IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 26, 2018, accepted December 4, 2018, date of publication December 10, 2018,
date of current version January 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885895

Chip Power Scaling in Recent CMOS Technology Nodes

GHAVAM G. SHAHIDI, (Fellow, IEEE)
IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
e-mail: shahidi@us.ibm.com

ABSTRACT This paper tracks the scaling of total chip power at constant frequency (i.e., energy-per-operation) through the last few CMOS nodes. The focus is on high-performance microprocessors. To evaluate the progression of chip power, Intel's Core-i7 (Intel's highest performance consumer microprocessor manufactured in the highest performance CMOS technology node) was used as the benchmark. Core-i7 has been manufactured for eight generations starting in the 45-nm node and continuing through the 14++ node. This paper argues that in the more recent nodes, the total chip power at constant frequency (energy-per-operation) has scaled much less than that of the earlier CMOS nodes. The early 14-nm technology exhibited particularly poor power scaling, and in fact, the technology was improved by increasing the device current and relaxation of the contacted gate pitch in 14++. Early product data in 10 nm points to issue in dropping the chip power (at constant frequency) relative to the previous node (14++), which may challenge the power-performance justification for scaling to the 10 nm node and beyond. Improving chip power scaling (energy-per-operation) in upcoming nodes is critical as a key part of the value proposition for continued CMOS scaling, especially as applied to high-performance microprocessors.

INDEX TERMS Computer performance, CMOS scaling, FinFET, Moore's law, MOSFET, power dissipation, scaling, technology node.

I. INTRODUCTION

Two key entitlements of CMOS scaling have been the linear shrink of about 30% per dimension per node [1], and performance-voltage scaling [2]. A key benefit of scaling has been power scaling, specifically the scaling of chip power at a given frequency (i.e. energy per operation) in each technology node. Because of the slow-down in frequency scaling, total chip power scaling has become more important and the value of scaling has shifted to enabling new applications (enabled by lower power), more cores, or new functions. This benefit is highlighted in Table 1. The greater the node-to-node power reduction, the greater the opportunity to add features and functions in next generation designs. This paper starts by looking at the historical value of chip power scaling per node, and then focuses on the more recent node-to-node values. It appears that the benefit of chip power scaling in recent CMOS nodes has been diminishing as compared to earlier nodes. This conclusion seems particularly evident in study of the early 14 nm generation. A similar trend appears to hold based on early data for the upcoming 10/7 nm nodes. The paper speculates about the causes of the poor chip power scaling and possible remedies.

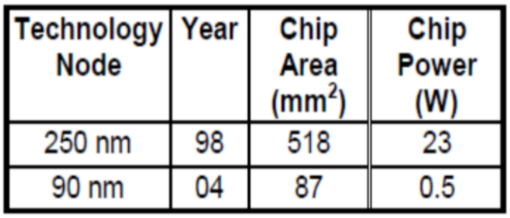
II. METHODOLOGY AND APPROACH

Total chip power, P_{Total} is can be approximated as $1/2\ fCV^2 + P_{Leakage}$, where f is the frequency and C is the effective switching capacitance, and $P_{Leakage}$ is the stand-by leakage power (power with the clock stopped). $P_{Leakage}$ is a function of total device widths on the chip, device off currents (i.e. thresholds), and operating voltage. In this study, the focus is on high performance CMOS, and specifically higher performance microprocessors. In recent high performance microprocessors, active power dominates, and the leakage power (i.e. deep sleep power) is about 5-20% of the total power [3], [4]. Node-to-node scaling of total chip power at a given frequency has two enablers. The first of these is the linear dimensional shrink in x, y, and z. A 30% linear shrink will reduce the capacitance and the leakage current or stand-by power by the shrink factor (~30% per node) and thus total chip power at constant frequency (i.e. active energy-per-operation, and the leakage current) will be reduced by 30% (assuming the same device characteristics, i.e. no reduction in supply voltage). The second major contributor to the energy-per-operation drop has been device scaling and the enhancements in the device that can be traded off for power reduction

TABLE 1. If energy-per-operation drops by p in node N+1 (relative to node N), then the number of cores (or device count) can be increased by 1/(1-p) in node N+1, while operating at the same chip power (as the previous node).

