7 Moore's law and the silicon revolution

> As I prepared for this event, I began to have serious doubts about my sanity. My calculations were telling me that, contrary to all the current lore in the field, we could **scale** down the technology such that **everything got better**: the circuits got more complex, they ran faster, and they took less power – WOW!
>
> Carver Mead[1]

Silicon and semiconductors

When we left the early history of computers in Chapter 2, we had seen that logic gates were first implemented using electromechanical relays – as in the Harvard Mark 1 – and then with vacuum tubes – as in the ENIAC and the first commercial computers. These early computers with many thousands of vacuum tubes actually worked much better and more reliably than many engineers had expected. Nevertheless, the hunt was on for a more dependable technology. After World War II, Bell Labs (Fig. 7.1) initiated a research program to develop solid-state devices as a replacement for vacuum tubes. The focus of the program was not on materials that were metals or insulators but on strange, "in-between" materials called *semiconductors*.

Fig. 7.1. An aerial view of AT&T Bell Labs in Holmdel, New Jersey. The building was designed by the architect Eero Saarinen and for forty-four years it was the home of an advanced research laboratory owned successively by Bell Telephone, AT&T, Lucent, and Alcatel.

In a solid, it is the flow of electrons that gives rise to electric currents when a voltage is applied. One of the great successes of quantum physics has been in giving us an understanding of the way in which different types of solids – metals, insulators, and semiconductors – conduct electricity. This quantum mechanical understanding of materials has led directly to the present technological revolution, with its accompanying avalanche of stereo systems, color TVs, computers, and mobile phones. A good conductor, such as copper, must have many *conduction electrons* that are able to move and thus constitute a current when a voltage is applied. By contrast, an insulator such as glass or carbon has very few conduction electrons, and little or no current flows when a voltage is applied. Semiconductors are solids that conduct electricity much better than insulators but much worse than metals. The elements germanium and silicon are two examples. The importance of silicon for computer technology is evident in the naming of California's "Silicon Valley," home to many of the earliest electronic component manufacturers (Fig. 7.2).

The properties of a solid depend not only on what element it is made of, but also on the way the atoms or molecules are stacked together. Many solid

Fig. 7.2. A Landsat photograph of Silicon Valley and San Francisco Bay. In 1971 journalist Don Hoefler ran a series of articles in *Electronic News* under the title "Silicon Valley USA" and the name caught on.

materials have their constituent atoms arranged in a regular array, like bricks in a wall but in three dimensions. This regular pattern of atoms is called a *crystal lattice*, and substances with such a structure are called *crystalline solids*. Arranging all the atoms in a regular array has a dramatic effect on the allowed energy levels for the atomic electrons. The way to understand the energy levels of such crystalline materials was discovered by a Swiss physicist named Felix Bloch. To find the allowed electron energy levels for any quantum mechanical system, you need to solve the Schrödinger equation – a mathematical formula as fundamental for the behavior of quantum objects as Newton's laws are for classical objects. Solving this equation for an electron in the potential of a positively charged nucleus leads to definite, isolated energy levels. For electrons in a potential corresponding to a regular lattice of positive ions, Bloch found that instead of isolated energy levels, the allowed energy levels merged into several "bands" of allowed energies. The discovery of such *energy band structures* provides the foundation for our understanding of the difference between metals, semiconductor, and insulators. Figure 7.3 shows typical allowed energy band structures for these three types of materials.

In a metal such as copper, the lowest energy band has many unfilled levels and the conduction electrons can move freely into empty levels, gaining energy when a voltage is applied and generating an electric current (Fig. 7.3a). At absolute zero, the coldest possible temperature (–273.15 °C), the energy levels in the bands would be filled up one electron at a time, according to the Pauli Principle to give the minimum energy state (see the quantum theory primer at the end of this chapter for more details). At room temperatures, the lattice ions have some

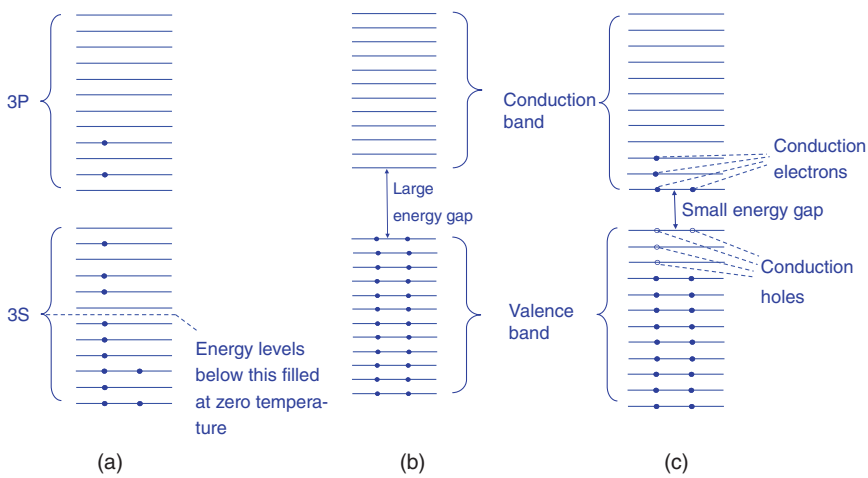


Fig. 7.3. Band structures of metals, insulators, and semiconductors. (a) Band structure of a typical metal like sodium. There are many unfilled energy levels in the "3S" valence band for the conduction electrons to occupy. At normal temperatures, only a few electrons will be excited into the almost empty "3P" band. (b) In an insulator, the valence band is full and the gap between the valence and conduction bands is too large for any significant number of electrons to jump across the gap with normal thermal energy distributions. As a result, an insulator conducts electricity very poorly, if at all. (c) In a semiconductor, the valence is almost full but there is only a small energy gap to the mostly empty energy levels in the conduction band. At normal temperatures, some of the electrons have enough thermal energy to be able to jump across this energy gap.

thermal kinetic energy corresponding to vibrations about their positions in the crystal lattice. As the conduction electrons move through the metal they can gain or lose energy in collisions with the lattice ions and with each other. Thus, instead of the conduction electrons filling up only the lowest energy levels in the band, some electrons will be *thermally excited* to higher levels in the band and even to higher bands. This has the effect of leaving some empty energy levels in the bottom of the lowest band.

For an insulator like carbon, the lowest energy band is full and there is a large energy gap to the next band (Fig. 7.3b). In this case, almost no electrons are able to gain enough energy from collisions to jump into the empty, higher energy band. When a voltage is applied to the material, there are therefore no empty levels close by for the electrons to be able to move to and gain energy, so the material acts as an insulator. The energy bands in a semiconductor are shown in Figure 7.3c. These materials have a similar band structure to an insulator with the bottom band filled, but the energy gap to the next band of energy levels is much smaller. At ordinary temperatures, some electrons are excited by thermal collisions into the upper conduction band. When a voltage is applied, the electrons in the upper band have plenty of empty states to move to and allow the electrons to gain energy. There will also be some empty states in the lower band that allow conduction. Thus semiconductors will conduct currents fairly easily, but their conductivity will depend strongly on temperature, in contrast to metals and insulators.

