Lithography-Induced Limits to Scaling of Design Quality

Andrew B. Kahng
ECE and CSE Depts., University of California at San Diego, La Jolla, CA USA 92093

Abstract

Quality and value of an IC product are functions of power, performance, area, cost and reliability. The forthcoming 2013 ITRS roadmap observes that while manufacturers continue to enable potential Moore's Law scaling of layout densities, the "realizable" scaling in competitive products has for some years been significantly less. In this paper, we consider aspects of the question, "To what extent should this scaling gap be blamed on lithography?" Non-ideal scaling of layout densities has been attributed to (i) layout restrictions associated with multi-patterning technologies (SADP, LELE, LELELE), as well as (ii) various ground rule and layout style choices that stem from misalignment, reliability, variability, device architecture, and electrical performance vs. power constraints. Certain impacts seem obvious, e.g., loss of 2D flexibility and new line-end placement constraints with SADP, or algorithmically intractable layout stitching and mask coloring formulations with LELELE. However, these impacts may well be outweighed by weaknesses in design methodology and tooling. Arguably, the industry has entered a new era in which many new factors – (i) standard-cell library architecture, and layout guardbanding for automated place-and-route; (ii) performance model guardbanding and signoff analyses; (iii) physical design and manufacturing handoff algorithms spanning detailed placement and routing, stitching and RET; and (iv) reliability guardbanding – all contribute, hand in hand with lithography, to a newly-identified "design capability gap". How specific aspects of process and design enablements limit the scaling of design quality is a fundamental question whose answer must guide future R&D investment at the design-manufacturing interface.

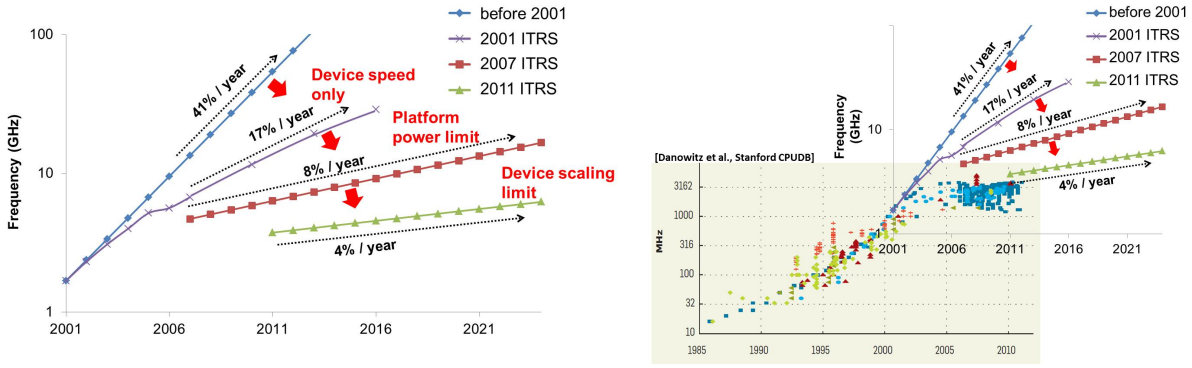
1. INTRODUCTION: SCALING CONTEXT

Moore's Law is, and has always been, a law of cost scaling. "Scaling of cost" is typically seen as the inverse of "scaling of value", and many proxies for "value" – transistor density, clock frequency, power efficiency, cycles per instruction, etc. – have been invoked in various contexts over the years.

In the *International Technology Roadmap for Semiconductors* (ITRS),[11] *maximum on-chip clock frequency* has been a key metric of value for the microprocessor (MPU) and system-on-chip (SOC) *system driver* product classes, which have been central to the last three decades of the industry's growth. Figure 1(a) shows how the ITRS frequency roadmap has evolved since 2001: (i) doubling of clock frequency per technology node (41%/year improvement), through aggressive pipelining in combination with device improvements, ends with the microarchitectural knob running out of steam, so that frequency scaling is limited to device speed improvements; (ii) the 17%/year improvement in device speed (CV/I metric) ends when product platform power limits (e.g., 150W for a desktop, or 3W peak for an application processor in a mobile phone) are reached; and (iii) the reduced 8%/year device speed improvement ends when device engineering cannot deliver corresponding drive currents at small geometries without prohibitive leakage currents. Figure 1(b) shows the ITRS frequency roadmap overlaid with data from the Stanford CPUDB repository.[39] One observation is that the trajectory of the MPU product class, which foreshadows the trajectory of future cost- and power-driven SOC products (i.e., the center of gravity on the logic side of the $200B semiconductor industry), is well-predicted by the ITRS roadmap. Another, more important, observation is that very little of the frequency scaling roadmap – in any of its pre-2001, 2001, 2007 or 2011 incarnations – has been driven by the patterning technologies which have traditionally been seen as the "heart" of the semiconductor roadmap.

Density (i.e., layout area per DRAM bit, SRAM bitcell, or logic gate) has been another key metric of value in the ITRS roadmap. Indeed, Moore's Law has often been equated with the $2\times$ scaling of transistor density per lithography node. In this context, the "heartbeat" of the roadmap is the Metal-1 (M1) or Metal-x (Mx) half-pitch, which decreases by 30% at each lithography node. (When $0.7\times$ scaling is achieved in both x and y dimensions, area scales by $0.7 \times 0.7 = 0.49$, so that density is doubled.) We may credit lithography and other patterning enablers with *geometric scaling*, that is, the scaling of physical dimensions of oxide thickness, channel length, gate pitch, etc. to improve density, performance and reliability. Historically, as geometric scaling has enabled doubling of transistor count in a constant die area with each successive technology node; this in turn has brought higher levels of integration, functionality and product value.

A crucial observation from product data in recent years is that transistor density in actual products *has not scaled* as would have been expected according to Moore's Law.[19] Figure 2(a) shows that even as lithography has delivered "available" Moore's Law scaling (i.e., geometric scaling) per the ITRS roadmap at least through the year 2013, "realized" transistor density scaling has *since 2007* slowed to 1.6× per node instead of the traditional 2× per node. The gap between "available" density scaling from patterning pitch, versus "realizable" scaling in actual products, is a clear challenge to the validity of Moore's Law. Explaining and deconstructing this *design capability gap* – as a consequence of reliability constraints, variability in process and operating conditions, signoff analysis pessimism, foundry models, design architectures, and, yes, lithography – is critical to future resumption of Moore's-Law scaling of value. Section 2 below notes example impacts of lithography on transistor density scaling in recent technology nodes.



(a) (b)

Figure 1. (a) Evolution of the ITRS frequency roadmap. (b) Overlay of the ITRS frequency roadmap with data from the Stanford CPUDB repository.

To compound the design capability gap, even geometric scaling of Mx pitch is now slowing. Daunting challenges to pitch scaling have arisen from resistivity and manufacturability of damascene copper interconnects, poor design-level ROI from new technologies with heavily restricted layout ground rules, and wide (pessimistic) parasitic extraction corners in multi-patterning technologies. Survey data, physical analysis of recent MPU and SOC products, and announced foundry offerings all point to a slowdown of Mx pitch scaling to a three-year cycle in the next two technology nodes, as opposed to the two-year cycle projected in the 2011 ITRS roadmap. This slowdown, depicted in Figure 2(b), is potentially not restricted to logic products alone. For example, the roadmap for contacted poly half-pitch in NAND flash products will likely have both a "2D" version (18*nm* in 2013, scaling to 12*nm* in 2022) and a "3D" version (64*nm* in 2013, scaling to 26*nm* in 2022) – either of which allows the product capacity to double every two years. The latter trajectory, like the slowdown of Mx pitch scaling, relaxes the requirements for patterning technologies, and potentially implies a lessening criticality of lithography (or, acknowledgment of risks and costs of EUV, quad-patterning, etc.) in the semiconductor roadmap.