Power Reduction	Cores in New Node
70%	>3.33
50%	>2
40%	>1.67
30%	**>1.43**
20%	>1.25
10%	>1.11

TABLE 2. SONY PlayStation 2 chip power in 250 nm node, and after migration through 3 nodes, in 90 nm node [7].



Technology Node	Year	Chip Area (mm^2)	Chip Power (W)
250 nm	98	518	23
90 nm	04	87	0.5

at a given frequency [2], i.e. running the chip at lower voltage, or reducing the effective device size beyond that dictated by dimensional scaling, or using low threshold devices and operating at lower voltage (at the price of higher leakage power) while maintaining the same frequency. One benefit of the P_{Total} formulation ($fCV^2 + P_{Leakage}$) is that the resistance of the device or of the BEOL resistance does not appear directly. This work uses total chip power at a given frequency to track the evolution of the total power (or total energy-per-operation) through node transitions. In this study, Intel's highest performance desktop and mobile processor, Core-i7, is used to evaluate the energy-per-operation through recent nodes., This processor family has data available for 8 design generations starting from 45 nm and continuing through the 14++ nodes [5]. For the total chip power, the thermal design power (TDP), which is the highest steady amount of power that the chip can dissipate while running applications, is used. To keep the number of cores and the amount cache the same throughout this study, the TDP-frequency of the 4 core/8 MB cache, or the 2 cores/4 MB, and the 6 cores/12 MB product families are all scaled to a normalized 4 core/8 MB case. The total number of processor core transistors has been fairly constant at about 800 million, and while the graphic engine device count has increased dramatically over many generations, the power of the graphic engine is only a few percentage points of the total power, especially at high frequencies. In the case of Intel, in moving products from node-to-node, they use a "Tick-Tock" approach [6], where the first product in the new node is a straight map of the previous generation ("Tick"), followed by the introduction of a new core and/or features in the new node ("Tock"). The early product (i.e. the Tick" family) allows evaluation of the impact of technology on a given product family. We plot all available data [5] for the Core-i7 at a given generation and node, and the data envelope represents the best that the technology can achieve.

III. RESULTS

Per classical scaling theory, for a dimensional scaling factor of α (typically 0.7x shrink factor per node transition), and potential (voltage) scaling factor of β (which has varied over a range from a 0.7x to a 1.0x reduction in voltage), at constant delay (frequency), the active energy per operation scales as $\alpha\beta^2$ [2], i.e. about a 66% drop in power at constant frequency per node for α and β set to about 0.7×. The leakage power scales as $\alpha\beta$ assuming device off current is kept the same from node to node (i.e. >50% in the above case).

In order to establish that it has been indeed feasible to observe the full scaling benefit in terms of total chip power reduction when migrating to the next node at the chip level, processor chip power data from earlier nodes was considered. The Sony PlayStation 2 chip was initially manufactured in a 250 nm node, and later moved through 3 shrinks (180 nm, 130 nm and 90 nm). Its power at constant frequency was dropped from 23 W to 0.5 W (Table 2) during this migration, which corresponds to an average factor of 0.72 drop in power per node, i.e. $(1\text{-}0.72)^3$, for α and β values of approximately 0.7× [7]. The slightly higher amount of power scaling at the product (microprocessor) level was achieved through a combination of technology and circuit optimization. To look at more recent nodes, Intel Core-i7 is used, starting with the first generation of design in 45 nm (1.2 V supply voltage), and the 2nd generation in 32 nm (1.0 V nominal supply voltage). The map of 45 to 32 nm (the "Tick" design) resulted in a power drop (at constant frequency) of 38%, i.e. i7-975 to i7-990X, where neither design included graphic cores (Fig. 1). In 2011, Intel introduced the "Tock" design, the Sandy Bridge family [8]. When operated at the same frequency as the preceding "Tick" design, Sandy Bridge offers a benchmark improvement of approximately 10% [9]. Considering the two product families in 32 nm, it is concluded that 32 nm technology lowered the chip power at constant frequency by about 50% (Fig. 1). This is in line with what is expected from the classical scaling.

