New Game, New Goal Posts: A Recent History of Timing Closure

Invited

Andrew B. Kahng
CSE and ECE Departments, UC San Diego, La Jolla, CA 92093 USA
abk@ucsd.edu

ABSTRACT

Timing closure is the most critical phase of modern system-on-chip implementation: without timing closure, there is no tapeout. Timing closure is the end result of (i) years of methodology development, script development, signoff recipe development, etc.; (ii) months of block- and top-level final physical implementation; and (iii) a last set of manual noise and DRC fixes, with a final signoff analysis and physical verification. Over the past decade, key aspects of the underlying process and device technologies, modeling standards, EDA tooling, design methodology, and signoff criteria have changed the nature of timing closure. This paper surveys such recent evolutions in timing closure and notes directions for near-term future evolutions.

Categories and Subject Descriptors

B.7.2 [**Hardware**]: INTEGRATED CIRCUITS—*Design Aids*

Keywords

Timing closure, signoff, IC implementation, IC physical design methodology

1. INTRODUCTION

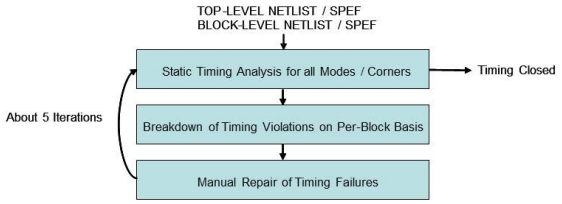
Timing closure immediately precedes final signoff and tapeout in modern system-on-chip (SOC) implementation. Requirements for timing closure, along with enablements and paths taken to reach this final state of the IC design, vary widely across companies and products. Whether a part is binned, whether it is in a cost- and/or low power-driven market, and many other considerations (lifetime, range of functional modes, maturity of target process, maturity of EDA tooling, etc.) all affect how timing closure is achieved today. In practice, timing closure melds (i) years of methodology development, script development, signoff recipe development, etc.; (ii) months of block- and top-level final physical implementation; and (iii) a last set of several hundred manual noise and DRC fixes, along with a final multi-day pass of full-chip signoff analysis and physical verification. A long-time physical design (PD) engineer might claim that timing closure in 16/14nm FinFET technology closely resembles timing closure of a decade ago in 65nm low-power planar bulk technology. Indeed, activities such as DRC and noise fixes, scripting of memory and clock/power distribution, etc. remain crucial to crossing the finish line. Yet, this five-node span has also seen major evolutions of underlying process and device technology, modeling standards, EDA tooling, design methodology, and signoff criteria – with further significant changes needed soon. This paper gives a personal overview of recent evolutions in the timing closure arena, along with some directions for near-term future evolutions.

DAC '15, June 07 - 11, 2015, San Francisco, CA, USA

http://dx.doi.org/10.1145/2744769.2747937.

1.1 Traditional View of Timing Closure

Figure 1, from the DAC Knowledge Center article of MacDonald [30], shows a recommended scope and main steps of (top-level) timing closure. The article dates from 2009-2010 and the 65nm-40nm node transition. The figure shows five iterations,[1] each of which involves static timing analysis, breakdown of timing failures, and manual repair of timing failures. It is expected that the top-level timing improves after each iteration. During the manual fix step in a given iteration, the PD engineer should apply simplest optimizations first; the recommended ordering in [30] is V_t-swap first, followed by gate sizing, buffer insertion, non-default routing rule (NDR) application, and useful skew.



Operations Permitted at Each Iteration (in order of preference)

Iteration 1: Vt Swap, Resizing, Buffer Insertion, NDR Changes, Useful Skew
Iteration 2: Vt Swap, Resizing, Buffer Insertion, NDR changes
Iteration 3: Vt Swap, Resizing, Buffer Insertion
Iteration 4: Vt Swap, Resizing
Iteration 5: Vt Swap

Violation Classes Addressed for Each Iteration (in order of priority)
(1) Electrical Rule Violations
(2) Noise Violations
(3) Setup Violations
(4) Hold Violations

Figure 1: Scope and main steps of timing closure, from [30].