Two Nobel Prizes: The transistor and the integrated circuit

Pure semiconductors are not in themselves of great practical importance. In metals almost every atom contributes one or more conduction electrons but in semiconductors only one atom in about a thousand million contributes an electron to conduct electricity. This apparent drawback has the great advantage that the conduction properties of semiconductors can be easily modified by introducing tiny amounts of additives called *impurity atoms* – at the level of around one atom in a million. Germanium and silicon both have four *valence electrons*, electrons in the atom's outermost shell that can be easily transferred to or shared with other atoms. The valence electrons fill up most of the states in the *valence band*, which lies below the almost-empty *conduction band*. If we introduce an impurity such as phosphorous, which has five valence electrons, into the pure semiconductor only four of these electrons are needed to maintain the crystal lattice structure. As a result there will be an electron left over that can easily be detached from the phosphorous atom and contribute to the conductivity. Similarly, if we introduce an impurity atom such as boron, with only three valence electrons, there will be one electron missing in the bonds that hold the lattice together. The missing electron creates a site that can capture electrons from filled states in the valence band, leaving empty states and allowing some conduction. These two situations are represented on the energy level diagram shown in Figure 7.4. The process of adding impurity atoms is called *doping*. Semiconductors that have been doped with phosphorus are called *n-type semiconductors*. The phosphorus atoms give rise to electron *donor* states just

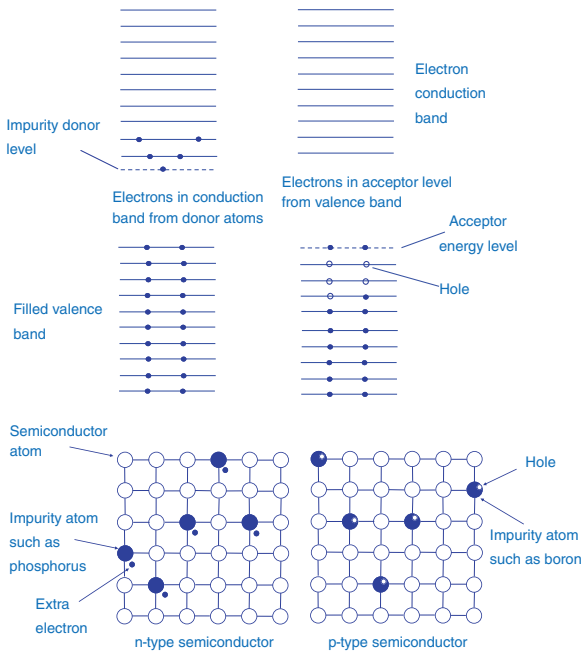


Fig. 7.4. Semiconductors doped with impurity atoms. (a) n-type semiconductor in which the impurity atoms have an extra electron. This results in the effective energy-level diagram shown here with a "donor level" just below the conduction band. (b) p-type semiconductor doped with impurity atoms with one fewer electron, resulting in electron "holes." The equivalent energy-level diagram has an empty "acceptor level" just above the valence band.

below the conduction band, and these electrons need only gain a small amount of energy to jump into the conduction band. Semiconductors doped with boron are called *p-type semiconductors*. The boron atoms give rise to electron *acceptor* states just above the nearly full valence band and at room temperatures electrons are readily excited into these levels. Compared to an undoped semiconductor, the boron impurity site is missing a negatively charged electron. This is equivalent to the p-type semiconductor having a positive charge compared to the undoped material. Conductivity in the nearly full valence band is possible because electrons can move into the unoccupied "hole" states. In a p-type semiconductor, instead of thinking of a negatively charged electron moving in response to a voltage, we can equally well think of a positively charged *hole* moving in the opposite direction. Because moving a negative charge to the left has the effect of increasing the charge on the right, we can alternatively think of the current as a flow of positively charged holes moving to the right.

Why is all this useful? Russell Ohl (B.7.1) at Bell Labs had discovered that p- and n-type semiconductors could be put together to form interesting semiconductor devices. The simplest device is the *p-n junction diode*, which prevents current from flowing in one direction but not the other. This p-n junction device is able to convert an alternating current into a unidirectional current – a property called *rectification*. The development of the p-n junction diode was the first step toward the invention of the transistor, a semiconductor device that could be used either to amplify a signal or to switch a circuit on or off. John Bardeen, Walter Brattain, and William Shockley (B.7.2) were awarded the 1956 Nobel Prize for physics for their invention of the transistor. The transistor was not discovered by accident – it was the culmination of an extensive research program at Bell Labs. As Bardeen later said in his Nobel Prize lecture: "The

B.7.1. Russell Shoemaker Ohl (1898–1987) was a researcher investigating the behavior of semiconductors at AT&T's Bell Labs in Holmdel, New Jersey. In 1939, Ohl discovered the "p-n junction" by which he was able to manipulate current flows. He also recognized the importance of using exceptionally pure semiconductor crystals to make repeatable and usable semiconductor diodes. His work with these devices led him to develop and patent the first silicon solar cells.

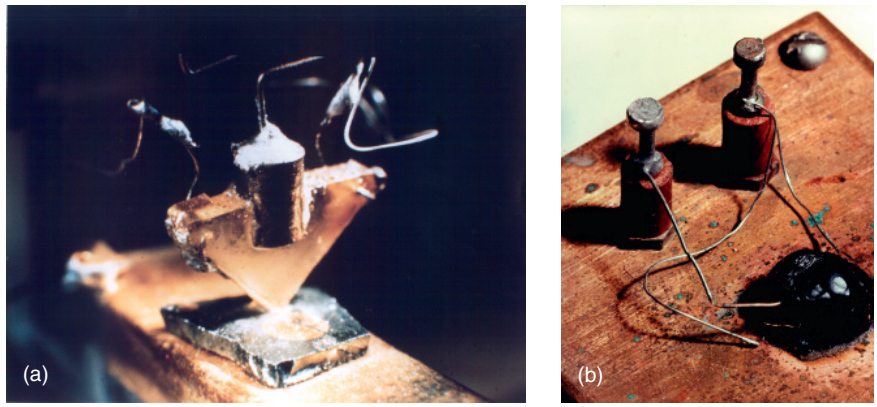


Fig. 7.5. The first transistors: (a) replica of the point-contact transistor invented by John Bardeen and Walter Brattain. The wedge of semiconductor that forms the base is about three centimeters on each side. (b) William Shockley's junction transistor consisted of a thin layer of n-type semiconductor sandwiched between two thicker regions of p-type material.

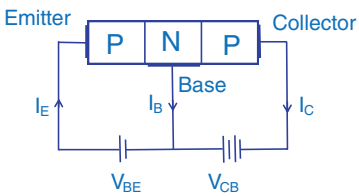
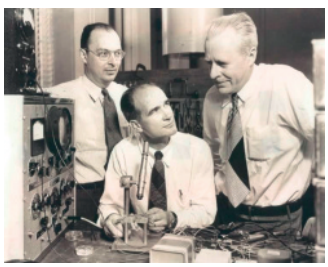


Fig. 7.6. Schematic of a PNP transistor illustrating the direction of currents (I) and voltages (V). The current to the emitter is the sum of the currents of the base and collector ($I_E = I_B + I_C$).

general aim of the program was to obtain as complete an understanding as possible of semiconductor phenomena, not in empirical terms, but on the basis of atomic theory."[2] A replica of Bardeen and Brattain's first transistor, called a *point-contact transistor*, is shown in Figure 7.5a. The two scientists first succeeded in observing amplification of a signal by this device on 24 December 1947. This discovery was followed in 1951 by Shockley's *p-n-p junction transistor* (Fig. 7.4b). This device turned out to be much more reliable and easier to manufacture than the point-contact transistor.

A junction transistor consists of a thin layer of n-type semiconductor sandwiched between two thicker regions of p-type material (Fig. 7.6). A transistor has three *electrodes*, conductors that can emit or collect electrons or holes, or that can be used to control the movement of the current through the device. Thus the current flowing in the electrode called the *collector* is controlled by a small current applied to the electrode called the *base*. In a p-n-p transistor, a large current through the high-resistance collector–base p-n junction can be controlled by a small current through the low-resistance base–emitter n-p junction (see Fig. 7.6). This action can be understood by a detailed consideration of the energy levels and the electron and hole currents across the two p-n junctions. The word *transistor* refers to this effect and comes from combining the two words *trans*fer and re*sistor.* The first commercial application of transistors was in hearing aids, followed soon after by the first portable transistor radio produced in 1955 by a company called Regency in Indianapolis. However, transistors also proved to be ideal for implementing the "on–off" binary logic of computers. Their speed and reliability, together with a large number of incredible engineering advances, have made them the basic ingredient of modern microelectronics.

One of the most important of the engineering advances leading to the development of the modern electronics industry was first envisioned by a British engineer named Geoffrey Dummer (B.7.3). He worked at the Telecommunications Research Establishment (later the Royal Radar Research Establishment), a research facility in Malvern, England. Dummer was an expert on the reliability of electronic components and was concerned about the performance of radar equipment under extreme conditions. He realized that it was both inefficient



B.7.2. The three inventors of the transistor, from left to right, John Bardeen (1908–91), William Shockley (1910–89), and Walter Brattain (1902–87). They were awarded the 1956 Nobel Prize for Physics. Bardeen went on to win a share of a second Nobel Prize for Physics for his work on the theory of superconductivity.

