Feature | Semiconductors | Devices

21 Jul 2020 | 15:00 GMT

A Better Way to Measure Progress in Semiconductors

It's time to throw out the old Moore's Law metric

By **Samuel K. Moore**

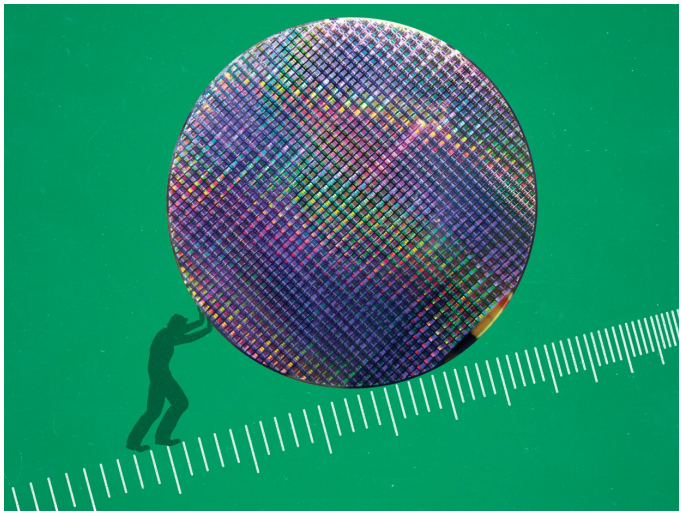


Photo-illustration: Edmon de Haro

One of the most famous maxims in technology is, of course, Moore's Law. For more than 55 years, the "Law" has described and predicted the shrinkage of transistors, as denoted by a set of roughly biannual waypoints called technology nodes. Like some physics-based doomsday clock, the node numbers have ticked down relentlessly over the decades as engineers managed to regularly double the number of transistors they could fit into the same patch of silicon.

When Gordon Moore first pointed out the trend that carries his name, there was no such thing as a node, and only about 50 transistors could economically be integrated on an IC.

But after decades of intense effort and hundreds of billions of dollars in investment, look how far we've come! If you're fortunate enough to be reading this article on a high-end smartphone, the processor inside it was made using technology at what's called the 7-nanometer node. That means that there are about 100 million transistors within a square millimeter of silicon. Processors fabricated at the 5-nm node are in production now, and industry leaders expect to be working on what might be called the 1-nm node inside of a decade.

And then what?

After all, 1 nm is scarcely the width of five silicon atoms. So you'd be excused for thinking that soon there will be no more Moore's Law, that there will be no further jumps in processing power from semiconductor manufacturing advances, and that solid-state device engineering is a dead-end career path.

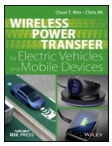
You'd be wrong, though. The picture the semiconductor technology node system paints is false. Most of the critical features of a 7-nm transistor are actually considerably larger than 7 nm, and that disconnect between nomenclature and physical reality has been the case for about two decades. That's no secret, of course, but it does have some really unfortunate consequences.

> **"Picking something that is agreed upon, even if imperfect, is more useful than the current node branding."**
> *—Michael Mayberry, Intel CTO*

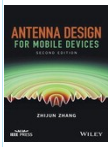
One is that the continuing focus on “nodes” obscures the fact that there are actually achievable ways semiconductor technology will continue to drive computing forward even after there is no more squeezing to be accomplished with CMOS transistor geometry. Another is that the continuing node-centric view of semiconductor progress fails to point the way forward in the industry-galvanizing way that it used to. And, finally, it just rankles that so much stock is put into a number that is so fundamentally meaningless.

Efforts to find a better way to mark the industry’s milestones are beginning to produce clearly better alternatives. But will experts in a notoriously competitive industry unite behind one of them? Let’s hope they do, so we can once again have an effective way of measuring advancement in one of the world’s largest, most important, and most dynamic industries.

Suggested Wiley-IEEE Reading



Wireless Power Transfer for Electric Vehicles and Mobile Devices



Antenna Design for Mobile Devices

So, how did we get to a place where the progress of arguably the most important technology of the past hundred years appears, falsely, to have a natural endpoint? Since 1971, the year the Intel 4004 microprocessor was released, the linear dimensions of a MOS transistor have shrunk down by a factor of roughly 1,000, and the number of transistors on a single chip has increased about 15-million-fold. The metrics used to gauge this phenomenal progress in integration density were primarily dimensions called the metal half-pitch and gate length. Conveniently, for a long time, they were just about the same number.

