Sagar Doshi, UID: 604901376
EE121B, Dis 1A

A Brief Review of Dennard Scaling in 2019

The seminal 1974 paper[1] co-authored by Robert H. Dennard and his fellow researchers at IBM set forth the foundational framework on MOSFET scaling that the semiconductor industry would strive to follow over the next three decades. At the time of the paper's publication, commercially available integrated circuits were on the order of 5 microns[2], but laboratories including Dennard's team were already exploring MOSFET technologies on the scale of 1 micron. The journal article advocates the use of ion-implantation as it demonstrates a small-scale device fabricated with this technique. Moreover, it provides doping and transport models for the fabricated device[1]. The enduring legacy of this paper, though, comes from the projections Dennard and his team made based upon the successful implementation of their 0.5μ channel-length device. Their categorical prediction of how device characteristics would change as MOSFETs scaled down can be summarized by the following canonical table.

Table 1
Scaling Results for Circuit Performance

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

Figure 1 – Table 1 from [1] describing scaling results for circuit performance. The key parameters to note are the power density (Power/Area) and power dissipation scaling factors.

Device length and width both scale down when creating smaller MOSFETs, so the total device area scales down by a factor of $\frac{1}{\kappa^2}$. Similarly, voltage and current both scale down with factors $\frac{1}{\kappa}$ meaning that $P = IV$ scales down with a factor of $\frac{1}{\kappa^2}$. The combination of these scaling characteristics means that the device's total power dissipation would remain in proportion with

its area as it scaled down. In other words, Dennard and his team predicted that the power density of MOSFET devices would remain constant as they scaled down[1].

Without this property, integrated circuits could not continue to reduce in size without simply melting the silicon substrate that they were manufactured on. Moore's Law—the doubling of transistors per chip generation—combined with the constant power density provided by Dennard scaling meant that the performance/watt of next generation chips also doubled. Up until around 2006[2], Dennard scaling meant that reducing MOSFET transistor size was enough to continue Moore's law and improve CPU speed and power performance. Leading up to this point, however, the scaling predictions set forth by the paper were steadily breaking down.

The original 1974 paper[1] listed that voltage would scale down by a factor of $\frac{1}{\kappa}$, but it ignored the effect of sub-threshold leakage on overall chip power[2]. By 2007, the threshold voltage (V_T) for MOSFET devices had been scaled to the point where sub-threshold leakage had increased from levels of $\sim 10^{-10}\,\frac{A}{mm}$ all the way to $\sim 10^{-7}\,\frac{A}{mm}$. This meant that further scaling down V_T would prove problematic. Moreover, the paper also assumed oxide thickness would scale down indefinitely as substrate doping increased unboundedly thus resulting in shorter channel lengths[2]. Even Intel's (now obsolete) 65nm generation technology had an SiO_2 thickness of just $1.2nm$, and simply put MOSFET designers were running out of atoms[2]. The doping problem would also require more creative approaches to MOSFET scaling as degenerate levels of doping result in degraded carrier mobility as well as increases in source and drain junction leakage current due to Zener tunneling[2]. As if these problems were not enough, the 1974 paper's less well-known predictions regarding interconnects were problematic as RC time constant delays among the "wires" that made up integrated circuits were becoming troublesome. While this issue was somewhat remedied by replacing aluminum interconnects with copper and

utilizing low-K dielectrics[2], solving the problem of interconnects at the nm scale is a burgeoning area of research. Figure 2 illustrates the growth of the power problem throughout the 90s (Pentium III was released in 1999).

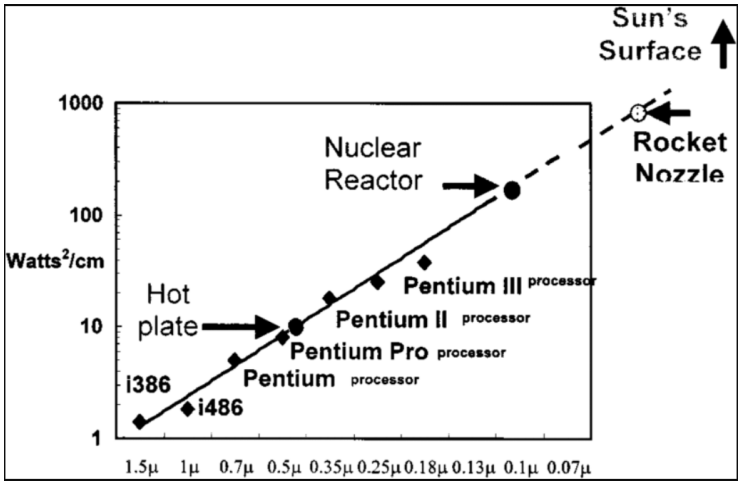

Figure 2 – Adapted from [3], plot of transistor length vs. power density. The key feature is that power density is on a logarithmic scale and approaching unsustainable levels.

Many of the previously mentioned problems have been addressed with new transistor technologies such as strained silicon, high-K dielectrics, metal gates and multiple-gate devices[2], and these innovations helped keep the spirit of Dennard scaling alive. However, the "killing blow" to Dennard scaling was that static power losses increased more rapidly as a proportion of overall power supplied to devices as MOSFET operating voltages dropped[4]. The reason for the increase in static power losses comes from the equation for static power loss in CMOS technologies[5] $P_s = \sum I_{leakage} \times V_{supply}$. If the leakage current is increasing as devices scale down, and the supply voltage is not scaling down at an adequate rate, then these static losses become very significant. In terms of processor speeds, it means that it is no longer feasible to just drive devices at higher and higher frequencies as the static power dissipation would lead to thermal runaway processes[5] and melt the CPU. Consequently, the clock speed of computer chips has not significantly jumped in the past two decades as processors in 2010 operated at

frequencies of $\sim 3GHz$ while Intel's 2019 i7 processors run at frequencies of $\sim 5GHz$[6]. Figure 3 illustrates an incorrect prediction of the increase of processor clock speeds made in a 2001 paper.