1.2 Context: Node Timing and Low Power

The recent evolution of timing closure is arguably the consequence of two "big-picture" trends.

The race to the end of the roadmap. Today's dominant business foundry-fabless framework (with equipment, IP and EDA also in the supply-chain picture), along with the huge costs of both technology development and design enablement, induces a "(death) race to the end of the roadmap". Those who cannot come up with the required investments (capex, design enablement) and/or successful product offerings (process nodes, application processors) drop out of the race.

A consequence of this race is that technology node timing has not slowed despite many near-term "red bricks" [16] in the semiconductor roadmap. Indeed, the timing of node enablement, measured by SPICE model stabilization, has been accelerating; this makes timing closure and signoff particularly challenging for an early-adopter fabless design house.[2]

[1] The number of iterations is a function of schedule (e.g., three weeks for the final pass permits five three-day repair and signoff analysis iterations).

[2] In recent nodes, the model convergence 'dance' between foundry and fabless customer has four basic stages. (1) From "paper models" to a v0.1 SPICE model, with only sparse R&D silicon data. (2) v0.5 SPICE model, supported by early process qualification vehicle and test-structure data, with preliminary binning data (and, possibly, tightening of global (SSG, FFG) corners. [Background: the SS corner includes global variation plus (on-die) mismatch, while the SSG "global corner" includes only the global variation (leaving on-die variation to path structure-aware AOCV,

[17] notes the immutability of basic time constants in the co-evolution of product design and manufacturing: (i) technology development, application market definition, and architectural and front-end design are O(years); (ii) RTL-to-GDS implementation and reliability qualification are O(months); (iii) fab latency, cycles of yield learning, design re-spins, and mask flows are O(weeks); (iv) process tweaks and design ECOs are O(days). Mismatches among these time constants are a root cause of model-hardware miscorrelation and model guardbanding, and and make acceleration of node enablement challenging if not unrealistic. Another observation is that by keeping its foot on the accelerator, the industry increases the pain from materials challenges (e.g., formation of damascene copper wires, nearing the "fundamental limit" of ∼14nm trench CD) and manufacturing variability in the middle-of-line (MOL) and back-end-of-line (BEOL). For example, lateness of EUV lithography is put into the spotlight by the cost and variability impacts of (self-aligned) double-/quadruple-patterning in advanced BEOL stacks.

The low-power grand challenge. "Mobility." "Big data, green datacenters, and the cloud." "The Internet of Things." All anticipated drivers for future growth in semiconductors share one critical requirement: low power. However, low-power design techniques (cf. [12] and [19]) – multiple supply voltages, multiple voltage domains, power and clock gating, DVFS, MTCMOS, multi-Lgate, etc. – increase the timing closure burden by adding complexity to analysis and/or optimization. Recent FinFET technologies (from the IDM 22nm node and the foundry 16/14nm node onward) offer enticing opportunities for voltage scaling and dynamic power reduction, but the wider ranges of supply voltages[3] vastly increase the number of signoff corners. Of particular note is the difficulty of multi-corner, multi-mode (MCMM) clock network synthesis in a regime where each of hundreds of scenarios has different clock insertion delay and timing constraints.

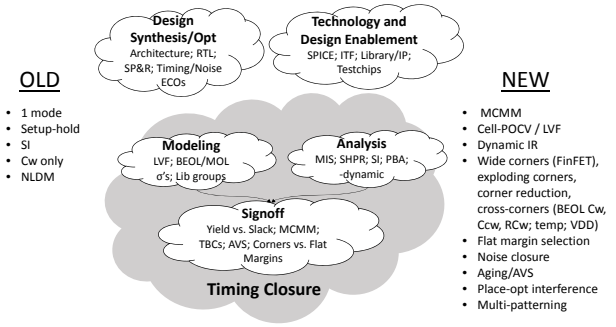


Figure 2: Timing closure (analysis, modeling, and signoff) and its context: design synthesis/optimization, and design/technology enablement. Also sketched: aspects of "Old" vs. "New".

