The Limits of Chip Technology

Vendors Continue to Improve Their Chips, But How High is Up?

By Don Lindsay, Carnegie Mellon University

Every year, the chip industry's products are faster and denser. This is good for its customers, and it keeps the vendors busy. But the price of memory is down to pocket change per megabyte, and serious compute power is being marked down at a Radio Shack near you. "Faster and denser" begins to seem inevitable, and of little interest. So something's faster—so what?

Dijkstra gave an excellent and pragmatic answer many years ago. He noted that a quantitative difference is also a qualitative difference, if the quantity has changed by an order of magnitude. Think about this example:

- 1 MPH is the speed of a baby crawling.
- 10 MPH is the speed of a top marathon runner.
- 100 MPH is the speed of a fast automobile.
- 1000 MPH is the speed of a fast airplane.

Driving is not only faster than running, it allows people to go places and do things they could not do on foot. Likewise, the airplane further extends our abilities and brings the world closer together. These orders of magnitude have made fundamental changes to society.

Dijkstra made his case two decades ago, when a thousand-line program was a pound of punched cards, and operators held big jobs (150 Kbytes!) until after 5 pm. Computers are indeed qualitatively different now. New capabilities have emerged, such as spreadsheets, real-time modeling, and graphical user interfaces. If technology continues to improve, future computers will have significant new abilities—some that no one has even thought about.

The classic "S" curve (see Figure 1) shows that a particular technology, after an initial development period, often progresses rapidly until it approaches certain fundamental limits. Once it reaches maturity, however, progress is much slower unless a new technology is found, breaking through the previous limitations. For example, propeller planes became limited once the tip speed of the propeller approached the speed of sound, but jet engines overcame this barrier. Today's silicon integrated circuits are in the middle stage, but when will we reach maturity? And will there be a new technology to jet-propel computers to even higher speeds?

Better Living Through Faster Silicon

Some observers pin their hopes on massively parallel machines containing thousands of fairly conventional CPUs. Another approach is to use superscalar processors that execute multiple instructions per clock cycle. Due to limitations in available software parallelism, however, neither of these techniques is likely to give us another order of magnitude in performance over current designs *(see **061606.PDF**)*. As a result, increases in the speed of the underlying devices (transistors) will be the major factor in determining processor performance by the end of this decade.

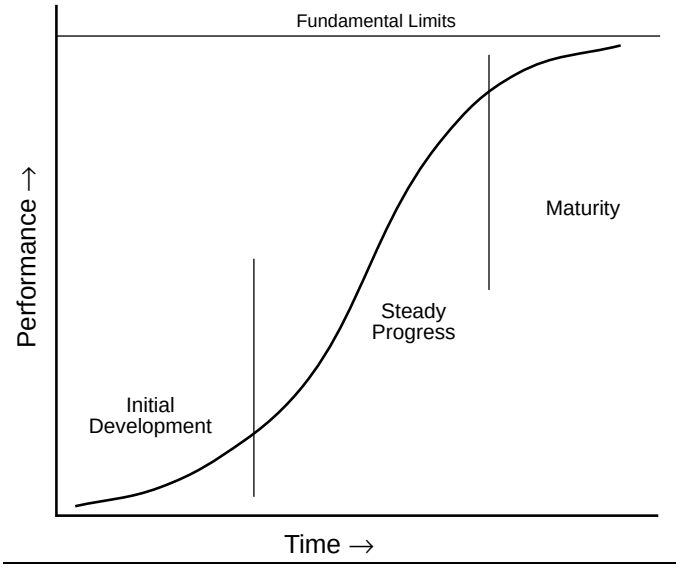

Figure 1. Theoretical technology "S" curve.