Fig. 7.7. Jack Kilby's first IC. Instead of making the components of the electronic circuit separately, Kilby incorporated a junction transistor, a capacitor, and resistances in the same piece of germanium. The device is 1/16 by 7/16 inches or 1.6 × 11.1 mm.

B.7.3. Geoffrey Dummer (1909–2002) was a researcher at the Royal Radar Research Establishment in Malvern, England. In the IC manufacturing world he was known as the "The Prophet of the Integrated Circuit." His vision for an IC was motivated by a desire to make electronic components more reliable.

and unnecessary to manufacture each of the components of an electronic circuit in separate pieces. If all these devices could be contained in the same piece of semiconductor, the circuit would be much smaller and more reliable. In May 1952, Dummer wrote:

> With the advent of the transistor and the work in semiconductors generally, it seems now possible to envisage electronic equipment in a solid block with no connecting wires. The block may consist of layers of insulating, conducting, rectifying and amplifying materials, the electrical functions being connected directly by cutting out areas of the various layers.[3]

Dummer's description was an amazingly accurate vision of a modern *integrated circuit*, or IC, a circuit etched or imprinted on a slice of semiconductor. But there was a long way to go to make such an IC an engineering reality.

The vital breakthrough was made in the summer of 1958 by an American electrical engineer named Jack Kilby. In the early 1950s, Kilby had worked on printed circuits, transistors, and the miniaturization of electronics, which were of great interest to the U.S. military. He then joined Texas Instruments, a semiconductor manufacturing company, to work "in the general area of microminiaturization."[4] Kilby arrived in the summer, just before most of the employees took their summer vacations:

> Since I had just started and had no vacation time, I was left pretty much in a deserted plant; so I began to think … I began to cast around for alternatives [to making circuits out of individual components] – and the monolithic [or solid circuit] concept really occurred to me during that two-week vacation period. I had it all written up by the time Willis [engineer Willis Adcock, who helped develop the silicon transistor] got back, and I was able to show him the sketches that pretty well outlined the idea – and the process sequence showing how to go about building it.[5]

By September 1958, Kilby had built the first working IC, all from a single piece of germanium (Fig. 7.7). The device was an *oscillator* (a circuit that generates a regular signal), containing a single junction transistor; a *capacitor* to store electrical energy; and several *resistors* to limit the electrical current – all made from a single piece of semiconductor. Kilby wired together the different components of the circuit by soldering tiny wires to his device. His version of the IC was limited by the difficulty of physically wiring up many components, but it was still a major breakthrough. Kilby was rapidly converted from electrical engineer to

B.7.4. Jack Kilby (1923–2005) and Robert Noyce (1927–90) display their medals at the first Draper Prize award in 1989. The citation for their award was "for their independent development of the monolithic integrated circuit." Kilby went on to gain a share of the 2000 Nobel Prize for Physics. In his Nobel Lecture he made an explicit acknowledgment of Noyce's contribution: "I would like to mention another right person at the right time, namely Robert Noyce, a contemporary of mine who worked at Fairchild Semiconductor. While Robert and I followed our own paths, we worked hard together to achieve commercial acceptance for integrated circuits. If he were still living, I have no doubt we would have shared this prize."[B1]

Fig. 7.8. Memorial sign for the site of the original Shockley Semiconductor Laboratory. In later life, Shockley became a controversial figure for his views on race and genetics. Despite this controversy it seems regrettable that an earlier plaque with Shockley's name on it has been replaced with this more "politically correct" version.

physicist by the Nobel Prize committee when he was awarded the Nobel Prize for physics in 2000 for his invention of the IC (B.7.4).

The beginnings of Silicon Valley

After his invention of the junction transistor, Shockley had a falling out with both Bardeen and Brattain. As a result, Bardeen left Bell Labs in 1951 to be a professor at the University of Illinois at Urbana-Champaign. In 1972, Bardeen won his second Nobel Prize for his role in developing the "BCS" (Bardeen, Cooper, and Schrieffer) theory of superconductivity, the only physicist to have been awarded two Nobel Prizes in physics. Shockley took leave of absence from Bell Labs in 1953 and, with the help of a Caltech friend, Arnold Beckman, set up the Shockley Semiconductor Laboratory, a division of Beckman Instruments in Mountain View, California, in 1955 (Fig. 7.8). Mountain View was near Palo Alto, Shockley's hometown and the location of Stanford University. Three of the first recruits to Shockley's company were physicists Robert Noyce and Jean Hoerni and chemist Gordon Moore. Sadly, Shockley was a terrible people manager and eventually alienated most of his employees. By the summer of 1957, Noyce, Hoerni, Moore, and five others – the "Traitorous Eight" – decided to leave Shockley and set up a company by themselves (B.7.5). With financing from the Fairchild Camera and Instrument Corporation, they formed Fairchild Semiconductor. While their building was under construction, their temporary workplace was a large garage in Palo Alto – now a tradition for Silicon Valley start-ups!

Two key innovations were needed to make the mass production of powerful and robust ICs a reality. At the time, the state of the art in transistors was the silicon *mesa transistor*, which had a tiny round plateau or "mesa" of silicon set on top of a base, also made of silicon (Fig. 7.9). Because the contacts on the mesa structure were exposed, these transistors were easily contaminated or damaged. Soon after Fairchild Semiconductor was established, the Swiss physicist Jean Hoerni (B.7.6) came up with a brilliant innovation, which now underpins all modern ICs. His idea was the *planar transistor*, where the mesa is now fully embedded into the silicon wafer, resulting in a completely flat transistor (Fig. 7.10). Hoerni also coated his device with a layer of silicon dioxide, which insulated and protected the transistor's contacts. The remaining obstacle to mass production was a method to electrically isolate the components in the silicon. This was solved in late 1958 by the Czech-born physicist Kurt Lehovec, who worked for the Sprague Electric Company in Massachusetts. He had heard of Kilby's IC and realized the importance of isolating the different components in the silicon. His solution was very simple: he proposed inserting back-to-back p-n junctions, or diodes, between the transistors in the silicon so that no current could flow in either direction. All these ideas came together in January 1959 when Noyce (B.7.7) developed a process for manufacturing ICs using Hoerni's planar transistors and Lehovec's p-n junctions. As Noyce said later:

> When this [the planar process] was accomplished, we had a silicon surface covered with one of the best insulators known to man, so you could etch holes through to make contact with the underlying silicon. Obviously, then,

B.7.5. A photograph of the Traitorous Eight – the founders of Fairchild Semiconductor who left Shockley's original Silicon Valley start-up. From left to right, they are Gordon Moore, Sheldon Roberts, Eugene Kleiner, Robert Noyce, Victor Gingrich, Julius Blank, Jean Hoerni, and Jay Last.

Fig. 7.9. In geology, a mesa is a flat-topped mountain, typically found in the U.S. Southwest as in this example in Monument Valley. In the same way, a semiconductor mesa transistor also rises above the surrounding semiconductor base with a height typically less than one micron.

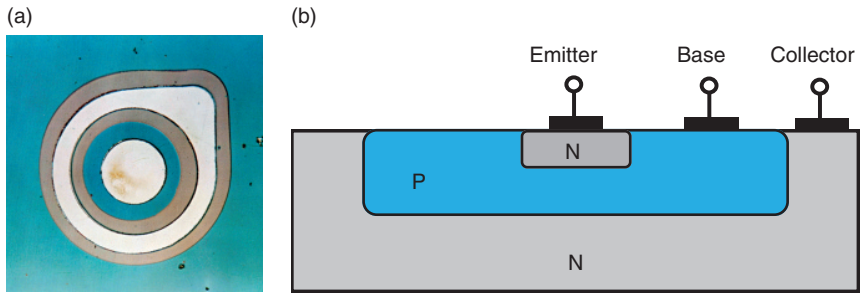


Fig. 7.10. (a) A view of Hoerni's planar transistor under the microscope. (b) A diagram showing the bowls of p-type semiconductor and n-type semiconductor together with connections for the emitter, base, and collector. The entire surface would be coated with silicon dioxide.

you had a whole bunch of transistors embedded in an insulating surface, and the next thing was that, instead of cutting them apart physically, you cut them apart electrically, added the other components you needed for circuits, and finally the interconnection wiring.[6]

There were several techniques, but the main one was, basically, to build back to back diodes [p-n junctions] into the silicon between any two transistors so that no current could flow between the two in either direction. The other element you needed was a resistor, and it was relatively simple to make a diode-isolated piece of silicon that acts as a resistor. You now had resistors and transistors, and could start building logic circuits, which you could interconnect by evaporating metal on top of the insulating layer. So it was a progressive buildup of bits and pieces of the technology to make the whole thing possible.[7]

B.7.6. Jean Hoerni (1927–94) was the inventor of the planar process that revolutionized chip production.

