

CHAPTER 14

Integrated Circuits (ICs)

THE FIRST INTEGRATED CIRCUITS

In the early days of semiconductors, transistors and other electronic components were available only in individual packages. These *discrete components* were laid out on a circuit board and connected by hand using separate wires. At that time, an electronic memory element capable of storing a single binary bit of data cost more than \$2. By comparison, in the early 1990s, enough logic gates to store 5000 bits of data cost less than a penny.¹ This vast reduction in price was primarily due to the invention of the *Integrated Circuit* (IC).²

A functional electronic circuit requires transistors, resistors, diodes, etc., and the connections between them. A *monolithic integrated circuit* (the *monolithic* qualifier is usually omitted) has all of these components formed on the surface layer of a sliver, or chip, of a single piece of semiconductor; hence, the term *monolithic*, meaning “seamless.” Although a variety of semiconductor materials are available, the most commonly used is silicon, and integrated circuits are popularly known as *silicon chips*. (Unless otherwise noted, the remainder of these discussions will assume integrated circuits based on silicon as the semiconductor.)

To a large extent, the demand for miniaturization was driven by the demands of the American space program. For some time, people had been thinking that it would be a good idea to be able to fabricate entire circuits on a single piece of semiconductor. The first public discussion of this idea is credited to a British radar expert, Geoffrey William Arnold (G.W.A.) Dummer (1909–2002), in a paper presented in 1952. However, it was not until the summer of 1958 that Jack St. Clair Kilby (1923–2005), working for Texas Instruments, succeeded in

¹I just did a search on the web as I pen these words. I found a 1-gigabyte memory stick for \$9.99, which equates to more than 8.5 million bits of memory per penny.

²In conversation, IC is pronounced by spelling it out as “I-C”.

fabricating multiple components on a single piece of semiconductor. Kilby's first prototype was a phase shift oscillator comprising five components on a piece of germanium, half an inch long and thinner than a toothpick. Although manufacturing techniques subsequently took different paths from those used by Kilby, he is still credited with the creation of the first true integrated circuit.

Around the same time that Kilby was working on his prototype, two of the founders of Fairchild Semiconductors—the Swiss physicist Jean Hoerni (1924–1997) and the American physicist Robert Noyce (1927–1990)—were working on more efficient processes for creating these devices. Between them, Hoerni and Noyce invented the *planar process*, in which optical lithographic techniques are used to create transistors, insulating layers, and interconnections on integrated circuits.

By 1961, Fairchild and Texas Instruments had announced the availability of the first commercial planar integrated circuits comprising simple logic functions. This announcement marked the beginning of the mass production of integrated circuits. In 1963, Fairchild produced a device called the 907 containing two logic gates, each of which consisted of four bipolar transistors and four resistors. The 907 also made use of *isolation layers* and *buried layers*, both of which were to become common features in modern integrated circuits.

During the mid-1960s, Texas Instruments introduced a large selection of basic “building block” ICs called the 54xx (“fifty-four-hundred”) series and the 74xx (“seventy-four-hundred”) series, which were specified for military and commercial use, respectively. These *jelly bean* devices each contained small amounts of simple logic. For example, a 7400 device contained four 2-input NAND gates, a 7402 contained four 2-input NOR gates, and a 7404 contained six NOT (inverter) gates.

TI's 54xx and 74xx series were implemented in *Transistor-Transistor Logic* (TTL). By comparison, in 1968, RCA introduced a somewhat equivalent CMOS-based library of parts called the 4000 (“four thousand”) series.

In 1967, Fairchild introduced a device called the *Micromosaic*, which contained a few hundred transistors. The key feature of the Micromosaic was that the transistors were not initially connected to each other. A designer used a computer program to specify the function the device was required to perform, and the program determined the necessary transistor interconnections and constructed the photo-masks required to complete the device. The Micromosaic is credited as the forerunner of the modern *Application-Specific Integrated Circuit* (ASIC),³ and also as the first real application of computer-aided design. In 1970,

³ASICs are discussed in more detail in *Chapter 17: Application-Specific Integrated Circuits (ASICs)*.

Fairchild introduced the first 256-bit static RAM, called the 4100, while Intel announced the first 1024-bit dynamic RAM, called the 1103, in the same year.⁴

One year later, in 1971, Intel introduced the world's first *microprocessor* (μP), the 4004, which was conceived and created by Marcian "Ted" Hoff, Stan Mazor, and Federico Faggin. Also referred to as a *computer-on-a-chip*, the 4004 contained only around 2300 transistors and could execute 60,000 operations per second.

AN OVERVIEW OF THE FABRICATION PROCESS

The construction of integrated circuits requires one of most exacting production processes ever developed. The environment must be at least a thousand times cleaner than that of an operating theater, and impurities in materials have to be so low as to be measured in parts per billion.⁵ The process begins with the growing of a single crystal of pure silicon in the form of a cylinder with a diameter that can be anywhere up to 300 mm.⁶ The cylinder is cut into paper-thin slices called *wafers*, which are approximately 0.2 mm thick (Figure 14.1).

The thickness of the wafer is determined by the requirement for sufficient mechanical strength to allow it to be handled without damage. The actual thickness necessary for the creation of the electronic components is less than $1\ \mu\text{m}$ (one-millionth of a meter). After the wafers have been sliced from the cylinder, they are polished to a smoothness rivaling the finest mirrors.

The most commonly used fabrication process is *optical lithography*, in which *Ultraviolet Light* (UV) is passed through a stencil-like⁷ object called a *photo-mask*, or just *mask* for short. This square or rectangular mask carries patterns formed by areas that are either transparent or opaque to ultraviolet frequencies (similar in concept to a black-and-white photographic negative) and the resulting image is projected

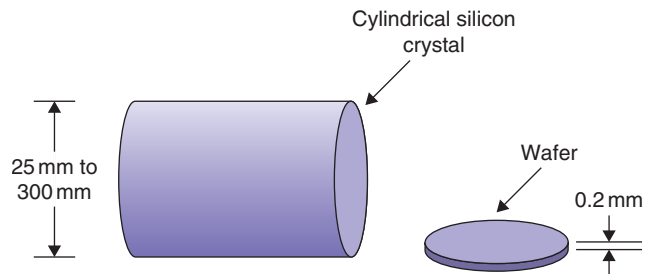


FIGURE 14.1
Creating silicon wafers.