Intel introduced the first Core i7 (the "Tick design) in 22 nm with a 0.9V supply voltage [10] (Ivy Bridge, the third generation Core i7) in 2012. Based on the early parts in 22 nm, it can be deduced that 22 nm offered about 23-27% lower chip power at constant frequency (dashed line in Fig. 2 at high frequencies). Chip application benchmarks of Ivy Bridge showed similar improvements over Sandy Bridge, and the gains in application benchmarks in Ivy Bridge tracked the gains in frequency over Sandy Bridge [11]. The fourth generation Core-i7, the Haswell family (the Tock design) was introduced a year later. Haswell had an evolutionary design and operated at approximately a 5-10% higher

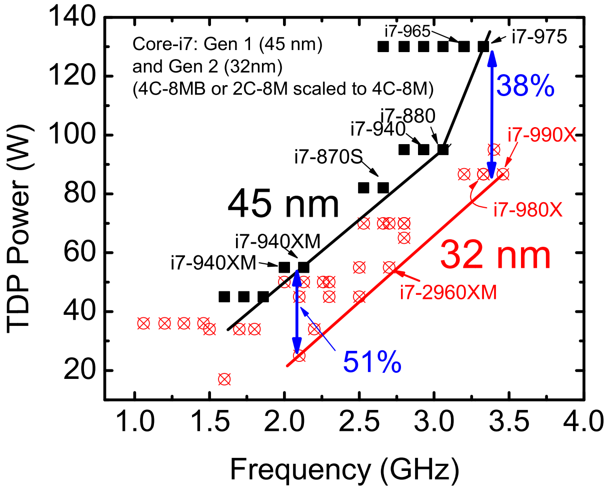


FIGURE 1. Power (TDP_vs. frequency for Intel Core-i7, generation 1 (45 nm) and generation 2 (32 nm).

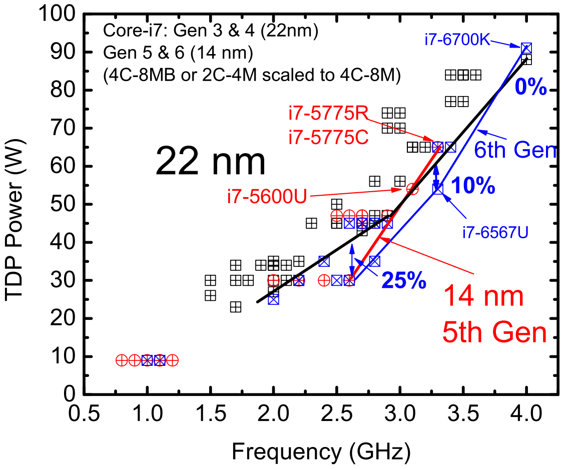


FIGURE 3. Power (TDP) vs. frequency for Intel Core-i7, generation 3 & 4 (22 nm) and generation 5, 6, (early 14 nm).

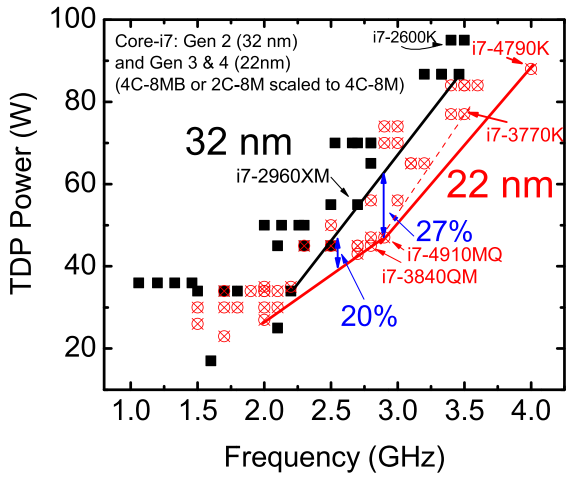


FIGURE 2. Power (TDP) vs. frequency for Intel Core-i7, generation 2 (32 nm) and generation 3 & 4 (22 nm).

