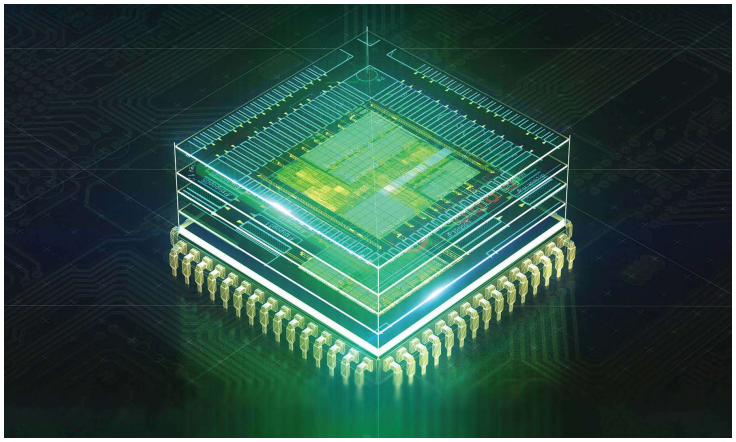
POINT OF VIEW

A Density Metric for Semiconductor Technology

By H.-S. PHILIP WONG
Stanford University, Stanford, CA 94305 USA
Taiwan Semiconductor Manufacturing Company, Hsinchu 300, Taiwan
KEREM AKARVARDAR
Taiwan Semiconductor Manufacturing Company, San Jose, CA 95134 USA
DIMITRI ANTONIADIS
Massachusetts Institute of Technology, Cambridge, MA 02139 USA
JEFFREY BOKOR
University of California at Berkeley, Berkeley, CA 94720 USA
CHENMING HU
University of California at Berkeley, Berkeley, CA 94720 USA
TSU-JAE KING-LIU
University of California at Berkeley, Berkeley, CA 94720 USA
SUBHASISH MITRA
Stanford University, Stanford, CA 94305 USA
JAMES D. PLUMMER
Stanford University, Stanford, CA 94305 USA
SAYEEF SALAHUDDIN
University of California at Berkeley, Berkeley, CA 94720 USA



Since its inception, the semiconductor industry has used a physical dimension (the minimum gate length of a transistor) as a means to gauge continuous technology advancement. This metric is all but obsolete today. As a replacement, we propose a density metric, which aims to capture how advances in semiconductor device technologies enable system-level benefits. The proposed metric can be used to gauge advances in future generations of semiconductor technologies in a holistic way, by accounting for the progress in logic, memory, and packaging/integration technologies simultaneously.

I. BACKGROUND

Starting in the 1960s, the semiconductor industry has used the lateral physical dimension of a key minimum-size feature (the transistor gate length) as a label to denote the progress from one generation of manufacturing technology to the next. This label, known as the node number, has shrunk from micrometer size in the 1980s to single-digit nanometer size today. Driven by competitive marketing in the

Digital Object Identifier 10.1109/JPROC.2020.2981715

most recent decade, this label has become decoupled from, and can be several times smaller than, the actual minimum gate length, while it also fails to convey other essential characteristics of the technology. Moreover, similar logic technologies from different semiconductor manufacturers have been branded with different node labels, thus creating further confusion. Perhaps equally important, the single-digit nanometer label for the next node in development (3 nm) is the size of only about a dozen atoms. This gives the false impression that the semiconductor technology will soon reach a barrier it cannot surmount. Yet, it is a foregone conclusion that the semiconductor industry will continue to make progress because there are still many ways to advance semiconductor technology beyond 2-D miniaturization and also because societal demand for more capable electronic systems is insatiable. Thus, it is time for the semiconductor industry to adopt a new metric that properly denotes semiconductor manufacturing technology advancements. The use of this new metric will allow industry, research institutes, academic researchers, students, funding agencies, and government policymakers to recognize, project, and plan for continued technological progress.