A further observation is that in the past decade, benefits from scaling to the next technology have given rise to a "20-20-20" world, where 20% speed, 20% power, and 20% (or slightly better) density scaling comprise the overall design benefit from a given next technology node. Examples of "20-20-20" include (i) TSMC from 40*nm* to 28*nm* to 20*nm*,[40–43] (ii) UMC from 40*nm* to 28*nm*,[44–46] (iii) Samsung from 45*nm* to 32*nm*,[36–38] and (iv) GLOBALFOUNDRIES from 20*nm* to 14*nm*.[31–33] In this context, the relative benefit of IC design technologies and other vehicles for scaling of product value can only increase. For example, *equivalent scaling* that achieves non-geometric enhancements of electrical performance (high-K metal gates, FinFET device architectures, etc.) can help bridge the value scaling gap.

Finally, we observe that the slowing of Mx pitch scaling introduces a *further gap* in Moore's-Law scaling. Specifically, if transistor counts continue to scale at 1.6× per node in order to deliver improved product value, then a slowdown of pitch scaling leads to an explosion of die area as shown by the purple line in Figure 2(d). Near-term compensation of the slowdown in density may be afforded by *design-based equivalent scaling* (DES) (Figure 2(c)) which, like equivalent scaling, achieves non-geometric enhancements of performance, density, and other key value metrics. Examples of DES

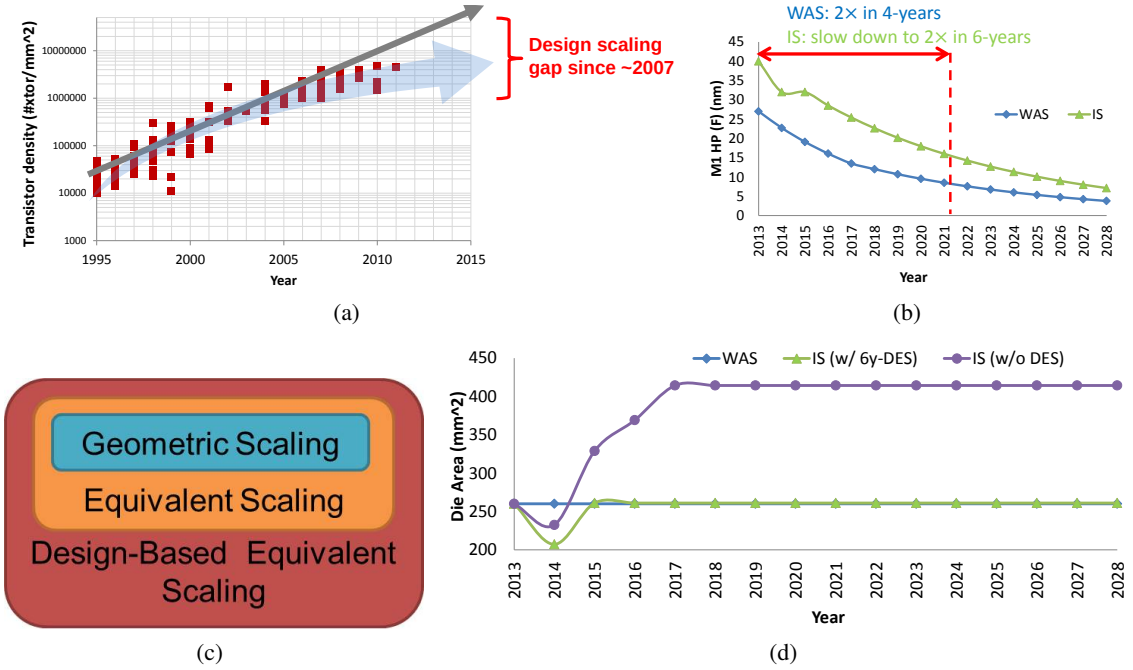


Figure 2. Technology scaling. (a) Scaling gap between 2× "available" density scaling and 1.6× "realizable" density scaling. (b) Slowdown of Mx pitch scaling. (c) Scaling taxonomy: geometric scaling, equivalent scaling, and design-based equivalent scaling. (d) Potential trajectory of design-based equivalent scaling of areal density to rescue Moore's Law for the next several technology nodes.

span error-correcting codes to improve memory reliability, double patterning-aware design techniques that reduce design guardband, clock gating, adaptive voltage and frequency scaling to reduce design margin, etc.

The green line in Figure 2(d) shows the potential impact of DES in a regime of slowed geometric scaling, as transistor counts continue to increase by 1.6× per node. It may be (optimistically) projected that for server and desktop processors (MPU), DES can recover one entire node of Moore's-Law scaling from 2013 to 2019; for processors in SOCs, DES can recover one node of scaling from 2013 to 2020. Put another way, DES can potentially scale down the area of logic overhead (wasted space in logic) to 0.63× its present levels over the next six years, so as to meet the 1.6× transistor density growth requirement and rescue Moore's Law over this near-term time frame. Section 3 below gives several examples to convey the immense potential scope, and benefits, of DES.

2. IMPACT OF LITHOGRAPHY

This section reviews example mechanisms by which lithography strongly impacts the scaling of design quality. The message here is that even as other technologies join lithography as critical enablers of Moore's Law, lithography still limits the incremental value that can be extracted from a new process node.

2.1. Impact of Multiple Patterning Lithography

It is now well-understood that double patterning lithography (DPL) provides a viable enablement for foundry 20*nm* and below nodes, given the delays or other insufficiencies of other technology options such as high refractive index materials, extreme ultraviolet (EUV), or e-beam lithography. In DPL, overlay introduces additional variability in both front-end-of-line (FEOL) and back-end-of-line (BEOL) by means of coupling capacitance variation. Moreover, two layout features patterned by DPL must be assigned to opposite colors (corresponding to different exposures) if their spacing is less than the minimum coloring spacing.[4] However, in any DPL technology there will exist layout configurations for which the layout features separated by less than the minimum color spacing cannot be assigned different colors. In such cases, at

least one feature must be split (i.e., "stitched") into two or more parts (for LELE DPL) or the layout must be changed. The overlay and/or layout constraints in DPL impact design quality in many ways.

To account for the chip-level impact of DPL overlay, Jeong et al.[14] present a resistance and capacitance (RC) extraction flow. The chip-level analysis flow decomposes each local interconnect layer into two masks and shifts the masks to model the overlay effects. The chip-level framework is used to analyze an *AES* encryption core (implemented with RTL obtained from *Opencores*[35] and the *Nangate 45nm* open cell library[34]). *P-DE/DP* and *N-DE/DP* refer to the double exposure or double patterning method with positive and negative photoresist processes, respectively. *P-SDP* and *N-SDP* refer to the "spacer is metal" and "spacer is dielectric" spacer double patterning methods, respectively. Figure 3 shows interconnect capacitance changes of the top 5,307 high-capacitance nets ($\geq 2fF$) in the design. In most cases, more than 10% increase or decrease of capacitances from the nominal capacitance values is observed. Such increases and decreases of capacitances will contribute to larger on-chip variations in timing analysis. Figure 4 shows the relative sensitivity of double patterning lithography options with respect to overlay. Both P-SDP and N-SDP are more sensitive than DE/DP with the same overlay, and the lower layer (M2), which uses smaller-dimension design rules, is more sensitive than higher layers with larger-dimension design rules.