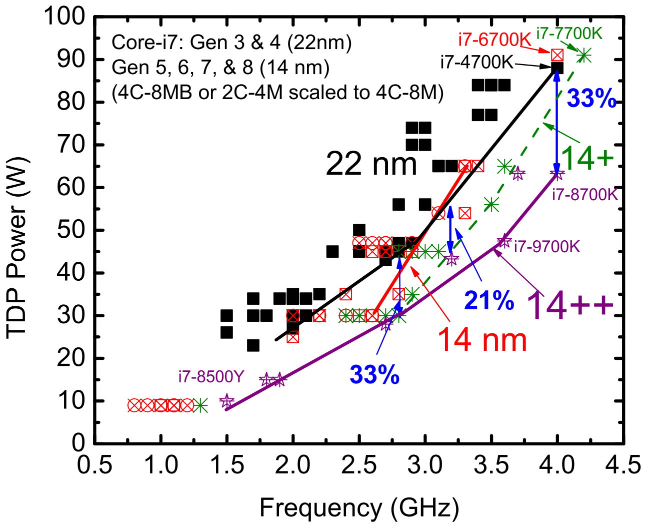


FIGURE 4. Power (TDP) vs. frequency for Intel Core-i7, generation 3 & 4 (22 nm) and generation 5, 6, 7, & 8 (14, 14+ & 14++ nm).

frequency (solid line in Fig. 2) with correspondingly improved application benchmarks as compared to the Ivy Bridge [12]. Fig. 2 provides a comparison of all Core-i7 designs in 32 nm and 22 nm; as seen in the figure, power reduction is in the range of 20-27% over the designs' operating frequency range.

Going to 14 nm (0.8 V nominal supply voltage) [13], for the first generation (the "Tick" design) Core-i7 (Broadwell, 5th generation of Core-i7 introduced in late 2014), targeted mostly the mobile market. The sixth generation Core i7, Skylake, was introduced next (Skylake switched to DDR4 from DDR3, i.e. faster memory access). For the both 5^{th} and 6^{th} generation designs there was little chip power drop when the chips ran in the high frequency range. Over the total frequency range, a chip power drop of 0%-25% (at constant frequency) was observed (Fig 3). For mobile parts there is more drop in power, probably due to better voltage scaling. The low gain in power-frequency at high frequencies was reflected in chip application benchmarks [14].

Intel introduced 14+ and 14++ nm nodes in 2017; these nodes deliver a higher performance technology as compared to 14 nm [15]. The corresponding Core i7 families are Gen 7 (Kaby Lake) and Gen 8 (Coffee Lake). Gen 7 is an optimized version of Skylake (i.e. more of a technology enhancement), and Coffee Lake is even more significant technology enhancement (14++) with higher core count, enabled by lower power per core. For the 14++ node, the chip power scaling improved in the range of 20%-33% (Fig. 4). Fig. 5 is a summary of the percentage drop in chip power at constant frequency at a given node when compared to the previous node, starting from the transition from 250 nm to 180 nm, and covering the more recent transitions (22 nm to 14++). The device spacing (i.e. contacted gate electrode pitch, CPP) in 14++ was increased to 84 nm from 70 in 14 nm.

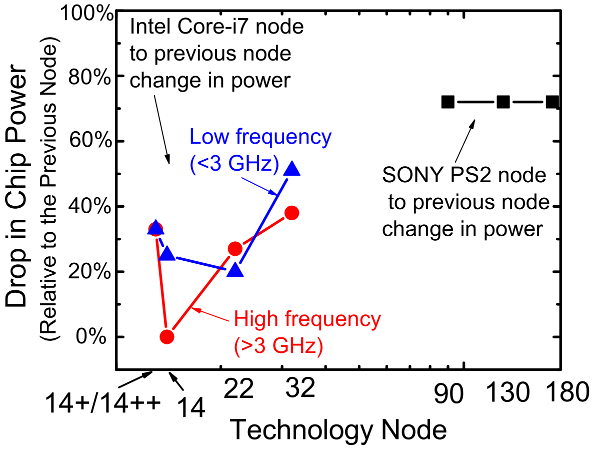


FIGURE 5. Change in chip power at constant frequency, relative to the previous node.