II. RATIONALE FOR A NEW METRIC

Semiconductor technology has advanced exponentially for more than five decades, driven by a self-fulfilling prophecy known as Moore's law [1]. In 1965, Gordon Moore observed that the number of transistors in an integrated circuit doubles with each new generation of technology. Since 1971 (the Intel 4004 microprocessor [2]), the transistor size has been shrunk down in the 2-D plane of the chip roughly by 1000-fold and the number of transistors on a single 2-D chip has increased by about 15 million times [3]. The metric that has been used to gauge this phenomenal progress in integration density has been primarily the minimum physical gate length of the transistors on the chip. This physical dimension (also known as the node number [4]) has been used as a label to characterize the semiconductor manufacturing technology.

The most advanced technology in high-volume manufacturing today is known as the 7-nm node, and the 5-nm node is slated to enter high-volume manufacturing within a year. As such, we will soon run out of nanometers for naming future generations of technologies.[1] This gives the false impression that semiconductor technology is reaching physical limits and will no longer be a contributor to future advancements in information technology and electronic systems. It is certainly true that the 2-D miniaturization will eventually reach a limit (the size of an atom, and likely well before that), and some argue that progress in 2-D miniaturization is already slowing [5]. At the same time, it is also true that continuous improvement of semiconductor technology can (and will) be obtained by many other means—new approaches that are already being investigated [6]–[11] (3-D integration being a prominent example) and new approaches that are yet to be invented.

Notably, since the mid-1990s, the node number that signifies the technology generation has been decoupled from the physical transistor gate length that was used to identify a technology. The adoption of "equivalent scaling" [12] since the 2000s has further decoupled the essence of a technology from physical dimensions inside a chip. More recently, design-technology co-optimization (DTCO) has played a significant role at each technology generation [13] and basically renders the node number a somewhat arbitrary name—similar to the model name of a computer—that is unrelated to the essential attributes of the technology [14]. Because the label that describes a technology node is decoupled from the essential attributes of that node, the semiconductor industry and its research and development community urgently need a simple and rational metric to better characterize future generations of semiconductor technologies that are increasingly complex and nuanced.

[1]The use of angstrom or picometer does not substantially change the picture as it further decouples the name from the essential attributes of the technology as the number approaches the size of atoms.

III. LOGIC, MEMORY, CONNECTIVITY (LMC) METRIC

Improved semiconductor device density directly translates into benefits for more advanced computing systems—the primary driver for progress in semiconductor technology. Thus, we propose the use of the following three-part number as a metric to gauge advancement of future semiconductor technologies: $[D_L, D_M, D_C]$, where D_L is the density of logic transistors (in #/mm^2), D_M is the bit density of main memory (currently the off-chip DRAM density, in #/mm^2), and D_C is the density of connections between the main memory and logic (in #/mm^2). As an example, today's leading edge technologies that are published in the literature [15]–[17] can be characterized by [38M, 383M, 12K]. As another example, 3-D stacking of multiple logic and memory dies can increase D_L, D_M, and D_C.

Fig. 1 shows the historical logic, memory, and interconnect density trends. In Fig. 1(a), transistor density is simply given by the number of transistors divided by the die area. However, more sophisticated