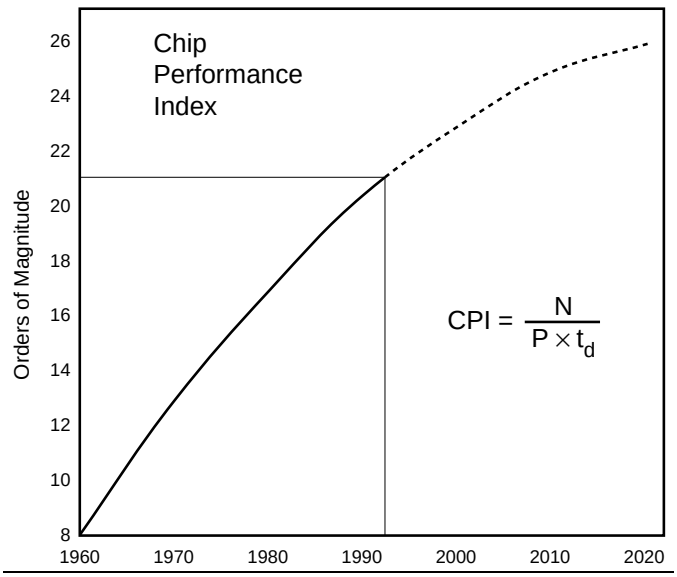

Figure 2. Chip performance index (from James Meindl, RPI).

So how far can we speed up these transistors? A recent study defined a Chip Performance Index, or CPI, as the transistor count of a chip divided by the power-delay product of each transistor. In room-temperature silicon, this CPI has gone up by 13 orders of magnitude (10^{13}) since 1960 (see Figure 2). Arguments from some basic principles suggest that CPI has 3 to 6 more orders to go, allowing a billion-transistor chip by the year 2000. Intel's projections are a bit more pessimistic (see Figure 3), but they still expect to see a 256-megabit DRAM by then.

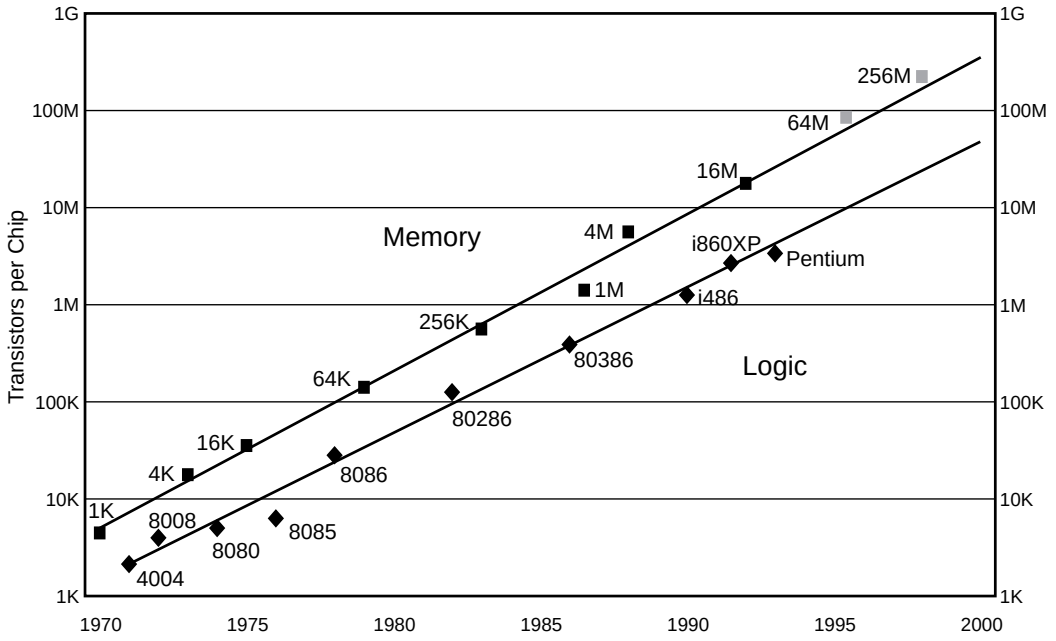

Figure 3. Transistor count for various memory and processor chips (from Intel).

It would be nice if this scaling could just continue forever. Unfortunately, those graphs get pretty silly by the year 2010—for example, they predict that gate oxide will be about one atom thick (see Figure 4). Clearly, the time will come when we'll need to try something more exotic. Silicon will be dominant in the coming decade, and will most likely remain important, but only a foolish forecaster would say more.