Noyce's patent for his version of the IC was filed in July 1959 (Fig. 7.11). His assembly of the key component technologies produced a design for a circuit that could be mass produced. Because ICs were made from tiny chips of silicon, they became known as *chips* or *microchips*. Fairchild started marketing a whole family of logic chips – the decision-making units of computers – in 1962.

Also in 1962, a new type of transistor debuted that could be more easily incorporated in mass-produced chips. This was the MOSFET, the *metal-oxide–semiconductor field effect transistor*. It was first successfully produced by John Atalla and Dawon Khang of Bell Labs in 1959. Bell Labs did not pursue the

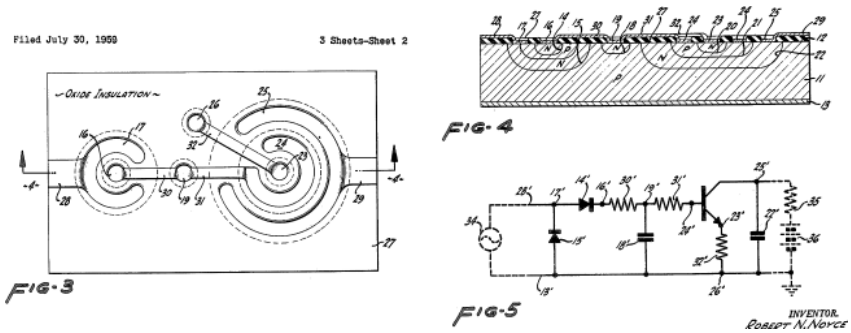
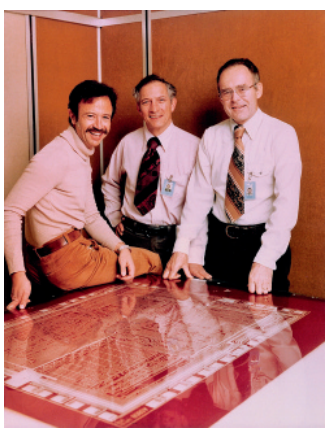


Fig. 7.11. Diagrams from Robert Noyce's patent for ICs. His design used Jean Hoerni's planar transistors with Kurt Lehovec's p-n junctions to isolate the different electrical components.

technology but Khang commented on its potential in a 1961 memo because of its "ease of fabrication and the possibility of application in integrated circuits."[8] It was left to two young engineers, Steven Hofstein and Frederic Heiman, at RCA Corporation's research laboratory in New Jersey to build an experimental IC using sixteen metal-oxide–semiconductor (MOS) transistors. Because of their small size and low power consumption, more than 99 percent of the microchips produced today use MOSFET transistors. Both p-type MOSFETs and n-type MOSFETs are employed in the dominant technology for constructing ICs, a method known as complementary metal-oxide–semiconductor (CMOS) technology.



B.7.7. A photograph of Andy Grove, Robert Noyce, and Gordon Moore who were responsible for making Intel such a successful chip manufacturer. Grove, a Hungarian-born engineer and businessman, was one of the evangelists behind Intel's relentless drive to pack more and more transistors on a chip. Grove's book *Only the Paranoid Survive* has become a classic in business management. Noyce's nickname was the "Mayor of Silicon Valley."

Chip development continued apace, with ever-increasing miniaturization and complexity. By 1967, chips were being produced that incorporated thousands of transistors. Although the initial steps toward IC development had little to do with funding from the U.S. military sector, the U.S. military and the aerospace community played a key role in improving the quality of chips and developing better techniques for their mass production. In the early 1960s, at the height of the Cold War, the U.S. Air Force needed to expand its Minuteman ballistic missile program (Fig. 7.12). It looked to ICs to replace the discrete electronic components and increase the computing power (Fig. 7.13). The air force wanted to produce missiles at a rate of "around six to seven missiles a week,"[9] and it ordered more than four thousand ICs per week from Texas Instruments, Westinghouse Electric Corporation, and RCA. The air force's insistence on more reliable components also forced suppliers to introduce "clean rooms" – facilities with little dust and other pollutants, adapted from those developed at Sandia National Laboratories in New Mexico for the assembly of atomic weapons.

This expansion of mass production of ICs drove down the costs for later consumer applications. Together the U.S. Air Force and Navy programs accounted for the entire $4 million IC market in 1962: by 1968, the U.S. government accounted for only 40 percent of a $300 million IC market. From 1962 to 1968 the average price of an IC dropped from more than $50 per microchip to about $2.

In 1959, Fairchild Camera and Instrument bought out the eight founders of Fairchild Semiconductor. The parent company then introduced a more rigid management style that triggered an exodus of the founders and other talent. From its beginning with Shockley's first semiconductor company in Mountain View, the diaspora from Fairchild Semiconductor led to the establishment of

Fig. 7.12. Minuteman missile silo, South Dakota, United States.

more than fifty IC companies in and around San Jose, California. Noyce and Moore were the last of the founders to leave Fairchild. In 1968, they set up a new company with the intention of specializing in memory chips. The name of their new company was Intel Corporation, short for *int*egrated *el*ectronics. At that time, the best random access memory (RAM) chip had a capacity of sixty-four bits, which was not sufficient to supplant magnetic core memory in computers. RAM refers to data storage that allows information to be accessed in any order, unlike storage on a magnetic tape, which can only access information in a linear fashion by moving along the tape. By storing frequently used or active files in RAM, the computer can access the data much faster from RAM than it can retrieve information from a hard disk – and very much faster and much more flexibly than from magnetic tape.

In 1968, Bob Dennard (B.7.8) from IBM patented a one-transistor design for *dynamic* RAM, or DRAM. In his design, a single bit of information can be stored in a memory cell consisting of one transistor and a tiny capacitor. This innovation simplified the design of memory chips and permitted a significant increase in memory capacity. Dynamic memory is so-called because the stored charge leaks slowly away and the memory cell needs to be regularly refreshed to maintain its contents. It is possible to design *static* RAM (SRAM) chips that do not need refreshing but this type of RAM needs more transistors to implement and is therefore more expensive. In 1970, Fairchild, now a competitor to Moore and Noyce's new company, brought out a 256-bit DRAM memory chip. This chip received much momentum by being chosen as the memory for an ambitious "parallel computer," the Illiac-IV, under construction at the University of Illinois. A parallel computer has many processing units and a programmer has to orchestrate the work of all of these units to solve a problem. Although the Illiac-IV computer had only limited success – parallel programming is still too hard – the project showed that semiconductor memory was a viable alternative to magnetic cores. By the end of 1970, Intel had responded to Fairchild's 256-bit chip by introducing the 1103, the first 1,024-bit DRAM chip, using a three-transistor design. In 1971 Intel's revenues were about $9 million; three years later these had almost tripled. The end of magnetic core memories was in sight.

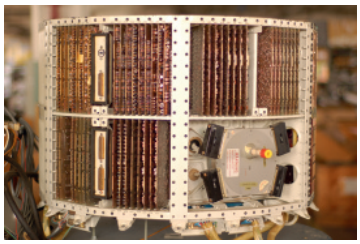
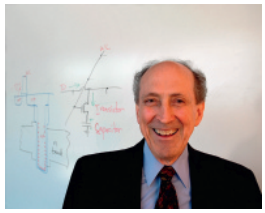


Fig. 7.13. Photograph of the Minuteman I missile guidance computer assembled from discrete electronic components: transistors (1,521), diodes (6,282), resistors (504), and capacitors (1,116). The computer was specially designed to survive and function in extreme conditions. In the Minuteman II, ICs manufactured by Texas Instruments replaced multiple transistor boards. The Minuteman II computer used about two thousand ICs.