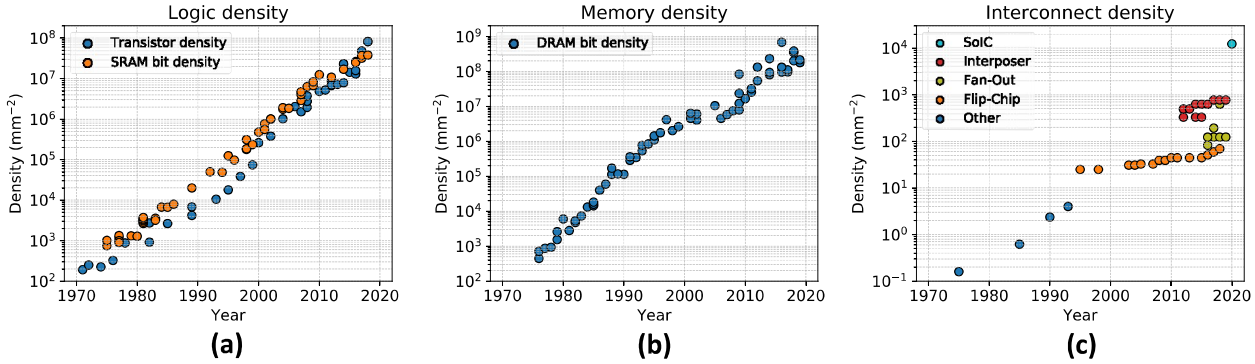


Fig. 1. *Historical density trends. (a) Transistor density (from https://en.wikipedia.org/wiki/Transistor_count) given by the number of transistors divided by chip area and SRAM bit density (data compiled by Fiona Wang of Stanford University[+]). (b) DRAM bit density (data compiled by Haitong Li of Stanford University[+]). (c) Interconnect density between memory and logic (data compiled by Doug Yu of TSMC). [+]The data is available at permanent link https://purl.stanford.edu/jj585np1768. Current, up-to-date data can be accessed at https://nano.stanford.edu/technology-integration-trend*

proposals to calculate the transistor density, such as using a weighted average of logic gate density (based on the frequency of use of various gates in a typical chip design), have been made as well [14]. Fig. 1 also shows that the number of transistors and SRAM bits per mm^2 have been following a very similar trend. The historical trend of dynamic random access memory (DRAM) bit density is shown in Fig. 1(b). The averaged improvement rate over the years is comparable to the D_L improvement. DRAM has been the main memory for computing systems and, as of now, represents the D_M component of the system metric. In a future scenario where alternative memory technologies (e.g., emerging nonvolatile memory [18]) are used, D_M can seamlessly correspond to the bit density of such new memory acting as the main memory. The interconnect density between (on-chip) logic and (off-chip) main memory can be considered as representative of logic to memory connectivity. Fig. 1(c) shows the density of off-chip interconnects associated with various packaging technologies. These density values can be used as proxies that correlate with the D_C component of LMC metric. However, to be general, D_C does not necessarily have to reflect connectivity to off-chip memory. In a scenario where the main memory can be integrated on-chip as the logic, the connectivity between logic and on-chip memory can become very important from a system perspective and can also be expressed with D_C. Compared to D_L and D_M, the progress in D_C has been characterized by somewhat discrete jumps as new technologies are introduced. The advancement of this memory to logic connectivity has gained significant momentum in the last decade through substantial improvements in the 2.5-D and 3-D packaging techniques [17].

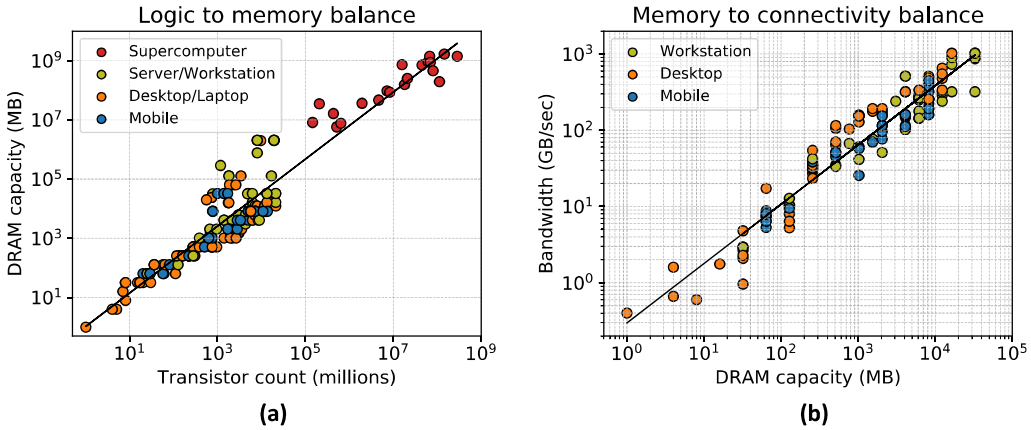


Fig. 2. *Balance of resources in computing systems. (a) DRAM capacity versus transistor count (data compiled by Mohamed Sabry Aly of Nanyang Technological University and Wei-Chen (Harry) Chen of Stanford University[+]). (b) Bandwidth versus DRAM capacity for GPUs (data compiled by Wei-Chen (Harry) Chen of Stanford University and Kerem Akarvardar[+]). These trends suggest that a balanced growth between logic, memory, and connectivity has been an implicit guide for computing system optimization. [+]The data is available at permanent link https://purl.stanford.edu/jj585np1768. Current, up-to-date data can be accessed at https://nano.stanford.edu/technology-integration-trend*