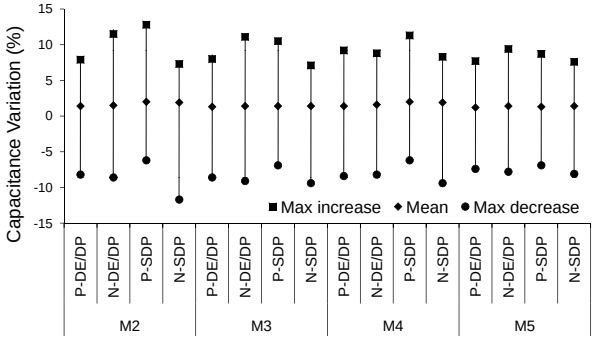


Figure 3. Capacitance changes (%) of high-capacitance nets ($\geq 2fF$) from 3σ overlay.

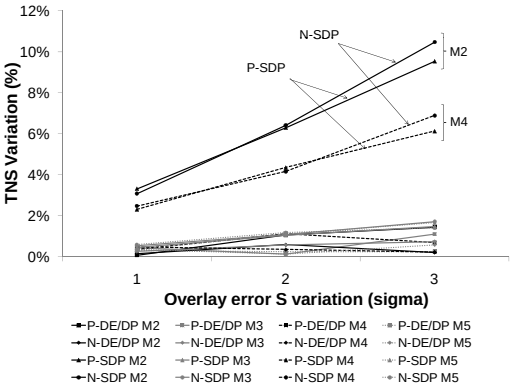


Figure 4. Total negative slack variation (%) from the nominal value in each double patterning lithography option with respect to overlay *S* variation.

Beyond the impact of overlay on circuit timing, DPL also incurs area penalty. Table 1, due to Brink,[1] shows that flip-flop (FF) area at the foundry N14 node with double and triple patterning technology is 84% and 59% of the reference flip-flop area at N20. Meanwhile, because of the increased resolution, single patterning with EUV can achieve a 50% area reduction. This implies that due to the constraints in double- or multi-patterning lithography, there is a 9% area overhead in a FF cell compared to EUV.

Table 1: Flip-flop area with different lithography options.[1]

	20*nm* node (reference)	14*nm* double patterning	14*nm* triple patterning	14*nm* node EUV
Area (um^2)	1.952	1.638	1.143	0.976
N14/N20	100%	84%	59%	50%

2.2. Impact of Double Patterning on Reliability

With respect to reliability mechanisms and design margins, *time-dependent dielectric breakdown* (TDDB) is becoming a critical issue since the electric field across dielectric increases as technology scales. Moreover, dielectric reliability is aggravated when interconnect spacings vary due to (vias and wires) mask misalignment in double patterning lithography. As shown in Figure 5, CD variation in double patterning leads to reduced insulator (dielectric) thickness, which increases

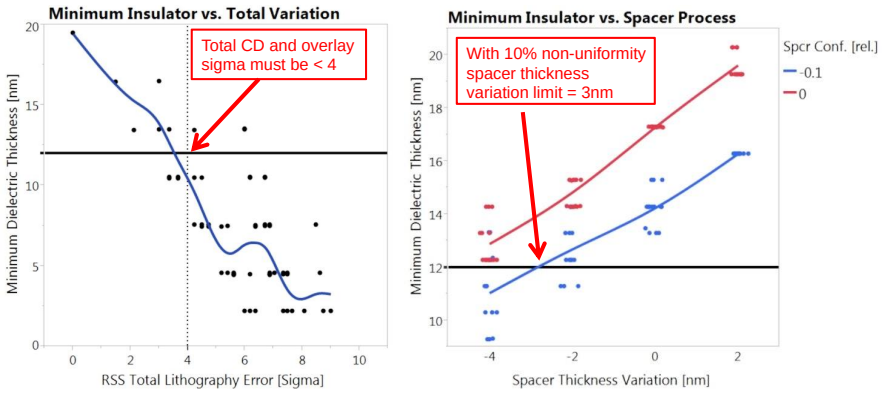


Figure 5. LELE (left plot) and SADP (right plot) double patterning variation increases TDDB risk due to reduced insulator (dielectric) thickness. Horizontal lines are at 12*nm*, assuming 1.2V and a 1.0MV/cm breakdown field.[7]

TDDB risk. Although dielectric reliability can be mitigated by a larger interconnect pitch, such a guardband leads to significant area overhead.

Chan and Kahng propose alternative approaches to reduce the design margin for TDDB in the double-patterned BEOL context, through (1) signal-aware TDDB reliability estimation and (2) post-detailed routing layout optimization.[3] First, conventional TDDB reliability estimation is based on a worst-case assumption in which each interconnect pair is under DC TDDB stress (i.e., each pair of adjacent wires always carries opposite logic signals). Such estimation is clearly very pessimistic. To reduce the pessimism, total stress time for interconnects is estimated using state probabilities (i.e., the probability that an interconnect has a logic state '1') that are available from simulation during the logic design phase of IC implementation. In particular, the state probability of all interconnects can be obtained from EDA tools through vectorless logic simulation. Given the state probabilities of two interconnects, one can calculate the worst-case *stress ratio*, α_{ij}, for each interconnect pair. The *stress ratio* is defined as the fraction of the time when a pair of interconnects have opposite logic signals:

$$\alpha_{ij} = \begin{cases} q_i + q_j, & \text{if } = (1-q_i) > q_j \\ (1-q_i) + (1-q_j), & \text{otherwise} \end{cases} \quad (1)$$

where q_i is the probability of the i^{th} interconnect to be logic '1'. Estimation of the maximum stress time using Equation (1) is less pessimistic compared to the estimation with DC TDDB stress. Second, the post-routing optimization improves TDDB reliability by local shifting of the edges of small wire segments so to enlarge the particular interconnect spacing (dielectric) that is at risk. Figure 6 shows an example of wires before and after the post-routing layout optimization. From the figure, we see that the via-to-wire spacing is increased by shifting the wire edges. As a result, the electric field across the dielectric is reduced to improve TDDB lifetime.

Table 2 shows that the signal-aware TDDB reliability analysis gives chip lifetime estimates that are 1.7 to 2.8 times the lifetime estimates obtained with a pessimistic DC stress assumption (both estimates obtained without layout optimization). This confirms that TDDB reliability is design-specific, i.e., dependent on the stress ratio of interconnect pairs in the design. All four designs studied exhibit a marked reduction of pessimism when signal-aware TDDB reliability estimation is used. Results in Table 2 also show that by applying the post-routing layout optimization method, we can improve chip TDDB lifetime by 9% to 10% (compared to the original layout). The improvement is slightly larger if edge shifting is allowed whenever there is no via located above the edges.