1.3 New Game, New Goal Posts

In many ways, *all* recent evolutions and near-term futures of timing closure are consequences of the above context. Some examples are the following.

- The "rise of the MOL and BEOL" with their dominant resistivity and variability impacts, as well the explosion of signoff corners (C-worst, Cc-worst, C-best, Cc-best, etc. for each additional double-patterning layer). A consequence is the importance of corner selection and signoff criteria (e.g., tightened corners, signoff at typical with flat margin, etc.) to maintain design productivity with minimal PPA and yield loss.

- The criticality of holistic margin reduction [20] [21] and relentless pursuit of margin recovery. It is now well-understood that margin is synonymous with overdesign, cost, and loss of competitiveness.[4] This drives interest in, e.g., higher-dimensional delay and slew modeling (cf. Liberty Variation Format (LVF) [32] [38]) or mask color-aware place-and-route and signoff. A notable open challenge is reduction of flat (aka 'fixed') margins that must be defined at so many signoff corners; this is difficult since such margins are intended to "model what cannot be modeled".[5]

- The rapid and near-universal adoption of adaptivity to (process, lifetime) variations in the form of monitor-enabled adaptive voltage scaling (AVS), as in [2]. AVS has been a true game-changer: it enables setup timing to be closed at typical corners (particularly when in a mature process), and forces product engineering / operations teams to decide the meaning of a "setup timing violation" when voltage can be increased to meet setup.[6]

- The need to use STA with path-based analysis (pba) with noise analysis enabled, as opposed to traditional graph-based analysis (gba), earlier in the PD flow. Pessimism reduction via use of pba has led to overheads in STA turnaround times, EDA license costs, and engineering compute infrastructure costs. In this light, there are interesting future interactions between the adoption of high-dimensional variability modeling standards such as LVF and a *lessened* need for pessimism reduction via pba.

All of these exemplify how timing closure has changed, resulting in a 'new game' with such 'new goalposts' as signoff at typical.[7] Figure 2 notes some of the 'old' vs. 'new' aspects of timing closure – spanning analysis, modeling, and signoff criteria – in the context of design optimization and design/technology enablements.[8] In what follows, Section 2 calls out newer timing closure challenges such as multi-input switching, BEOL corner proliferation with multi-patterning, and placement-sizing interferences. Section 3 then notes several near-term mitigations for these challenges. Section 4 concludes with potential futures for timing closure.

cell-aware POCV or (cell, load, slew)-aware LVF modeling mechanisms [38]). Cross-corners (FSG, SFG) are increasingly required as well, e.g., for signoff of clock distribution.] (3) v0.8 SPICE model, with solid global corners, incorporation of layout-dependent effects, updated mismatch modeling, reliability models and data, etc. At this point, the design house has taped out any number of test chips that separately support characterization of key IPs, model-hardware correlation, and yield learning. (4) v1.0 SPICE model and volume production-readiness (verified yield of large SRAMs, reliability metrics satisfied, etc.). The historical ∼18-month interval between stages (1) and (4) has been decreasing, with competitive pressure driving a proposed reduction to ∼12 months at the foundry N10 node. At the same time, tighter corners (e.g., tighter than SSG for setup paths on which sufficient statistical averaging is likely) may be on offer to foundry customers earlier than might be expected from historical rates of process maturation.

[3] For example, core supply voltage for logic may be scaled across a range of 0.46V to 1.25V in foundry 16/14nm, with separate rails and voltage ranges for memories (in active and retention modes) and analog circuits. SOC designs will continue to see an explosion of voltage, power and clock domains; the latter already number in the thousands for leading-edge products.

