IC Manufacturing Drives CPU Performance

Don't Be Misled—Not All 0.8 Micron Processes are Created Equal

By Linley Gwennap and Alisa Scherer

This is the first in an occasional series of articles discussing IC manufacturing. Future articles will look at the cost of building chips, compare different vendors' manufacturing capabilities, and look at packaging issues.

Alisa Scherer is a CPU design engineer at Hewlett-Packard.

The process of manufacturing an integrated circuit is somewhat arcane, even for many of those who design them. For those who don't, the only thing that is certain is that everything keeps changing. When the 8086 was first built, its transistors were about 3 microns in length. Just a dozen years later, the 486 uses a 0.8-micron technology that results in transistors that are 80% smaller in area. This amazing trend drives ongoing improvements in CPU performance and memory density. Fortunately for computer users, it shows no sign of stopping.

Since each IC manufacturer creates its own manufacturing process, there is no "standard" 0.8-micron process. Furthermore, each vendor moves to the next generation at its own pace. Thus, at any given time, chip vendors each use a different IC process, and these differences can have a significant effect on the performance of microprocessors (and other chips) built using these processes. Understanding these differences provides insight into the overall microprocessor market.

Types of IC Processes

The first popular IC technology was TTL, which uses *bipolar* transistors. TTL devices are relatively fast and are easy to build. Because a TTL gate draws a certain current even when it isn't switching, it has a relatively high power consumption. A variation called ECL uses a continuous current sink in each gate, drawing even more power. This current sink, however, is like an idling engine, allowing the gate to switch very quickly when needed. Both TTL and ECL use similar bipolar manufacturing processes.

CMOS—like its predecessor, NMOS—requires a different manufacturing process. These circuits, which use field-effect transistors (FETs), use power only when they are active and switching, and even then typically use less power than bipolar designs. CMOS gates are also smaller than bipolar gates, allowing designers to pack more logic onto a CMOS chip. One drawback is that FETs are slower than bipolar transistors, particularly when driving signals from one chip to another, or driving large on-chip loads.

Nearly all popular microprocessors today are designed in CMOS, primarily because of the greater density. CMOS allows an entire processor to be implemented on a single chip, alleviating the need to drive high-speed signals from chip to chip. Although ECL circuits are inherently faster than CMOS, ECL processors are usually spread across multiple chips, increasing cost and reducing performance due to the inter-chip communication. CMOS processors are generally smaller, cheaper, easier to cool, and have only slightly lower performance than similar ECL processors.

For example, Bipolar Integrated Technology (BIT) designed a MIPS-architecture processor—the 66-MHz R6000—using ECL, but by the time it began shipping, other vendors were delivering 40-MHz R3000 chips using lower-cost CMOS technology. The R6000, arguably the most successful ECL microprocessor to date, had poor yields and limited sales; BIT is now out of the processor business.

A more recent advance combines the density of CMOS with the speed of bipolar logic. BiCMOS uses a hybrid process to place bipolar and CMOS devices on the same chip. For a typical processor, most of the transistors use CMOS to save space and power, but bipolar drivers speed up key areas (such as cache sense amps and ALUs) and drive signals with large loads (such as clocks and off-chip buses). Too many bipolar circuits, however, waste space and increase power usage.

Because of the extra process steps required to include both types of devices, BiCMOS chips cost more to manufacture than comparable CMOS or ECL chips. Texas Instruments' SuperSPARC was the first commercial BiCMOS microprocessor, and Intel's Pentium also takes advantage of this technology. At around $1000 each, these chips have a high enough price to pay for the added process cost.

Nearly all chips today are based on silicon, but other materials can also be used to build either FETs or bipolar devices. Gallium arsenide, or GaAs ("gas"), is the leading alternative, as it can provide faster circuits with lower power than silicon. Unfortunately, GaAs chips are currently expensive and hard to build. Today, they are most popular for small gate arrays and in military applications, which are willing to pay more for GaAs' high performance and radiation tolerance.