2.3. Impact of Minimum Implant Area

A third example of how lithography impacts design quality arises with the implant steps used to define active areas for multiple threshold voltage (V_t) transistors. As advanced process nodes move to smaller feature sizes, the geometric constraints on layout that arise from limits of patterning technology remain constant, and thus increasingly limit physical designs. Of particular note at the foundry 20*nm* node and below are *minimum implant area* (MinIA) design rules that

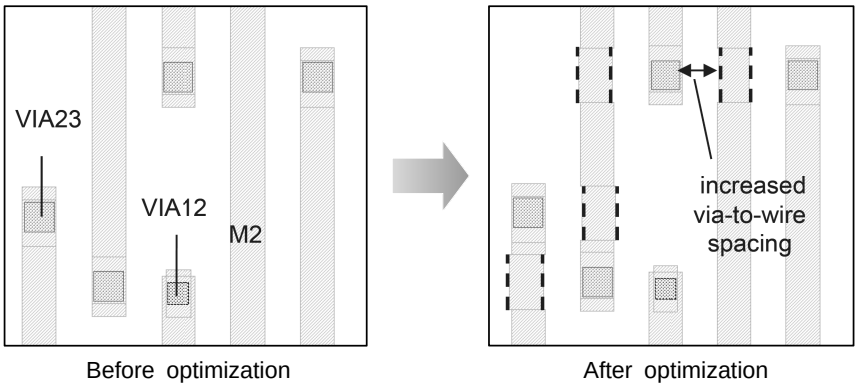


Figure 6. Example of BEOL layout modification. The black dotted lines indicate edges of wire segments that are shifted (locally) to increase via-to-wire spacings and improve TDDB reliability.

Table 2: Chip lifetime (TDDB reliability), normalized to lifetime before layout optimization and with DC stress assumption.

	DC stress			Design-specific stress ratio		
	no opt.	shift edges when there is no via		no opt.	shift edges when there is no via	
		above or below	below only		above or below	below only
AES	1.000	1.087	1.099	1.696	1.846	1.865
JPEG	1.000	1.085	1.097	2.146	2.333	2.359
MPEG2	1.000	1.087	1.102	2.763	3.017	3.052
SPARC_EXU	1.000	1.089	1.100	1.964	2.138	2.158
average	1.000	1.087	1.099	2.142	2.334	2.359

come into effect in multi-V_t standard cell-based designs. Traditional timing- and routability-driven placement of cells with multiple V_t values can result in a small island of a given V_t that cannot meet the MinIA layer rule. The smaller cell that cannot meet the MinIA rule should be abutted to cells with the same V_t, so as to form a wider implant layer polygon. That is, a narrow cell cannot be sandwiched between different-V_t cells as shown in Figure 7. We note that the impact of the MinIA rule can be huge when the portion of small-width cells in the netlist is large; this is a common scenario especially in cost-driven, low-power mobile IC products since the multi-V_t technology context is essential to achieve low-leakage, high-performance design implementations.[10]

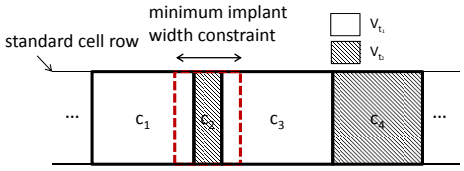


Figure 7. An example of the minimum implant area violation. The dotted line indicates the minimum width constraint of the implant layer. The cell instance c_2 (V_{t_2}) violates the constraint as it is narrow and sandwiched by two cells (c_1 and c_3) that have a different V_t (V_{t_1}).

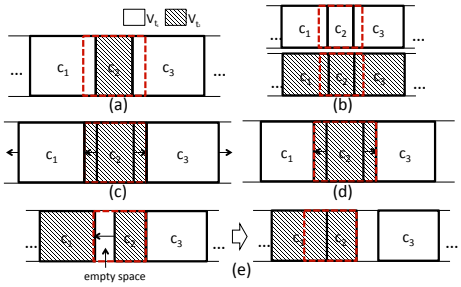


Figure 8. Available fixing approaches for MinIA rule violation. A given violation, depicted in (a), can be fixed by using (b) V_t-swapping, or (c) moving a neighbor cell, or (d) downsizing a neighbor, or (e) moving the narrow cell.

The (minimum width and spacing) design rules for implant layers have not been critical before, as cell sizes have been

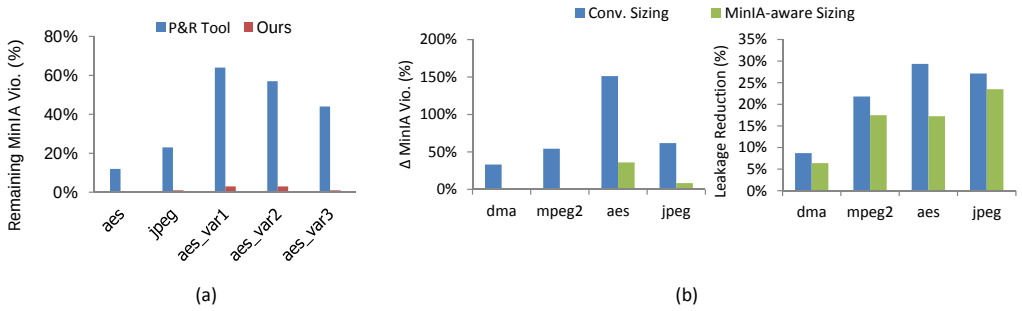


Figure 9. Results of the MinIA-aware placement and sizing heuristics: (a) placement results of a commercial tool and the proposed heuristic, (b) sizing results of the conventional sizing algorithm and the proposed sizing algorithm.

large enough to cover these minimum rules. Hence, up until recent technology generations, placement and gate sizing/V_t-swapping methods in commercial electronic design automation (EDA) tools have not had to consider these rules, as any legal cell placement would be correct by construction with respect to the MinIA criterion. However, as cell sizes have continued to shrink in advanced process nodes, even as the wavelength used in 193*i* optical lithography remains constant, the MinIA rule has become larger than the minimum width of standard cells (e.g., INV$\times$1 cell). MinIA rules constrain cell placement starting with the foundry "20*nm*" (64*nm* minimum metal pitch) node, due to the minimum pattern size limits and cell library (diffusion layer layout) strategies. The minimum width constraint of implant layers changes the traditional placement and post-layout gate sizing problems. That is, in addition to existing constraints such as timing, power and area, additional *geometric* information of cells must be considered.

Kahng and Lee propose a method to optimize power under the minimum implant area constraint with placement perturbation and gate sizing/V_t-swapping.[24] They suggest two heuristics to solve the new placement and sizing problem considering MinIA rules, based on four ways to fix a MinIA violation shown in Figure 8. Figure 9 (a) shows the remaining Δ MinIA violation after applying a commercial EDA tool and the proposed placement heuristic. The commercial tool fixes only 36% of MinIA violations in the worst case (i.e., 64% of violations remain), and 74% on average, across the testcases studied. By contrast, the proposed new heuristic significantly reduces the number of MinIA violations (97% in the worst case, 99% on average). Figure 9 (b) shows sizing results of the conventional sizing algorithm and the proposed sizing heuristics considering MinIA rules. The proposed approach achieves comparable leakage reduction results to the conventional method while minimizing the MinIA violations.