[4] Katz [42] notes that margin is rapidly becoming scarce across next-generation products in many sectors: IoT, mobile, communications, etc. Not only do products aggressively push the envelope of complexity, performance, power and cost, but there is an increasingly direct punishment from the market for trading away (spec, yield, time-to-market) for padding of margins. (See [15] for an early analysis of "cost of guardband".)

[5] There are clear opportunities to detangle e.g., PLL jitter, CTS jitter, foundry-dictated jitter margin and dynamic IR drop margin – all of which are swept under a single jitter margin rug. Methodology for frequency-aware hold margin definition, or compensation for SPICE model accuracy changes across PVT corners (particularly extreme super-overdrive and super-underdrive corners), can also provide benefits. There is a dependency here on improved model-hardware (signoff to silicon) correlation.

[6] Redmond [45] notes that AVS changes the goal from "ensuring timing is met under every case" to "accurately modeling delay". Further, AVS removes a "DC component" of timing margin, allowing signoff analyses to focus on remaining margin components; this has lessened impacts of mode-corner proliferation.

[7] Lutkemeyer [43] makes the excellent observation that while the game is indeed new (e.g., slacks now reported at a confidence tail of the slack distribution, affording an approximate statistical analysis), the goalposts are actually 'old' in that STA tools and timing closure still center on absolute slack violations (as opposed to yield losses). Unfortunately, sigmas are unstable, and committed sigmas are difficult to obtain from the silicon provider. Longevity of the timing slack 'goal post' might also result from PD teams' need to have a clear timing closure finish line.

[8] Only fragments of this picture can be discussed here. Yet, my hope is that this paper can sketch a "lower bound" on what must be considered by a design team as it establishes its plan of record signoff and timing closure methodology when moving to 20nm or below.

2. NEW TIMING CLOSURE CHALLENGES

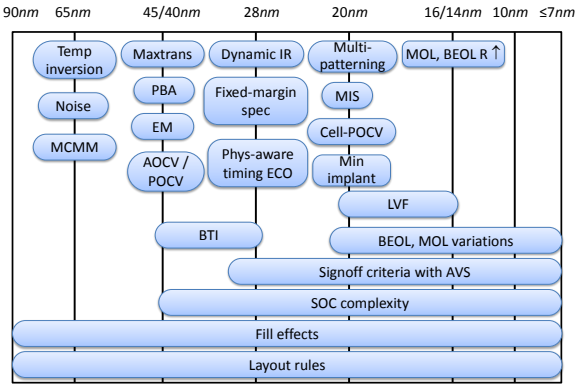


Figure 3: Evolution of timing closure care-abouts.

Figure 3 shows a sampling of timing closure concerns, mapped against technology nodes. This section samples "what is new" in timing closure.

2.1 Multi-Input Switching

Conventional timing libraries consider only single-input switching (SIS) in gate delay models, even though multi-input switching (MIS) – when more than one input switches at given time – can significantly change arc delay. Figure 4(b) shows SPICE-calculated [36] MIS and SIS arc delays for a NAND2 standard cell in a foundry 28nm FDSOI library; the cell has a FO3 load, as shown in Figure 4(a). In the simulation, a ramp transition is made at IN, and delay is calculated on the arc from IN to the NAND2 output. For MIS, a ramp transition is made at IN1 with the same switching direction and slew time as IN. The IN1 arrival time offset with respect to arrival time of IN is swept to find the minimum arc delay, which is taken as the MIS delay. For SIS, IN1 is set to VDD and the arc delay is taken as SIS delay. Both nominal (0.9V) and 80% of nominal supply voltage values are used. The figure shows that MIS delay can be less than ∼50% of SIS delay when the input is falling (and, more than ∼10% greater than SIS delay when the input is rising); the MIS delay reduction is critical to model correctly in hold signoff. The recent paper of Lutkemeyer [26] describes improvements to simple derating approaches which are now being implemented in commercial STA products; however, gaps in the modeling standards such as Liberty [38] still exist [43].