Figure 1 compares various types of IC processes. Both bipolar devices and FETs can be built in either silicon or GaAs. CMOS provides the best density and lowest power consumption but has the slowest transistors. Bipolar silicon, such as TTL or ECL, can be faster but uses more power. Bipolar GaAs is best for small but very

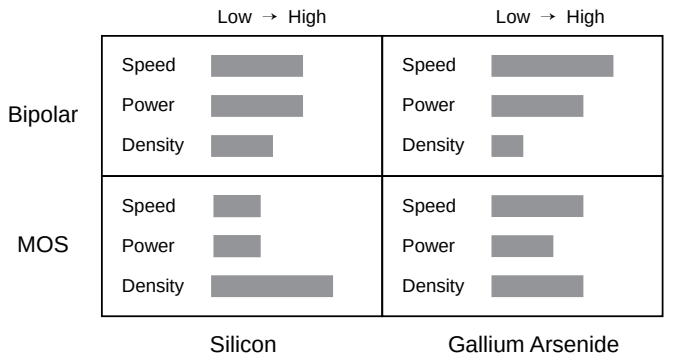

Figure 1. A relative comparison of basic IC process types shows that silicon MOS, typically CMOS, provides the highest density and lowest power, but others have faster transistors.

fast circuits. When built with FETs, GaAs circuits can provide similar performance to ECL but use less power, making them ideal for fast designs of limited complexity. Compared to CMOS, GaAs circuits use more power and are also less dense.

Building a CMOS Chip

Companies generally describe their CMOS process with a single number such as "0.8 micron." Like any single-number metric, this can be misleading. The original intent was to show the size of a transistor built in that process. When transistors were a few microns in size, no one worried about minor variations in the length or electrical characteristics. But with today's sub-micron technology, these factors become important.

Let's take a simplified look at how a modern chip is built. Once the engineers have completed their design, they create a set of drawings, one per layer, that show how the signals will be routed (much like a PC board) and the location of the actual transistors. These drawings are converted into a set of *masks* that contain a photographic image of each layer.

The chip starts as a blank silicon wafer. For each layer, the surface of the wafer is coated with the appropriate material for that layer (polysilicon, insulating silicon dioxide, or aluminum for metal layers) and then with an organic material known as *resist*. The wafer is then exposed to short-wavelength light through the appropriate mask, removing the resist where there are blank areas in the mask.

"Dry" plasma is used to etch the wafer where the resist has been removed, but the resist protects the rest of the wafer. At this point, the exact pattern of the mask has been etched into the chip. To complete the layer, the remaining resist is washed away with a solvent.

The sequence to produce one layer is called a *mask step*. Some steps replace the plasma etch with exposure to an ion beam ("ion implantation"), which creates diffusion regions in the silicon. Other steps use a "negative" resist that allows material to be added where the resist is not present. Modern CMOS chips require up to 18 mask steps, and BiCMOS chips need a few more.

Inside a Transistor

Figure 2 shows a typical MOS transistor as if one sliced a wafer through the middle and looked at it on edge. Of course, one would have to look very carefully to see anything at all; the height of the built-up area in the center of the figure is under one micron, or 1/1000 of a millimeter. The crossing of the polysilicon trace and the diffusion region forms the transistor. One major factor in the switching speed of this transistor is the *channel length* shown in the figure and referred to as L. This is the distance that electrons must traverse when the transistor is "on;" the shorter it is, the faster the electrons can complete their journey. When a vendor says its process is "0.8 micron," they are usually referring to this length.

Unfortunately, there are different ways to measure the channel length. One is to measure the width of a trace on the mask. This measurement is called the drawn length, or L_{DRAWN}. Another way is to quote the actual distance the electrons must travel, known as the effective length or L_{EFF}. The effective length is typically 10%–20% less than the drawn length; for example, a process with an L_{DRAWN} of 0.8 micron might have an L_{EFF} of 0.65 micron.

The effective length takes into account factors in the manufacturing process ("out diffusion") that cause the physical channel to be slightly smaller than the image on the mask. It also includes electrical field effects that help the electrons jump from one diffusion region to the other. Although vendors sometimes disagree on how to measure L_{EFF}, when used properly it is a better performance metric than L_{DRAWN}. Vendors usually quote drawn channel length, but because of these differences, it is prudent to ask which measure they are using.