Note that D_C is the interconnection density between the main memory and the logic circuits, regardless of whether such connection is made by 2.5-D, 3-D, monolithic 3-D integration, or other techniques that may be developed in the future.

These three components of the system metric contribute to the overall speed and energy efficiency of computing systems. Historical data in Fig. 2 show a correlated growth in logic, memory, and connectivity, which suggests a balanced increase of D_L, D_M, and D_C for the decades to come. This balance is implicit in computer architectures [6] and allows the improvement of overall system performance in an optimal fashion. Fig. 2(a) shows the DRAM capacity versus the number of transistors for computing systems of various degrees of complexity, from mobile/desktop processors all the way up to the world's fastest supercomputers. We note that the aforementioned logic to memory balance holds across eight orders of magnitude change in the transistor count and main memory capacity, the slope of the best fitting line to data being close to 1.

Providing adequate connectivity (bandwidth) between main memory and logic is essential; otherwise, the speed and energy efficiency of computing systems would be severely limited by memory access. This memory access challenge is already evident in today's computational workloads and systems [6]. The widespread adoption of high-bandwidth memory (HBM) [19] despite its relatively high cost indicates how crucial the connectivity is. Indeed, the historical bandwidth versus memory capacity trend for desktop GPUs in Fig. 2(b) also suggests a balanced growth between the memory capacity and the bandwidth, which is proportional to the number of physical connections (bus width) between logic and memory. As clock frequencies saturate due to power limitations [20], bandwidth improvements may increasingly rely upon the density of connections between logic and memory. Depending on the system design and cost–performance tradeoff, the density of physical connections between logic and main memory, D_C, varies by orders of magnitude: from printed circuit boards to interposers, then to chip-to-wafer and wafer-to-wafer direct bonding, and finally to ultradense interlevel vias in a monolithic 3-D integrated chip stack in the future.

IV. BENEFITS TO THE SEMICONDUCTOR INDUSTRY

This more comprehensive LMC density metric [D_L, D_M, D_C] can be used to capture the essential technical attributes of semiconductor technologies that are becoming increasingly complex and nuanced. While companies may continue to use their preferred labels to market their technologies, the LMC density metric can serve as a common language to gauge technology advances among semiconductor manufacturers for their customers and other parties to facilitate clear communication. This metric accounts for the benefits that come from the *integration* of logic, memory, and connectivity into a system. In addition to being consistent with historical trends and our intuition about computing systems, the LMC density metric is applicable and extensible to future logic, memory, and packaging/integration technologies.

Technology providers and researchers may address one or more of the components of the LMC metric. Companies providing end products, such as domain-specific hardware accelerators, may choose to mention all three components to describe the specific logic, memory, and packaging technology that are marshaled to build the latest model of their products. This three-pronged metric directly connects the device technology advances to system-level benefits in a comprehensive fashion while acknowledging the synergy between various components. For instance, a semiconductor technology featuring 3-D packaging that stacks multiple logic and memory dies would have a commensurately increased D_L and D_M, thus showcasing the progress versus another possible product employing the same logic and memory technologies but not featuring 3-D die stacking. Similar to technology companies, businesses, consumers, and government agencies are likely to find this more comprehensive description of the state of a given semiconductor technology useful and convenient. Most importantly, the use of this LMC density metric takes the semiconductor industry out of the quandary of using the vanishing nanometer as a label to describe advancements in semiconductor technology that will remain very important to society for a very long time to come.

