Journal of
Micro/Nanolithography,
MEMS, and MOEMS

Nanolithography.SPIEDigitalLibrary.org

Retrospective on VLSI value scaling and lithography

Michael L. Rieger

SPIE.

Michael L. Rieger, “Retrospective on VLSI value scaling and lithography,” *J. Micro/Nanolith. MEMS MOEMS* **18**(4), 040902 (2019), doi: 10.1117/1.JMM.18.4.040902.

Retrospective on VLSI value scaling and lithography

Michael L. Rieger*
Consultant, Skamania, Washington, United States

Abstract. In recent decades, the rate of shrinking integrated-circuit components has slowed as challenges accumulate. Yet, in part by virtue of an accelerating rate of cleverness, the end-user value of new semiconductor processes steadily advances. On top of the miniaturization benefits delivered by optical lithography, value is boosted by innovations in wafer processing, mask synthesis, materials and devices, microarchitecture, and circuit design. Focusing on three decades of microprocessor data enables quantification of how innovations from those domains have contributed over time to integrated-circuit "value scaling" in terms of performance, power, and cost. At some point, lateral shrinking will end altogether and the kinds of ingenuity emerging from those domains may provide clues for how very large-scale integration value creation will advance beyond that point.
[DOI: 10.1117/1.JMM.18.4.040902]

Keywords: Moore's law; Dennard scaling; VLSI design and manufacturing; lithography trends.

Paper 19068V received Jul. 31, 2019; accepted for publication Nov. 4, 2019; published online Nov. 26, 2019.

1 Introduction

Reflecting on Moore's law, Moore[1] summed up a remarkable feature of scaling: "by making things smaller, everything gets better simultaneously." Each new generation of silicon provided a profusion of value: chips captured more functionality at higher performance, with lower power per function, and at lower cost per circuit. System reliability improved as well. Although steady progress in lithography-driven miniaturization provided the foundation for this progress, Moore also noted that engineering "cleverness," accounted for a substantial share of those gains. Dennard[2] codified a set of MOSFET scaling rules (Table 1) for achieving certain electrical benefits.

The first three rules are prescriptions for scaling, and the remaining items capture beneficial electronic properties from scaling. Rule (3)—reduce switching voltage in proportion to the geometric shrink factor—is what is widely known as "Dennard scaling" where the main benefits are to reduce power per circuit (rule 7) and to maintain constant power density (rule 8) as circuit area shrinks. Adapting these rules to complementary metal-oxide-semiconductor (CMOS) logic circuits, where average current is proportional to voltage and switching frequency f, power dissipation becomes CV^2f. For example, scaling by $\kappa = 1.4$ (0.7 shrink, k), increasing clock frequency by 40% (matching 0.7 shorter delay, rule 6), and scaling voltage by 0.7, delivers 40% higher performance, at half power per circuit ($CV^2f = 1.4^{-1} \times 1.4^{-2} \times 1.4$). This example, in the form of "value scaling" terms used in this paper, gives 1.4× performance gain, 2× power reduction, and 2× cost decrease per circuit assuming no change in silicon areal manufacturing cost. The 0.7 shrink with Dennard scaling provides an overall 5.6× total value scaling per circuit—performance per Watt, per dollar.

Although the rate of geometric scaling has varied over the years and new process nodes were introduced unevenly over time, the long-term trend of doubling components (transistors) per chip every 2 years has held remarkably consistent. Prior to the 350-nm node, transistor density doubled about every 3 years and component count was boosted with increasing die sizes. For a period from the mid-1990s through the early 2000s, transistor density doubled about every year and a half, and from the 90-nm node to the present logic-transistor density has nearly doubled every 2 years, on average.

The semiconductor fabrication data that follow throughout this paper has been anchored to year of microprocessor chip introduction. For the purpose of normalizing progress-comparisons over different time-frames, the term "generation" will refer to any 2-year time period throughout this paper—which roughly corresponds to the introduction rate for new process nodes. Node names, traditionally anchored to transistor gate lengths, do not accurately capture overall improvement rates over time, and they have become unreliable indicators of process capabilities more recently.

2 Optical Lithography, Pitch, and Density

Geometric scaling is geared to the resolving power of optical printing tools. Figure 1, solid line, plots the progress of optical lithography printer resolution in terms of illumination wavelength divided by numerical aperture λ/NA. The absolute minimum pitch for resolving line patterns is $0.5\,\lambda/\mathrm{NA}$ (dotted line, Fig. 1), and for point sources the pitch limit is $0.61\,\lambda/\mathrm{NA}$, (the Rayleigh criterion). Data points in Fig. 1 plot reported very large-scale integration (VLSI) minimum pattern pitches where minimum pitch, as the term is used in the lithography community, is the minimum spacing period for layout features—the inverse of the number of features per unit length. Component density is driven mainly by minimum pattern pitch. Minimum feature sizes, lines, and spaces, are nominally around half the minimum pitch, but smaller features or spaces can be realized with lithography and process tricks while keeping pitch constrained to the optical limits. Data points for transistor physical gate lengths, which typically are less than the half-pitch length, are also plotted in Fig. 1.

*Address all correspondence to Michael L. Rieger, E-mail: mike@mlrieger.com

Table 1 Dennard's scaling rules (numbering added). Note that Dennard's scaling factor κ (e.g., 1.4) is the inverse of the "shrink" factor more commonly used (e.g., 0.7).

(1)	Device dimension, t_{ox}, L, W	$1/\kappa$
(2)	Doping concentration, N_a	κ
(3)	Voltage, V	$1/\kappa$
(4)	Current, I	$1/\kappa$
(5)	Capacitance, $\varepsilon A/t$	$1/\kappa$
(6)	Delay time/circuit, VC/I	$1/\kappa$
(7)	Power dissipation/circuit, VI	$1/\kappa^2$
(8)	Power density, VI/A	1

Prior to the 1990s, design-pattern pitch remained somewhat larger than the available resolution of lithography tools, and pitch reduction progressed at ∼0.8 shrink every 2 years, roughly tracking the rate of printer-resolution improvements. Starting in the mid-1990s pitch scaling accelerated to ∼0.62 scaling every 2 years until pitch dimensions caught up with tool λ/NA. From 90 nm onward, layout pitches closely tracked λ/NA of optical tools which continued to improve resolution by an average rate of around 0.8 every 2 years. Printer progress for resolution stalled in 2012 after the introduction of the most advanced deep ultraviolet (DUV) tool: argon fluoride (ArF) 193 nm λ laser illumination, 1.35 NA immersion. Printer development for smaller wavelengths (e.g., 157 nm) was abandoned when it became clear that no smaller wavelength with a refractive system could compete with immersion technology at 193 nm. Extreme ultraviolet (EUV) lithography tools (λ 13.5 nm and NA 0.33) use radically different mirror-based optics. Originally forecast for the 65-nm node, EUV is just now ramping up production for the 7- and 5-nm nodes. (Note that at EUV's 2019 introduction into high-volume manufacturing, its λ/NA figure is fairly consistent with the long-term 0.8/2-year printer resolution scaling trend.)