⁴Memory devices are discussed in *Chapter 15: Memory ICs*.

⁵If you took a bag of flour and added a grain of salt, this would be impure by comparison.

⁶When the first edition of this tome hit the streets in 1995, the maximum wafer diameter was 200 mm. By 2002, leading manufacturers were working with 300 mm diameter wafers, which are relatively standard at the time of this writing. By 2012, it is expected that a shift will have started to use 450 mm diameter wafers.

⁷Just in case you were wondering, the term *stencil* comes from the Middle English word *stanseld*, meaning "adorned brightly."

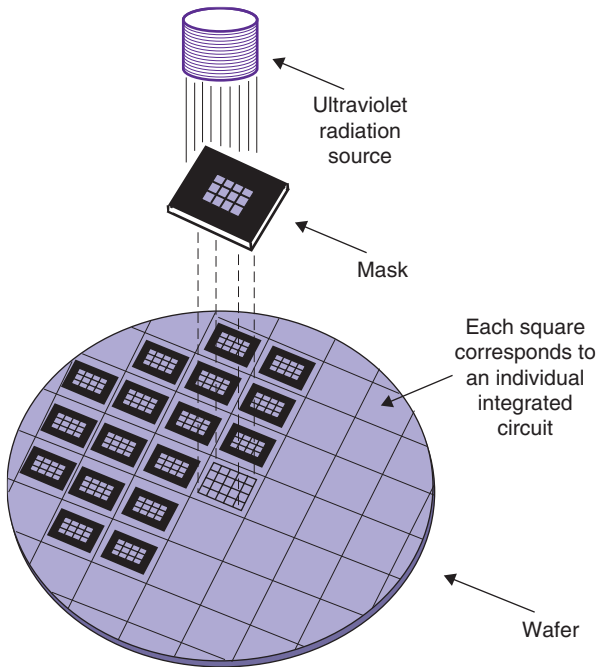


FIGURE 14.2
The opto-lithographic step-and-repeat process.

onto the surface of the wafer. By means of some technical wizardry that we'll consider in the next topic, we can use the patterns of ultraviolet light to "grow" corresponding structures in the silicon. The simple patterns shown in the following diagrams were selected for reasons of clarity; in practice, a mask can contain hundreds of millions (sometimes billions) of fine lines and geometric shapes (Figure 14.2).

Each wafer can contain hundreds or thousands of identical integrated circuits. The pattern projected onto the wafer's surface corresponds to a single integrated circuit, which is typically in the region of $1\text{ mm} \times 1\text{ mm}$ to $10\text{ mm} \times 10\text{ mm}$, but some chips are $15\text{ mm} \times 15\text{ mm}$, and some are even larger. After the area corresponding to one integrated circuit has been exposed, the wafer is moved and the process is repeated until the same pattern

has been replicated across the whole of the wafer's surface. This technique for duplicating the pattern is called a *step-and-repeat process*.

As we shall see, multiple layers are required to construct the transistors (and other components), where each layer requires its own unique mask. Once all of the transistors have been created, similar techniques are used to lay down the tracking (wiring) layers that connect the transistors together.

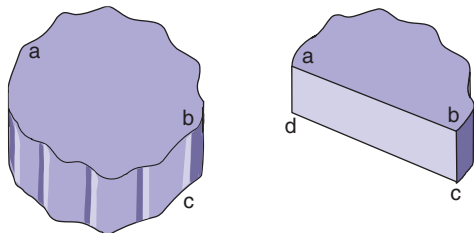


FIGURE 14.3
Small area of silicon somewhere on the wafer.

A SLIGHTLY MORE DETAILED LOOK AT THE FABRICATION PROCESS

To illustrate the manufacturing process in more detail, we will consider the construction of a single NMOS transistor occupying an area far smaller than a speck of dust. For reasons of electronic stability, the majority of processes begin by lightly doping the entire wafer to form either N-type or, more commonly, P-type silicon. However, for the purposes of this discussion, we will assume a process based on a pure silicon wafer (Figure 14.3).

Assume that the small area of silicon shown here is sufficient to accommodate a single transistor in the middle of one of the integrated circuits residing

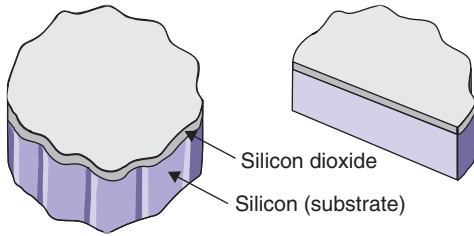


FIGURE 14.4
Grow or deposit a layer of silicon dioxide.

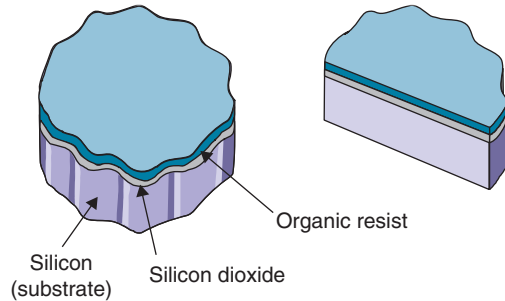


FIGURE 14.5
Add a layer of organic resist.

somewhere on the wafer. During the fabrication process, the wafer is often referred to as the *substrate*, meaning “base layer.” A common first stage is to either grow or deposit a thin layer of silicon dioxide (glass) across the entire surface of the wafer by exposing it to oxygen in a high-temperature oven (Figure 14.4).

After the wafer has cooled, it is coated with a thin layer of organic resist,⁸ which is first dried and then baked to form an impervious layer (Figure 14.5).

A mask is created and ultraviolet light is applied. The ionizing ultraviolet radiation passes through the transparent areas of the mask into the resist, silicon dioxide, and silicon. The ultraviolet breaks down the molecular structure of the resist, but does not have any effect on the silicon dioxide or the pure silicon (Figure 14.6).

