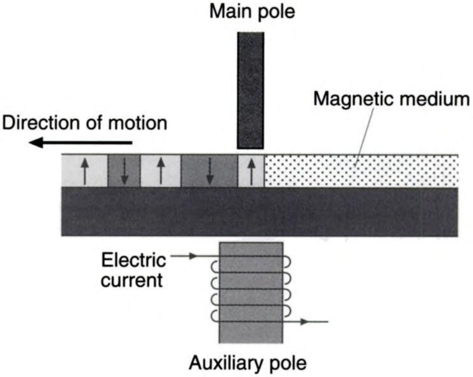


FIGURE 2.9 In perpendicular recording, the preferred direction of the magnetic moments in the magnetic storage medium is perpendicular to the plane of the film. Recording in such a medium can be achieved by means of two magnetizable heads, one on each side of the film. A signal-bearing current induces magnetization of the so-called auxiliary pole (here the lower head), and this in turn magnetizes the main pole (upper head), which lies close to the magnetic medium. The combined influence of both poles is sufficient to orient the magnetization of the storage medium.

SCIENCE / CHEMISTRY

One of *Choice*'s Outstanding Academic Books of 1998

MADE TO MEASURE

NEW MATERIALS FOR THE 21ST CENTURY

PHILIP BALL

Made to Measure introduces a general audience to one of today's most exciting areas of scientific research: materials science. This book is written in the same engaging manner as Ball's popular book on chemistry, *Designing the Molecular World*, and it links insights from chemistry, biology, and physics with those from engineering as it outlines the various areas in which new materials will transform our lives in the twenty-first century.

"Let me state up front that *Made to Measure* . . . is an outstanding book. Written for the general reader, it will also greatly appeal to specialists. If you are a solid-state physicist, chemist, materials scientist, engineer, science policy maker, or keen amateur scientist, then sell your shirt to buy it."

—COLIN HUMPHREYS, *New Scientist*

"Philip Ball offers a panorama of 1,001 new materials for the next century. . . . His survey would make a good textbook for an introductory course in materials science. For the rest of us, the sheer range of examples is impressive."

—JON TURNEY, *Financial Times*

"Philip Ball writes about the very modern science of materials. . . . [He] is full of fascinating insights and, especially on the photonic side of things, he really opens the reader's eyes. . . . [His] book is the first to be entirely devoted to this field. That task has been very well accomplished, and the book is warmly recommended."

—ROBERT W. CAHN, *European Journal of Physics*

Philip Ball is an associate editor for physical sciences with *Nature*. He contributes regular articles on all fields of science to the academic and popular press. He is also Writer in Residence with the Chemistry Department of University College, London, and the author of *Designing the Molecular World* (Princeton), and *The Self-Made Tapestry: Pattern Formation in Nature*.

Cover illustration: Photovoltaic roofing tiles in a traditional Japanese design. (Courtesy of Sanyo Electric Co., Ltd., Japan)

PRINCETON PAPERBACKS

ISBN 0-691-00975-9

90000

9 780691 009759