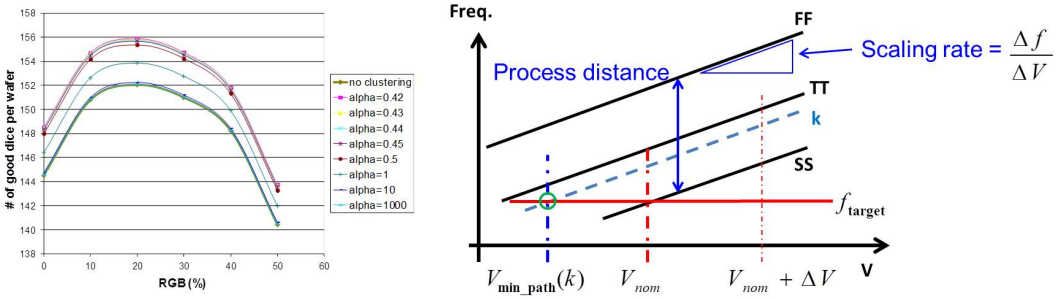
3. DESIGN-BASED EQUIVALENT SCALING

In the ITRS roadmap,[11] *geometric scaling* indicates downscaling of dimensions, which includes thinner gate oxide, narrower poly gate channel length, denser lines and spaces, etc. However, the slowdown in scaling of the Mx pitch introduces a hiatus in traditional Moore's Law scaling. Because transistor density scales at 1.6$\times$ every node, it leads to an explosion in the die area. As noted in Section 1 above, *design-based equivalent scaling* (DES) is now imperative to maintain Moore's-Law scaling of value through novel design techniques, tools and methodologies. Particularly in the near term, DES will be needed to compensate the slowdown of pitch scaling and prevent die area explosion. In this section, we describe several examples that illustrate the tremendous breadth and potential value of DES.

3.1. Guardband Reduction and Adaptivity

A first example of DES consists of new techniques to reduce design margin (guardband), optionally supported by monitoring and *adaptivity* mechanisms built into the design. In a highly motivating study of design guardband impact on design outcomes, Jeong et al.[12] show that 40% reduction in library model guardband can lead to typical reductions of 13% in standard-cell area, 12% in routed wirelength, and 28% in SP&R (synthesis, placement and routing) tool turnaround time for both 90*nm* and 65*nm* designs. (These improvements are quite substantial, especially in a "20-20-20 world".) Further, up to 4% increase in the number of good dice per wafer can be achieved with a 20% guardband reduction. This increase comes without any assumption of improved manufacturing capability (i.e., reduced process variation); it simply reflects a sweet spot in the tradeoff between more raw die per wafer (due to smaller die area with reduced guardband)

and more parametric yield loss (due to design signoff with reduced guardband). Figure 10 (a) shows the change in the number of good dice per wafer over different guardband reductions and defect clustering parameters. The design yield is maximized at around 20% of guardband reduction and decreases afterward. Further, the trend is not affected by the clustering of defects.



(a) (b)

Figure 10. (a) Change in number of good dice per wafer, versus guardband reduction (%) and defect clustering parameters. Die area = 1 cm^2. (b) Illustration of process-aware voltage scaling.

Given the cost of margin, a recent focus in IC product implementation has been the recovery of excess margin allocated for worst-case process variation. To this end, many *adaptive voltage scaling* (AVS) techniques have been proposed. AVS techniques use on-chip monitors or lookup tables (LUTs) to find the minimum supply voltage for given frequency requirements; this enables "signoff at typical", with slow (worst-case) silicon being compensated by available voltage scaling. Chan and Kahng[2] study the voltage scaling characteristics of digital circuits and propose tunable monitoring circuits for process variation. Figure 10 (b) illustrates the basic idea of *process-aware voltage scaling*, with performance modeled as a linear function of the supply voltage. By tuning cell types and pass-gate resistances in monitoring circuits, the monitors can be applied to arbitrary circuits and at any process variation. The proposed methodology achieves 30 mV supply voltage reduction as compared to other AVS techniques.

3.2. Reduced Pessimism in Analysis Flows

A second example of DES highlights the opportunity to improve design analysis tools, such that pessimism in analysis of timing and power is reduced. In timing signoff for leading-edge SOCs, even few-picosecond timing violations will not only increase design turnaround time, but also degrade design quality (e.g., through power increase from insertion of extra buffers). Conventional flip-flop timing models have fixed values of setup/hold times and clock-to-q (c2q) delay, with some advanced "setup-hold pessimism reduction" (SHPR) methodologies[28] exploiting multiple setup-hold pairs in the timing model. Kahng and Lee propose to use multiple timing models to give more flexibility at timing path boundaries, thus recovering significant "free" margins and reducing the number of timing violations that require unnecessary fixes.[23] They exploit a *flexible flip-flop timing model* that captures the three-way tradeoff among setup time, hold time and c2q delay (see Figure 11 (a)), so as to reduce pessimism in timing analysis of setup- or hold-critical paths. A sequential linear programming optimization for multiple corners is used to selectively analyze setup- or hold-critical paths with less pessimism. Further improvements are made based on partitioning of timing paths according to different modes. Kahng and Lee demonstrate that the proposed method can improve worst setup/hold slack metrics over conventional signoff methods, using a set of open-source designs implemented in a 65*nm* foundry library. Figure 11 (b) shows the resultant setup slack after applying the conventional method based on a fixed flip-flop timing model (*conventional*) and the proposed method (*proposed*). With the proposed method, setup time violations are removed in most of designs.

3.3. Exploit Bimodality

A third example of DES points out ways in which IC design methodology can potentially "make lemonade from lemons", with respect to the uncorrelated, bimodal CD variation that is a consequence of LELE double patterning. Figure 12 shows a bimodal CD distribution for 32*nm* technology measured from 24 wafers processed by LELE double patterning.[5]

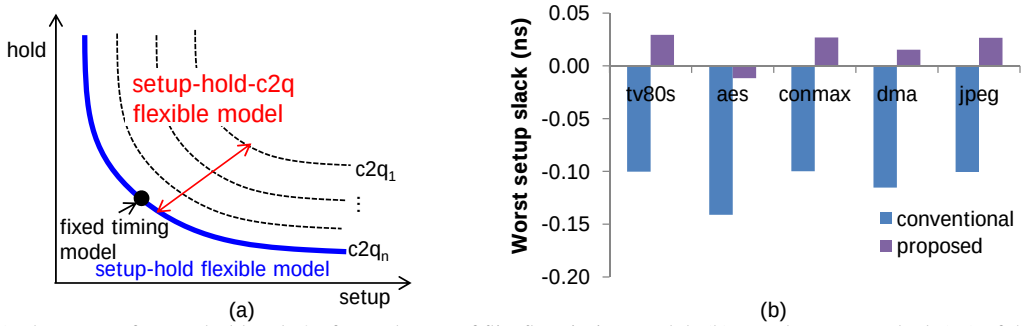


Figure 11. (a) The space of setup, hold and c2q for each type of flip-flop timing model. (b) Resultant setup slack (*ns*) of the conventional timing analysis (*conventional*) and the proposed methodology (*proposed*).

The bimodal distribution can be modeled as two CD groups with independent mean and sigma values. We refer to the different CD distributions as corresponding to the colorings *M1* or *M2* (i.e., mask exposures) of the gate polys in a cell layout. The existence of two independent CD populations in a design takes away the presumptions of spatial correlation in corner-based timing analysis. Interestingly, this bimodal distribution can be *exploited* to reduce datapath delay variation by alternating the coloring of the cells in a given datapath.[13] Figure 13 shows delay variations of a 16-stage inverter chain, normalized to mean values. Here, only four (out of 2^{16}) path colorings are studied: (i) M1-only, (ii) M1-M2-M1-... alternation, (iii) M2-M1-M2-... alternation, and (iv) M2-only. Alternative coloring of the cells along a timing path reduces the covariance among the cell delays, which leads to smaller total delay variations.[13]

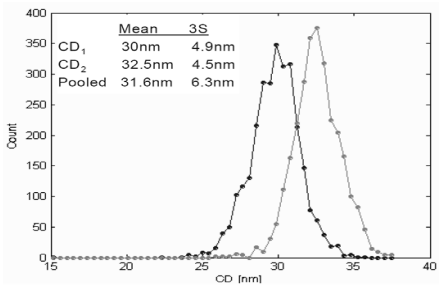


Figure 12: Bimodal CD distribution.