As was previously noted, the small area of the mask shown here is associated with a single transistor. The full mask for a high-end integrated circuit can comprise hundreds of millions (sometimes billions) of similar patterns.⁹ After the area under the mask has

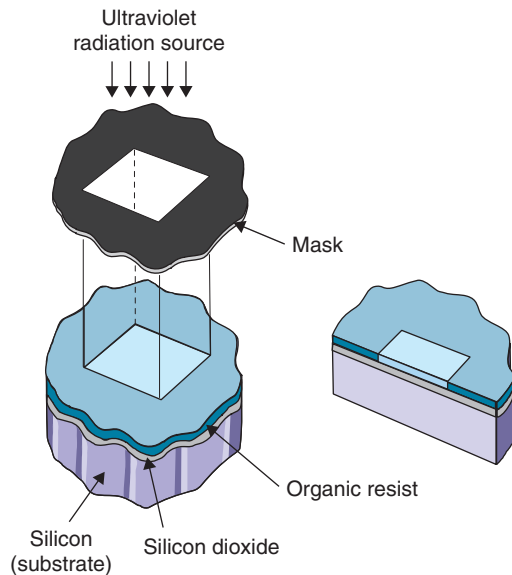


FIGURE 14.6
Use ultraviolet light to degrade exposed resist.

⁸The term *organic* is used because this type of resist is a carbon-based compound, and carbon is the key element for life as we know it.

⁹For the purpose of these discussions, I’m thinking of high-end digital integrated circuits containing tens or hundreds of millions of logic gates. It is, of course, possible to have simpler devices containing much fewer elements. In fact, only a couple of days ago as I pen these words, I was chatting with an analog designer who had just designed a chip containing only nine painstakingly handcrafted transistors.

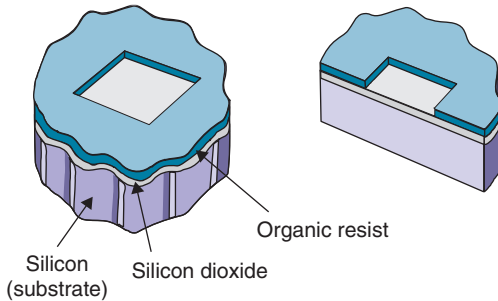


FIGURE 14.7
Dissolve the degraded resist with an organic solvent.

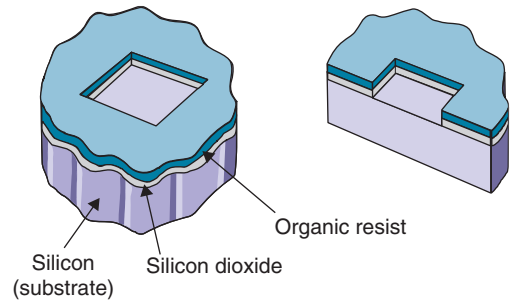


FIGURE 14.8
Etch the exposed silicon dioxide.

been exposed, the wafer is moved, and the process is repeated until the pattern has been replicated across the entire surface of the wafer, once for each integrated circuit. The wafer is then bathed in an organic solvent to dissolve the degraded resist. Thus, the pattern on the mask has been transferred to a series of corresponding patterns in the resist (Figure 14.7).

A process by which ultraviolet light passing through the transparent areas of the mask causes the resist to be degraded is known as a *positive-resist* process; *negative-resist* processes are also available. In a negative-resist process, the ultraviolet radiation passing through the transparent areas of the mask is used to cure (harden) the resist, and the remaining uncured areas are then removed using an appropriate solvent.

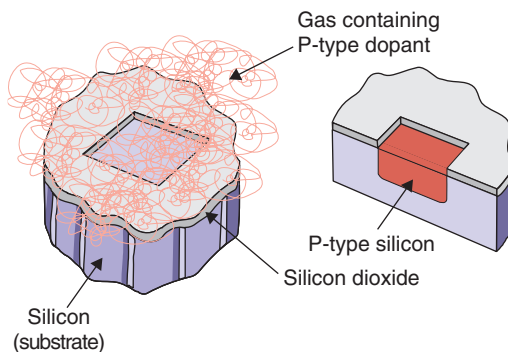


FIGURE 14.9
Dope the exposed silicon.

After the unwanted resist has been removed, the wafer undergoes a process known as *etching*, in which an appropriate solvent is used to dissolve any exposed silicon dioxide without having any effect on the organic resist or the pure silicon (Figure 14.8).

The remaining resist is removed using an appropriate solvent. Next, the wafer is placed in a high-temperature oven where it is exposed to a gas containing the selected dopant (a P-type dopant in this case). The atoms in the gas diffuse into the substrate, resulting in a region of doped silicon (Figure 14.9).¹⁰

The remaining silicon dioxide layer is removed by means of an appropriate solvent that doesn't affect the silicon substrate (including the doped regions).

¹⁰In some processes, diffusion is augmented with *ion implantation techniques*, in which beams of ions (which were introduced in *Chapter 2: Atoms, Molecules, and Crystals*) are fired at the wafer to alter the type and conductivity of the silicon in selected regions.

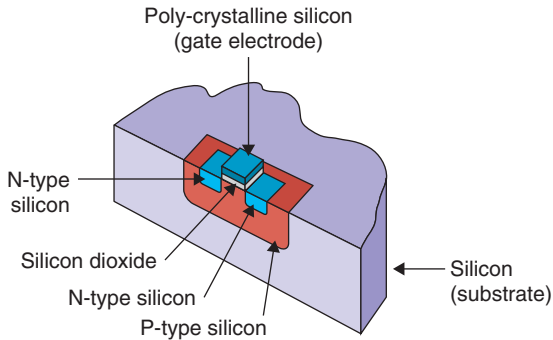


FIGURE 14.10
Add N-type diffusion regions and the gate electrode.

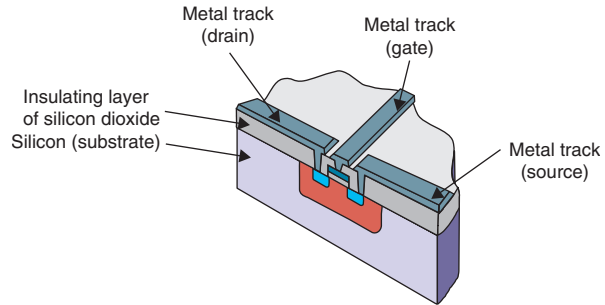


FIGURE 14.11
Add a layer of metal tracks.

Additional masks and variations on the process are used to create two N-type diffusion regions, a gate electrode, and a layer of insulating silicon dioxide between the substrate and the gate electrode (Figure 14.10).