Although optical resolving power remained stalled with ArF immersion, pitch scaling was energized with resolution enhancement techniques (RETs) (orange-shaded region, Fig. 1), and then again by multipatterning processes (pink-shaded region). In general, resolving random, or "free-form," pattern graphics is relatively straightforward where pattern pitches are larger than λ/NA. As feature dimensions approach and then fall below $0.5\,\lambda/\mathrm{NA}$, their imaged shapes become increasingly distorted from optical proximity effects.[5] Optical proximity correction (OPC) is a computation that predicts those distortions and synthesizes photomask layout patterns with "inverse distortions" to counteract these effects. Widespread OPC deployment began around year 2000 with the 150- to 130-nm nodes, when minimum feature sizes dipped below $0.5\,\lambda/\mathrm{NA}$.

When pattern pitches fall below λ/NA, image contrast degrades. In this pitch regime, to effectively capture binary images—that is, to achieve crisp demarcation of features and spaces in the photosensitive resist layer—requires so-called RETs[6] to improve contrast and depth of focus. Some of these methods involve adding special layers on the photomask, phase-shift masks, to control the phase of light rays passing through various features. Other RETs involve tailoring the printer illumination optics, such as in source-mask optimization, to control the diffraction patterns emanating from mask

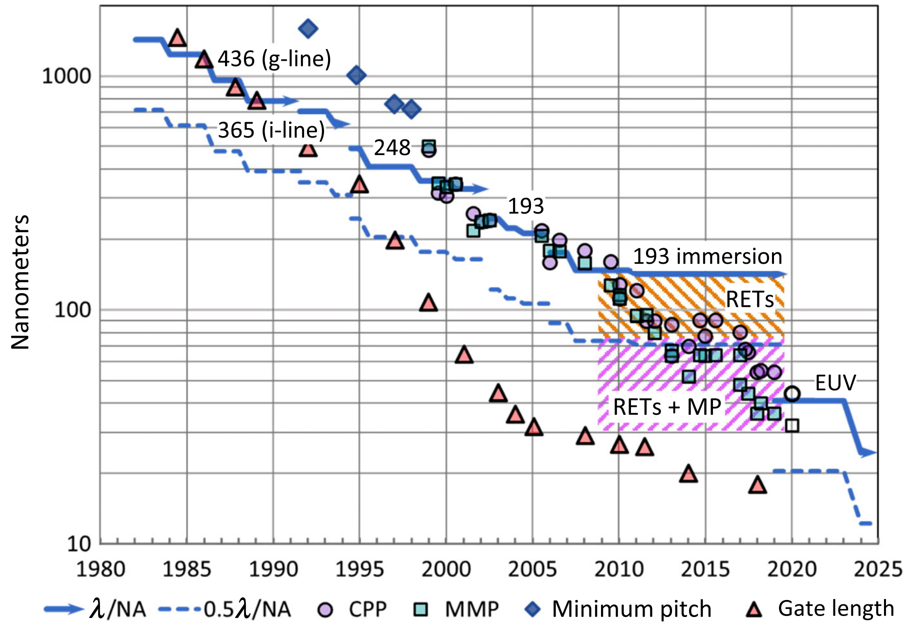

Fig. 1 Optical resolving power and VLSI minimum geometry over time. Lambda λ is the illumination wavelength, and NA is the sine of the lens angular aperture times the index of refraction for the coupling medium. CPP is contacted poly pitch, referring to end-to-end transistor spacing, and MMP is minimum metal pitch. Historical λ and NA values from Matsuyama,[3] most pitch and gate length values from Wikichip.[4]

features which then interfere with each other to define images in resist. Typically, in combination with OPC functions, RETs involve computational lithography software to synthesize a unique mask layout pattern for each design layout pattern. With most RET approaches, certain spatial frequencies are necessarily diminished to improve image contrast, and increasingly strict constraints (design rules) are put on design-layout shapes and configurations to ensure their printability with the remaining spatial frequencies.[7] It is not possible to resolve pitches below $0.5\,\lambda/\mathrm{NA}$ and, as pattern pitches approach that limit, design, and mask layouts are constrained to resemble uniform arrays of features. Over the past decade, RETs have roughly doubled the useful pitch-resolving capabilities of optical tools, as indicated with the orange shaded region in Fig. 1.

A second set of innovations, called multipatterning,[8] enable printing pattern pitches below optical limits. A single layout is decomposed into a set of relaxed-pitch mask patterns, and the original layout image is reconstructed with separate exposures of those masks combined with the aid of etch and deposition processes. One type of multipatterning, called "litho-etch," involves interlacing the design layout features into two or more masks, each with relaxed pitch in their partial patterns. The final silicon structure is built up with a sequence of lithography-then-etch (LE) steps. (For example, a process involving three such steps is denoted LELELE or LE^3.)

Another multipatterning method, self-aligned double patterning (SADP), uses deposition and etch processes to create line features with a spacer material along the perimeters of resist features. Applying SADP to a resist pattern of parallel lines, for example, produces line patterns at twice the line density. The rendered line width of the spacer everywhere is fixed to a value defined by deposition and etch processes. Functional features are realized with a subsequent mask exposure and a process that trims away unwanted lines, or that blocks unwanted spaces. SADP can be applied sequentially. Applying SADP on top of a previous SADP treatment on parallel lines quadruples rendered line density, reducing pitch by a factor of 4. The latter is called self-aligned quadruple patterning, and there are self-aligned methods that produce other pitch division factors as well.