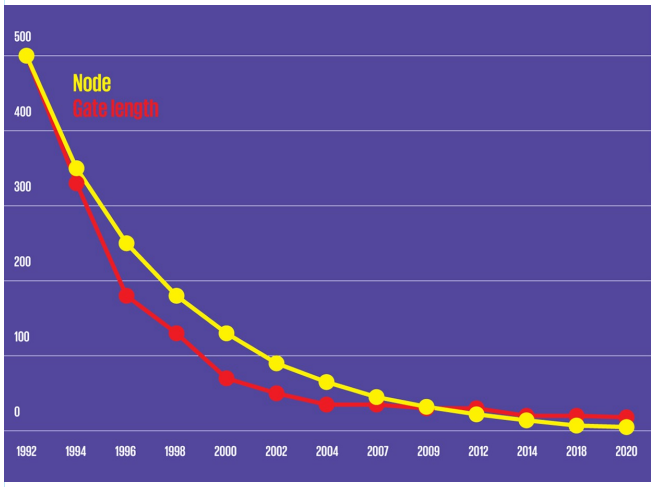
Metal half-pitch is half the distance from the start of one metal interconnect to the start of the next on a chip. In the two-dimensional or “planar” transistor design that dominated until this decade, gate length measured the space between the transistor’s source and drain electrodes. In that space sat the device’s gate stack, which controlled the flow of electrons between the source and drain. Historically, it was the most important dimension for determining transistor performance, because a shorter gate length suggested a faster-switching device.

In the era when gate length and metal half-pitch were roughly equivalent, they came to represent the defining features of chip-manufacturing technology, becoming the node number. These features on the chip were typically made 30 percent smaller with each generation. Such a reduction enables a doubling of transistor density, because reducing both the x and y dimensions of a rectangle by 30 percent means a halving in area.

Using the gate length and half-pitch as the node number served its purpose all through the 1970s and ’80s, but in the mid-1990s, the two features began to uncouple. Seeking to continue historic gains in speed and device efficiency, chipmakers shrank the gate length more aggressively than other features of the device. For example, transistors made using the so-called 130-nm node actually had 70-nm gates. The result was the continuation of the Moore’s Law density-doubling pathway, but with a disproportionately shrinking gate length. Yet industry, for the most part, stuck to the cadence of the old node-naming convention.

The Meaningless Technology Node

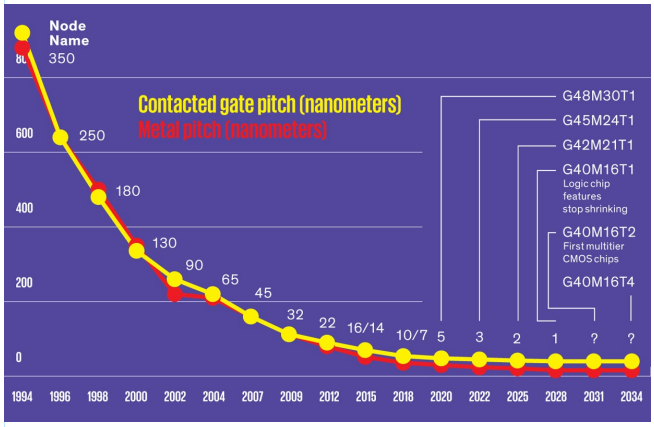
Before the mid-1990s, logic technology nodes were synonymous with the gate length of the CMOS transistors they produced. Actual gate lengths shrunk faster for a while, then stopped shrinking.



Sources: Stanford Nanoelectronics Lab, Wikichip, IEEE International Roadmap for Devices and Systems 2020

The GMT Method

Limits of Lithography: The most advanced lithography technology, extreme ultraviolet lithography, relies on light with a wavelength of 13.5 nanometers. That means chip features will soon stop shrinking. Chipmakers will have to turn to monolithic 3D integration, adding tiers of devices, to keep density increases coming in silicon CMOS . The GMT method tracks this by stating the size of the two most crucial features, contacted gate pitch and metal pitch, as well as the number of tiers.



Sources: Stanford Nanoelectronics Lab, IEEE International Roadmap for Devices and Systems 2020

Developments in the early 2000s drove things further apart, as processors ran up against the limitations of how much power they could dissipate. Engineers found ways to keep devices improving. For example, putting part of the transistor's silicon under strain allows charge carriers to zip through faster at lower voltages, increasing the speed and power efficiency of CMOS devices without making the gate length much smaller.