In the original MOS technologies, the gate electrode was metallic: hence, the “metal-oxide semiconductor” appellation. In modern processes, however, the gate electrode is formed from *poly-crystalline silicon* (often abbreviated to *poly-silicon* or even just *poly*), which is also a good conductor.

The N-type diffusions form the transistor’s *source* and *drain* regions (you might wish to refer back to *Chapter 4: Semiconductors (Diodes and Transistors)* to refresh your memory at this point). The gap between the source and drain is called the *channel*. To provide a sense of scale, the length of the channel in one of today’s state-of-the-art technologies is measured in a few tens of billionths of a meter (see also the discussions on *device geometries* later in this chapter).

Another layer of insulating silicon dioxide is now grown or deposited across the surface of the wafer. Using lithographic techniques similar to those described above, holes are etched through the silicon dioxide in areas in which it is desired to make connections, and a *metallization layer* of interconnections called *tracks* (you can think of them as wires) is deposited (Figure 14.11).^{11,12}

¹¹In the early days, the tracks were formed out of aluminum, because this didn’t react with (diffuse into) the insulating silicon dioxide. As the structures on chips got smaller and the tracks got thinner and narrower, their resistance increased; eventually, aluminum could no longer do the job. Thus, modern chips use copper interconnect; this requires more steps to isolate the copper from the silicon dioxide, but it’s worth it because copper is a much better conductor than aluminum.

¹²Silver is the best electrical and thermal conductor of any metal (followed by copper and then by gold), but it’s even harder to work with than copper when it comes to building silicon chips.

The end result is an NMOS transistor; a logic 1 on the track connected to the *gate* terminal will turn the transistor ON, thereby enabling current to flow between its *source* and *drain* terminals. An equivalent PMOS transistor could have been formed by exchanging the P-type and N-type diffusion regions. By varying the structures created by the masks, components such as resistors and diodes can be fabricated at the same time as the transistors. The tracks are used to connect groups of transistors to form primitive logic gates and to connect groups of these gates to form more complex functions.

An integrated circuit contains three distinct levels of conducting material: the *diffusion* layer at the bottom, the *polysilicon* layers in the middle, and the *metallization* layers at the top. In addition to forming components, the diffusion layer may also be used to create embedded wires. Similarly, in addition to forming gate electrodes, the polysilicon may also be used to interconnect components. There may be several layers of polysilicon and multiple layers of metallization, with each pair of adjacent layers separated by an insulating layer of silicon dioxide. The layers of silicon dioxide are selectively etched with holes that are filled with conducting metal and are known as *vias*; these allow connections to be made between the various tracking layers.

Early integrated circuits typically supported only two layers of metallization. The tracks on the first layer predominantly ran in a “North-South” direction, while the tracks on the second predominantly ran “East-West.”¹³ As the number of transistors increased, engineers required more and more tracking layers. The problem is that when a layer of insulating silicon dioxide is deposited over a tracking layer, you end up with slight “bumps” where the tracks are (like snow falling over a snoozing polar bear—you end up with a bump).

After a few tracking layers, the bumps are pronounced enough that you can’t continue. The answer is to re-planarize the wafer (smooth the bumps out) after each tracking and silicon dioxide layer combo has been created. This is achieved by means of a process called *Chemical Mechanical Polishing* (CMP), which returns the wafer to a smooth, flat surface before the next tracking layer is added. Using this process, high-end silicon chips can support up to ten tracking layers.

¹³In 2001, a group of companies announced a new chip interconnect concept called *X Architecture* in which logic functions on chips are wired together using diagonal tracks (as opposed to traditional North-South and East-West tracking layers). It is claimed that this diagonal interconnect strategy can increase chip performance by 10% and reduce power consumption by 20%. I know of a couple of chips that have been fabricated using this technology, but it has not yet gained widespread adoption.

AN INTRODUCTION TO THE PACKAGING PROCESS

There are a wide variety of different packaging techniques. We'll start by considering one of the simplest packaging styles (one that's easy to understand) and then we'll consider some slightly more complex techniques. (We'll also look at some really sophisticated packaging technologies in *Chapter 20: Advanced Packaging Techniques*.)

In the case of our simple packaging technique, relatively large areas of aluminum or copper called *pads* are constructed at the edges of each integrated circuit for testing and connection purposes. Some of the pads are used to supply power to the device, while the rest are used to provide input and output signals to the components in the chip ([Figure 14.12](#)).

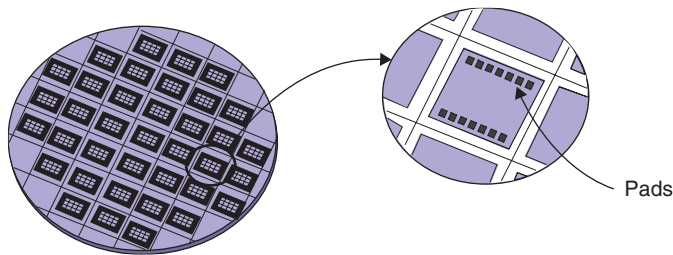


FIGURE 14.12

Power and signal pads.

In a step known as *overglassing*, the entire surface of the wafer is coated with a final *barrier layer* (or *passivation layer*) of silicon dioxide or silicon nitride, which provides physical protection for the underlying circuits from moisture and other contaminants. One more lithographic step is required to pattern holes in the barrier layer to allow connections to be made to the pads. In some cases, additional metallization may be deposited on the pads to raise them fractionally above the level of the barrier layer. Augmenting the pads in this way is known as *silicon bumping*.

The entire fabrication process requires numerous lithographic steps, each involving an individual mask and layer of resist to selectively expose different parts of the wafer.

The individual integrated circuits are tested while they are still part of the wafer in a process known as *wafer probing*. An automated tester places probes on the device's pads, applies power to the power pads, injects a series of signals into the input pads, and monitors the corresponding signals returned from

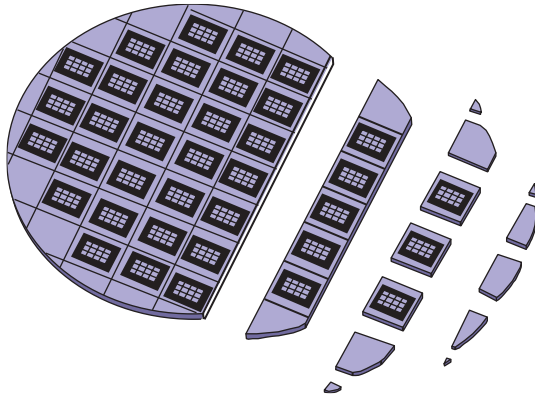


FIGURE 14.13
Die separation.

the output pads. Each integrated circuit is tested in turn, and any device that fails the tests is automatically tagged with a splash of dye for subsequent rejection. The *yield* is the number of devices that pass the tests as a percentage of the total number fabricated on that wafer.