RETs in concert with multipatterning have more than quadrupled pattern pitch capabilities to date (orange and pink shaded regions, Fig. 1) and they have advanced pitch scaling at an average rate of 0.8 per 2-years over the past decade. That rate is consistent to the 0.65 scaling rate of carefully optimized static random-access memory (SRAM) bit cell footprint areas (Fig. 2), where $\sqrt{0.65}$ corresponds to 0.8 average pitch reduction. A shrink factor of 0.8 corresponds to a density increase of 56%, yet logic-transistor density continued to nearly double every 2 years (Fig. 3, solid curve), recently exceeding 100 million transistors per mm^2. The dashed curve (Fig. 3) shows how pitch scaling, including that from RETs and multipatterning, alone would have driven transistor density.

The gap between pitch scaling and more rapid density growth, shaded area in Fig. 3, captures the effects of equivalent scaling[9] innovations—also termed hyperscaling,[10] and scaling boosters.[11] These innovations to date have advanced transistor densities nearly threefold on top of pitch-reduction progress. Significant gains came from introducing the finFET transistor, where transistor channel width is flipped to the vertical dimension, thus shrinking the space required to achieve drive current. Additional gains were from layout-packing efficiencies driven by other process innovations such as strained silicon to increase current density and thereby reducing gate widths and by stacking contacts on top of gates instead of alongside them.

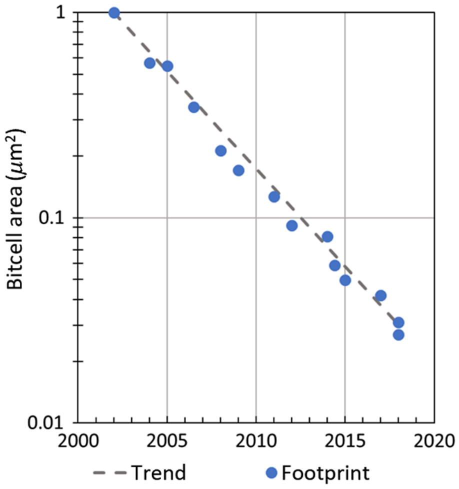

Fig. 2 SRAM bit cell area.[4] Trend is an average 0.65 shrink of bit cell area every 2 years.

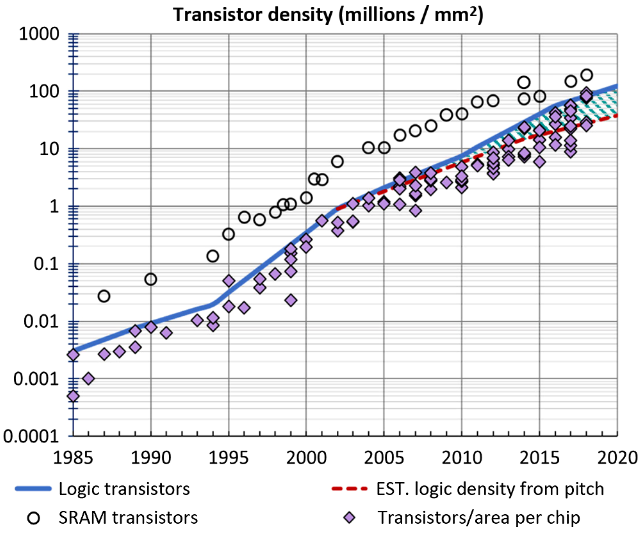

Fig. 3 Density of logic transistors (solid line) has advanced on average by 2× per generation. Dotted line is estimated density from pitch scaling alone. SRAM transistor densities, derived from bit cell area,[4] assumes 6-transistors per cell. Transistor density data points were obtained by dividing reported total transistors by reported die area[12] per chip.

3 Performance

Geometric scaling alone improves performance by shrinking capacitive circuit load. Smaller transistors have lower gate capacitance. Regardless of shrink factor, capacitance for dense VLSI interconnection wires is roughly constant per unit length[13]—$0.2\ fF/\mu\mathrm{m}$—and those loads are scaled down by shorter distances between connections for shrunken circuits. Scaling MOSFETS uniformly in gate-length and width does not change transistor drive current, which is proportional to channel width divided by channel length. Logic switching delay, VC/I (Dennard rule 6), is thus shortened by the shrink factor. Dennard voltage-scaling minimally affects switching delay because transistor drive current falls proportionally to voltage. The time needed to charge a load to a lower voltage with proportionally less current remains about the same. In short, geometric scaling drives shorter switching delays by reducing load capacitance, which improves circuit performance proportionally to the shrink factor *k*.

In the late 1990s, when pitch scaled at 0.62 per 2-year cycle, gate lengths were being shortened (Fig. 1) at an even faster clip of 0.54—which increased transistor drive current by about 15% per generation. This combined with the 0.62 reduction of load capacitance from pitch scaling nearly halved switching delays, which enabled 2-year clock frequency increases of 1.8× per generation (Fig. 4). In this time-frame microprocessor, single-thread performance, as measured by SPECint2006®, gained 2.3× per generation (Fig. 5). SPEC CPU® 2006 provides suites of benchmark workloads for measuring computer performance.[15] The performance score is the ratio of a reference completion time to the completion time of the target CPU. Though retired in 2017 and replaced with SPEC CPU 2017®, a substantial time span of comparative data is available[14] based on SPECint2006.

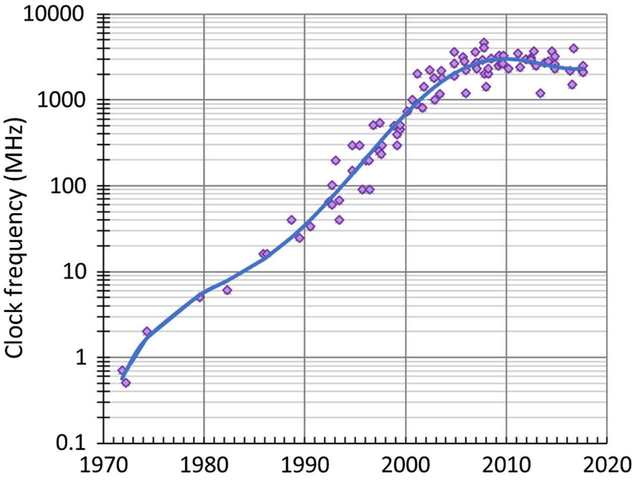

Fig. 4 Microprocessor clock frequency data[14] and trend.