Things got even stranger as current-leakage problems necessitated structural changes to the CMOS transistor. In 2011, when Intel switched to FinFETs at the 22-nm node, the devices had 26-nm gate lengths, a 40-nm half-pitch, and 8-nm-wide fins.

The industry's node number "had by then absolutely no meaning, because it had nothing to do with any dimension that you can find on the die that related to what you're really doing," says Paolo Gargini, an IEEE Life Fellow and Intel veteran who is leading one of the new metric efforts.

There's broad, though not universal, agreement that the semiconductor industry needs something better. One solution is simply to realign the nomenclature with the sizes of actual features important to the transistor. That doesn't mean going back to the gate length, which is no longer the most important feature. Instead, the suggestion is to use two measures that denote a real limit on the area needed to make a logic transistor. One is called the contacted gate pitch. This phrase refers to the minimum distance from one transistor's gate to another's. The other vital metric, metal pitch, measures the minimum distance between two horizontal interconnects. (There is no longer any reason to divide metal pitch in half, because gate length is now less relevant.)

These two values are the "least common denominator" in creating logic in a new process node, explains Brian Cline, a principal research engineer at Arm. The product of these two values is a good estimate of the minimum possible area for a transistor. Every other design step—forming logic or SRAM cells, blocks of circuits—adds to that minimum. "A good logic process with well-thought-out physical design characteristics will enable the least degradation" from that value, he says.

Gargini, who is chairman of the IEEE International Roadmap for Devices and Systems (IRDS), proposed in April that the industry "return to reality" by adopting a three-number metric that combines contacted gate pitch (G), metal pitch (M), and, crucially for future chips, the number of layers, or tiers, of devices on the chip (T). (IRDS is the successor to the International Technology Roadmap for Semiconductors, or ITRS, a now-defunct, decades-long, industry-wide effort that forecast aspects of future nodes, so that the industry and its suppliers had a unified goal.)

"These three parameters are all you need to know to assess transistor density," says Gargini, who also led ITRS.

The IRDS road map shows that the coming 5-nm chips have a contacted gate pitch of 48 nm, a metal pitch of 36nm, and a single tier—making the metric G48M36T1. It doesn't exactly roll off the tongue, but it does convey much more useful information than "5-nm node."

As with the node nomenclature, the gate pitch and metal pitch values of this GMT metric will continue to diminish throughout the decade. However, they will do so more and more slowly, reaching an endpoint about 10 years from now, at current rates of progress. By that time, metal pitch will be nearing the limits of what extreme-ultraviolet lithography can resolve. And while the previous generation of lithography machines managed to cost-effectively push well past the perceived limits of their 193-nm wavelengths, nobody expects the same thing will happen with extreme ultraviolet.

"Around 2029, we reach the limit of what we can do with lithography," says Gargini. After that, "the way forward is to stack.... That's the only way to increase density that we have."

That's when the number of tiers (T) term will start to become important. Today's advanced silicon CMOS is a single layer of transistors linked together into circuits by more than a dozen layers of metal interconnects. But if you could build two layers of transistors, you might nearly double the density of devices at a stroke.

For silicon CMOS, that's still in the lab for now, but it shouldn't be for long. For more than a decade, industrial researchers have been exploring ways to produce "monolithic 3D ICs," chips where layers of transistors are built atop one another. It hasn't been easy, because silicon-processing temperatures are usually so high that building one layer can damage another. Nevertheless, several industrial research efforts (notably at Belgian nanotech research firm Imec, France's CEA-Leti, and Intel) are developing technology that would build the two types of transistors in CMOS logic—NMOS and PMOS—one on top of the other.

Upcoming nonsilicon technology could go 3D even sooner. For example, MIT professor Max Shulaker and his colleagues have been involved in the development of 3D chips that rely on tiers of carbon-nanotube transistors. Because you can process these devices at relatively low temperatures, you can build them up in multiple tiers much more easily than you can with silicon devices.

Others are working on logic or memory devices that can be built within the layers of metal interconnect above the silicon. These include micromechanical relays and transistors made from atom-thin semiconductors such as tungsten disulfide.

About a year ago, a prominent group of academics got together on the campus of the University of California, Berkeley, to come up with their own metric.