The completed circuits, known as *die*,¹⁴ are separated by marking the wafer with a diamond scribe and fracturing it along the scribed lines (much like cutting a sheet of glass or breaking up a Kit Kat® bar) (Figure 14.13).

Following separation, the majority of the die are packaged individually. Since there are almost as many packaging technologies as there are device manufacturers, we will initially restrain ourselves to a low-end “cheap-and-cheerful” process. First, the die is attached to a metallic lead frame using an adhesive (Figure 14.14).

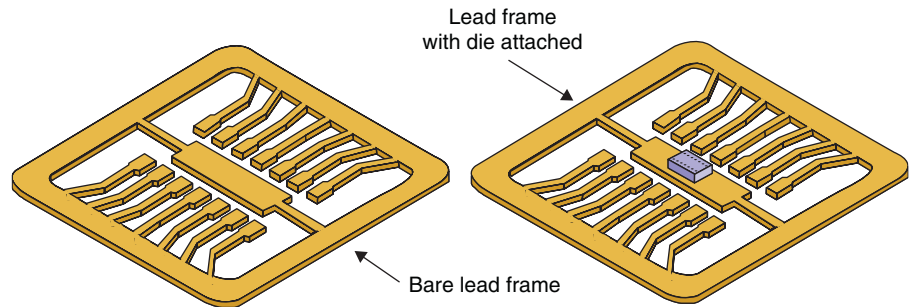


FIGURE 14.14
The die is attached to a lead frame.

One of the criteria used when selecting the adhesive is its ability to conduct heat away from the die when the device is operating. An automatic wire-bonding tool connects the pads on the die to the leads on the lead frame with wire bonds finer than a human hair.¹⁵ The whole assembly is then encapsulated in a block of plastic or epoxy (Figure 14.15).

¹⁴The plural of *die* is also *die* (in much the same way that *herring* is the plural of *herring* as in “a shoal of herring”).

¹⁵Human hairs range in thickness from around 0.07 mm to 0.1 mm. A hair from a typical blond lady’s head is approximately 0.075 mm (three-quarters of one-tenth of a millimeter) in diameter. By comparison, integrated circuit bonding wires are typically one-third this diameter, and they can be even thinner.

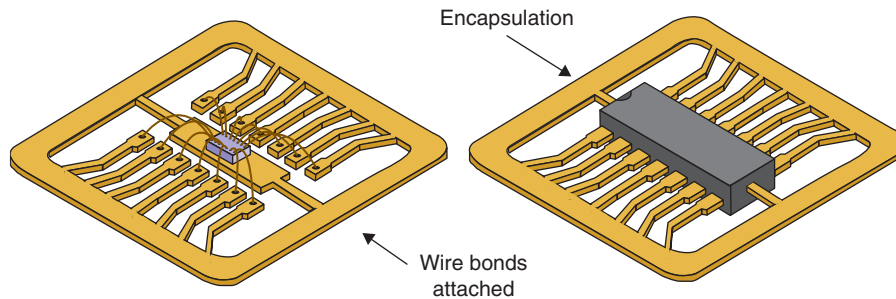


FIGURE 14.15
Wire bonding and encapsulation.

A dimple or notch is formed at one end of the package so that the users will know which end is which. The unused parts of the lead frame are cut away and the device's leads, or pins, are shaped as required; these operations are usually performed at the same time (Figure 14.16).

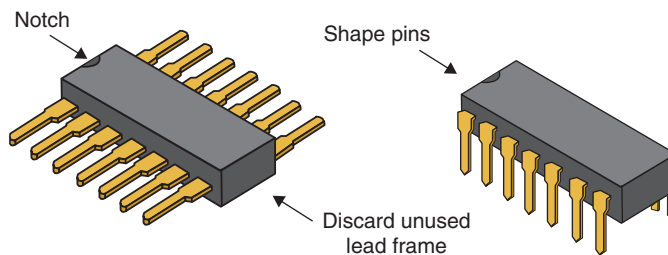


FIGURE 14.16
Discard unused lead frame and shape pins.

An individually packaged integrated circuit consists of the die and its connections to the external leads, all encapsulated in the protective package. The package protects the silicon from moisture and other impurities and helps to conduct heat away from the die when the device is operating.

As was previously noted, there is tremendous variety in the size and shape of packages. A rectangular device with pins on two sides, as illustrated here, is called a *Dual In-Line* (DIL) package or a *DIP*.¹⁶ A standard 14-pin packaged device is approximately 18 mm long by 6.5 mm wide by 2.5 mm deep, and has 2.5-mm spaces between pins. An equivalent *Small Outline Package* (SOP) could be as small as 4 mm long by 2 mm wide by 0.75 mm deep, and have 0.5-mm

¹⁶Invented at Fairchild in 1965, DIL packages were the mainstay of the industry through the 1970s and 1980s. Their use started to decline in the 1990s as *Surface Mount Technology* (SMT) gained in popularity, but you can still find new components from some manufacturers in these packages to this day; for example, Microchip Technology (<http://www.microchip.com>) continue to provide many of their latest-and-greatest PIC[®] microcontrollers in both DIL and SMT packages.

spaces between pins. Other packages can be square and have pins on all four sides, and some have an array of pins protruding from the base.

The shapes into which the pins are bent depend on the way the device is intended to be mounted on a circuit board. The package described above has pins that are intended to go all the way through the circuit board using a mounting technique called *Lead Through Hole* (LTH). By comparison, the packages associated with a technique called *Surface Mount Technology* (SMT) have pins that are bent out flat, and which attach to only one side (surface) of the circuit board [an example of this is shown in *Chapter 18: Printed Circuit Boards (PCBs)*].