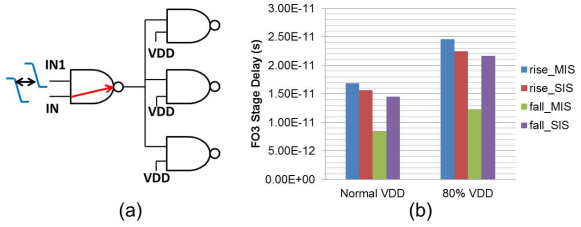


Figure 4: (a) Setup of 28nm FDSOI NAND2 cell with FO3 load for SPICE-based studies. (b) Arc rise and fall delays with MIS and SIS.

2.2 BEOL Multi-Patterning Impacts

Sub-20nm BEOL (and MOL) layers are not only highly resistive, but the variations of line geometry due to multi-patterning and/or planarization steps have significant RC impacts. Foundry plans-of-record for 10nm and below incorporate self-aligned multiple-patterning (SAMP) for pitch scaling and protection against overlay error impacts. However, SAMP induces complex layout restrictions (via placement, unidirectional Mx routing) which challenge detailed routing and cell library design – and, ultimately, density and value. Further, increased BEOL variability, seen on more metal layers, significantly impacts timing closure [14] [9]. Figure 5(a) gives a

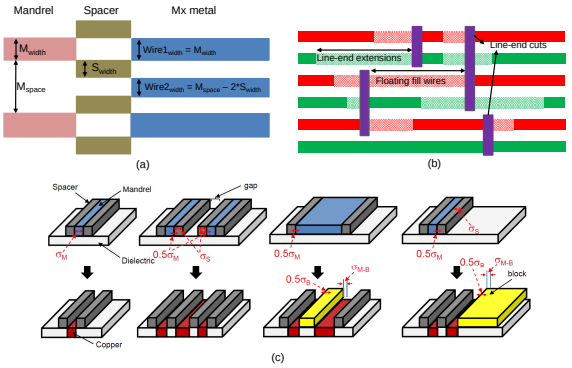


Figure 5: (a) Schematic view of SADP process. (b) Line-end extensions and floating fill wires induced by cut-mask restrictions. (c) Four possible patterning solutions for a BEOL wire in SID-type SADP [13]: (i) both line edges defined by mandrel edges ($\sigma^2 = \sigma_M^2$); (ii) both line edges defined by spacer edges ($\sigma^2 = \sigma_M^2 + 2\sigma_S^2$); (iii) one line edge defined by mandrel edge and the other edge defined by block edge ($\sigma^2 = (0.5\sigma_M)^2 + \sigma_{M-B}^2 + (0.5\sigma_B)^2$); and (iv) one line edge defined by spacer edge and the other edge defined by block edge ($\sigma^2 = (0.5\sigma_M)^2 + \sigma_S^2 + \sigma_{M-B}^2 + (0.5\sigma_B)^2$).

schematic of a self-aligned double-patterning (SADP) process:[9] changes to mandrel width and spacing will change wire width and ground/coupled capacitances. This variation becomes more severe with self-aligned quadruple patterning (SAQP). To compensate corner rounding and pattern fidelity loss in the (line-end) cut mask step, restricted (rectangular) cut mask shapes are used; this forces metal line-end extensions and floating fill wires (Figure 5(b)) which again impact signal timing by unpredictably increasing grounded and coupling capacitances for a given net. Further, Figure 5(c) shows how in the "spacer is dielectric" (SID) form of SADP, σ of a wire segment's CD can vary according to whether the segment is formed as mandrel, as gap (between spacers), etc. Below 20nm, implementation, signoff and physical verification tools must co-own (and agree on the analysis and mitigation of) this variability mechanism.

2.3 Corner Super-Explosion

There are several obvious root causes of the "combinatorial explosion" of views at which timing must be closed for a complex SOC: (i) a plethora of functional (scenario-based, overdrive, underdrive) and test (scan, at-speed, BIST) modes; (ii) Cw, Ccw, Cb, RCw, RCb, etc. corners per each double-patterned layer in the BEOL stack; (iii) 20+ power domains, with many 'cross-corner' analyses forced by asynchronous interfaces between domains that can independently scale supply voltage. In this context, the central engineering team that chooses a subset of PVT corners and constraints for timing closure has enormous influence on the balance between product quality, design effort, and schedule. Yet, some factors in the 'corner super-explosion' are unavoidable.