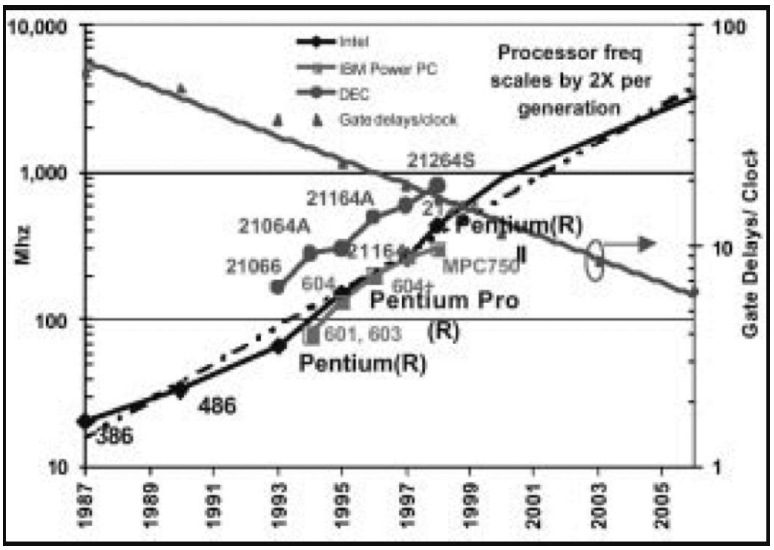

Figure 3 – Adapted from [3], plot of processor frequency vs. year. The plot's prediction of CPU frequency doubling every generation has been proven false due to the breakdown of Dennard scaling.

What many incorrectly dubbed the death of Moore's law[4] in the first decade of the 2000s is more accurately described as the end of Dennard scaling. Moreover, it is 2019 and Moore's law has been alive and well for the past decade. The number of transistors on a chip has continued to increase as advances in lithography are made (Samsung is purportedly manufacturing 5 nm technology[7]), but the increase in CPU speed has come from utilizing a greater number of transistors in parallel rather than simply driving transistors faster. The last decade has been the era of multi-core processors[6], and leveraging parallelism has enabled Moore's law to continue on despite Dennard scaling ending.

However, another bottleneck has appeared in the design of high-performance CPUs. Newer generations of multi-core chips are reaching the limits of how much parallelism can be used to speed up performance, and many transistors are starting to go unused—a phenomena known as dark silicon[8]. Just as the industry was able to overcome the end of Dennard scaling,

new technologies such as optical transistors[9] and state-changing metal oxides[10] are being developed in hopes of overcoming the next challenge in improving device performance. Moreover, there has been a recent resurgence in computer architecture design as application-specific based hardware such as Google's TPU chips for machine learning have optimized task-specific CPU performance far beyond the current benefits of Moore's Law. New approaches to computing such as quantum processors may also push the boundaries of modern computing. While Dennard scaling may have ended years ago, the notion of developing new technologies to advance the semiconductor industry and carry on Moore's law has endured.

Work Cited

[1] Dennard, R.h., et al. “Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions.” *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, 1974, pp. 256–268., doi:10.1109/jssc.1974.1050511.

[2] Bohr, Mark. “A 30 Year Retrospective on Dennard's MOSFET Scaling Paper.” *IEEE Solid-State Circuits Newsletter*, vol. 12, no. 1, 2007, pp. 11–13., doi:10.1109/n-ssc.2007.4785534.

[3] Ronen, R., et al. “Coming Challenges in Microarchitecture and Architecture.” *Proceedings of the IEEE*, vol. 89, no. 3, 2001, pp. 325–340., doi:10.1109/5.915377.

[4] Mcmenamin, Adrian. “The End of Dennard Scaling.” *Cartesian Product*, 15 Apr. 2013, cartesianproduct.wordpress.com/2013/04/15/the-end-of-dennard-scaling/.

[5] Sarwar, Abul. “CMOS Power Consumption and Cpd Calculation.” Texas Instruments, 1997.

[6] “Intel® Core™ i7 Processors.” *Intel*, Intel Corporation, www.intel.com/content/www/us/en/products/processors/core/i7-processors.html.

[7] Mu-Hyun, Cho. “Samsung Develops EUV 5nm Chip Process.” *ZDNet*, ZDNet, 16 Apr. 2019, www.zdnet.com/article/samsung-develops-euv-5-nanometre-chip-process/.

[8] Esmaeilzadeh, H., et al. “Dark Silicon and the End of Multicore Scaling.” *IEEE Micro*, vol. 32, no. 3, 2012, pp. 122–134., doi:10.1109/mm.2012.17.

[9] “World's First Ultrafast All-Optical Room Temperature Transistor.” *IBM Research Blog*, IBM, 5 June 2019, www.ibm.com/blogs/research/2019/05/ultrafast-optical-room-temperature-transistor/.

[10] Fingas, Jon. “IBM Turns Metal Oxides into Non-Volatile Chips through Liquid Currents.” *Engadget*, IBM, 19 July 2019, www.engadget.com/2013/03/21/ibm-turns-metal-oxides-into-non-volatile-chips-through-currents/.