Moore's Law: Another 50 Years?

Randal E. Bryant

In April, 1965, *Electronics* magazine published Gordon Moore's article "Cramming more components onto integrated circuits." There he predicted the semiconductor industry would double the number of devices integrated onto a single chip every year. He later modified this prediction to doubling every two years, and this prediction became known as "Moore's Law." Although not a law in any physical, moral, or legal sense, it has become the defining metric for a $300 billion industry that generates the building blocks of information technology systems.

Back in 1965, Moore's company, Fairchild Semiconductor, used its industry-leading semiconductor technology to fabricate four Nand gates on a single chip, requiring around 50 components (chips of that era contained transistors, resistors, and diodes). Fifty years later, the Apple A8 processor, which powers the iPhone 6, has over two billion transistors, almost exactly matching the 25 doublings (3.4×10^7) predicted by Moore's Law. The impact of having such powerful technology available at such a low price has been dramatic. In the eight years since their introduction, for example, smartphones produced by Apple and its competitors are in the hands of around one-fourth of the world's population.

After 50 years of Moore's Law, it's instructive to ask the question "Can Moore's Law hold for another 50 years?" In the year 2065, will it be possible to manufacture an integrated system containing 60 quadrillion (6×10^{16}) components? That number seems impossibly large to us today, but so would a modern microprocessor to those who built systems out of individual Nand gates. In the spirit of the iPhone, this analysis focuses on small, low-power devices for consumer products and the Internet of Things. The high manufacturing volumes for these products will continue to draw the bulk of industry's attention, motivating it to make the huge research and development investments required to drive technology forward. This focus rules out quantum computing and other technologies that require cryogenic operation.

Over the past 50 years, Moore's Law has been applied specifically to the fabrication of planar transistors on silicon wafers using photolithographic methods. Progress in this technology can be quantified by a *linear spacing metric* l_s, defined as $\sqrt{A/N}$ where A is the area of the chip, and N is the number of transistors. The A8 processor, having $A = 89\,\mathrm{mm}^2$, has $l_s = 211\,\mathrm{nm}$. (See the appendix for a guide to sub-millimeter units of length.) Note that the A8 is fabricated with what is described as a $20\,\mathrm{nm}$ process, defined in terms of the smallest feature that can be patterned on a chip. The linear spacing is around 10 times greater than this minimum feature size to accommodate the full transistor structure, its connections to power, ground, and signal wires, and to provide isolation between adjacent transistors. In the early days of the semiconductor industry, Moore's Law improvements were obtained both by increasing the chip size and by decreasing l_s. Recent chips for specialized servers have reached sizes of nearly $700\,\mathrm{mm}^2$, but those in portable devices are typically less than $100\,\mathrm{mm}^2$ ($1\,\mathrm{cm}^2$) to meet packaging constraints and to keep costs low. Tracking a Moore's Law growth in components per chip for these systems therefore requires decreasing l_s, as well as the minimum feature size, by a factor of around 1.4 (approximately $\sqrt{2}$) every two years.

Chips are manufactured using a series of steps that involve depositing and removing materials on a silicon wafer, with the regions of deposition and removal being controlled by photolithographic masks. Although the process is highly sophisticated and complex, and it requires a fabrication facility that costs billions of dollars, it can still be characterized as a constant-time process, manufacturing a chip containing $O(N)$ components in $O(1)$ steps.[1] This efficiency is the key reason that such complex systems can be manufactured at such low cost.

[1] Steps involving photolithography form the bottlenecks in chip fabrication. These require scanning the wafer with a *stepper*, exposing a fixed region (around $9\,\mathrm{cm}^2$) at a time.

The Moore's Law trend clearly cannot be sustained for another 50 years by fabricating transistors on planar surfaces. Fitting 6×10^{16} transistors onto a $1\,\mathrm{cm}^2$ chip would require achieving linear spacing $l_s = 41\,\mathrm{pm}$. That's smaller than the distance separating the two atoms in a hydrogen molecule and over 13 times smaller than the spacing between atoms in a silicon crystal. Indeed, most industry experts predict that existing semiconductor fabrication methods will no longer be viable once feature sizes drop below around $5\,\mathrm{nm}$, yielding chips with a maximum of 50–100 billion transistors.