The microprocessor and Moore's law

The first electronic calculator was called the ANITA and was produced in 1961 by the Bell Punch Company in the United Kingdom. It used discrete transistors and was about the size of a typewriter. Later in the 1960s, Texas Instruments built a calculator using ICs. Some were logic circuits for doing the calculations, others were RAM, and still other circuits provided *read-only memory* (ROM) for the operating system and subroutine libraries. The contents of ROM can be accessed and read but cannot be changed. When cheap pocket versions appeared in the 1970s, the traditional slide rule of the engineer rapidly disappeared. At the beginning of NASA's Apollo space program in 1963, the guidance computer of the lunar module was constructed from about five thousand logic chips. On the last Apollo mission in 1975, one astronaut carried a Hewlett Packard HP-65 pocket calculator more powerful than the spacecraft's guidance computer.



B.7.8. A photograph of Robert Dennard from IBM. On the white board behind him is a sketch of his one-transistor DRAM cell. This led to the widespread availability of cheap semiconductor memory chips. His clear statement of the consequences of the physics behind Moore's law is known as *Dennard scaling*.

In the summer of 1969, the Japanese calculator manufacturer Busicom asked Intel to design a set of chips for their new range of *programmable calculators*, which could be programmed much like a computer. The Japanese engineers had a design that involved twelve logic and memory chips, each with several thousand transistors. Ted Hoff, the Intel engineer assigned to the project, thought that this was not a very efficient solution to their problem. Instead, Hoff suggested developing a general-purpose logic chip that, like the central processor of a computer, could be programmed to perform any logical task (B.7.9). Together with some RAM and ROM, and a chip to control input and output, Hoff's solution needed only four chips to be designed instead of the original twelve. Intel engineers Stan Mazor and Frederico Faggin, together with Busicom engineer Masatoshi Shima, implemented the design (Fig. 7.14). This was the first *microprocessor* sold as a component – an IC that can perform all of the functions of a computer's central processing unit (CPU). At first Intel was not sure of the market for its microprocessor (called the 4004) because the company thought that there was too little demand for calculators. As it turned out, any machine that handled and manipulated information or controlled a complex process was a potential market for the microprocessor.

The Intel 4004 microprocessor was launched in 1971. It contained more than two thousand transistors and measured 1/8 inch by 1/16 inch a side. This single chip had about the same computing power as the original ENIAC computer. Intel introduced a more powerful microprocessor called the 8080 in 1974. The 8080 led to a host of new applications and, as we shall see in the next chapter, to the creation of the personal computer. By the early 1980s Intel had more than $1 billion in sales: twenty years later, the global market for microprocessors was more than $40 billion. The vast majority of microprocessor chips are used in "embedded" applications – such as in washing machines, cookers, elevators, airbags, cameras, TVs, DVD players, and mobile phones. Automobiles and planes are increasingly reliant on microprocessors, as are the many of the infrastructure systems vital to the functioning of a modern city.

B.7.9. The three inventors of Intel's first microprocessor, Ted Hoff, Stan Mazor, and Frederico Faggin. Hoff was Intel employee number 12; he and colleagues Mazor and Faggin were awarded the 2010 U.S. National Medal for Technology and Innovation. Faggin credits Masatoshi Shima, an engineer from the Japanese company Busicom, for help with the detailed design work for the 4004.

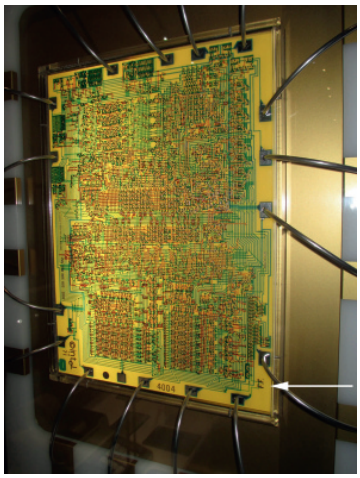


Fig. 7.14. The Intel 4004 microprocessor with Frederico Faggin's initials. This giant model with a 128× magnification is on display at the Intel Museum.

In 1965, Moore (B.7.10) wrote an article for the thirty-fifth anniversary issue of *Electronics* magazine entitled "Cramming More Components onto Integrated Circuits," in which he noted that the complexity of ICs had been doubling every year since 1962. He made the bold prediction that this rate of progress would continue for another decade (Fig. 7.15a), and he speculated that the eventual impact of such chips would be enormous, not only for industry but also for individual consumers: "Integrated circuits will lead to such wonders as home computers – or at least terminals connected to a central computer – automatic controls for automobiles, or portable communications equipment."[10]

Moore made his predictions more than a decade before Steve Jobs and Stephen Wozniak produced the first mass-market personal computer and sixteen years before the appearance of the IBM PC. Caltech engineering professor Carver Mead (B.7.11) dubbed Gordon Moore's prediction *Moore's law*, but it took a long time for Moore to get used to using the name! In 1975, Moore updated his law by suggesting that a doubling of the complexity of ICs every two years was more realistic. Nowadays, Moore's law is usually stated as a doubling of the number of transistors on a chip every eighteen to twenty-four months. This rapid, year-on-year decrease in size of transistors and the corresponding increase in complexity have continued for more than thirty-five years (Fig. 7.15b). Moore reviewed the status of his law again in 1995, at a time when the Intel Pentium microprocessor contained nearly five million transistors. His conclusion was: "The current prediction is that this is not going to stop soon."[11] Today, nearly fifty years after his initial prediction, there are now devices with more than one billion transistors.

The doubling of complexity embodied in Moore's law occurs primarily because each generation of semiconductor fabrication facility decreases the minimum feature size on the chip so that the individual transistors can be made smaller. What was not clear in 1965 when Moore made his prediction was whether *quantum tunneling* would prove to be a major limitation on how small the transistors could be made. Quantum tunneling is the process in quantum mechanics in which a quantum particle can tunnel through a barrier in a way that would be impossible for a classical particle. Moore asked Carver Mead at Caltech for advice on this problem. The results of Mead's investigation were stunning. This is how Mead described the first public presentation of the results of his analysis:

B.7.10. Gordon Moore with Robert Noyce. Moore has a BSc degree in chemistry from UC Berkeley and a PhD from Caltech. He joined Shockley Semiconductor Laboratory in California in 1956 before leaving to found Fairchild Semiconductor Corporation as one of Shockley's Traitorous Eight. He and cotraitor Noyce later left Fairchild to create Intel in 1968. Moore is probably best known for his observation, originally made in 1965, that the number of transistors on ICs would continue to double every year. Although this increase has slowed to a doubling in eighteen to twenty-four months, Moore's law has held true for nearly fifty years and is the foundation of the astounding IT revolution we see around us.

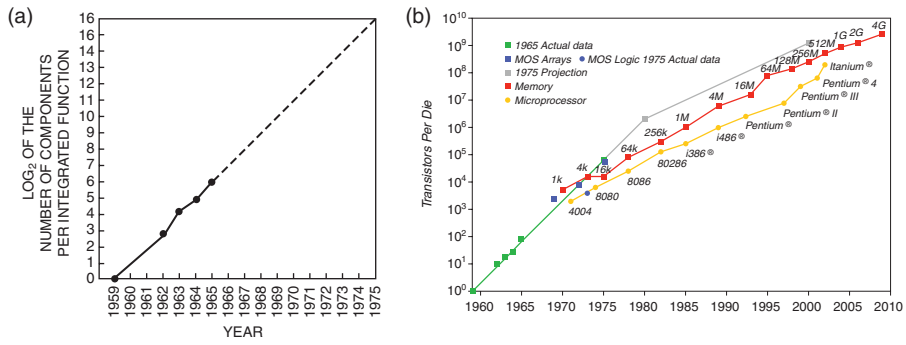


Fig. 7.15. Moore's law. (a) Gordon Moore's original prediction. (b) Moore's law still working after nearly fifty years.