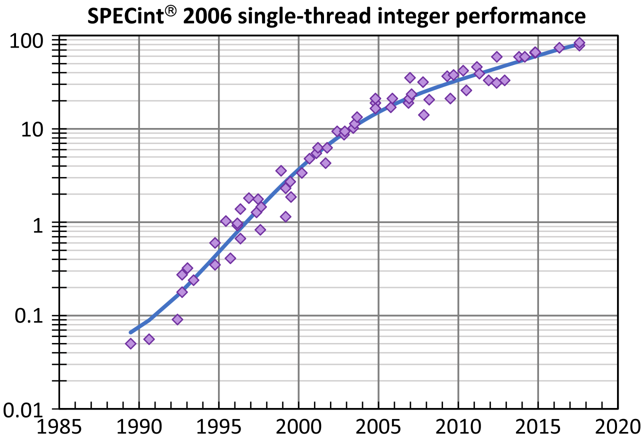

Fig. 5 Single-threaded performance. Data compiled by Rupp.[14]

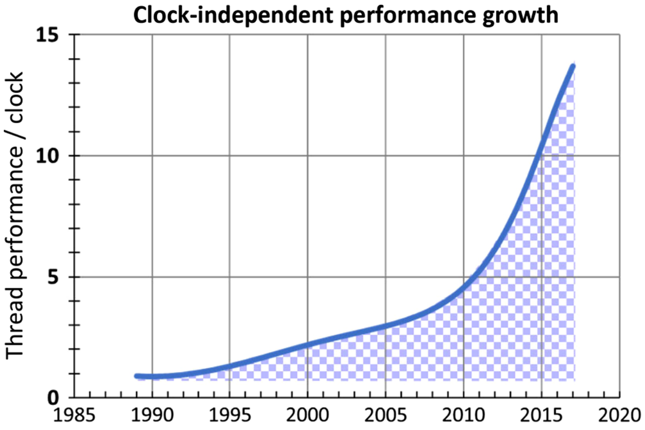

Fig. 6 Thread performance per clock cycle.

Increased performance per clock cycle, plotted in Fig. 6, is from performance-enhancing innovations in design.[16] The trend of increasing processor clock frequency stopped in the late 2000s primarily to dampen escalating power density (Sec. 4). Another factor limiting clock frequency was that the RC time constant (delay) of interconnect did not shorten with shrink because resistance increases with thinner wires.

With stalled clock frequency, performance gains from architecture innovation accelerated to a rate of 1.4× per generation (Fig. 6). Those innovations involved extensive architecture and design cleverness,[16] including pipelining, branch prediction, out-of-order execution, more cache memory, and hierarchic cache architectures to keep fast memory more localized to computation. A downside of this type of performance acceleration is its increased complexity and thus its need for more area. Fred Pollack, Intel, observed that the performance gain from this complexity is roughly proportional to the square root of the increase in logic area (Pollack's rule).[17] For example, a 40% increase in the performance from architectural innovations may cost twice the area.

By the late-2000s, chipmakers revamped their architectural approaches for performance acceleration and began integrating many CPU cores into each processor chip. For parallelizable computing tasks, multicore architectures accelerate cycle-time performance. Overall throughput for nonparallelized, concurrent computations is boosted as well. With all cores working full-speed area roughly scales with throughput gain. A benefit more compelling than driving raw throughput performance is that multicore architectures can be leveraged for improved performance per Watt, as discussed in Sec. 5.

4 Energy and Power

Energy is dissipated in CMOS circuitry by dynamic and static mechanisms. Most dynamic energy is spent charging and discharging load capacitance when circuits are switching states from low to high voltage, and vice versa. (There is also a small amount of dynamic energy wasted through crowbar leakage current when coupled pMOS and nMOS transistors briefly conduct concurrently during state transitioning.) Leakage paths waste energy continuously. A formerly grim leakage path, gate leakage coming from electron tunneling between transistor gate and channel, accelerates exponentially with thinner gate-oxide films. The magnitude of this effect has been quelled for now with high-κ dielectric materials, which allow thicker gate insulator films. A significant remaining leakage path is subthreshold leakage which, because of inherently imperfect switching properties of MOSFETs, worsens exponentially with (threshold) voltage scaling.

The energy dissipated in switching a logic state is ½CV^2 J, where C is load capacitance and V is the switching, or "swing," voltage. Figure 7, solid line, plots the energy dissipated in switching a four-fan-out (FO4) inverter between logic states as scaling and technology evolved over time. The curve was constructed by taking relative CV^2 trends[18,19] and anchoring them to the reported energy[20] for a 65-nm inverter. The dashed line is an estimate of the contribution of geometric pitch scaling by lowering capacitance proportionally to the shrink factor k (Dennard rule 5). Materials and process (low κ dielectrics) have reduced capacitance further, but most of the nonshrink related energy reduction—Fig. 7, shaded gap between dotted and solid lines—has come from lowering swing voltages. Dennard scaling was most closely followed in the period from the early 1990s down to the point when rising subthreshold leakage currents became unacceptable in the early 2000s. Voltage scaling all but stopped until the introduction of the finFET transistor in 2012. The finFET, or double-gate[21] transistor is a better switch than planar FETs and, with its tightened subthreshold leakage, it enabled another incremental lowering of swing voltage.

From 1990 to present, energy per circuit element, per state-change has fallen by 2000×. A little more than half that progress (50×) comes from reducing capacitance by miniaturizing dimensions, driven by lithographic pitch scaling. Materials innovation provided additional capacitance reduction. Dennard scaling accounted for slightly more than 25× as voltages fell from the pre-1990 standard 5 V to just under 1 V today. Making tiny circuits operate effectively at higher speeds and with decreasing voltages was facilitated by key innovations[22] in process and device technologies, such as copper interconnect (1997), strained silicon (introduced in the mid-2000s to increase transistor drive current after gate length scaling began to stall), high κ metal gate (2007), low κ insulators for interconnect, and finFET (2012), to name a few prominent examples.