The only way to sustain the increasing levels of integration predicted by Moore's Law, then, is to fabricate devices in three dimensions. This has been a goal for semiconductor research for many years; limited forms are just now appearing in commercial products. Suppose, for example, a device could make use of a full square meter of surface area. Then integrating 6×10^{16} components would require $l_s = 4.1\,\mathrm{nm}$, far smaller than can be achieved today, but still plausible. Of course, a square-meter chip would not be very usable as part of a portable system, but imagine that this square meter could be fabricated as a device with 10,000 logical layers, each $1\,\mu\mathrm{m}$ thick and having an area of $1\,\mathrm{cm}^2$. (Each logical layer would comprise 10–25 physical layers to create the components and the wires connecting them.) This 10,000-layer structure would then yield a cube, $1\,\mathrm{cm}$ on a side.

One important concern for 3-d manufacturing is cost. Moore himself recognized that greater levels of integration are desirable only if they decrease the cost per component. Using photolithography to pattern each layer in a 3-d structure would increase the fabrication cost in proportion to the number of layers. More generally, if 3-d fabrication requires photolithography for each layer, the process will require $O(N^{1/3})$ steps to create a system with N components. Currently, the cost to manufacture a $1\,\mathrm{cm}^2$ chip is around \$10, including the amortization of the fabrication facility. Increasing this cost by a factor of 10,000, or even by 100, would put the technology out of the viable range for consumer products. One approach to reducing the per-layer cost is to have one set of photolithographic steps pattern multiple layers. This approach is now being used in flash memory manufacturing, fabricating as many as 48 layers of memory cells at a time. Alternatively, it may be possible to avoid photolithography altogether by using the nanoscale operation of chemical processes to generate devices via self assembly.

Keeping the system within the strict power budget required for portable applications also presents a major challenge. For example, current cellphones draw a maximum of around 4 watts, due to both limited battery capacity and the need to avoid becoming too hot to touch. The human brain is often held up as model for low-power computing, with its 86 billion neurons consuming around 25 watts. Is it possible to create a system with 700,000 times more components that draws 16% of the power? There is some hope—for example, the chemical processes that drive DNA transcription and replication require around 10^9 times less energy per operation than does the firing of a neuron. Both neuronal firing and DNA transcription have low power requirements in part because they operate at much lower frequencies than do semiconductor chips. It's not clear how to harness these processes in computing devices, nor whether they would yield sufficient performance.

Looking for guidance to the laws of physics, Landauer's Principle states that any logically irreversible manipulation of information requires at least $kT \ln 2 \approx 3 \times 10^{-21}$ Joules per bit. Although much of the computation performed by a processor could be done reversibly, any system has bounded storage, and so it must keep destroying information as it receives new data. But, even for a system having a data input rate of one petabyte (10^{15} bytes) per second, this principle yields a lower bound of only 24 microwatts.

Achieving another 50 years of Moore's Law progress will be a huge challenge, far more so than what was required in its first 50 years. The components must use different materials and be based on fundamentally different principles than the silicon transistors that have served so well thus far. Extending into three dimensions will require novel approaches to photolithography, or dispensing with photolithography

altogether, in order to keep costs low. Running 6×10^{16} components from a portable energy source will require improvements in power efficiencies that are hard to imagine. Designing, manufacturing, and programming such systems will require major advances along many fronts. Still, there does not seem to be any fundamental physical principle that makes such technology impossible.

It's natural to ask the question "What possible applications warrant a system with 6×10^{16} components?" Isn't there some limit to the world's appetite for computing power? History has shown that the only limits in this regard are our imaginations. No futurist or science fiction writer of 1965 came close to recognizing the power that today's portable electronics, Internet, and cloud computing provide to a major fraction of the world's population. The best strategy is to forge ahead in creating the most advanced technology possible and to trust that those who come after us will put it to good use.

Appendix: Understanding sub-millimeter units of length

The following diagram illustrates the range of length units discussed in this paper, with the goal of providing the reader some intuition as to how small these sizes really are. Note that the logarithmic scale, with sizes ranging over nine orders of magnitude.