Initially, these exotic new technologies will struggle—witness the long, hard road that GaAs (gallium arsenide) has traveled. There will be niches that allow new technologies to gain a foothold; some customers always want speed at any cost. Real-world simulations, signal processing and analysis, missile-guidance systems, and many other applications require vast amounts of computation. Another niche is in high-radiation or high-temperature environments, such as those found in space or in deep oil wells, that offer more extreme conditions than normal silicon circuits can handle.

How Fast Can We Go?

The first question is how to measure speed. The usual ways are:

1) A theoretical value
2) A simulated value
3) An oscillator frequency
4) A ring oscillator frequency
5) The delay of a gate into no load
6) The toggle frequency of a flip-flop
7) The average gate delay in a real system
8) An interchip delay through a bi-directional line
9) A CPU clock rate
10) The time to a solution in the customer's hands

Of course, these differ greatly, and one often isn't told which kind of number is being discussed. Types 3–10 at least have the virtue of having been measured, but type 5 might as well not have been, since eliminating the load may increase the "speed" by as much as a factor of 5 or 10. Some technologies are much more sensitive to loading than others.

A technology may also do things for "free," for example, provide both the output and its complement, or permit a wire-OR (where outputs are simply joined together). Those tricks came from ECL, and the pass transistor came to us from NMOS. More recently, there have been circuits that directly perform a parity calculation or a 1-bit full add. Noise sensitivity is another issue, since noise reduction might require half of the chip's pins in some technologies. The final system speed may depend heavily on whether the designer takes advantage of these technology quirks.

The highest speeds usually are of type 3, a simple oscillator. Someone measured an InAs/AlSb resonant tunneling diode at a stupendous 712 GHz. Unfortunately, this type of circuit is of no immediate use for CPUs—in fact, the diode was *aimed* at the test equipment rather than attached to it. Similarly irrelevant to computers is a superconducting Josephson junction oscillator running at 1000 GHz.

Ring oscillators are more indicative of logic performance, since they consist of an odd number of inverters connected in series. Several labs have built CMOS ring oscillators with inverter delays of about 50 ps. Laboratory GaAs beats this with 10-ps delays at room temperature and 5 ps when cold. GaAs flip-flops have toggled at 18 GHz

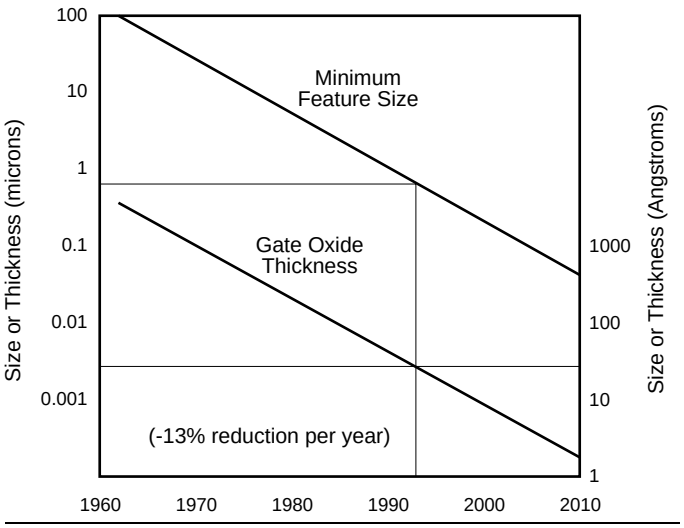

Figure 4. Scaling of MOS parameters over time (from Intel).

(55 ps) and the latest Japanese supercomputer claims to have "70-ps" GaAs circuits.