It's important to remember that the example discussed above reflects a very simple packaging strategy for a device with very few pins. By 2002, some integrated circuits (and their packages) had as many as 1000 pins; by 2008, some high-end components were available with 2000 pins or more. This multiplicity of pins requires a very different packaging approach. For example, the pads on the die are no longer restricted to its periphery, but are instead located over the entire face of the die. A minute ball of solder is then attached to each pad, and the die is flipped over and attached to the package substrate (this is referred to as a *flip-chip* technique). Each pad on the die has a corresponding pad on the package substrate, and the package-die combo is heated so as to melt the solder balls and form good electrical connections between the die and the substrate ([Figure 14.17](#)).

Eventually, the die will be encapsulated in some manner to protect it from the outside world. The package's substrate itself may be made out of the same

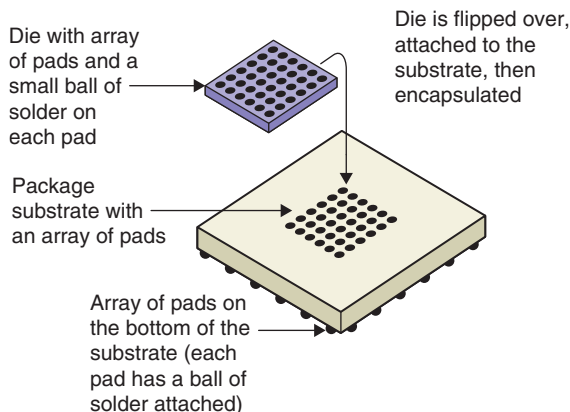


FIGURE 14.17

A flip-chip ball grid array packaging technique.

material as a printed circuit board, or out of ceramic, or out of some even more esoteric material. Whatever its composition, the substrate will contain multiple internal wiring layers that connect the pads on the upper surface to pads (or pins) on the lower surface. The pads (or pins) on the lower surface (the side that actually connects to the circuit board) will be spaced much further apart—relatively speaking—than the pads that connect to the die. At some stage the package will have to be attached to a circuit board. In one technique known as a *Ball Grid Array* (BGA), the package has an array of pads on its bottom surface, and a small ball of solder

is attached to each of these pads. Each pad on the package will have a corresponding pad on the circuit board, and heat is used to melt the solder balls and form good electrical connections between the package and the board.

Modern packaging technologies are extremely sophisticated. Some ball grid arrays have pins spaced 0.3 mm (one-third of a millimeter) apart! In the case of *Chip-Scale Package* (CSP) technology, the package is barely larger than the die itself. In the early 1990s, some specialist applications began to employ a technique known as *die stacking*, in which several bare die are stacked on top of each other to form a sandwich. The die are connected together and then packaged as a single entity.

As was previously noted, there are a wide variety of integrated packaging styles. There are also many different ways in which the die can be connected to the package. We will introduce a few more of these techniques in future chapters.¹⁷

INTEGRATED CIRCUITS VERSUS DISCRETE COMPONENTS

The tracks that link components inside an integrated circuit have widths measured in fractions of a millionth of a meter, and lengths measured in millimeters. By comparison, the tracks that link components on a circuit board are orders of magnitude wider, and have lengths measured in tens of centimeters. Thus, the transistors used to drive tracks inside an integrated circuit can be much smaller than those used to drive their circuit board equivalents, and smaller transistors use less power. Additionally, signals take a finite time to propagate down a track, so the shorter the track, the faster the signal.

A single integrated circuit can contain hundreds of millions of transistors (see also the topic titled *How Many Transistors?* later in this chapter). A similar design based on discrete components would be tremendously more expensive in terms of price, size, operating speed, power requirements, and the time and effort required to design and manufacture the system. Additionally, every solder joint on a circuit board is a potential source of failure, which affects the reliability of the design. Integrated circuits reduce the number of solder joints and, hence, improve the reliability of the system.

¹⁷Additional packaging styles and alternative mounting strategies are presented in *Chapter 18: Printed Circuit Boards (PCBs)*, *Chapter 19: Hybrids*, and *Chapter 20: Advanced Packaging Techniques*.

In the past, an electronic system was typically composed of a number of integrated circuits, each with its own particular function (say a microprocessor, some peripheral functions, some memory devices, etc.). For many of today's high-end applications, however, electronics engineers are combining all of these functions on a single device, which may be referred to as a *System-on-Chip* (SoC).

DIFFERENT TYPES OF ICs

The first integrated circuit—a simple phase-shift oscillator—was constructed in 1958. Since that time, a plethora of different device types have appeared on the scene. There are far too many different integrated circuit types for us to cover in this book, but some of the main categories—along with their approximate dates of introduction—are shown in [Figure 14.18](#).¹⁸

Memory devices (in particular SRAMs and DRAMs) are introduced in *Chapter 15: Memory ICs*; programmable integrated circuits (PLDs, CPLDs, FPGAs, and CSSPs) are presented in *Chapter 16: Programmable ICs*; and *Application-Specific Integrated Circuits* (ASICs) are discussed in *Chapter 17: Application-Specific Integrated Circuits* (ASICs).

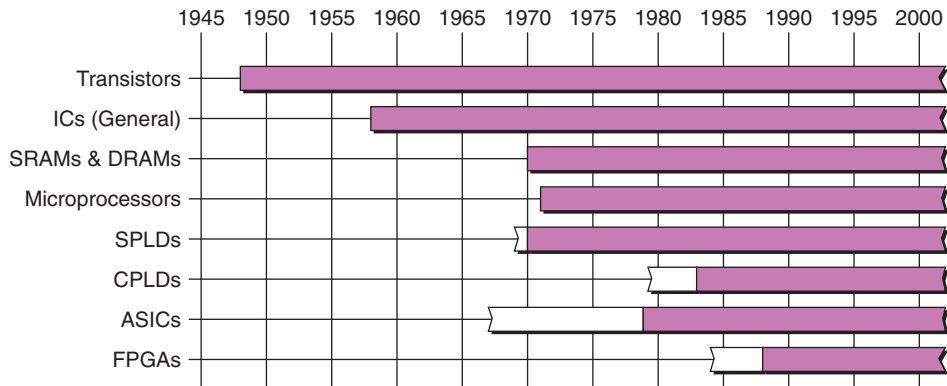


FIGURE 14.18

Timeline of device introductions (dates are approximate).