The informal group included some of the biggest names in semiconductor research. In attendance at that June 2019 meeting were all three of the Berkeley engineers credited with the FinFET: Chenming Hu, Tsu-Jae King Liu, and Jeffrey Bokor. Bokor is chair of electrical engineering at the university. Hu is a former chief technology officer at Taiwan Semiconductor Manufacturing Co. (TSMC), which is the world's largest semiconductor foundry, and he was awarded the IEEE Medal of Honor this year. Liu is dean of the college of engineering and sits on the board of directors at Intel. Also present from Berkeley was Sayeef Salahuddin, a pioneer in the development of ferroelectric devices.

From Stanford University, there was H.-S. Philip Wong, a professor and corporate research vice president at TSMC, Subhasish Mitra, who invented a key self-test technology and codeveloped the first carbon-nanotube-based computer with Wong, and James D. Plummer, a former board member at Intel and the longest serving dean of engineering at Stanford. TSMC researcher Kerem Akarvardar and MIT's Dimitri Antonidis joined later.

They all had the sense that their field was becoming less attractive to top students, particularly U.S. students, says Liu. The logic behind that conviction seemed straightforward: If you saw a field where advances were unlikely just 10 years from now, why would you spend four to six years training for it? This perceived lack of attraction for top students was coming when "we actually need more and more innovative solutions to continue to advance computing technology," she says.

This mix of experts sought a metric that would erase the node's doomsday-clock vibe. Crucially, this metric should have no natural endpoint, they decided. In other words, numbers should go up with progress rather than down. It also had to be simple, accurate, and relevant to the main purpose of improving semiconductor technology—more capable computing systems.

To that end they wanted something that did more than describe just the technology used to make the processor, as the IRDS's GMT metric does. They wanted a metric that took into account not just the processor but also other key performance-impacting aspects of the entire computer system. That may seem overly ambitious, and perhaps it is, but it jibes with the direction computing is beginning to go.

Crack open the package of an Intel Stratix 10 field-programmable gate array, and you'll find much more than an FPGA processor. Inside the package, the processor die is surrounded by a range of "chiplets," including, notably, two high-bandwidth DRAM chips. A small sliver of silicon etched with a dense array of interconnects links the processor to the memory.

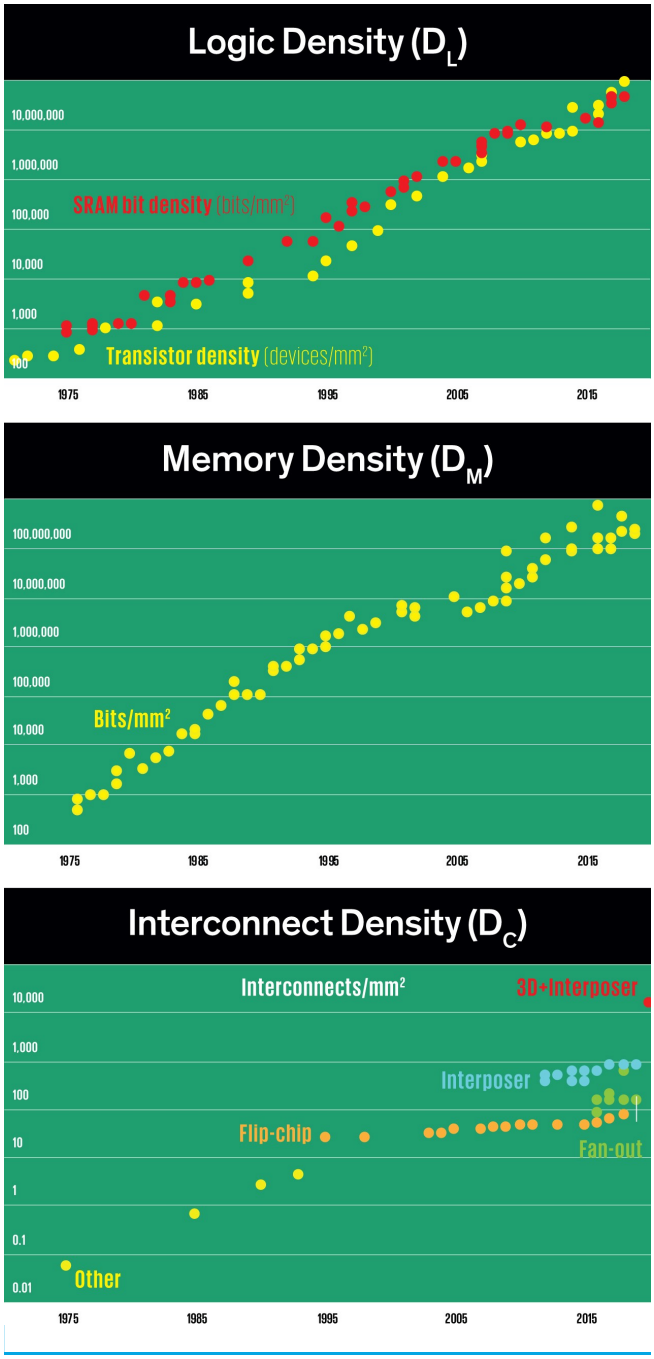
At its most basic, a computer is just that: logic, memory, and the connections between them. So to come up with their new metric, Wong and his colleagues chose as parameters the density of each of those components, calling them D_L, D_M, and D_C. Combining the subscripts, they dubbed their idea the LMC metric.