Transistors Go Ballistic

All of these seem slow when researchers talk about 1.9-ps ballistic transistors. These devices, still in the conceptual stage, are based on a remarkably simple idea. When we put a voltage across a device region, electrons begin to travel at the device's *drift velocity*. However, the drift velocity of an electron is actually a time average; the electron is accelerated by a field, but then is slowed by bumping into the crystal lattice structure of the transmission medium. (In technical terms, these bumps cause the electron to be "scattered" or "emit a phonon.") Like a person making their way through a crowded room, the electron keeps starting and stopping as it moves across the device. The average distance between scattering events is referred to as the *mean free path*. If we scale the device until it is smaller than one mean free path, then an electron might traverse the region ballistically, much faster than before.

One problem with this idea is that room-temperature silicon would need a very short channel length—about 0.1 micron. GaAs, on the other hand, can go ballistic with larger channels, and other materials may do even better. To perform best, these devices may ultimately use a *heterostructure*, which is a layered combination of different materials, that can "launch" electrons into the device region. These various possibilities make 1-ps gates seem very possible.

Connecting the Dots

The next problem is connecting these tiny ultra-fast devices. One problem is that wires don't scale well; resistance increases as the cross-section becomes smaller. As frequencies increase, other electrical characteristics degrade signal propagation. Around 1 GHz, chip designers may turn to *microstrip* interconnection, a sort of miniature flat coaxial cable, for cleaner signals, although this technique requires more chip area than traditional interconnect. Even microstrip, however, is only good to about 10 GHz (100 ps), which is bad news for the builders of 1-ps gates.

There are a few ways out. We could, for instance, design chips without long wires, but this would be difficult and less efficient. Another possibility is superconducting microstrip, which *can* deliver 1-ps signals. The "old" superconductivity required temperatures below 20 degrees absolute, which typically required liquid helium. The new "high-temperature" (high-T_C) superconductors are less well understood but dramatically easier to cool, since they require only liquid nitrogen (LN) temperatures. (Liquid helium is a hundred times more expensive than liquid nitrogen, and sixty times worse at cooling to boot.)

Computer Systems Chill Out

If the wires are superconducting then the logic is also going to be cold. Many people assume that cold or cryogenic systems cannot possibly invade their desk—why, think of the cost! Think of the huge, clanking, unreliable refrigerator! This sort of argument has been losing force as cryogenic systems move out of the lab in a big way. The US military, for example, has fielded some 50,000 cryogenic systems. Most are infra-red detectors, but this is actual field equipment kept at LN temperatures.

Most circuits can be safely immersed in liquid nitrogen, which is available in every city in the world as a by-product of the oxy-acetylene welding business. LN is cheap enough and safe enough that it is reasonable to let the vapor simply escape into the room. There is also ongoing effort to build compact low-vibration space-rated refrigerators that provide LN temperatures and even much cooler liquid-helium temperatures. With a high-speed network, the cryogenic system can be tucked away in a closet while serving users with X-terminals.

Still not convinced? All right, let's just chill the chips a little. We might do it with, for instance, the free-piston Stirling engine, to which Intel happens to own the patent rights. This device, which can keep a 35-watt chip at –50° C, occupies about a hundred cubic inches, a bit smaller than a two-quart milk carton. Early work using "cryo-CMOS" shows that it has some nice properties at this temperature. It's faster, for one thing, due to a higher drift velocity. It has low noise, lower resistance, and no electromigration. Some experts believe there are potential long-term reliability problems at these temperatures, particularly at higher (5V) voltages. If these problems can be overcome, a 50% performance boost at –50° C is reasonable. Up to 2.5× has been observed with liquid-nitrogen cooling.

So why not build superconducting transistors?

Fujitsu has built a transistor with a superconducting emitter, and Sanyo built one with a superconducting base. A more common idea has been to use Josephson junctions, and in the "old" superconductivity, this was followed to the point of demonstrating a (somewhat limited) 8-bit, 1.1-GHz microprocessor. More recent thinking has led to several kinds of *quantum flux transfer* devices, such as the RSFQ (Rapid Single Flux Quantum). These devices represent a binary "one" by a single quantum of magnetic flux, with "zero" indicated by no quantum. (This is not voltage-level logic but rather pulse logic, as used in ENIAC, the first computer.) A 45-GHz RSFQ shift register and a 100-GHz counter have been built, and 145-GHz toggle frequencies have been demonstrated. More recently, high-T_C RSFQ devices have been announced using the new superconductor materials.