Another important speed parameter is gate oxide thickness, also shown in Figure 2. As with the channel length, smaller is better. Typical values are 100 to 300 Å (angstroms), which is equivalent to 0.01 to 0.03 microns; this layer is only a few hundred atoms thick! Although the gate oxide thickness is less frequently quoted, it is as important as channel length in determining the speed of transistors built in a particular process.

Connecting the Transistors

Metal layers (not shown in Figure 2) are used to connect circuit blocks by routing signals around the chip. They also provide power and ground throughout the chip. Although they do not directly affect the switching time of a transistor, the metal width and spacing can limit the overall clock frequency by determining the time needed to send critical signals from one part of the chip to another. Metal layers also typically control how many

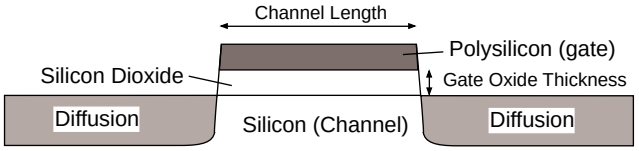

Figure 2. Side view of MOS transistor shows a thin layer of oxide isolating the poly layer from the channel. The voltage applied to the poly controls the current from one diffusion region to the other.

transistors can be packed onto a chip, particularly for microprocessors and other chips that have a large number of interconnections. The most basic question regarding metal layers is how many there are; today's processes typically have three.

The size of these metal layers is also important. In this case, the normal measurement quoted is the minimum distance, or *pitch*, between two adjacent metal traces. Figure 3 shows some typical metal traces. The square areas are contacts, or *vias*, to layers above or below (similar to vias on a PC board), and are a bit wider than the trace itself. The minimum pitch can be slightly different depending on whether it is measured between contacts (Figure 3a) or not (Figure 3b).

Like channel length, metal pitch may vary by a few tenths of a micron depending on which of these measurements is used. Metal pitch is larger than channel length—typically 2 to 3 microns in current processes. The third metal layer often has a much larger pitch of 4 to 8 microns. This layer is generally used for routing power and ground, and sometimes clock signals.

Other factors have a lesser effect on the overall design, although they are important to the chip designer. The resistance and capacitance of each layer helps determine how fast a transistor can drive a signal across the chip. Resistance is usually given as Ω/❑ (ohms per unit square area) or Ω/mm (ohms per millimeter) of length at minimum pitch. Capacitance is given as pF/mm (picofarads per millimeter) of length, also at minimum pitch.

The Move to Lower Voltages

In general, CMOS circuits are happy with supply voltages from 2V to 12V, while TTL is restricted to a well-regulated 5V supply. Most CMOS microprocessors are optimized for a 5V supply, however, to simplify the design of mixed CMOS and TTL systems. Some newer processor chips are being designed for lower-voltage operation, typically 3V or 3.3V, for various reasons.

Chips for portable systems have led the way to lower supply voltages. This is because power consumption of a CMOS chip is proportional to the square of its voltage. When running a system from batteries, every milliwatt of power must be used wisely. The IC process itself can also reduce power consumption; smaller transistors require less power to switch. Another power-reduction technique is lowering the clock frequency, reducing the amount of switching and thus the power required. Lower clock rates, however, result in reduced performance.

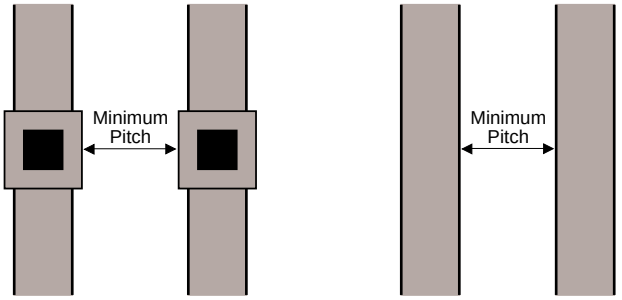

Figure 3. The metal pitch determines how closely signal traces can be spaced. (Top view of chip shown.)