Acknowledgment

The authors would like to acknowledge the DARPA 3DSoC project, the National Science Foundation (NSF) E3S Science and Technology Center, and the Semiconductor Research Corporation (SRC) JUMP ASCENT Center. The authors are grateful to the following colleagues and graduate students for their contributions to this article: Wei-Chen (Harry) Chen, Haitong Li, Shengjun (Sophia) Qin, and Ching-Hua (Fiona) Wang from Stanford University; Jin Cai, Don C. L. Chen, Benson Chiang, Chih-Hang Tung, Chuei-Tang Wang, and Douglas Yu from Taiwan Semiconductor Manufacturing Company (TSMC); and Mohamed M. Sabry Aly from Nanyang Technological University. They would also like to thank Min Cao, Godfrey Cheng, Michael Wu, and Kevin Zhang from TSMC; Supratik Guha from The University of Chicago; and Thomas N. Theis from Utopus Insights for the discussions and critical reading of the article. ■

REFERENCES

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Proc. IEEE*, vol. 86, no. 1, pp. 82–85, Jan. 1998.

[2] *Intel 4004*. [Online]. Available: https://en.wikipedia.org/wiki/Intel_4004

[3] *Moore's Law—The Number of Transistors on Integrated Circuit Chips*. [Online]. Available: https://en.wikipedia.org/wiki/Moore%27s_law#/media/File:Moore's_Law_Transistor_Count_1971-2018.png

[4] Accessed: Mar. 21, 2020. [Online]. Available: https://en.wikipedia.org/wiki/International_Technology_Roadmap_for_Semiconductors

[5] T. N. Theis and H.-S. P. Wong, "The end of Moore's law: A new beginning for information technology," *Comput. Sci. Eng.*, vol. 19, no. 2, pp. 41–50, Mar. 2017.

[6] M. M. S. Aly *et al.*, "The N3XT approach to energy-efficient abundant-data computing," *Proc. IEEE*, vol. 107, no. 1, pp. 19–48, Jan. 2019.

[7] S. Salahuddin and S. Datta, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.*, vol. 8, no. 2, pp. 405–410, Feb. 2008.

[8] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature Nanotechnol.*, vol. 10, no. 3, pp. 191–194, Mar. 2015.

[9] D. Monroe, "Neuromorphic computing gets ready for the (really) big time," *Commun. ACM*, vol. 57, no. 6, pp. 13–15, Jun. 2014.

[10] D. C. H. Yu, "Wafer level system integration for SiP," in *IEDM Tech. Dig.*, Dec. 2014, Paper 27-1.

[11] G. Hills *et al.*, "Modern microprocessor built from complementary carbon nanotube transistors," *Nature*, vol. 572, no. 7771, pp. 595–602, Aug. 2019.

[12] M. Bohr, "The new era of scaling in an SoC world," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2009, pp. 23–28.

[13] G. Yeric, "Moore's law at 50: Are we planning for retirement?" in *IEDM Tech. Dig.*, Dec. 2015, pp. 1–8, Paper 1.1.

[14] M. Bohr. *Let's Clear Up the Node Naming Mess*. [Online]. Available: https://newsroom.intel.com/editorials/lets-clear-up-node-naming-mess/#gs.ywolh7

[15] T. Song *et al.*, "A 7 nm FinFET SRAM using EUV lithography with dual write-driver-assist circuitry for low-voltage applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 198–200.

[16] K. C. Chun *et al.*, "A 16 Gb LPDDR4X SDRAM with an NBTI-tolerant circuit solution, an SWD PMOS GIDL reduction technique, an adaptive gear-down scheme and a metastable-free DQS aligner in a 10nm class DRAM process," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 206–208.

[17] M.-F. Chen, F.-C. Chen, W.-C. Chiou, and D. C. H. Yu, "System on integrated chips (SoIC) for 3D heterogeneous integration," in *Proc. IEEE 69th Electron. Compon. Technol. Conf. (ECTC)*, May 2019, pp. 594–599.

[18] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable DRAM alternative," in *Proc. Int. Symp. Comput. Archit. (ISCA)*, 2009, pp. 2–13.

[19] *JEDEC Standard High Bandwidth Memory (HBM) DRAM Specification*, Standard JESD235A, 2015.

[20] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 10–14.