For example, Figure 6(b) illustrates the temperature reversal effect: when the supply voltage is lower than the temp reversal point V_{tr}, the gate is slower at low temperature (e.g., -30°C). On the other hand, when the supply voltage is higher than V_{tr}, the gate is slower at high temperature (e.g., 125°C). Thus, when the signoff voltage is near V_{tr}, both low and high temperature corners must be checked.

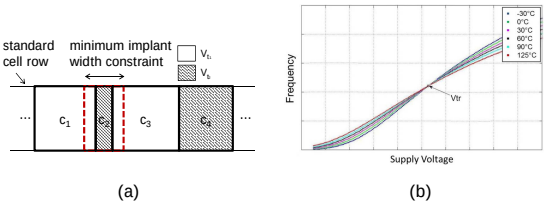
Gate-wire balance is another design consideration that makes different timing paths critical at different PVT corners. With increase of supply voltage, gate delay decreases much faster than wire delay. For example, at the foundry 20nm node, supply voltage scaling from 0.7V to 1.2V might reduce gate delay by ∼50%, while wire delay (say, 100μm on M3) reduces

[9]The mandrel pattern is defined by a mask in the first lithography process and the sidewall spacer is formed with deposition. The mandrel pattern is then selectively removed and the cut mask covers part of spacers in the second lithography process. The substrate is then etched with the cut mask and the remaining spacers (which are not covered by cut masks), and the etched trench is filled with conductive material.

by only ∼2%. Further, while temperature increase always leads to increased wire resistance and delay, its impact on gate delay is uncertain due to the temperature reversal effect.[10] Therefore, to manage clock skew variation and/or fix timing violations (without ping-pong effects) across multiple modes and/or corners, it is increasingly important to comprehend gate-wire delay balancing on clock and data paths.

2.4 Placement-Sizing Interferences

At foundry 20nm and below, new "interferences" arise between post-layout optimization and P&R. Notably, *minimum implant area* (MinIA) constraints[11] imply that post-detailed routing V_t-swap is no longer independent of detailed placement, and can force ECO place and route changes; see Figure 6(a). (This weakens or even obviates the strategy in Figure 1.) The work of [24] proposes heuristics to fix MinIA violations and reduce power with gate sizing, while minimizing placement perturbations that potentially create new timing violations. The proposed methods substantially reduce (by up to 100%) the number of MinIA violations while satisfying timing/power constraints, compared to recent versions of commercial P&R tools. This being said, more complex (intra- and inter-row) cell placement constraints starting at the foundry 10nm node will further intertwine the historically separate tasks of P&R and post-route optimization.



(a) (b)

Figure 6: (a) An example of the minimum implant area (MinIA) violation. The dotted line indicates the minimum width constraint of the V_{t_2} implant layer. The placement of the cell instance c_2 (V_{t_2}) violates the MinIA rule as it is narrow and sandwiched by two cells (c_1 and c_3) that have a different V_t (V_{t_1}). (b) Illustration of temperature reversal effect.

3. NEAR-TERM MITIGATIONS

This section gives a sampling of potential near-term improvements to timing closure enablement and methodology.

3.1 Variation Modeling For STA

The history of timing delay and slew calculation, along with timing variation modeling, traces back to simple lumped-C interconnect models, Elmore's bound on delay in RC trees, the O'Brien-Savarino pi model, k-factor PVT derating, TLF and Liberty NLDM tables, CCS and ECSM current-source models, and onward to more recent variation-aware gate delay and slew models (AOCV, POCV and LVF).