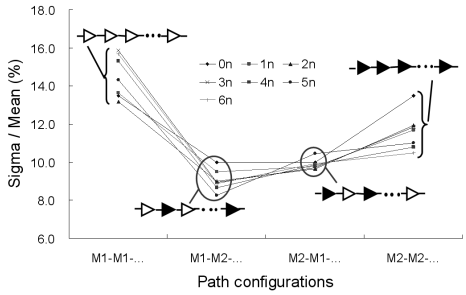


Figure 13. Relative delay variation σ/μ (%) over all process corners.

Similarly, we can exploit bimodality in BEOL interconnects with LELE double patterning, by judicious selection of stitching locations. Figure 14 shows the impact of stitching location on circuit delay. Stitching location is denoted by an index from 1 to 21 which corresponds to equally-spaced discrete locations from source to sink. In particular, stitching location = 1 (resp. = 21) means that the stitching location is immediately after the driver (resp. immediately before the receiver), and the entire interconnect is assigned to Color 2 (resp. Color 1). If the stitching location = 11, the driver-side half is Color 1 and the receiver-side half is Color 2. Results in Figure 14 show that DPL interconnect has less delay variation compared to the single patterning lithography (SPL) case, due to the averaging effect of DPL interconnects. Further, stitching around the middle of interconnect leads to minimal delay variation (long interconnect). This is expected because the capacitance variation of interconnect is minimal when the portions of Color 1 and Color 2 are equal (for DPL). Note that for all testcases, minimum $3\sigma/mean$ is attained when stitching location is slightly shifted towards the driver side. This is because circuit delay is more sensitive to RC changes on the driver side, due to the resistance shielding effect. Resistance shielding implies that driver-side capacitance has more contribution to RC delay than receiver-side capacitance. As a result, the stitching location shifts slightly toward the driver side to balance the effective RC of interconnects with Color 1 and Color 2.

Similar to exploitation of bimodality, other methods have been proposed that seek to use averaging effects to optimize interconnects. Kahng et al. propose an idea is to reduce the worst-case Miller coupling by offsetting the inverters on adjacent lines as shown in Figure 15.[20] With offset inverter locations, any worst-case simultaneous switching on a neighbor

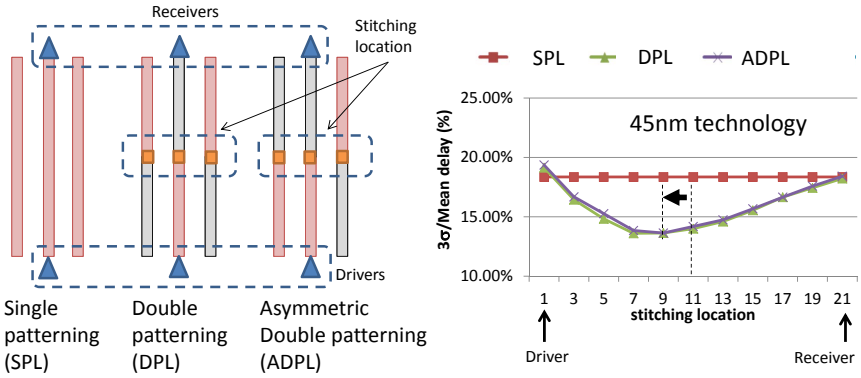


Figure 14: Relative stage delay variation σ/μ (%) with different stitching locations.

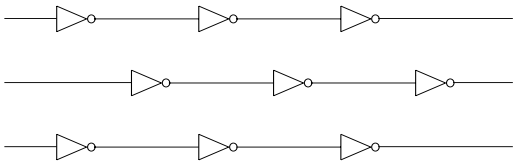


Figure 15. Reduction of worst-case Miller coupling by offsetting inverters. Inverters on the left and right neighbors are at phase = 0.5.

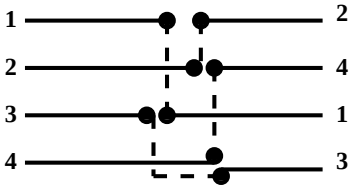


Figure 16. Swizzling routing (metal layer) to minimize delay uncertainty. Dotted lines denote use of the vertical (orthogonal) routing layer and circles denote vias. Note that wire 3 is routed on the same horizontal track as wire 4, but is drawn slightly below 4 for clarity.

line persists only for half of each period between consecutive inverters, and furthermore becomes best-case simultaneous switching for the other half of the period. They claim that the proposed new repeater offset technique can reduce worst-case cross-chip delays by over 30% in current technologies. Gupta and Kahng suggest a routing only layout solution – *swizzling* – which reduces worst-case coupling delay for long parallel wires such as in wide on-chip global buses.[8] A general method is given for construction of good swizzling patterns, and up to 31.5% reduction in worst-case delay and 34% reduction in delay uncertainty is obtained using empirically determined, optimal swizzling patterns as shown in Figure 16.

3.4. Bridges from System Design to IC Implementation

A fourth example of DES focuses on opportunities for improved communication between system designers and IC implementation teams, or between IC design and IC manufacturing. A mantra of "what if we knew (what they know) ...?" is common to a number of potential new, high-value bridges – e.g., "what if we knew a target error rate?" or "what if we knew the operating scenarios and duty cycles?", etc. For instance, if chip implementation teams know likely workloads as well as the error-tolerance of a design (that is, how well the design can detect and correct errors), then it is possible to minimize power for a target error rate through timing slack redistribution based on functional information. Or, if operating scenarios and duty cycles of each operating scenario are known, then lifetime energy can be correctly minimized in the regime of dynamic voltage and frequency scaling (DVFS). We briefly describe these examples of new bridges in the following.

Recovery-driven design ("what if we knew workloads and error-tolerance ..."). Conventional CAD methodologies optimize a processor module for correct operation and prohibit timing violations during nominal operation. *Recovery-driven design* is a design approach that optimizes a processor module for a target timing error rate instead of correct operation.[21] The target error rate is chosen based on how many errors can be gainfully tolerated by a hardware or software error resilience mechanism.

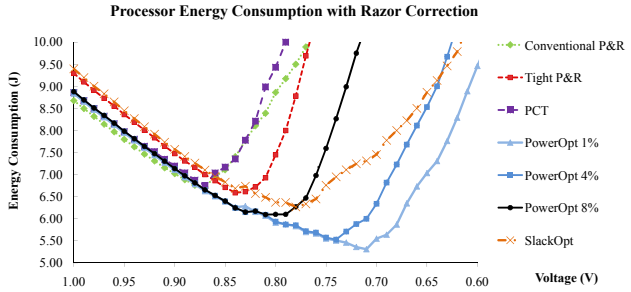


Figure 17. The benefit of designing a processor to produce errors, then correcting them with an error tolerance mechanism, versus designing for correctness and then relaxing the correctness guarantee, can be significant. Results are shown for processors that employ Razor.