In 1968, I was invited to give a talk at a workshop on semiconductor devices at Lake of the Ozarks. In those days you could get everyone who was doing cutting-edge work in one room, so that the workshops were where all the action was. I had been thinking about Gordon Moore's question, and decided to make it the subject of my talk. As I prepared for this event, I began to have serious doubts about my sanity. My calculations were telling me that, contrary to all the current lore in the field, we could **scale** down the technology such that **everything got better**: the circuits got more complex, they ran faster, and they took less power – WOW! That's a violation of Murphy's law that won't quit! But the more I looked at the problem, the more I became convinced that the result was correct, so I went ahead and gave the talk, to hell with Murphy! That talk provoked considerable debate, and at the time most people didn't believe the result. But by the time the next workshop rolled around, a number of other groups had worked through the problem for themselves, and we were pretty much in agreement. The consequences of this result for modern information technology have, of course, been staggering.[12]

The basic scaling principles underlying Moore's law were first described in papers in 1972 by Carver Mead and Bruce Hoeneisen of Caltech and by Robert Dennard and colleagues at IBM. But it was a paper from Dennard in 1974 that laid out the astonishing result – now called *Dennard scaling* – most clearly for the industry. This showed that shrinking the geometry and reducing the supply voltage led to both power reduction and performance improvement. In summary, Dennard scaling said that reducing the length, width, and gate oxide thickness of transistor features by a constant k results in transistors that are k^2

B.7.11. The citation for the 2002 award of the U.S. National Medal of Technology to Caltech professor Carver Mead, reads as follows: "For his pioneering contributions to microelectronics that include spearheading the development of tools and techniques for modern integrated circuit design, laying the foundation for fabless semiconductor companies, catalyzing the electronic design automation field, training generations of engineers that have made the United States the world leader in microelectronics technology, and founding more than twenty companies."[B2]

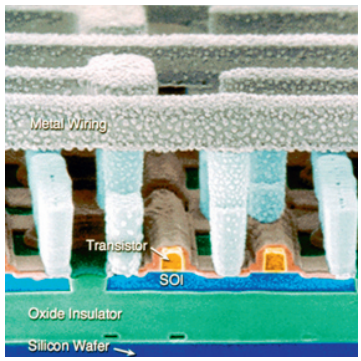


Fig. 7.16. Scanning Electron Microscope image of sub-100nm transistor developed at IBM.

times smaller, k times faster, and dissipate k^2 less power. IBM's MOS memory chips had a *feature size* of five microns at the time. Dennard and his colleagues projected shrinking the feature size to a fraction of a micron. The use of CMOS technology has allowed IBM to shrink the minimum feature dimension to well below 0.1 micron and enabled IBM to release the Power7 processor in 2010 with 1.2 billion transistors fabricated using a forty-five nanometer process (Fig. 7.16).

As chips grew smaller, not only could more complex chips be designed but also more of them could be produced on a single silicon wafer for the same cost. Moore's law has now held true for nearly fifty years and has been the engine for the vast growth in computing and information processing devices. It was 1970 when Intel produced the first 1,024 bit (1 kilobit) DRAM chip (Fig. 7.17). Only a year later, the first microprocessor, the Intel 4004, was produced with more than two thousand transistors etched in circuits ten microns wide. Just twenty-five years later, in 1995, the industry was producing DRAM chips with sixty-four million bits (64 megabit) and microprocessors like the Pentium with more than four million transistors and a minimum feature size of 0.35 microns. By the turn of the millennium, the industry had moved on to one thousand million bit (1 gigabit) DRAMs and to microprocessors such as the Pentium 4 with more than forty million transistors and a minimum feature size of 0.18 microns. By 2010, the minimum feature size was down to thirty-five nanometers and Intel, AMD, and Nvidia were producing chips with several billion transistors. There will soon be chips that are capable of storing a hundred thousand million bits – more bits than there are stars in our galaxy!

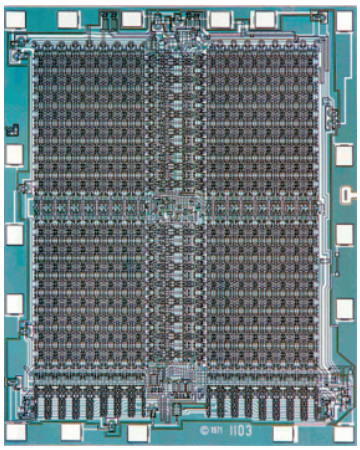


Fig. 7.17. Intel's revolutionary 1103 memory chip. One of the first customers was Xerox PARC where Chuck Thacker and Butler Lampson used the temperamental chip for the memory of famous Alto personal computer (see next chapter).

Powerful computers are now needed to design each new generation of chips: literally, we are using our present-day computers to design the next generation of computers. The international silicon industry produces the "Semiconductor Roadmap" that examines the engineering and design challenges required to keep on the track of Moore's law. Although the charts of the Semiconductor Roadmap boldly carry forward Moore's law many years into the future, there are many significant technical problems to be solved along the way. We will look at some possible solutions in Chapter 15.

Finally, in addition to these technical challenges, there is an economic one: the sheer cost of building the manufacturing facility for each new generation of chips (Fig. 7.18). Arthur Rock, one of the investors who helped Moore and Noyce raise funding to start Intel, is credited with Rock's law, which states: "A very small addendum to Moore's Law which says that the cost of capital equipment to build semiconductors will double every four years."[13] It was this spiraling cost of fabrication facilities that led Morris Chang, a Taiwanese engineer and entrepreneur, to pioneer the concept of a silicon foundry – essentially a "fab-for-hire." Companies can do their own chip design and then pay the foundry to manufacture their chips. Chang set up the Taiwan Semiconductor Manufacturing Company (TSMC) in 1987. It is now the world's largest foundry with more than $13 billion in revenue. According to James Plummer, Dean of the School of Engineering at Stanford University:

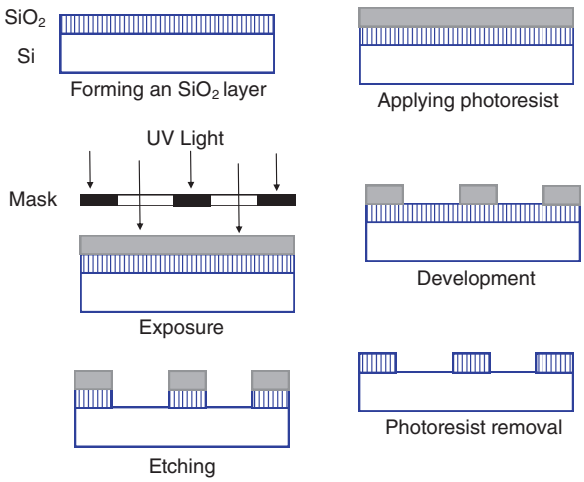


Fig. 7.18. ICs are produced by a complicated process called photolithography. The technique is similar to traditional photography but in this case we use a silicon wafer instead of a film of emulsion.

> Morris Chang completely changed the landscape of the semiconductor industry. He enabled start-ups to start with a few million dollars rather than a few hundred million. That makes a huge difference.[14]

Jen-Hsun Huang, a co-founder of Nvidia, credits TSMC with enabling all sorts of creative ideas in areas such as networking, consumer electronics, computers, and automotive technology to be turned into successful companies because "the barriers to getting your chips built, to realizing your imagination, disappeared."[15]

The end of the free lunch: parallel computing and the multicore challenge

As Moore's law predicted, designers and manufacturers have delivered smaller, faster chips requiring less power for more than four decades. But we have now reached the length scale at which the transistor's gate insulator is only a few atoms thick. Because the transistor is not a perfect switch, it leaks some current even when it is in the turned off state. As the transistor size decreases, if we continue to scale down the voltage, the current leakage increases exponentially. To keep this leakage under control, the voltage can no longer be scaled down with the dimensions of the chip. We can still shrink the size of the transistors and place more of them on a chip, but they will not be much faster than current generation transistors because the insulating silicon dioxide layer cannot get thinner and the power consumption of the chips limits our ability to clock the chip as fast as we could. As a result, chip architects have developed *multicore* architectures with multiple CPUs integrated on a single chip. Dual-core chips have been in widespread use for some time now, and quad-core chips are increasingly common. Eight-core chips are now available and the industry is experimenting with chips containing tens or even hundreds of cores (Fig. 7.19).