Together, improvements in D_L, D_M, and D_C are prime contributions to the overall speed and energy efficiency of computing systems, especially in today's age of data-centric computing, according to the originators of the LMC metric. They have plotted historical data showing a correlation between the growth in logic, memory, and connectivity that suggests a balanced increase of D_L, D_M, and D_C has been going on for decades. This balance is implicit in computer architectures, they argue—and, strikingly, it holds true for computing systems of various degrees of complexity, from mobile and desktop processors all the way up to the world's fastest supercomputers. This balanced growth suggests that similar improvements will be needed in the future, says Wong.

The LMC Method

An alternative to the node metric, called LMC, captures a technology's value by stating the density of logic (D_L), the density of main memory (D_M), and the density of the interconnects linking them (D_C).

Source: H.-S. Philip Wong et al., "A Density Metric for Semiconductor Technology," Proceedings of the IEEE, April 2020



In the LMC metric, D_L is the density of logic transistors, in number of devices per square millimeter. D_M is the density of a system's main memory in memory cells per square millimeter. And D_C is the connections between logic and main memory, in interconnects per square millimeter. If there are multiple tiers of devices or a 3D stack of chips, the entire volume above that square millimeter counts.

D_L is perhaps the most historically familiar of the three, as people have been counting the number of transistors on a chip since the first ICs. While it sounds simple, it's not. Different types of circuits on a processor vary in density, largely because of the interconnects that link the devices. The most dense part of a logic chip is typically the SRAM memory that makes up the processor's caches, where data is stored for fast, repeated access. These caches are large arrays of six-transistor cells that can be packed closely together, in part because of their regularity. By that measure the highest value reported for D_L so far is a 135-megabit SRAM array made using TSMC's 5-nm process, which packs in the equivalent of 286 million transistors per square millimeter. In the proposed nomenclature, that'd be written 286M.

But blocks of logic are more complex, less uniform, and less dense than the SRAM that's embedded in them. So judging a technology on SRAM alone might not be fair. In 2017, then Intel senior fellow Mark Bohr advocated a formula that uses weighted densities of some common logic cells. The formula looks at the transistor count per unit area for a simple and ubiquitous two-input, four-transistor NAND gate and for a common but more complex circuit called a scan flip-flop. It weights each according to the proportion of such small gate and large cells in a typical design to produce a single transistors-per-square-millimeter result. Bohr said at the time that SRAM is so different in its density that it should be measured separately.

Internally, AMD uses something similar, according to AMD senior fellow Kevin Gillespie. If a metric doesn't take into account how devices are connected, it won't be accurate, he says.

Another possibility, separately suggested by several experts, would be to measure the average density across some agreed-upon, large block of semiconductor intellectual property, such as one of the widely available processor designs by Arm.

Indeed, Arm abandoned its attempts at a single metric in favor of extracting the density of functional blocks of circuitry from complete processor designs, according to Arm's Cline. "I don't think there is a one-size-fits-all logic density metric for all hardware applications" because the diversity of different types of chips and systems is too great, he says. Different types of processors—CPUs, GPUs, neural network processors, digital signal processors—have different ratios of logic and SRAM, he points out.

In the end, the LMC originators chose not to specify a particular way of measuring D_L, leaving it for debate in the industry.

Measuring D_M is a bit more straightforward. Right now, main memory generally means DRAM, because it is inexpensive, has a high endurance, and is relatively fast to read and write from.