Advanced on-chip variation (AOCV) delay derating tables have been mainstream since the 40nm foundry node. The AOCV table lookup comprehends stage counts of launch path, capture path, and datapath as well as spatial extents (e.g., bounding box diagonal) of clock and data circuit elements (extreme variations are assumed to be less when paths have more stages, or are spread over a smaller region). However, the methodology essentially assumes that all gates are identical and identically loaded. Parametric on-chip variation (POCV, or cell-based POCV) is another methodology to capture per-cell relative variation margin. It improves on AOCV in that stage counts are no longer needed; rather, σ^2 terms are accumulated over a given path [43]. A nascent advance in variation modeling methodology is the Liberty Variation (Variance) Format (LVF) [32] [38], which represents slew- and load-dependent delay, slew and constraint variation per timing arc. (Where the POCV variation model has "one number per cell", LVF is fundamentally different in that it provides "one number per load-slew combination per cell".) [32] and other studies suggest that LVF-based timing analysis has greater accuracy than AOCV/POCV with respect to Monte Carlo SPICE results.[12] The advantage of LVF over previous standards can also be seen in its ability to handle the well-known non-Gaussian distribution of path delay under process variation (Figure 7), via separate delay σ values for late- and early-mode analyses. It may be concluded that LVF-based timing analysis (guiding optimization) of 'true' timing-critical paths offers potential major improvements over OCV-based STA for future timing closure methodology. As noted earlier, there is a possible design turnaround time benefit as well, in that LVF-based closure and signoff can hold back the encroachment of expensive path-based analysis into the PD flow.

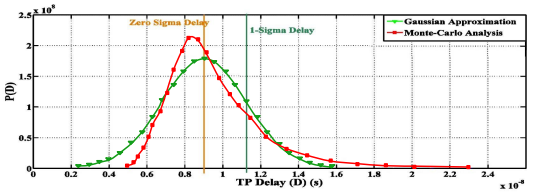


Figure 7: Asymmetry of Monte Carlo path delay distribution, showing the "setup long tail" and motivating separate σ values in the timing model to support late (setup) vs. early (hold) analyses. The zero-sigma delay is the nominal delay. Adapted from [27].

While following the above trajectory, the industry has also for over a decade flirted with full statistical static timing analysis (SSTA). Although SSTA is a 'holy grail' used in production at IBM, is seems to remain perpetually in the future.[13] Another flirtation, Sensitivity SPEF (SSPEF) for statistical modeling of interconnect, seems to have recently dropped by the wayside, leaving BEOL variations as a major hole in signoff enablement (see the discussion of "tightened BEOL corners" below).

3.2 Tightened BEOL Corners

As noted above, BEOL layers at foundry 20nm and below have become major sources of variation. Typically, this is accounted for by signoff using homogeneous, "conventional BEOL corners" (CBCs), such as Cw, Ccw, RCw, Cb, etc. Chan et al. [2] point out the inherent pessimism of signing off with worst-case conditions for all layers, since the per-layer variations are not fully correlated. To quantify pessimism of a given CBC Y_{CBC} in the analysis of a given timing path j, [2] defines a pessimism metric α_j as shown in the following equations. The statistical 3σ worst delay is denoted by $3\sigma_j$ (of course, any other number of sigmas could be used as a delay criterion), and $d_j(Y)$ denotes the delay of path j at corner Y. Note that small values of α imply large pessimism of the conventional BEOL corner for setup analysis.

$$\alpha_j = 3\sigma_j / \Delta d_j(Y_{CBC}) \quad (1)$$

$$\Delta d_j(Y_{CBC}) = [d_j(Y_{CBC}) - d_j(Y_{typ}))] \quad (2)$$

$$Y_{CBC} \in \{Y_{cw}, Y_{cb}, Y_{rcw}, Y_{rcb}\} \quad (3)$$

[10] Scalability of device performance across voltage is also exposed across corners. E.g., at low voltage, critical paths are gate-dominated (net delays comprising only 2-5% of path delay) and may also be dominated by HVT devices. For this case (and, for shorter driven wires), the Cw BEOL corner is dominant. On the other hand, at high voltage, critical paths are wire-dominated (net delays comprising 30-50% of path delay) and may be dominated by LVT devices. For this case (and, for long driven wires), the RCw BEOL corner is dominant. Pruning of corners is difficult!