Performance improvement must now come from writing software that uses multiple cores together to solve a problem. For many types of applications it is

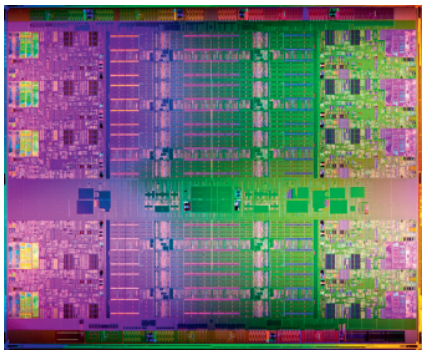


Fig. 7.19. The Intel Xeon E7 processor contains ten cores and a total of 2.6 billion transistors.

easy to make use of the parallelism of multiple cores. Such problems are said to exhibit *embarrassingly obvious* parallelism and, although obvious, this type of parallel computing application is very common. It is much more difficult to get speed-up by parallelizing a single application and then distributing the required computation over the multiple cores. Figure 7.20 illustrates three common types of parallelism.

One challenging problem that chip designers now face is the problem of *dark silicon*. This is the fact that although we will be able to make devices with many more transistors than today's chips, we will not be able to power all of them simultaneously due to power density limitations on the chip. Engineers are now actively looking for new ways to reduce power consumption in their chip designs. Multicore chips – requiring parallel programming – are therefore only a short-term solution to the problem of providing more performance per chip. In April 2005, Moore stated in an interview that it was clear that his law cannot be sustained indefinitely:

> In terms of size [of transistors] you can see that we're approaching the size of atoms which is a fundamental barrier, but it'll be two or three generations before we get that far – but that's as far out as we've ever been able to see. We have another 10 to 20 years before we reach a fundamental limit. By then they'll be able to make bigger chips and have transistor budgets in the billions.[16]

Unless we can come up with some radically new processor technologies, the end of Moore's law is in sight! In Chapter 15 we look at some possible solutions.

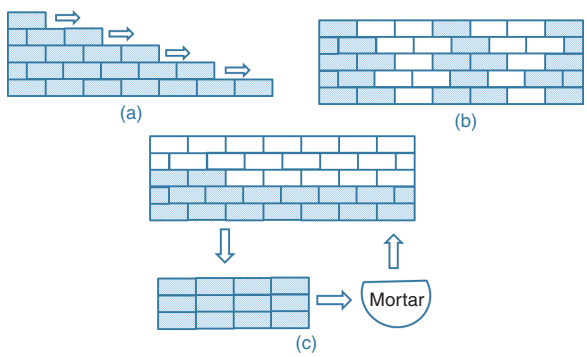


Fig. 7.20. Parallel Computing Paradigms illustrated by Fox's wall construction analogy: (a) "pipeline" parallelism, (b) "domain" parallelism, (c) "task" parallelism. (a) Pipeline parallelism in which each bricklayer is responsible for laying one row of bricks. Obviously the bricklayer for the second row cannot start until the first bricklayer has laid some bricks. Similarly, when the first bricklayer has finished the top row of bricks, the others are still finishing. This is a good analogy for the parallelism used in vector supercomputers (see section on supercomputers at the end of this chapter). (b) In domain parallelism, each bricklayer is responsible for a given section of the wall. Obviously at the edges of these domains the two bricklayers need to coordinate their activity. This is a good analogy for the parallelism used in distributed memory microprocessor-based parallel supercomputers. (c) Task parallelism is where each bricklayer is free to collect a brick and put it anywhere in the wall. It is a good analogy for the very common "embarrassingly obvious" parallelism of many types of application.

Key concepts

- Metals, insulators, and semiconductors
  - Band theory
  - Doped semiconductors: p-type and n-type
- Transistors
  - Point-contact and junction transistors
  - Mesa and planar transistors
- Integrated circuits
  - MOSFET transistors and CMOS technology
  - Random access memory: DRAM, SRAM, and ROM
- Moore's law
  - Microprocessors
  - Dennard scaling
  - Fabrication facilities and silicon foundries
- Multicore chips
  - Parallel computing
  - Dark silicon

A cartoon from Gordon Moore's original 1965 paper on Moore's law. Remarkably, Moore envisaged home computers being sold as a commodity long before the microprocessor gave birth to the personal computer (see next chapter).

A quantum theory primer

In the first half of the twentieth century, our understanding of matter underwent a profound revolution with the advent of quantum theory. Although a deep understanding is not needed for a reader's comprehension of this book, this section summarizes the essence of quantum theory that now underpins all of modern physics. This primer is not intended as a substitute for learning more about quantum theory but will be helpful in our understanding of the semiconductor materials that are central to the modern computer industry and of attempts to develop a new type of "quantum computer."

Although it is only about one hundred years old, quantum theory helped settle a scientific debate about the nature of light that began in the seventeenth century. Was light best described as a stream of particles, as Isaac Newton claimed, or was light some form of wave motion, as the Dutch physicist Christiaan Huygens had proposed? In 1801, the English physicist Thomas Young demonstrated that when two rays of light meet, they form a series of bright and dark bands called an *interference pattern*. Since these patterns are characteristic of waves – like ripples on a pond – the nature of light seemed to be settled. Then in 1921, Albert Einstein won the Nobel Prize for his explanation of the "photoelectric effect," the emission of electrons by a metal when exposed to light. Einstein found that light is made up of particle-like packets of energy called photons.

The theory of quantum mechanics emerged in the 1920s, pioneered by physicists such as Werner Heisenberg, Erwin Schrödinger, and Paul Dirac. Quantum mechanics provided successful "explanations" – in terms of its predictions – of the behavior of light, electrons, atoms, and nuclei in the microscopic world of atoms and nuclei (see Appendix 1). But there is a price to pay for this success: objects like photons and electrons behave in an essentially quantum mechanical way. All we can know about their motion is described by the evolution of a "probability wave." The wave equation discovered by Schrödinger describes how the probability wave for a quantum object – usually represented by the Greek letter ψ – evolves with time. We can only observe probabilities and, according to quantum theory, it is the square of this wave amplitude ψ that gives us the probability that we will observe the object at any given place and time.

Despite this emphasis on probability and uncertainty – epitomized by Heisenberg's famous "uncertainty principle" – quantum mechanics is the only theory capable of making accurate predictions for systems of atomic sizes or smaller. In addition, it is the very certainties of quantum mechanics that are responsible for the existence of the different chemical elements we see around us! According to quantum theory, electrons bound to an atom can only have certain energies. We can see how this comes about as follows. The problem of finding the allowed energies of an electron in an atom is analogous to finding the allowed energy levels of a charged particle in a potential well. In real life, we have to solve the Schrödinger wave equation in three-dimensional space but we can get some idea of the quantum solution for an atom by considering the problem of finding the allowed energy levels of an electron confined to a one-dimensional box. In the classical world, the electron in a box could have any energy; in the quantum world, the wave patterns must match the dimensions of the well – like finding the allowed wavelengths for a violin string. This means that only certain energies are allowed for the electron in a box (Fig. 7.21).

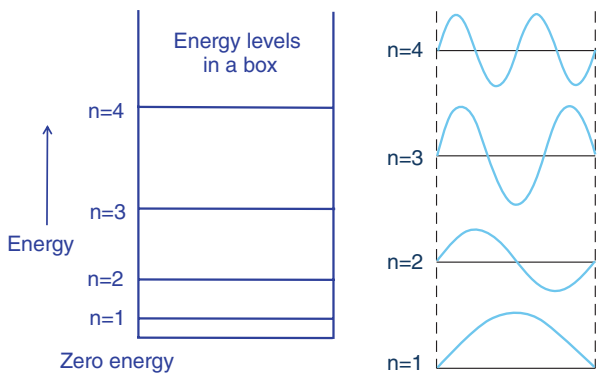


Fig. 7.21. Energy levels and wave functions for electrons confined to a box. (a) Energy levels for a quantum particle in a box. The energy levels are labeled by the quantum number "n." (b) This shows how the corresponding wave forms fit into the box. The wave function has to be zero at the edges.