Conversely, desktop system designers want to maximize performance by pushing up the clock rate while packing many transistors on a chip. To make transistors smaller and faster, the gate-oxide layer (see Figure 2) must be very thin and becomes, in effect, delicate. With gate-oxide thicknesses below 150 Å, as is typical in half-micron processes, a 5V supply can cause the gate oxide to break down over time, destroying the effectiveness of the transistor. A single failed transistor among millions will usually render a processor chip unusable.

Many vendors are reducing the supply voltage for their half-micron processes to avoid this problem. The lower voltage has the added benefit of decreasing the power consumption as well. Although desktop systems don't have a problem supplying power to the processor chip, the chip must dissipate all of that power in the form of heat. As chip designers take advantage of faster transistors to push CPU clocks to 100 MHz and beyond, cooling the chip becomes a problem. The Alpha 21064 chip, for example, would give off a scorching 67 Watts with a 5V supply, but moving to a 3.3V supply reduces the heat to a merely toasty 30 Watts at 200 MHz.

Unfortunately, reducing the power supply decreases the electrical current available to overcome on-chip resistance and capacitance when driving signals across the chip. This lower current results in slower signals; clock rates for a 3.3V part can be anywhere from 40% to 80% of the speed of a 5V part, depending on various process parameters. Vendors that design their manufacturing process to work at either voltage can achieve results near the high end of that range.

Despite best efforts, 3.3V operation will result in reduced performance, but gate-oxide reliability and cooling issues are forcing this change in most half-micron processes. Even some previous-generation processes are being adjusted for low-power applications.

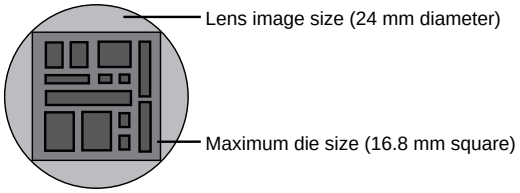

Figure 4. The diameter of the imaging lens limits the maximum diagonal dimension of an individual die.

Die Sizes Continue to Grow

The maximum size of an individual chip, or *die*, has also increased over time, although not quite at the rate of other factors. As described previously, each layer of a chip is created by exposing it to a photographic image of its mask. To create features that are less than 1/1000 of a millimeter in size, this image must be focused with incredible precision. Each die is individually exposed through a very expensive lens, and the maximum die size is limited by the size of this lens. Since most chips are square (or rectangular) and the lens is round, the maximum die area is much smaller than the lens area, as shown in Figure 4. About ten years ago, most chips were under 10 mm on a side; today, with improvements in optical technology, processors such as Pentium and SuperSPARC are over 16-mm square, pushing the limits of a state-of-the-art 24-mm lens.

Microprocessor designers have several reasons for wanting a larger die size. Circuits on the same chip can communicate faster than circuits on different chips because they are closer together, improving performance. Also, combining circuits onto one chip reduces the number of expensive chip packages needed and cuts down on board space and assembly costs. Cost tradeoffs must be made carefully, since both larger dice and larger packages are more expensive to manufacture.

Another reason for moving to larger chips is to allow room for more wire-bonding pads. Every signal that goes on or off the chip requires its own pad. In addition, dozens of pads are needed for power and ground to supply the needed current. These pads are typically placed in a "ring" around the edge of the chip. Because the pad is used to attach a physical wire to the die, pad size has remained fairly constant even as transistors have become much smaller.

Since modern processors use multiple 64-bit buses to move data into and out of the chip, it becomes difficult to fit all the required pads around the edge of a small chip. Processor designers are thus forced to use larger die sizes to accommodate all the required pads. These chips are *pad-limited*. Many popular single-chip processors today are pad-limited, or close to it. For these designs, the pad size and pad pitch (distance between pads) is a significant aspect of the IC manufacturing process. Most processes today use a 125-micron pad pitch, but some are moving to a 100-micron pitch.