TABLE 3. Fin pitch, height, and contacted gate pitch (CPP) for the technologies used for Intel Core-i7 (or planned use in 10 nm).

Node	Fin Pitch	Fin Height	CPP
22	60	34	90
14	42	42	70
14++	42	42	84
10	34	46	54

Larger device pitch (CPP) helps with lower gate to contact (PC-CA) capacitance and can result in higher strain, but can also lead to larger chip and higher metallization capacitance (i.e. higher C in BEOL). In summary, one can notice that in the last few nodes, despite the drop in nominal voltage, device improvements (I_{on} per device footprint) and full dimensional chip shrinks, the historical chip power (energy-per-operation) scaling has not been achieved. This is especially true for early 14 nm. One part of this may be driven by the fact that capacitance was not scaled as expected: high C may have been caused partly by the increase in fin height from 34 nm in the 22nm node, to 42 nm in the 14nm node (Table 3). Another source of this issue may have been the need to keep the operational voltage higher than its nominal value, especially for high performance microprocessors (i.e. closer to 1.0 V as opposed to 0.7-0.8 V) to achieve better performance. To shed further light on the possible cause of decreased power benefit from scaling, the performance gain per node is evaluated using two available metrics. One metric is the "turbo-mode" frequency, as a measure of the maximum frequency at which a part (or a core) can run (i.e. when only one core is running at high frequency to maximize single thread performance, while the other cores run much slower, subject to thermal constraints). Using similar methodology as described previously (Fig. 6 and Fig. 7) to track the turbo frequency throughout recent nodes (summary in Fig. 8), one notices that the turbo-frequency at the same power (as a measure of performance) takes a step down for designs realized in the early 14 nm node. With the introduction of 14+ and 14++ (and the higher I_{on}), the turbo frequency improves as compared to 22 nm. As another metric, we look at the current per actual device perimeter. For FinFET technologies, it is customary to report the nFET and pFET current per device (finfet) footprint (i.e current per fin pitch), but probably a better metric is to look at the current per actual fin perimeter (i.e. $2{*}\mathrm{Fin}_{\mathrm{Height}} + \mathrm{Fin}_{\mathrm{width}}$) since the device gate C is proportional to the actual perimeter (Fig. 9) [10], [13], [15], [16]. It is noticed that in early 14 nm, the current per fin perimeter drops as compared to 22 nm. The drop in current per actual width in 14 nm may explain some of the poor performance and power-per-operation in 14 nm; if the current drops, then one needs to operate at a higher voltage to get to the same frequency, and thus the fCV^2 will increase. Later with the introduction of 14+ and 14++ nm nodes, the current drive improved, and therefore one could drop the voltage at a given frequency, and that can partially explain the improvement in power-per-operation scaling.

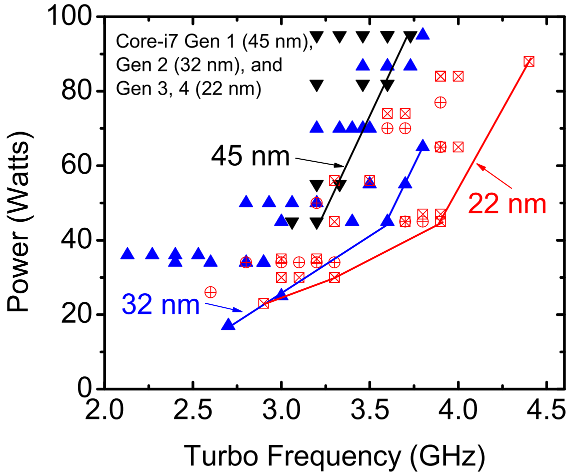


FIGURE 6. Power vs. turbo frequency (as a measure of the best that technology can achieve) for 45, 32, and 22. Notice the continuous improvement in turbo frequency across nodes.