The other two ingredients that we need from quantum theory to gain an understanding of the Periodic Table of the elements are *electron spin* and the *Pauli Exclusion Principle*. Electrons have "spin," somewhat like a spinning top, but unlike a classical top, an electron can only exist in one of two spin states, called "up" and "down." Pauli's Exclusion Principle says that only one electron is allowed in each quantum state. This means that in our electron potential box (Fig. 7.22) we can only put two electrons in the lowest energy state, called the *ground state*: one electron with spin up and the other with spin down. If we want to add another electron to the box, we have to give it more energy and place it in the next energy level – called the *first excited state*.

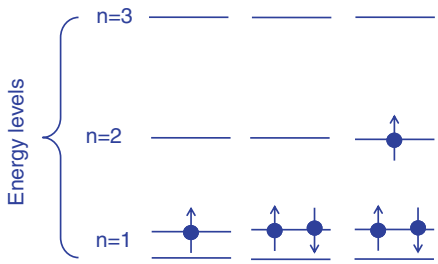


Fig. 7.22. Electrons in a box. The electrons can only fill up the energy levels according to Pauli's Exclusion Principle, which states that only one electron is allowed to occupy a quantum state. Each quantum level can therefore accommodate two electrons – one with spin up and one with spin down. The n = 1 level can therefore only accommodate two electrons: the next electron must go into the more energetic, first "excited" state, n = 2.

It is the exclusion principle – insisting that electrons have to occupy distinct quantum states – that gives the stability and volume of ordinary matter. As Richard Feynman says: "It is the fact that electrons cannot all get on top of each other that makes tables and everything else solid."[17] The exclusion principle applies to all "matter-like" quantum objects such as electrons, protons, and neutrons. For "radiation-like" objects such as photons, the exclusion principle does not apply, and we can put as many photons as we like into the same quantum state. This has led to amazing applications such as lasers and superconductivity.

Armed with these fundamental quantum concepts, we are now able to explain the difference between metals, semiconductors, and insulators. In a later chapter we will see how these quantum ideas are being used to build a new type of quantum computer.

Supercomputers

Cray computers were the first commercially successful supercomputers. They were designed and constructed in the small remote town of Chippewa Falls in Wisconsin. As we have seen in Figure 2.20, fetching data from main memory takes many more clock cycles than the number of cycles required to perform an operation on data that is already in the registers. For programs that involve operations on *vectors* – a one-dimensional array of numbers – Seymour Cray (B.7.12) realized that it would be possible to set up a pipeline that can hide much of the latency caused by fetching the data from the main memory and getting a new instruction for each operation. For example, if we need to multiply two vectors, we need to apply the same operation on each pair of elements of the vectors. We can therefore arrange things so that while the CPU is multiplying the first two elements of the two vectors, the computer is already fetching the next two elements to be multiplied. This is an implementation of pipeline parallelism as illustrated in Figure 7.20a. Roger Hockney, a British computer scientist who pioneered serious benchmarking of parallel computers, characterized vector supercomputers in terms of a parameter he called "$n_{1/2}$" – this is the length of the vector that was required for the supercomputer to reach half of its maximum speed. It is essentially a measure of the distance of main memory from the registers in terms of cycles. The success of the Cray-1 was due to its much smaller $n_{1/2}$ than its rival supercomputers. This meant that it was able to achieve high speeds on problems requiring much shorter vector operations.

B.7.12. The name of Seymour Cray (1925–96) has become inseparable from supercomputers. Cray was one of the first to recognize that maximizing the speed of moving data between the processor and memory and hiding memory latency using a vector pipeline was the key to achieving high-performance computing speeds. He is pictured here next to a Cray-1 supercomputer.

Today, the parameters of Cray-1 look almost laughably slow but in the 1970s the Cray-1 was at the cutting edge, with a clock speed of 80 megaHertz (MHz) and an 8 megabyte (Mbyte) main memory (Fig. 7.23). To minimize signal delays, the frame of the computer was bent into a C shape, thus enabling the use of shorter wires. The speed was eighty million operations per second on floating point numbers or eighty megaflop/s (Mflop/s). The cooling and power distribution required many ingenious solutions. The Cray-1 used liquid Freon instead of water for cooling, and copper plates between circuit boards. No wonder Cray joked that becoming a plumber was the peak of his career:

> I made square cabinets with glass doors and aluminum and imitation walnut trim [CDC 7600] and cylindrical cabinets [CDC 8600 and CRAY-1]. After a while, I got tired making cabinets and so I decided I needed to go into a new profession. Now, I am into plumbing [CRAY-2, 3]. It is a lot less prestige, but I am making even more money.[18]

Fig. 7.23. A photograph of two Cray-1 supercomputers at Lawrence Livermore National Laboratory. The upholstered bench around the computing tower hides the power supply and the intricate network of Freon cooling pipes. The supercomputer was sometimes referred to as "the most expensive seat in the world."

Cray supercomputers were expensive and only automotive firms and large national laboratories could afford them. The first Cray-1 was installed in Los Alamos in 1976. The applications were weapon simulations, weather prediction, and cryptoanalysis. Each new Cray machine was shipped with a case of Leinenkugel's beer, also a product of Chippewa Falls. In the automotive industry, they were used for the first car-crash simulations.

Such simulations have now become an essential element of car design with detailed crash simulation using models of the driver and passengers.

By the 1980s microprocessors based on new transistor technology (CMOS) started to appear as alternative building blocks of supercomputers. Cray was rather skeptical and when he was asked whether he had considered building the next generation of Cray computers on the new components he famously said "If you have a heavy load to move. What would you rather use a pair of oxen or hundred of chicken."[19]

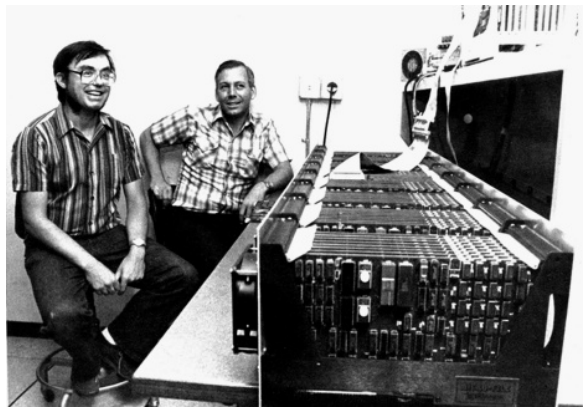


Fig. 7.24. Geoffrey Fox and Chuck Seitz pictured next to the Caltech Cosmic Cube machine. This parallel computer became operational in October 1983 and contained sixty-four nodes each with 128 kilobyte memory. A computing node was made up from an Intel 8086 processor with an 8087 coprocessor for fast floating-point operations. The nodes were linked together in a so-called hypercube topology – several cubes connected together – for minimizing communication delays between nodes. The size of the computer was only six cubic feet and drew less than a kilowatt of power. The Cosmic Cube and its successors represented a serious challenge for the Cray vector supercomputers. Today all the highest-performing machines use a distributed memory message-passing architecture similar to the Caltech design.

However, in the early 1980s, Geoffrey Fox and Chuck Seitz at Caltech put together a parallel computer called the Cosmic Cube. In essence, this was a collection of IBM PC boards, each with an Intel microprocessor and memory, connected together by a so-called *hypercube network*. The importance of the Cosmic Cube experience was that Fox and Seitz demonstrated for the first time that it was both feasible and realistic to use such "distributed memory" parallel computers to solve challenging scientific problems. In this case, programmers need to exploit domain parallelism as shown in Figure 7.20b. Instead of the latency caused by filling and emptying a vector pipeline, the overhead in such distributed memory programs comes from the need to exchange information at the boundaries of the domains since the data is subdivided among the different nodes of the machine. This style of parallel programming is called *message passing*.

Today, all the highest performance computers use such a distributed memory, message passing architecture, albeit with a variety of different types of networks connecting the processing nodes. Instead of the Cray-1's eighty Mflop/s peak performance, we now have distributed memory supercomputers with top speeds of teraflop/s and petaflop/s, while gigaflop/s performance is now routinely available on a laptop! The supercomputing frontier is to break the exaflop/s barrier and there are now U.S., Japanese, European, and Chinese companies taking on this challenge. In answer to Cray's sarcastic comment about hundreds of chickens, Eugene Brooks III, one of the original Cosmic Cube team at Caltech, characterized the success of commodity-chip based, distributed memory machines as "the attack of the killer micros."[20]