But Wait, There's More

A technology is at risk if its future depends on the success of a single line of development. Luckily, there are many promising threads, and it is unlikely that they will *all* fail to develop. One possibility is building devices out of silicon carbide. This material can be oxidized like normal silicon but also provides radiation hardness, higher breakdown voltages, high-temperature operation, and low temperature sensitivity. Or perhaps we will just use a layer of silicon carbide or germanium in a silicon heterostructure.

Another interesting substance is diamond. The discovery of a surprisingly simple fabrication technique has greatly decreased the cost of diamond circuits, and now diamond-film technology is receiving heavy funding. This crystal has wonderful properties. For starters, it conducts heat five times better than copper. (Already, diamond heat sinks have been announced, and molded diamond is for sale.) Diamond is also radiation hard, and its drift velocity is significantly higher than that of silicon. It has 20–50× the breakdown voltage, so insulation layers could be thin and field strengths could be high. Diamond VLSI could be enormously attractive, and the raw material—carbon—is very cheap.

If these new materials don't pan out, perhaps we could replace traditional transistors with quantum devices. These devices acknowledge that below 0.2 micron or so, quantum effects (such as tunneling) are unavoidable. So, they treat quantum effects as a feature, rather than a bug. Some device proposals *require* extremely small dimensions, because they communicate to the next device by tunneling, or because they are planned around the resonant wavelength that an electron is equivalent to. (In quantum mechanics, particles exhibit wave properties, such as interference. An electron in room-temperature GaAs has a wavelength of 0.02 micron, and tunneling distances are on the order of 0.005 micron.) The 712-GHz resonance mentioned above was measured from just such a device. Several research groups have 1000 GHz as their next target, and a full-adder circuit was recently reported, so this technology is certainly something to watch.

New packaging may also be required. Although multichip modules (MCMs) are fairly new for production parts, they are already reaching their limits in research projects, and to go much further requires a 3-D approach. Several research groups have glued SRAM chips into stacks, with interconnect traces down one face of the stack. One of these groups thins the chips first to get more compact stacks. Beyond this is the possibility of stacking things that are *really* thin—on the order of a few microns—producing a chunk of silicon resembling a tiny multi-layer PC board using vias for interchip communication.

Some argue that, as manufacturers try to add more layers to their chips, cumulative thermal stress and defect rates will force them to use some type of stacking arrangement. This would require new technologies to align and attach the stacked devices, as well as to test the individual pieces before final assembly. If these technologies are available, stacking several parts is the next logical step. The use of vias would allow the number of interchip connections to be increased from hundreds to thousands, increasing bandwidth without higher frequencies. Perhaps the entire processor/memory core could be placed into a single package.

Initial research into this stacking technology appears promising. Several groups have removed films from a GaAs wafer that are about 1 cm^2 and just 1 micron thick. They have then bonded these films, which contained typical transistors, to silicon and other substrates and verified that the transistors were still functional, with relatively good yield. Others are attempting to duplicate this feat starting with silicon wafers.

Another completely different approach is optical logic. These are devices that transmit and operate on light pulses instead of electrical pulses. Single devices have been built at well over 1000 GHz, along with a 350-GHz ring oscillator. Light pulses can be transmitted through free space, eliminating the need for wiring. Unfortunately, today's optical devices are physically large and are unlikely to ever match the size of silicon transistors. Totally new architectural designs, such as thousand-stage pipelines, may be required for this technology to achieve competitive computing power.

Now How Much Would You Pay?

It's been said that technology always moves more slowly than you expect over a year or two, but faster than you expect over five or ten years. With that in mind, it is not bold to predict the 256-Mbit DRAM, and the 2-GHz processor, by the year 2000. I also predict that won't be the end. The technology progress curve will begin to flatten out, but it won't be over. Stay tuned. ♦