Dynamic power dissipated per circuit element is usually expressed[13] as $W = \alpha CV^2 f$, where f is the clocking frequency (Hz) C is circuit load, V is a swing voltage, and α is the activity factor—the fraction of the clocking frequency, in which circuit elements switch on average. Dynamic power density is expressed as $\mathrm{W/cm^2} = \alpha CV^2 f \times (\mathrm{elements/cm^2})$. The solid curve (Fig. 8) estimates the power density trend by multiplying together the CV^2 trend (Fig. 7), the clock frequency trend (Fig. 4), the logic transistor density trend (Fig. 3, converted to transistors/cm^2), and a correction factor $\kappa_c = 0.17$ to visually align the curve to pre-65-nm data points. The curve projects the dynamic power dissipated if circuits from 65 nm designs—the point at which the CV^2 trend is anchored—were scaled only, with no other modifications. Accounting for leakage power, without taking design interventions into account, would steepen the post-65-nm total power above the dynamic power trend projection shown in Fig. 8. Empirical data points in Fig. 8 were calculated from reported processor chip peak power divided by die area.[12]

Within the decade of (near) Dennard scaling, power-density growth was tempered but not flattened (as predicted by Dennard rule 8). Until the mid-2000s power densities increased rapidly, eventually peaking at ~100 W/cm^2 at the 65-nm node. (For comparison, a household incandescent light bulb dissipates about 70 W/cm^2 over its filament surface area[24].) Beyond 100 W/cm^2 it is very difficult to remove enough heat from the die to prevent overheating and at the time there was widespread concern that escalating power would put an end to VLSI scaling. Yet accelerated innovation in design and architecture saved the day, as indicated by the expanding gap between predicted power and the trend of actual power data—shaded region, Fig. 8. Density and performance continued to advance after 65 nm, and though $CV^2 f \times$ density continued to rise, the actual power density trend for high-performance chips flattened, with

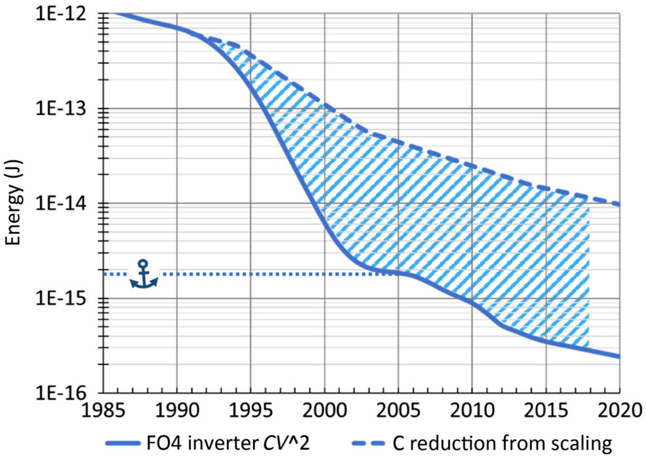

Fig. 7 Solid curve plots energy to switch an FO4 inverter in circuit. Dashed line is the estimated contribution from geometry to lowering capacitance. The relative trend of CV^2 is from Bohr[18] and ITRS[19] and anchored to an inverter energy value 1.72×10^{-15} for 65 nm technology from Stillmaker.[20]

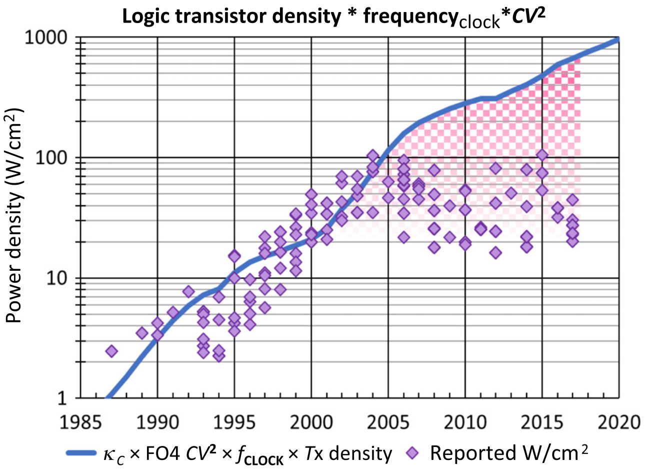

Fig. 8 Estimated trend for dynamic power density calculated from the product of logic transistor density, clock frequency, inverter CV^2 trends, and a correction coefficient κ_c of 0.17. Data points were calculated from reported peak power per chip, divided by die area.[12]

most chips remaining well below 100 $\mathrm{W/cm^2}$ ever since (Fig. 8).

5 Performance and Power

Chip designers have developed a vast number of power-saving optimizations and algorithms, from circuit design to system architecture.[25] A key principle leveraged by many of those methods is that dynamic power rises or falls by voltage squared while transistor delay time scales more linearly with voltage. This affords a design trade-off where small swing-voltage adjustments can achieve significant power savings at the cost of modest performance loss. For example, compared to a single CPU core, two low-voltage cores operating in parallel at half the speed can deliver the same throughput performance, but with substantially lower combined power dissipation. Note for this example, area is being used to achieve power savings. Subthreshold leakage power similarly can be managed in design with trade-offs between leakage and performance by determining swing voltages and by choosing transistor threshold voltages from a selection of V_{TH} options provided for transistors in advanced processes. One generally applicable design guideline is to target an optimum ratio around 2:1 for dynamic to static power.[25]

Considerable system power reduction is from dynamic power-management techniques involving real-time adjustments to regional switching frequencies and voltage. Functional blocks can be slowed down or turned off as needed to prevent chip overheating. A forcibly shut-down block is called dark silicon, and slowed area is sometimes called dim silicon. Dark and dim silicon impacts overall chip performance and thereby can raise system costs, depending on application. User-impact is significant if these slowdowns occur frequently in applications with high average CPU utilization, for example in server farms where computers are working at full speed continuously. On the other hand, these slowdowns rarely occur—or are rarely noticed by the user when they do occur—in general-purpose personal computing or in any application with low-average CPU utilization.

A flourishing architectural approach, heterogeneous multiprocessing, involves augmenting general-purpose, sequential-instruction (von Neumann) processing with specialized hardware processors.[27] Hardware processors dedicated to specific types of tasks can improve performance per watt by 10 to 100× or more compared to conventional, serial processing.[28,29] A prominent example of such a specialized processor is the graphics processor unit (GPU) originally tailored to render 3-D graphics for real-time animation. A main feature of a GPU is its array of thousands of compact arithmetic engines to support massively parallel computations. Applications for GPUs have expanded from display processing to other applications involving large-vector math, including signal and image processing, and neural network inference and training. GPUs are now integrated within general-purpose processor chips, and support for GPUs in operating systems has become mainstream.

6 Cost in Brief

From the beginning, steadily increasing VLSI processing complexity—more layers and more processing steps per layer—increased effort and complexity for subsequent generations. Improving productivity delivered by advancing manufacturing equipment, including periodic transitions to larger silicon wafers, largely tempered cost increases.[30] Gradually improving wafer yields nearly canceled remaining cost rises, and areal manufacturing cost for yielded chips rose very slowly over the long term.