IV. DISCUSSIONS – TOWARD 10 nm

Looking at the high performance microprocessors (Intel Core-i7), that use high performance CMOS, it appears that the benefit of chip power scaling (at the same frequency) has been diminishing, as compared to earlier CMOS nodes. This seems particularly true of the most recent 14 nm generation. The situation was improved by the introduction of enhanced 14 nm nodes (i.e. 14+ and 14++) albeit with increased device pitch. A number of companies have reported results regarding the 10 nm/7 nm nodes. In case of Intel (probably the best reported 10 nm technology so far [15]), looking at the 10 nm data points (nominal supply of 0.7 V) shown in Fig. 8, we observe a degradation in drive current per perimeter in the 10 nm node devices vs. that of 14++ (similar to the first reported 14nm vs. 22 nm) devices, and an increase in fin height (from 42 to 46-50 nm). Poor current drive and

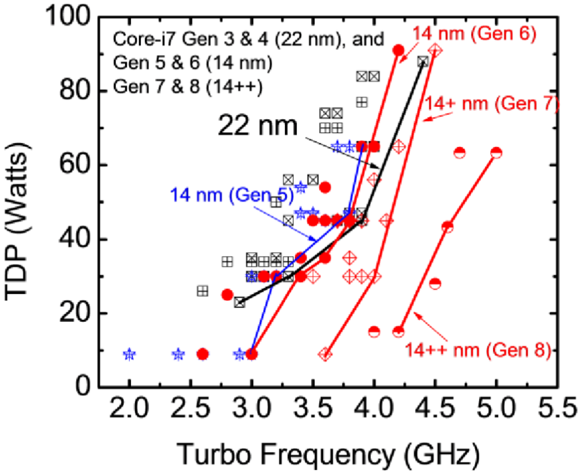


FIGURE 7. Power vs. turbo frequency (as a measure of the best that technology can achieve) for 22, 14, 14+ and 14++ nm nodes. 14 nm resulted in slower frequency compared to 22 nm for high frequency parts. The relative turbo frequency improved with the introduction of 14+ and 14++ nodes.

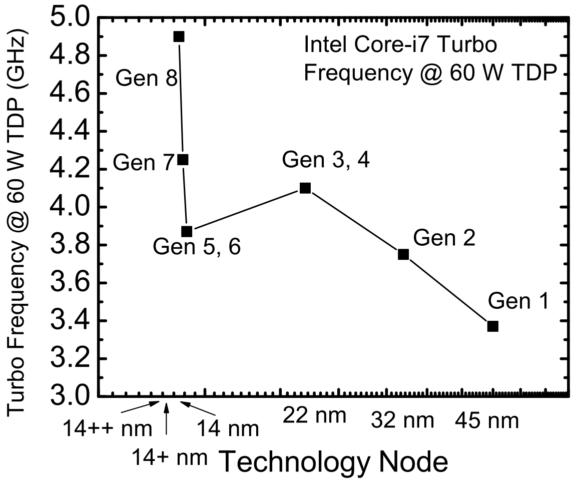


FIGURE 8. Evolution of turbo frequency on Intel Core-i7 from 45 nm through 14++ (at 60 W).

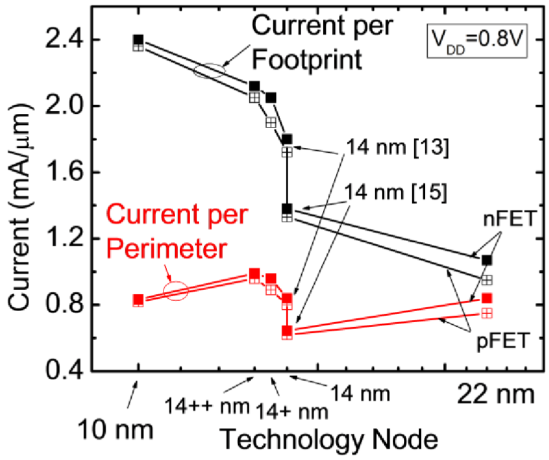


FIGURE 9. Current per device footprint (black line), and current per actual Fin perimeter (red line), at 0.8 V & I_{off} = 10 nA/ïĄm, for Intel FinFET technologies. In both cases, the top line corresponds to nFET and the bottom line to pFET. When the current was only available at 0.7 V (i.e. first report on 14 nm, or 10 nm), device transconductance was used to extrapolate current to 0.8 V.