¹⁸The white portions of the timeline bars in [Figure 14.18](#) indicate that although early incarnations of these technologies may have been available, they perhaps hadn't been enthusiastically received during this period. For example, Xilinx introduced the world's first FPGA as early as 1984, but many design engineers didn't really become interested in these little rapscallions until the late 1980s.

TTL, ECL, AND CMOS

Transistors are available in a variety of flavors called *families* or *technologies*. One of the first to be invented was the *Bipolar Junction Transistor* (BJT), which was the mainstay of the industry for many years. If bipolar transistors are connected together in a certain way, the resulting logic gates are classed as *Transistor-Transistor Logic* (TTL). An alternative method of connecting the same transistors results in logic gates classed as *Emitter-Coupled Logic* (ECL).

Another family called *Metal-Oxide Semiconductor Field-Effect Transistors* (MOSFETs) were invented some time after bipolar junction transistors. *Complementary Metal-Oxide Semiconductor* (CMOS) logic gates are based on NMOS and PMOS MOSFETs connected together in a complementary manner.

Some integrated circuits use a combination of technologies; in the case of *Bipolar CMOS* (BiCMOS), for example, the function of every primitive logic gate is implemented in low-power CMOS, but the output stage of each gate uses high-drive bipolar transistors.

Finally, gates fabricated using the gallium arsenide (GaAs) semiconductor as a substrate are faster than their silicon equivalents, but they are expensive to produce, and so are used for specialized applications only.

CORE SUPPLY VOLTAGES

Toward the end of the 1980s and the beginning of the 1990s, the majority of digital integrated circuits were based on a 5.0-volt supply. However, increasing usage of portable personal electronics, such as notebook computers and cellular telephones, began to drive the requirement for devices that consume and dissipate less power.

One way to reduce power consumption is to lower the supply voltage; thus, by the mid to late 1990s, the most common supplies were 3.3 volts for portable computers and 3.0 volts for communication systems.

The core supply voltage continued to fall over the years [by *core* we mean the supply voltage used to drive the internals of the silicon chip; the input/output (I/O) pins may use higher voltages]. Leading-edge devices were using 2.5 volts by 1999, 1.8 volts by 2000, 1.5 volts by 2001, and 1.2 volts by 2003. At the time of this writing in 2008, some cutting-edge devices have core supply voltages as low as 0.9 volts.

EQUIVALENT GATES

One common metric used to categorize a digital integrated circuit is the number of logic gates it contains. However, difficulties may arise when comparing devices, as each type of logic function requires a different number of transistors. This leads to the concept of an *equivalent gate*, whereby each type of logic function is assigned an equivalent gate value, and the relative complexity of an integrated circuit is judged by summing its equivalent gates.

Unfortunately, the definition of an equivalent gate can vary, depending on whom one is talking to. A reasonably common convention is for a 2-input NAND to represent one equivalent gate. A more esoteric convention defines an ECL equivalent gate as being “*one-eleventh the minimum logic required to implement a single-bit full-adder,*” while some vendors define an equivalent gate as being equal to an arbitrary number of transistors based on their own particular technology. The best policy is to establish a common frame of reference before releasing a firm grip on your hard-earned lucre.

The acronyms SSI, MSI, LSI, VLSI, and ULSI represent *Small-, Medium-, Large-, Very-Large-, and Ultra-Large-Scale Integration*, respectively. By one convention, the number of gates represented by these terms are: SSI (1–12), MSI (13–99), LSI (100–999), VLSI (1000–999,999), and ULSI (1,000,000 or more).¹⁹

DEVICE GEOMETRIES

Integrated circuits are also categorized by their *geometries*, meaning the size of the structures created on the substrate. For example, a 1- μm ²⁰ CMOS device would have structures that measure one-millionth of a meter. The structures typically embraced by this description are the width of the tracks and the length of the channel between the source and drain diffusion regions; the dimensions of other features are derived as ratios of these structures.²¹

¹⁹Truth to tell, not many folks use these terms anymore. They’ve become increasingly irrelevant as designers have moved away from using simple *jelly bean* components and chips have increased in complexity into the hundreds of millions (or billions) of transistors.

²⁰The “ μ ” symbol stands for *micro* from the Greek *micros*, meaning “small,” (hence, the use of μP and μC as abbreviations for microprocessor and microcontroller, respectively). In the metric system, μ stands for “one millionth part of,” so 1 μm means “one-millionth of a meter.”

²¹I’m simplifying things here; there are different ways of measuring things depending on the type of component; for example, memory chips versus devices containing general-purpose logic gates.

Each new geometry may be referred to as a *technology node* or a *process node*. Geometries are continuously shrinking as fabrication processes improve. In 1990, devices with 1- μm geometries were considered to be state of the art, and many observers feared that the industry was approaching the limits of manufacturing technology, but geometries continued to shrink regardless as illustrated in Table 14.1.

Around the time we reached the 0.5- μm technology node, it became common to refer to anything below this point as *Deep-Submicron* (DSM). Later, at some point that wasn't particularly well defined (or was defined differently by different people) we moved into the realm of *Ultra-Deep-Submicron* (UDSM). However, no one tends to use these terms anymore, because we've now travelled so deep into the rabbit hole that we've run out of meaningful qualifying words.

Table 14.1 Evolution of Technology Nodes

Year	Node
1990	1.00 μm
1992	0.80 μm
1994	0.50 μm
1996	0.35 μm
1997–1998	0.25 μm
1999–2000	180 nm
2001	130 nm
2001	100 nm
2003	90 nm
2005–2006	65 nm
2008	45 nm
2008	40 nm
2009–2010*	32 nm
2011–2012*	22 nm
2013–2014*	16 nm