With chip yields plateauing at acceptable levels and with no adoption to larger wafers (450 nm), areal costs for 300 nm wafer processes have steadily risen since the 130-nm node. Information extracted graphically from an Intel presentation[31] shows cost per $\mathrm{mm^2}$ increasing first at about 15% per generation after 130 nm, then accelerating to a 30% to 35% increase/generation after the 22-nm node. That inflection likely captures the cost impact of process complexity for multipatterning as additional layers are subjected to it; translating to a 15% additional cost per generation, approximately.

Escalating nonrecurring engineering costs impact VLSI chip cost and value, the amortized impact of which depends on production volumes. For example, VLSI design costs for microprocessors and large systems on chip from the 28- to the 10-nm node have been rising at a rate between 35% to 50% per generation,[32,33] with 10-nm design costs estimated to be in the range \$100 to \$300 M. Calculating design cost per transistor gives design-productivity improvement rates between 33% to 50% per generation (assuming a doubling number of devices per design generation). A similar analysis for photomask-set costs[30,35] (prices) from 130 nm (\$450 to \$700 K) to 28 nm (\$2 to \$3 M) reveals a 33% generational decrease in mask-set costs per transistor—despite increasing mask complexity from OPC and RETs. With multipatterning, mask-set costs have accelerated and recent price estimates[30] suggest those costs have been increasing faster than component density.

7 Summary

The value-generation aspects of scaling are summarized below for two distinct time periods to compare recent value drivers to those of the past. The approach is to measure rates of improvement in three axes—performance, power, and cost per circuit (area)—and attribute those improvements to innovations from particular technology domains. Table 2 rolls up the generational values from scaling in the 1995 to 2000 time frame, representing a 5-year snapshot within the decade or so when Dennard scaling was in full swing. Table 3 summarizes value generation for the more recent 2010 to 2017 time frame. Methodology for obtaining scaling figures was to take the net change from each value contributor within those periods and translate those ratios to compounding 2-year values.

Scaling values in these tables were derived from trend lines presented in this report, as indicated in the notes. The addition of the term for clock frequency is for translating energy (CV^2) values in the preceding rows to power in the subtotaled product. The clock also constrains maximum power as, for well-designed logic, no transistor will switch more than once per clock cycle. The area-penalty worst-case bound in Table 2 is estimated from Pollack's rule, where the 23% performance increase from architecture translates to about 50% extra area. Over the last decade designers have pulled back from single-thread architectural complexity[16] with more performance and power leverage coming from multicore architecture. Performance and power gains may not be achieved at the same time everywhere, and Pollack's

Table 2 Two-year average value growth by technology contribution for the period 1995 to 2000. Cost and power entries are inverted to make figures for positive benefit >1. Net benefits per circuit are computed by multiplying factors together.

Value contribution	Performance	1/power	1/cost
Lateral scaling			
Lithography pitch	1.61[a]	1.61[a]	2.60[b]
Accelerated gate shrink	1.15[c]		
Process, device, and materials			
Dennard scaling		2.25[d]	
Areal cost			0.87[e]
System and circuit architecture	1.23[f]		0.7[g] to 1.0
Impact of clock f on power		0.55[h]	
Subtotal $\prod$	2.28	1.99	1.58 to 2.26

[a] Impact of pitch reduction on capacitance: shorter switching delay, lower energy dissipated per state transition (Fig. 7).
[b] Density doubling every 17 months (Fig. 3).
[c] Increased transistor drive current from accelerated gate length shortening (Fig. 1).
[d] CV^2 energy reduction not provided by geometric scaling (Fig. 7).
[e] Assumes a 15% generation increase in areal processing costs, not accounting for wafer size differences.
[f] Increased single-thread performance relative to clock frequency (Fig. 6).
[g] Accounts for a plausible range of total die area penalties for architectural performance enhancements (Pollack's rule).
[h] Converting energy to power from 1.8× clock frequency increase per generation (Fig. 4).

rule again is used in Table 3 to estimate the area impact of duplicating processors and to account for dark and dim silicon.

The combined improvement gains in performance, power reduction, and cost reduction from the Dennard scaling era, Table 1, indicate net value scaling—performance per Watt, per dollar—between 7.2 and 10.3× per circuit every 2 years. The same calculation for more recent progress in Table 3 gives a 2.4 to 4.7× value increase every 2 years. The data on which those figures are based are for general-purpose microprocessors and they do not capture the benefits certain applications have enjoyed with specialized processors, such as GPUs. For those applications, architecture-driven power and performance figures are likely far better.

Taking the best-cost scenarios in Tables 2 and 3, the relative contributions by technology domain to VLSI value-growth are summarized in their respective time frames in Fig. 9. A significantly growing share of value scaling in recent times is from innovation in circuit design and system architecture and also from wafer processing and devices. The diminished contribution from lithography in the latter time frame is mainly from the difference in shrink rates, from 0.62 to 0.8 (per 2-years), and it is also driven down by the cost penalty for multipatterning on increasing numbers of layers.

Table 3 Two-year average value growth by technology contribution for the period 2010 to 2017.

Value contribution	Performance	1/power	1/cost
Lateral scaling			
Optical pitch	1	1	1
RET and multipatterning	1.25[a]	1.25[a]	1.56[b]
Multipatterning excess cost			0.87[c]
Process, device, and materials			
Dennard scaling		1.11[d]	
Hyperscaling (density)			1.21[e]
Base areal cost			0.87[f]
System and circuit architecture	1.40[g]	1.27[h]	0.5[i] to 1.0
Impact of clock f on power		1.07[j]	
Subtotal $\prod$	1.75	1.89	0.71 to 1.43

[a] Impact of pitch reduction on capacitance: shorter switching delay, lower energy dissipated per state transition.
[b] Density from lithography pitch reduction.
[c] Excess cost, 15%/generation, for increasing use of multipatterning.
[d] CV^2 energy reduction not accounted in geometric scaling (Fig. 7).
[e] Increased density from finFET and other process-driven compaction (hyperscaling).
[f] Assumes a 15% generational increase in areal processing costs (300-mm wafer), excluding multipatterning cost.
[g] Single thread performance gain over clock frequency (Fig. 6).
[h] Architecturally driven reduction of actual power from projected fCV^2 trend (Fig. 8).
[i] Application-specific penalty range for excess area from architecture and for temporally unusable area from "dark" or "dim" silicon states.
[j] Accounting as a benefit the pull-back of average clock frequencies (Fig. 4).