A DRAM cell consists of a single transistor that controls access to a capacitor that stores the bit as charge. Because the charge leaks out over time, the cells must periodically be refreshed. These days the capacitor is built in the interconnect layers above the silicon, so density is influenced not just by the size of the transistor, but by the geometry of the interconnects. The highest D_M value the LMC group could find in the published literature came from Samsung. In 2018, the company detailed DRAM technology with a density of 200 million cells per square millimeter (200M).

DRAM may not always hold its position as main memory. Alternative memory technologies such as magnetoresistive RAM, ferroelectric RAM, resistive RAM, and phase-change RAM are in commercial production today, some as memory embedded on the processor itself, and some as stand-alone chips.

Providing adequate connectivity between main memory and logic is already a major bottleneck in today's computational systems. Interconnects between processor and memory, what D_C measures, have historically been created by package-level technology rather than chipmaking technology. Compared with logic density and memory density, D_C has improved much less steadily over the decades. Instead there have been discrete jumps as new packaging technologies are introduced and then refined. The last decade has been particularly eventful, as single-die systems-on-chip (SoCs) have begun to give way to collections of chiplets bound tightly together on silicon interposers (so-called 2.5-D systems) or stacked in 3D arrangements. A system using TSMC's System on Integrated Chips 3D chip-stacking technology had the highest published D_C at 12,000 interconnects per square millimeter (12K).

However, D_C need not necessarily connect logic to a separate memory chip. For certain systems, main memory is entirely embedded. For example, Cerebras Systems' machine-learning megachip relies entirely on SRAM embedded in proximity with its logic cores on a single massive slab of silicon.

The LMC originators suggest that a system combining the best of all three parameters—D_L, D_M, and D_C—would be described [260M, 200M, 12K].

The time is long gone when a single number could describe how advanced a semiconductor node is, argues Intel CTO Michael Mayberry. However, he does like the idea of having a comprehensive system-level metric, in principle. "Picking something that is agreed upon, even if imperfect, is more useful than the current node branding," he says.

He'd like to see the LMC expanded with an additional level of detail to specify what's being measured and how. For example, regarding the D_M value, Mayberry says that it might need to specifically relate to memory that is within the same chip package as the processor it serves. And what classifies as "main memory" may need fine-tuning as well, he adds. In the future, there may be multiple layers of memory between the processor and data-storage devices. Intel and Micron, for example, make 3D XPoint memory, a type of nonvolatile system that occupies a niche between DRAM and storage.

A further criticism is that both a density-based metric like LMC and a lithography-based one like GMT are a step away from what customers of foundries and memory chipmakers want. "There's area [density], but there's also performance, power, and cost," says AMD's Gillespie. Each chip design makes trade-offs around those four axes, to the point that "there is no single number that can ever capture how good a node is," adds Mayberry.

"The most important metric for memory and storage is still cost per bit," says Gurtej Singh Sandhu, senior fellow and vice president at the world's No. 3 DRAM maker, Micron Technologies. "Several other factors, including various performance metrics based on specific market applications, are also closely considered."

There's also a faction that argues that a new metric isn't even needed at this point. Such measures are "useful really only in applications dominated by scaling," says Gregg Bartlett, senior vice president for engineering and quality at GlobalFoundries, which ended its pursuit of a 7-nm process in 2018. "There are only a few companies manufacturing in this space and a limited number of customers and applications, so it is less relevant to the vast majority of the semiconductor industry." Only Intel, Samsung, and TSMC are left pursuing the last few CMOS logic nodes, but they are hardly bit players, generating a big fraction of global semiconductor manufacturing revenue.

Bartlett, whose company is not in that group, sees the integration of CMOS logic with specialized technologies, such as embedded nonvolatile memory and millimeter-wave radio, as more crucial to the future of the industry than scaling.

But there's no doubt that continued scaling is important for many semiconductor consumers. And the originators of the LMC metric and of the GMT metric both feel a sense of urgency, though for different reasons. For Wong and the LMC supporters, the industry needs to make clear its long-term future in an era when transistor scaling is less important so that they can recruit the technical talent to make that future happen.

For Gargini and the GMT backers, it's about keeping the industry on track. In his view, without the synchronization of a metric, the industry becomes less efficient. "It increases the probability of failure," he says. "We have 10 years" until silicon CMOS stops shrinking entirely. "That's barely sufficient" to produce the needed breakthroughs that will keep computing going.

This article appears in the August 2020 print issue as "The Node is Nonsense."