One popular hardware-based scheme for recovery-driven design, i.e., error detection and correction, is circuit-level timing speculation.[6, 30] Circuit-level timing speculation-based techniques detect errors by sampling the same computation twice – once using the regular clock and again using a delayed clock – then comparing the two outputs. When the outputs do not match, an error is signaled. Correction involves treating the delayed clock output as the correct output. Razor[6] and EDS[30] are well-known examples of circuit-level timing speculation.

Figure 17 compares the energy consumption of a recovery-driven processor that has been designed and optimized for Razor against the power consumption of processors designed for other objectives, such as gradual slack or path constraint tuning (PCT), and against processors that have been designed for correctness but use the traditional Razor methodology to save energy. Figure 17 demonstrates that the minimum energy is indeed achieved by a processor that is designed to produce errors that can be gainfully tolerated by Razor. By designing the processor for the error rate target at which Razor operates most efficiently, the range of voltage scaling is extended from 0.84V (for the best "designed for correct operation" processor) to 0.71V (for the processor designed for an error rate of 1%), affording an additional 19% energy reduction. In today's "20-20-20 world", such potential gains can no longer be ignored.

**DVFS ("what if we knew scenarios and duty cycles ..."). ** Dynamic voltage and frequency scaling (DVFS) is a popular energy reduction technique that allows a hardware design to reduce average power consumption while still enabling the design to meet a high-performance target when necessary. To conserve energy, many DVFS-based embedded and mobile devices often spend a large fraction of their lifetimes in a low-power mode.[29] However, DVFS designs produced by conventional multi-mode CAD flows tend to have significant energy overheads when operating outside of the peak performance mode, even when they are operating in a low-power mode. A dedicated core can be added for low-energy operation, but has a high cost in terms of area and leakage.

Kahng et al.[22] present a *context-aware DVFS design* flow that considers the intrinsic characteristics of the hardware design, as well as the operating scenario – including the relative amounts of time spent in different modes, the range of performance scalability, and the target efficiency metric – to optimize the design for maximum energy efficiency. They also present a *selective replication-based DVFS design* methodology that identifies hardware modules for which context-aware multi-mode design may be inefficient and creates dedicated module replicas for different operating modes for such modules.

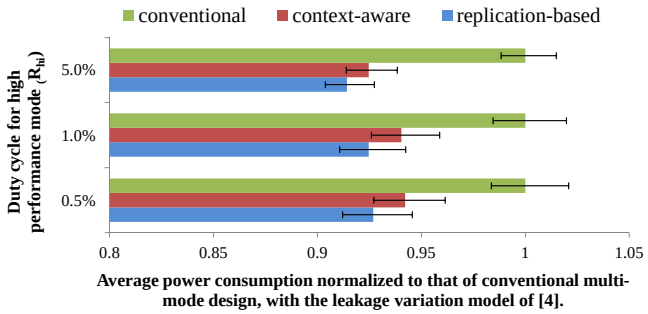


Figure 18: Normalized average power consumption with leakage variations (R_{hi}= {0.5%, 1%, 5%}, X= 5).

Figure 18 compares average power consumption for the processor observed during variation analysis for different multi-mode design styles. Error bars show the min and max average power observed during Monte Carlo analysis. The context-aware design methodology reduces the average power consumption by up to 7.5%. Applying the replication-based technique can further reduce the power by approximately 1%. For designs with low duty cycle at the high performance mode (R_{hi}), where leakage power variations impact total power more significantly, we see that variations impact average power savings by less than 2%.

4. CONCLUDING THOUGHTS

Lithography and fundamental patterning technologies have for decades enabled the continued scaling of semiconductor technology. Today, however, the "heartbeat of the roadmap" – Mx pitch scaling – is slowing due to many reasons that are not directly related to patterning. These reasons, spanning material properties, variability and design margins, electrical performance, and design tool limitations, have reduced the *design benefits* of recent technology nodes, to the point where the industry lives in a "20-20-20 world". In this context, realities of cost and risk force aggressive exploration of 3D scaling (for NAND flash and multi-die integration) and heterogeneous integration (More Than Moore, and beyond-CMOS) paths for future semiconductor products. Lithography and patterning have been joined by 3D scaling, deposition, etch, planarization, next-generation interconnect materials, etc. as first-class enablers of the continuation of Moore's Law.

In sub-28*nm* foundry nodes with multi-patterning in the BEOL, extraction of design value from process technology depends on improved comprehension of the ever-deepening interactions between patterning and design. Three critical challenges are as follows. (1) *Getting signals out.* While FEOL layers can continue geometric scaling with known methods (e.g., regular layout patterned with grids and cut masks, with SAQP and/or DSA looming on the horizon), getting the signals into and out of transistors is increasingly challenging. With the difficulty of scaling contact pitch and resistance, MOL layers have been introduced to access FEOL pins at the cost of process and design complexity. Notably, since the first-level contact and Metal 1 rules interact with the MOL layers, standard-cell layouts are becoming much more complex. Further, area density scaling depends on acceptable evolution of pitches and 'gear ratio' in *two* dimensions: Mx pitch, and contacted poly pitch. (2) *Metal 1 pitch scaling.* Although double-patterning is already a mainstream solution, it brings well-known issues of throughput and cost, layout decomposition, overlay-induced systematic variation, etc. Moreover, double-patterning is already approaching its pitch limit. Since continued scaling of Metal 1 pitch is critical to cell-level shrink, Metal 1 pitch selection must consider (i) scaling and patterning flexibility requirements (across multiple BEOL layers) versus cost and design quality, (ii) actual standard-cell area and design-level area shrink, and (iii) design rule interactions among FEOL, MOL and Metal 1. (3) *Pitch matching.* The gain of actual area scaling is also dictated by pitch matching requirements between metal layers for optimal via density (i.e., flexibility of routing choices for auto-routing tools), in balance with current delivery, variability, cost and other "design value" considerations. The open question is: What combination of standard-cell architecture and metal pitches throughout the BEOL stack will enable *cost-effective* continuation of Moore's-Law scaling?

Finally, several challenges lie purely in the realm of design technology, e.g., how to reduce pessimism and margins in signoff analyses, and how to optimize designs for power- and cost-efficiency over wide ranges of operating conditions and modes. Solving these challenges – and realizing the full potential of *design-based equivalent scaling* – will be essential to achieving high-value products in coming technology nodes.

5. ACKNOWLEDGMENTS

Many thanks are due to Tuck-Boon Chan, Wei-Ting Jonas Chan, Ilgweon Kang, Hyein Lee, Jiajia Li and Siddhartha Nath for their invaluable help with this paper. Thanks are also due to coauthors of the various works that have been cited as exemplars of design-based equivalent scaling, for fruitful discussions and collaborations over the years.

REFERENCES

1. M. V. D. Brink, "Continuing to Shrink: Next-Generation Lithography – Progress and Prospects", *Proc. ISSCC*, 2013, pp. 20-25.
2. T.-B. Chan and A. B. Kahng, "Tunable Sensors for Process-Aware Voltage Scaling", *Proc. ICCAD*, 2012, pp. 7-14.