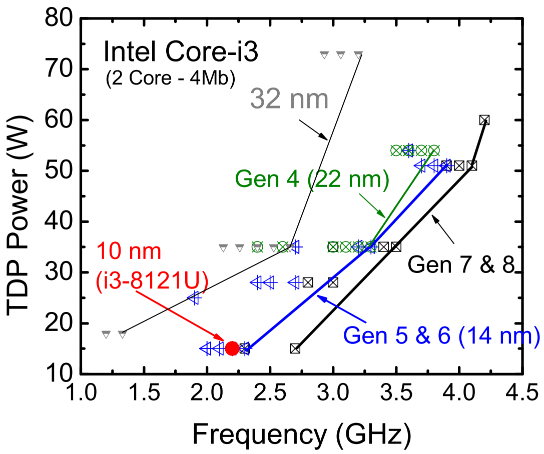


FIGURE 10. Power (TDP) vs. frequency for Intel Core-i3, through 32 nm, 22 nm, 14 nm (Gen 5 &6), 14+/14++ (Gen 7 & 8) and the first 10 nm part.

increase in fin height were some of the probable causes of poor power scaling in early14 nm. Thus it will not be surprising if the first generation of 10 nm parts faces similar chip power (energy-per-operation) scaling challenges as did the first generation of 14nm parts. In fact, in May2018 Intel announced their first 10 nm device, the Core i3-8121U processor [5]. Comparing the 10 nm processor to Core-i3s in earlier generations of technology (Fig. 10), it is observed that in terms of TDP power vs. frequency, Core-i3 behaves very much like Core-i7. In particular, early 14 nm node parts had limited power-frequency benefit with respect to 22 nm, and it was with the introduction of the 14+ and 14++ nodes that the chip power-frequency improved. Looking at the 10 nm part (i3-8121U), its power-frequency behavior is comparable to the early 14 nm Core-i7 parts. Based on this single data point, it seems that early 10 nm is behaving like the early 14 nm generation in terms of total chip power (energy-per-operation) scaling.

The focus of the paper has been chip power scaling trends. Part of the challenge in chip power scaling is likely due to the poor scaling of the device capacitance: Bulk FinFETs have parasitic capacitance caused by the non-conductive bottom of the fin which is heavily doped to stop the punch-through current, and also due to the non-scaling of the fin height. Alternative devices have been proposed that may result in better capacitance scaling from node to node [17]. It is hoped that the industry will be able to come up with a similar improvements in 10 nm as it did in 14 nm, and drive the chip power scaling of the future nodes toward historical values, especially as it relates to high performance microprocessors.

ACKNOWLEDGMENT

The author is grateful to Dimitri Antoniadis and Pouya Hashemi for technical discussions, and to the reviewers of the paper for very constructive feedback.

REFERENCES

[1] G. E. Moore, ''Progress in digital integrated electronics,'' in *IEDM Tech. Dig.*, vol. 21, Dec. 1975, pp. 11–13.

[2] G. Baccarani, M. R. Wordeman, and R. H. Dennard, ''Generalized scaling theory and its application to a 1/4 micrometer MOSFET design,'' *IEEE Trans. Electron Devices*, vol. 31, no. 4, pp. 452–462, Apr. 1984, doi: 10.1109/T-ED.1984.21550.