7.1 *Looking Ahead*

Dimensional scaling remains a powerful value multiplier. A shrink factor of k shortens CMOS switching delays by k, reduces energy by k, and potentially cuts cost up to k^2—which altogether approaches a k^{-4} compounded benefit. This drives robust efforts and investment in furthering lithographic shrink, such as with EUV technology. EUV high-volume deployment is just beginning and it is too early to

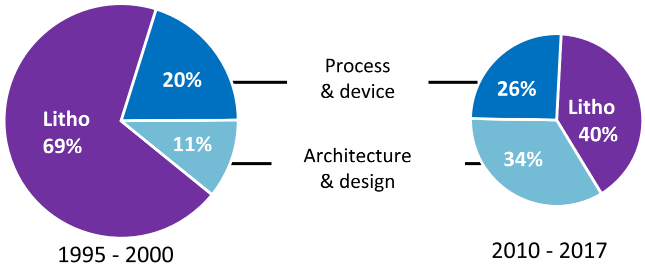

Fig. 9 Relative contributions to value-scaling rates over different time frames. "Litho" is pitch-driven lateral scaling, and it includes the cost of DUV multipatterning.

determine how it will affect lithography costs down the road, and how it might impact future scaling rates. Lateral shrinking may be slowed or stopped by other limits before lithographic capabilities are exhausted.[36]

Regardless whether pitch scaling ends or not, there remains another important value scaler for lithography and process innovation: reducing component variation. For conventional logic circuits removing variation tightens design margins required to account for worst cases, which translates to performance, power, (and yield) gains. To support analog circuits, lower variation amplifies value by improving circuit accuracy, precision, and signal-to-noise ratios. Value for logic circuits diminish as variation approaches zero, but value for analog circuits increases with each fractional reduction of variation.

Many innovations in process and device technologies are on deck.[37,38] Improved MOSFET transistor architectures are emerging, such as nanowire[39] to boost switching performance beyond that of finFET for power and performance, and nanosheet[40] to provide planar-like design flexibility and to improve upon (slightly) the switching characteristics of finFET. New types of integrated memory technologies such as memristors and spintronic-based magnetic RAM are emerging.[37] There is headroom for hyperscaling (scaling boosters) to further leverage the vertical dimension for increased density and performance including, for example, vertical and stacked transistors,[41] and backside power delivery.[42] Growing innovation in 3-D packaging technology[43] is compounding density. Die can be stacked and connected with through silicon vias to improve interconnection bandwidth[44] and to lower data-transfer energy, and it allows heterogenous mixes of die made with alternative process technologies.[45]

As density advances, power minimization remains a prime objective for architecture and circuit innovation. We should expect to see a growing number of integrated heterogenous processors dedicated to more specialized applications. There is burgeoning interest in new computation paradigms involving analog circuit and device properties to deliver orders of magnitude enhanced performance and power over conventional logic. These include neuromorphic computing,[46] quantum computing, and other innovations that leverage nonbinary electronic properties, such as memristor-based array multiplication.[47] Deployment of power-saving circuit methods such as adiabatic switching[48] or resonant energy recovery[49] may increase.

8 Conclusion

Recent rates of value scaling are half that for the period of Dennard scaling but, if you take into account the entire 50-year history of VLSI scaling, today's progress does not look bad at all. Prior to the early 1990s, when transistor density doubled about every 3 years (a shrink factor of 0.79 per 2 years—about the same as today), and before Dennard scaling, 2-year value scaling in terms of performance per Watt, per dollar was only 2.9×—from 40% performance gain, 21% power decrease, and 38% cost savings. But this is a somewhat incomplete comparison as it ignores tremendous integration value gained by eliminating large numbers of bulky, expensive, discrete components in those early years. Still, it may be that the decade or so of Dennard scaling was an anomaly in the big picture, and things are now settling back down to "normal." The main difference between now and then is an expanding proportion of value growth that is coming from architecture and circuit design and from process and device technologies.

Acknowledgments

Thanks to my colleague in retirement, Yan Borodovsky, for many helpful insights, critiques, and for suggesting reference materials. Thanks also to Jamil Kawa, Synopsys, for his review of this paper from a VLSI design perspective.