3. T.-B. Chan and A. B. Kahng, "Post-Routing Back-End-of-Line Layout Optimization for Improved Time-Dependent Dielectric Breakdown Reliability", *Proc. SPIE*, 2013, 8684, 86840L.
4. M. Drapeau, V. Wiaux, E. Hendrickx, S. Verhaegen and T. Machida, "Double Patterning Design Split Implementation and Validation for the 32nm Node", *Proc. SPIE*, 2007, 652109-1-652109-15.
5. M. Dusa, J. Quaedackers, O. F. A. Larsen, J. Meessen, E. van der Heijden, G. Dicker, O. Wismans, P. de Haas, K. van Ingen Schenau, J. Finders, B. Vleeming, G. Storms, P. Jaenen, S. Cheng and M. Maenhoudt, "Pitch Doubling Through Dual-Patterning Lithography: Challenges in Integration and Litho Budgets", *Proc. SPIE*, 2007, 65200G-1-65200G-10.
6. D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner and T. Mudge, "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation", *Proc. MICRO*, 2003, pp. 7-18.
7. D. Fried, "Lithography Challenges Threaten the Cost Benefits of IC Scaling", *Tech Design Forum*, 2013, http://www.techdesignforums.com/practice/technique/lithography-challenges-1xnm-threaten-scaling/
8. P. Gupta and A. B. Kahng, "Wire Swizzling to Reduce Delay Uncertainty Due to Capacitive Coupling", *Proc. VLSID*, 2004, pp. 431-436.
9. IMEC (Interuniversity Microelectronics Centre), "Cell Architecture Assessment – Layout, Parasitics, Variability", *IMEC Scientific Report*, 2012, pp 1-7.
10. ITRS Low-Power Design Technology Roadmap, Design Chapter Table DESN14, 2011, http://www.itrs.net/Links/2011ITRS/2011Tables/Design_2011Tables.xlsx
11. ITRS 2011. http://www.itrs.net/Links/ITRS2011/Home2011.htm
12. K. Jeong, A. B. Kahng and K. Samadi, "Quantified Impacts of Guardband Reduction on Design Process Outcomes", *Proc. ISQED*, 2008, pp. 790-897.
13. K. Jeong and A. B. Kahng, "Timing Analysis and Optimization Implications of Bimodal CD Distribution in Double Patterning Lithography", *Proc. ASPDAC*, 2009, pp. 486-491.
14. K. Jeong, A. B. Kahng and R. O. Topaloglu, "Assessing Chip-Level Impact of Double-Patterning Lithography", *Proc. ISQED*, 2010, pp. 122-130.
15. A. B. Kahng, "The Road Ahead: Shared Red Bricks", *IEEE Design and Test of Computers* 19(2) (2002), pp. 70-71.
16. A. B. Kahng, "The Road Ahead: The Cost of Design", *IEEE Design and Test of Computers* 19(4) (2002), pp. 136-137.
17. A. B. Kahng, "The Road Ahead: Scaling: More than Moore's Law", *IEEE Design and Test of Computers* 27(3) (2010), pp. 86-87.
18. A. B. Kahng, "The Road Ahead: Roadmapping Power", *IEEE Design and Test of Computers* 28(5) (2011), pp. 104-106.
19. A. B. Kahng, "Design Capability Gap", *UCSD CSE Department technical report* CS2013-1002, 2013.
20. A. B. Kahng, S. Muddu, E. Sarto and R. Sharma, "Interconnect Tuning Strategies for High-Performance ICs", *Proc. DATE*, 1998, pp. 471-478.
21. A. B. Kahng, S. Kang, R. Kumar and J. Sartori, "Recovery-Driven Design: Exploiting Error Resilience in Design of Energy-Efficient Processors", *IEEE Trans. on CAD* 31(3) (2012), pp. 404-417.
22. A. B. Kahng, S. Kang, R. Kumar and J. Sartori, "Enhancing the Efficiency of Energy-Constrained DVFS Designs", *IEEE Trans. on VLSI* 21(10) (2013), pp. 1769-1782.
23. A. B. Kahng and H. Lee, "Margin Recovery with Flexible Flip-Flop Timing", *Proc. ISQED*, 2014,
24. A. B. Kahng and H. Lee, "Minimum Implant Area-Aware Gate Sizing and Placement", *Proc. GLSVLSI*, 2014, to appear.
25. A. B. Kahng, C.-H. Park, X. Xu and H. Yao, "Layout Decomposition for Double Patterning Lithography", *Proc. ICCAD*, 2008, pp. 465-472.
26. A. B. Kahng and V. Srinivasan, "Big Chips", *IEEE Micro* 31(4) (2011), pp. 3-5.
27. T. Ragheb, S. Chan, A. Yeung, R. Monga, H. Mau, V. Ramasubramanian and R. Trihy, "Double Patterning-Aware Extraction Flows for Digital Design Signoff in 20/14nm", *SNUG*, 2013, pp 1-13.
28. E. Salman, A. Dasdan, F. Taraporevala, K. Kucukcakar and E. G. Friedman, "Exploiting Setup-Hold-Time Interdependence in Static Timing Analysis", *IEEE Trans. on CAD* 26(6) (2007), pp. 1114-1125.

29. A. Shye, B. Ozisikyilmaz, A. Mallik, G. Memik, P. Dinda, R. Dick and A. Choudhary, "Learning and Leveraging the Relationship Between Architecture-Level Measurements and Individual User Satisfaction", *Proc. ISCA*, 2008, pp. 427-438.
30. J. W. Tschanz, K. Bowman, S.-L. Lu, P. Aseron, M. Khellah, A. Raychowdhury, B. Geuskens, C. Tokunaga, C. Wilkerson, T. Karnik and V. De, "A 45nm Resilient and Adaptive Microprocessor Core for Dynamic Variation Tolerance", *Proc. ISSCC*, 2010, pp. 282-283.
31. GLOBALFOUNDRIES, http://www.globalfoundries.com/home
32. GLOBALFOUNDRIES, *GF 14nm XM (eXtreme Mobility) process*. http://www.globalfoundries.com/technology-solutions/leading-edge-technologies
33. GLOBALFOUNDRIES, *GF 20nm LPM (low power mobility) process*. http://www.globalfoundries.com/technology-solutions/leading-edge-technologies/20lpm
34. NANGATE. http://www.nangate.com
35. OPENCORES. http://www/opencores.org
36. Samsung Electronics Inc., Foundry Solution, http://www.samsung.com/global/business/semiconductor/foundry
37. Samsung Electronics Inc., *Samsung 32nm HKMG (high-k metal gate) process*. http://www.samsung.com/global/business/semiconductor/foundry/process-technology/32-28nm
38. Samsung Electronics Inc., *Samsung 45nm process*. http://www.samsung.com/global/business/semiconductor/foundry/process-technology/40-45nm
39. Stanford CPUDB. http://cpudb.stanford.edu
40. Taiwan Semiconductor Manufacturing Company, http://www.tsmc.com/english/default.htm
41. Taiwan Semiconductor Manufacturing Company, *TSMC 20nm process*. http://www.tsmc.com/english/dedicatedFoundry/technology/20nm.htm
42. Taiwan Semiconductor Manufacturing Company, *TSMC 28nm process*. http://www.tsmc.com/english/dedicatedFoundry/technology/28nm.htm
43. Taiwan Semiconductor Manufacturing Company, *TSMC 40nm process*. http://www.tsmc.com/english/dedicatedFoundry/technology/40nm.htm
44. United Microelectronics Corporation, http://www.umc.com/English/
45. United Microelectronics Corporation, *UMC 28nm process*. http://www.umc.com/English/process/a.asp
46. United Microelectronics Corporation, *UMC 40nm process*. http://www.umc.com/English/process/b.asp