*“Finger-in-the-air” predicted dates

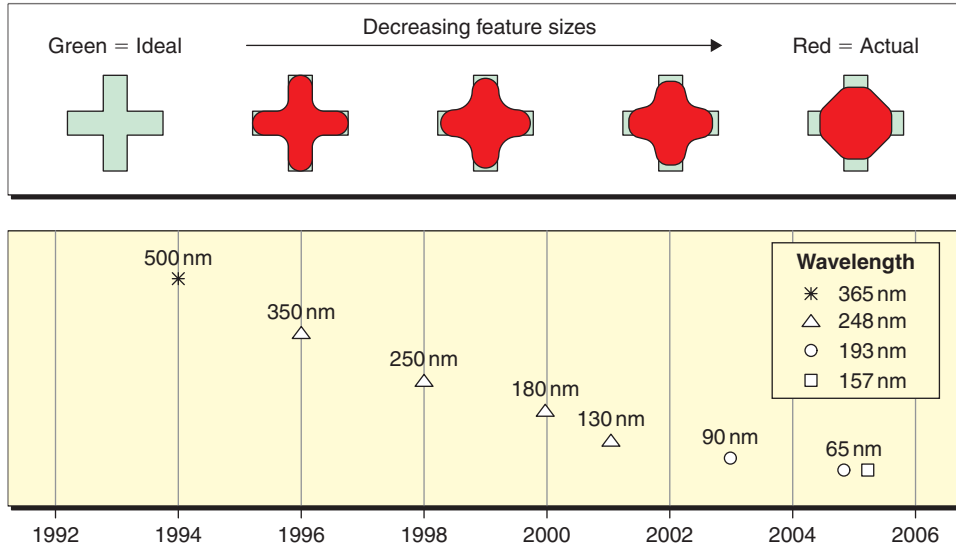
With devices whose geometries were $1\ \mu\text{m}$ and higher, it was relatively easy to talk about them in conversation. For example, one might say “I’m working with a one-micron technology.” But things started to get a little awkward when we dropped below $1\ \mu\text{m}$, because it’s a bit of a pain to have to keep on saying things like “zero-point-one-three microns.” For this reason, sometime around the end of the 20th century, it became common to talk in terms of “nano,” where one *nano* (short for *nanometer*) equates to one one-thousandth of a micron; that is, one-thousandth of one-millionth of a meter. Thus, when referring to a 130-nm ($0.13\ \mu\text{m}$) technology, instead of mumbling “zero-point-one-three microns,” you could now proudly proclaim “one-hundred and thirty nano.” Of course, both of these mean exactly the same thing, but if you want to talk about this sort of stuff, it’s best to use the vernacular of the day and present yourself as hip and trendy as opposed to an old fuddy-duddy from the last millennium.

While smaller geometries result in lower power consumption and higher operating speeds, these benefits do not come without a price. Logic gates implemented in nanometer technologies exhibit extremely complex timing effects, which make corresponding demands on designers and design tools. Additionally, all materials are naturally radioactive to some extent, and the materials used to package integrated circuits can spontaneously release alpha particles. Devices with smaller geometries are more susceptible to the effects of noise, and the alpha decay in packages can cause corruption of the data being processed by deep-submicron logic gates. Nanometer technologies also suffer from a phenomenon known as *subatomic erosion* or, more correctly, *electromigration*, in which the structures in the silicon are eroded by the flow of electrons in much the same way as land is eroded by a river.

WHAT COMES AFTER OPTICAL LITHOGRAPHY?

Although new techniques are constantly evolving, technologists can foresee the limits of miniaturization that can be practically achieved using optical lithography. These limits are ultimately dictated by the wavelength of ultraviolet radiation.

In fact, the features (structures) on a silicon chip are now smaller than the wavelength of the light used to create them (Figure 14.19). If we assume that the green geometric shape shown in this illustration is the ideal (desired) form, then this is the shape that would be generated by the design tools. The problem is that, if this shape were to be replicated as-is in the photo-mask, then the corresponding form appearing on the silicon would drift farther and farther from the ideal with the decreasing feature sizes associated with the newer technology nodes.


FIGURE 14.19

Technology nodes versus the ultraviolet wavelengths used to create them.

The way this is currently addressed in conventional design flows is for the manufacturing group to post-process the design files with a variety of *Resolution Enhancement Techniques* (RET), such as *Optical Proximity Correction* (OPC) and *Phase Shift Mask* (PSM). For example, they may modify the original design files by augmenting existing features or adding new features—known as *Sub-Resolution Assist Features* (SRAF)—so as to obtain better printability. One way to visualize this is that if the manufacturing group (actually, their tools) knows that the features will be distorted by the printing (imaging) process in certain ways, they can add their own distortions in the “opposite direction” in an attempt to make the two distortions cancel each other out.

The technology has now passed from using *Standard Ultraviolet* (UV) to *Extreme Ultraviolet* (EUV), which is just this side of soft X-rays in the electromagnetic spectrum. One potential alternative is true X-ray lithography, but this requires an intense X-ray source and is considerably more expensive than optical lithography. Another possibility is *electron beam lithography* (often abbreviated to *e-beam lithography*), in which fine electron beams are used to draw extremely high-resolution patterns directly into the resist without a mask. Electron beam lithography is sometimes used for custom and prototype devices, but it is much slower and more expensive than optical lithography. Thus, for the present, it would appear that optical lithography will continue to be the mainstay of mass-produced integrated circuits.

HOW MANY TRANSISTORS?

The first integrated circuits typically contained around six transistors. By the latter half of the 1960s, devices containing around 100 transistors were reasonably typical.

In the first half of 2002, Intel announced its McKinley microprocessor—an integrated circuit based on a 130-nano (0.13- μm) process, containing more than 200 million transistors. And by the summer of 2002, Intel had announced a test chip based on a 90-nano (0.09- μm) process that contained 330 million transistors.

On May 19, 2008 (a few days ago at the time of this writing) a company called Altera (<http://www.altera.com/>) announced a new family of FPGAs (see *Chapter 16: Programmable ICs*), the largest member of which contains 2.5 billion transistors! At this level of processing power, we will soon have the capabilities required to create Star Trek–style products, like a universal real-time language translator.

MOORE'S LAW

You can't read a technical paper these days without blundering into some mention of Moore's Law, so it behooves us to explain this here. In 1965, Gordon Moore (1929–) (who was to cofound Intel Corporation in 1968) was preparing a speech that included a graph illustrating the growth in the performance of memory ICs. While plotting this graph, Moore realized that new generations of memory devices were released approximately every 18 months, and that each new generation of devices contained roughly twice the capacity of its predecessor.

This observation subsequently became known as *Moore's Law*, and it has been applied to a wide variety of electronics trends. These include the number of transistors that can be constructed in a certain area of silicon (the number doubles approximately every 18 months), the price per transistor (which follows an inverse Moore's Law curve and halves approximately every 18 months), and the performance of microprocessors (which again doubles approximately every 18 months).