References

1. G. E. Moore, "Lithography and the future of Moore's law," *Proc. SPIE* **2440**, 2–17 (1995).
2. R. H. Dennard et al., "Design of ion-implanted MOSFETs with very small physical dimensions," *IEEE J. Solid-State Circuits* **9**, 256–268 (1974).
3. T. Matsuyama, "The lithographic lens: its history and evolution," *Proc. SPIE* **6154**, 615403 (2006).
4. "Wiki semiconductor and computer engineering," https://en.wikichip.org/wiki/WikiChip (2019).
5. C. Mack, *Fundamental Principles of Optical Lithography: The Science of Microfabrication*, p. 419, John Wiley & Sons Ltd., London (2007).
6. A. Wong, *Resolution Enhancement Techniques in Optical Lithography*, V. Tt 47 (Book 47), SPIE Tutorial Texts in Optical Engineering, SPIE Press, Bellingham, Washington (2001).
7. M. Rieger, "Communication theory in optical lithography," *J. Micro/Nanolithogr. MEMS MOEMS* **11**(1), 013003 (2012).
8. "Multiple patterning," *Wikipedia*, https://en.wikipedia.org/wiki/Multiple_patterning (2019).
9. ITRS, "2012 Update overview," https://www.dropbox.com/sh/49tu7ip2lsf4922/AAALux-uU0oD70A6-1QTV46za/2012Chapters?dl=0&preview=2012Overview.pdf&subfolder_nav_tracking=1.
10. K. Mistry, "10-nm hyper scaling," Intel, 2017, https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/03/Kaizad-Mistry-2017-Manufacturing.pdf.
11. S. Jones, "IMEC technology forum 2018– the future of scaling," https://semiwiki.com/semiconductor-services/semiconductor-advisors/7543-imec-technology-forum-2018-the-future-of-scaling/ (2018).
12. "Transistor count," *Wikipedia*, https://en.wikipedia.org/wiki/Transistor_count (2019).
13. N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, 4th ed., p. 184, 217, Addison-Wesley, Boston, Massachusetts (2011).
14. K. Rupp, "42 years of microprocessor trend data," https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/ (accessed 2019).
15. Standard Performance Evaluation Corporation, https://www.spec.org/cpu2006/ (2019).
16. S. Borkar and A. Chien, "The future of microprocessors," *Commun. ACM* **54**(5), 67–77 (2011).
17. F. Pollack, https://en.wikipedia.org/wiki/Pollack's_rule (2019).
18. M. Bohr, "Silicon technology leadership for the mobility era," *Intel Developer Forum 2012*, Slide 9, 2012, https://www.intel.com/content/dam/www/public/us/en/documents/presentation/silicon-technology-leadership-presentation.pdf.
19. ITRS, "Table PIDS2 high-performance (HP) logic technology requirements," 2012, https://www.dropbox.com/sh/49tu7ip2lsf4922/AABhXdcbtynWLj5FHukRDqlha/2012Tables?dl=0&preview=PIDS_2012Tables.xlsx&subfolder_nav_tracking=1.
20. A. Stillmaker and B. Baas, "Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm," *Integration* **58**, 74–81 (2017).
21. D. Hisamoto et al., "FinFET—a self-aligned double-gate MOSFET scalable to 20 nm," *IEEE Trans. Electron Devices* **47**(12), 2320–2325 (2000).
22. M. Bohr, "Moore's law in the innovation era," *Proc. SPIE* **7974**, 797402 (2011).
24. "For a 60-watt 120-volt lamp, the uncoiled length of the tungsten filament is usually 22.8 inches (580 mm), and the filament diameter is 0.0018 inches (0.046 mm)," *Wikipedia*, https://en.wikipedia.org/wiki/Incandescent_light_bulb (2019).
25. J. Rabaey, *Low Power Design Essentials*, p. 102, Springer Science+Business Media, LLC, New York (2009).
27. S. Mittal, "A survey of techniques for architecting and managing asymmetric multicore processors," *ACM Comput. Surv.* **48**(3), 1–38 (2016).
28. N. Goulding-Hotta, "GreenDroid: an architecture for the Dark Silicon Age," in *17th Asia and South Pac. Des. Autom. Conf.* (2012).
29. L. Gwennap, [as reported in], "Chip designers are sidestepping Moore's law slowdown with specialized processors," https://venturebeat.com/2016/09/27/as-moores-law-slows-chip-designers-focus-on-specialized-processors/ (2016).

30. K. Flamm, “Measuring Moore’s law: evidence from price, cost, and quality indexes,” Tech. Rep., p. 4, 36, National Bureau of Economic Research (2018).
31. W. Holt, [as reported in], “Intel sees path to extend Moore’s law to 7nm,” 2014, https://forwardthinking.pcmag.com/none/329835-intelsees-path-to-extend-moore-s-law-to-7nm.
32. H. Jones, International business strategies, [as reported in], “Big trouble at 3nm,” https://semiengineering.com/big-trouble-at-3nm/ (2018).
33. H. Jones, International business strategies, [as reported in], “How 5G differs from previous network technologies,” https://semiengineering.com/how-5g-differs-from-previous-network-technologies/ (2018).
34. B. Albing, “NRE Costs & Analog Integration,” https://www.planetanalog.com/nre-costs-analog-integration/ (2013).
35. C. Mack, “Fifty years of Moore’s law,” *IEEE Trans. Semicond. Manuf.* **24**(2), 202–207 (2011).
36. J. Kawa, “Beyond CMOS,” in *ACM Int. Workshop Timing Issues Specif. and Synth. Digital Syst.*, 2019, https://www.tauworkshop.com/2019/slides/Beyond%20CMOS%20%20TAU%20rev%202.0.pdf.
37. International Roadmap for Devices and Systems (IDRS), “Beyond CMOS,” IEEE, 2017, https://irds.ieee.org/editions/2017/beyond-cmos.
38. A. Hellemans, “Nanowire transistors could keep Moore’s law alive,” *IEEE Spectrum*, 2013, https://spectrum.ieee.org/semiconductors/devices/nanowire-transistors-could-keep-moores-law-alive.
39. K. Bourzac, “Nanosheets: IBM’s path to 5-nanometer transistors,” *IEEE Spectrum*, 2017, https://spectrum.ieee.org/nanoclast/semiconductors/devices/nanosheets-ibms-path-to-5nanometer-transistors.
40. J. Ryckaert et al., “The complementary FET (CFET) for CMOS scaling beyond N3,” in *2018 IEEE Symp. VLSI Technol.* (2018).
41. S. Moore, “Buried power lines make memory faster,” *IEEE Spectrum* (2019) https://spectrum.ieee.org/nanoclast/semiconductors/devices/buried-power-lines-make-memory-faster.
42. P. Wong, “What will the next node offer us?” Keynote talk, *Hot Chips Symp.*, August 2019, http://www.hotchips.org/hc31-keynotes-available-to-all/.
43. “Samsung Newsroom,” October 2019, https://news.samsung.com/global/samsung-electronics-develops-industrys-first-12-layer-3d-tsv-chip-packaging-technology.
44. P. Alcorn, “Intel Lakefield 3-D Foveros hybrid processors…,” tom’sHardware, August 2019, https://www.tomshardware.com/news/intel-lakefield-foveros-3d-chip-stack-hybrid-processor,40205.html.
45. R. Uhlig, “Intel’s Pohoiki Beach,” https://newsroom.intel.com/news/intels-pohoiki-beach-64-chip-neuromorphic-system-delivers-breakthrough-results-research-tests/#gs.qzkmnv (2019).
46. F. Cai et al., “A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations,” *Nat. Electron.* **2**, 290–299 (2019).
47. P. Teichmann, *Adiabatic Logic: Future Trend and System Level Perspective*, Vol. **34**, Springer Series in Advanced Microelectronics, Springer, Netherlands (2011).
48. I. Bezzam, “Power reductions with energy recovery using resonant topologies,” Engineering PhD Thesis, Santa Clara University (2015).

Michael L. Rieger retired from Synopsys in 2017, where his most recent position was chief technologist for the Silicon Engineering Group. He cofounded Precim Corp. (1993) and developed and commercialized model-based optical proximity correction software. Previous positions include technical director, ETEC Corp., engineering and marketing management at ATEQ Corp., and a director of computer-graphics and image processing research at Tektronix. He is a graduate of Dartmouth College (1972) and Stanford (1974), He has 24 US patents, has authored more than 60 technical papers, and he is a senior member of SPIE and life member of IEEE.