IBM has solved this problem for its PowerPC 601 by placing the pads on top of the other circuitry using a fourth layer of metal. Since the pads are arranged in a grid instead of a ring, there is plenty of room for them even on a relatively small chip. An added advantage is eliminating the 10%–15% of the die area usually devoted to the pad ring. The downside is the added cost of the extra metal layer.

The bugaboo of large chips is low yield. Despite the best efforts of the manufacturing team, many chips are ruined by a speck of dust, a minute impurity in the silicon, or a small glitch in creating a layer. These *defects* are usually scattered randomly across the surface of a wafer. The larger the die, the more likely it is to have a random defect. Furthermore, a single wafer can hold fewer large chips than small chips. As a result, increasing the chip size reduces the number of chips per wafer while increasing the percentage of bad chips. This double whammy drives up the cost of the good chips and provides a counter pressure on designers to keep their chips small. (We will look at yield issues in more detail in part 2.)

Putting It All Together

In determining the speed of an IC process, the two most important figures are effective gate length (L_{EFF}) and gate oxide thickness. These figures allow a designer to estimate the speed of the transistors. Metal capacitance and resistance are second-order parameters that help determine the speed of long on-chip connections. The capacitance and resistance of the polysilicon and diffusion layers are less importance because critical signals are typically routed with metal layers.

All things being equal, a designer will choose a higher voltage process (up to 5V) to improve switching speed. If power consumption is an issue (as in a portable system) or if cooling is a problem, the designer may have to accept a lower voltage. Most new designs are being forced to a 3V or 3.3V supply by the limitations of half-micron processes' thinner gate oxides.

Another key processor design issue is density. When packing large numbers of transistors onto a fixed-size die, the most important metrics are the minimum metal pitch of the first and second metal layers. Metal pitch is more important than transistor size, since most of the area of a typical microprocessor is limited by interconnections and not the gates themselves. Metal pitch usually limits the size of standard cells, memory cells, and most custom logic. Smaller transistors (i.e., with a smaller channel length) also improve density but only in those areas where metal routing is not the critical factor. As new processes decrease the gate length, they must also reduce the metal pitch, or the size of the chip will not decrease as much.

Another density factor is the presence of additional metal or interconnect layers. These layers will only be

useful if automatic routing programs can use them intelligently (many can't) or if the routing is done by hand. Stacked vias, if allowed, can also improve the density.

Future IC Manufacturing Trends

Today, a state-of-the-art CMOS process for CPU production has a drawn channel length of about 0.6 micron, resulting in an effective gate size of about 0.5 micron. Contacted metal pitch for the first two layers is around 2.5 microns and a third metal layer is provided with a much larger pitch. DEC, MIPS, and IBM are the first microprocessor vendors to deliver chips meeting these specifications, but the rest of the industry should follow by early 1994.

Each new generation of CMOS processes has historically brought a doubling of circuit speed while reducing circuit area by 40%–50%. All of these benefits come from reducing the linear feature size by about 20%. In the past, these generations have occurred about every 18 months. Future generations will continue to bring improvements regularly, although the pace may slow as the task of shrinking the transistors and other features becomes more difficult.

The current 0.8-micron processes are the last to support a 5V supply. Next-generation gate oxides are too thin to hold up at this voltage. Although half-micron processes can sustain 4V operation, many vendors are making the jump to 3.3V to avoid changing again later.

Although CMOS is currently dominant, both Intel and TI expect to ship hundreds of thousands of BiCMOS processors this year. CMOS will continue as the process of choice for low-cost, high-volume microprocessors in the foreseeable future. Some vendors will use CMOS for future high-end processors as well, but those with the volume to support a BiCMOS process may use it to boost the performance of high-priced products.

In the more distant future, new techniques may be required to advance IC technology *(see **070104.PDF**)*. By the end of the decade, we expect to see gate lengths under 0.2 micron, processor clock rates over 500 MHz, and single-chip microprocessors with 50 million transistors. Because chip-to-chip communication will not improve at the same rate, the processor designer's challenge will be to select the right mix of high-speed memory and functional blocks to use those transistors most effectively.♦