[3] Ian Cutress. *The Intel 9th Gen Review: Core i9-9900K, Core i7-9700K and Core i5-9600K Tested*. Accessed: Oct. 2018. [Online]. Available: https://www.anandtech.com/show/13400/intel-9th-gen-core-i9-9900k-i7-9700k-i5-9600k-review/21

[4] H. Wong. *A Comparison of Intel's 32nm and 22nm Core i5 CPUs: Power, Voltage, Temperature, and Frequency*. Accessed: Oct. 2012. [Online]. Available: http://blog.stuffedcow.net/2012/10/intel32nm-22nm-core-i5-comparison

[5] *8th Generation Intel Core i7 Processors*. Accessed: Oct. 2018. [Online]. Available: https://ark.intel.com/ and https://ark.intel.com/products/series/122593/8th-Generation-Intel-Core-i7-Processors

[6] (2014). *Intel Tick-Tock Model*. [Online]. Available: https://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html

[7] K. Kutaragi, ''Toward future computer entertainment systems,'' in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2006, pp. 40–56, doi: 10.1109/ISSCC.2006.1696032.

[8] M. Yuffe, E. Knoll, M. Mehalel, J. Shor, and T. Kurts, ''A fully integrated multi-CPU, GPU and memory controller 32nm processor,'' in *IEEE ISSCC Dig. Tech. Papers.*, Feb. 2011, pp. 264–266, doi: 10.1109/ISSCC.2011.5746311.

[9] A. L. Shimpi. *The Sandy Bridge Review: Intel Core i7-2600K, i5-2500K and Core i3-2100 Tested*. Accessed: Jan. 2011. [Online]. Available: https://www.anandtech.com/show/4083/the-sandy-bridge-review-intel-core-i7-2600k-i5-2500k-core-i3-2100-tested

[10] C. Auth *et al.*, ''A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors,'' in *Proc. Symp. VLSI Technol.*, Jun. 2012, pp. 131–132, doi: 10.1109/VLSIT.2012.6242496.

[11] A. L. Shimpi and R. Smith. *The Intel Ivy Bridge (Core i7 3770K) Review*. Accessed: Apr. 2012. [Online]. Available: https://www.anandtech.com/show/5771/the-intel-ivy-bridge-core-i7-3770k-review

[12] A. L. Shimpi. *The Haswell Review: Intel Core i7-4770K & i5-4670K Tested*. Accessed: Jun. 2013. [Online]. Available: https://www.anandtech.com/show/7003/the-haswell-review-intel-core-i74770k-i54560k-tested

[13] S. Natarajan *et al.*, ''A 14nm logic technology featuring 2^{nd}-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size,'' in *IEDM Tech. Dig.*, Dec. 2014, pp. 3.7.1–3.7.3, doi: 10.1109/IEDM.2014.7046976.

[14] Ian Cutress. *The Intel 6th Gen Skylake Review: Core i7-6700K and i5-6600K Tested*. Accessed: Aug. 2015. [Online]. Available: https://www.anandtech.com/show/9483/intel-skylake-review-6700k-6600k-ddr4-ddr3-ipc-6th-generation

[15] R. Brain. *14nm Technology Leadership*. Accessed: Mar. 2017. [Online]. Available: https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Ruth-Brain-2017-Manufacturing.pdf

[16] C. Auth, ''A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, self-aligned quad patterning, contact over active gate and cobalt local interconnects,'' in *IEDM Tech. Dig.*, Dec. 2017, pp. 29.1.1–29.1.4, doi: 10.1109/IEDM.2017.8268472.

[17] R. Muralidhar, R. H. Dennard, T. Ando, I. Lauer, and T. Hook, ''Advanced FDSOI device design: The U-channel device for 7 nm node and beyond,'' *IEEE J. Electron Devices Soc.*, vol. 6, pp. 551–556, Feb. 2018, doi: 10.1109/JEDS.2018.2809587.

GHAVAM G. SHAHIDI received the B.S., M.S., and Ph.D. degrees in electrical engineering from MIT. In 1989, he joined the IBM Thomas J. Watson Research Center, where he initiated the SOI Research Program which resulted in the first use of SOI in mainstream CMOS. In 2000, he led the development of multiple CMOS technology generations as the Director of high-performance logic development at IBM's Semiconductor Research and Development Center. From 2003 to 2015, he was involved in CMOS technology scaling down to single digit nodes as the Director of silicon technology at the IBM Research Division. He is currently with IBM Research. He is an IBM Fellow and a Foreign Member of the National Academy of Engineering.