[11] Implant (active) layers, which define regions for ion implantation, determine the threshold voltage (V_t) of transistors. Traditional timing- and routability-driven placement of cells with multiple V_t values, as well as subsequent sizing and V_t-swap optimization steps, can create a small island of a given V_t that violates the MinIA rule.

[12] [43] points out that the relative margining approach of AOCV/POCV will not provide any margin for a variation hotspot which has nominal delays close to zero.

[13] The litany of practical barriers to SSTA adoption includes (i) the complexity of deployment; (ii) the improbability of foundries committing to statistics; and (iii) the lack of benefit over emerging standards such as LVF that overcome 'relative margining' limitations of AOCV and POCV variation-aware modeling standards.

Figure 8(a) shows the α scaling factors of a set of setup-critical paths, at the Cw corner (Y_{cw}) and at the RCw corner (Y_{rcw}). A red dot is a path which has a larger delta delay at the Cw corner relative to the typical (nominal) corner, and a blue dot is a path which has a larger delta delay at the RCw corner. The left plot shows that some paths have $\alpha > 1$, meaning that the Cw corner actually underestimates the delay increment under variation compared to the statistical analysis. However, these paths have $\alpha < 1$ at the RCw corner, i.e., are "dominated" by the RCw analysis. These results imply that we must sign off at both corners to capture the impact of interconnect variation. But, only paths that do not have large delay increments (relative to nominal delay) at either corner are not pessimistically treated at one corner or the other.

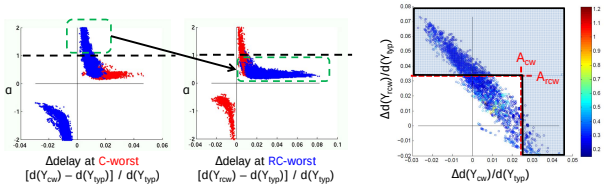


Figure 8: (a) Pessimism metric α_j of different critical paths. The left plot shows analysis at the Cw corner: the x-axis gives Δdelay of paths, i.e., $\Delta d_j(Y_{cw})$ normalized to the nominal path delay $d_j(Y_{typ})$, while the y-axis gives α_j. The right plot shows analogous values for the same paths at the RCw corner. Paths with small Δdelay and large α in the left plot (green dashed box) have large Δdelay and small α in the right plot. (b) Thresholds A_{rcw} and A_{cw} can identify paths amenable to signoff with tightened BEOL corners (TBCs).

Figure 8(b) shows how paths with small Δdelay at both the Cw and RCw corners also have large α. Thus, by setting Δdelay thresholds A_{rcw} and A_{cw}, one may identify paths (blue-shadowed region) that can be signed off safely with tightened BEOL corners (TBCs). As reported in [2], this reduction of pessimism in the BEOL corner methodology substantially reduces timing violations and fix/closure effort.

3.3 AVS-Aware Margin Definition

Over multi-year product lifetimes, adaptive voltage scaling (AVS) is applied to compensate performance degradation (V_t shift) of circuits due to bias temperature instability (BTI) aging. However, this creates a chicken-egg loop in the determination of signoff criteria, since increasing supply voltage (to compensate aging-induced performance degradation) itself accelerates the aging mechanism. Understanding this loop, for purposes of establishing design signoff criteria, has significant implications: (i) underestimation of aging increases lifetime energy consumption due to higher than expected supply voltage levels; and (ii) overestimation of aging increases layout area due to more pessimistic gate sizing to meet performance specifications at signoff. The work of [1] analyzes this chicken-egg dependency and proposes a methodology for aging-aware signoff in an AVS-enabled system; the authors further quantify the power and area overheads due to improper selection of signoff corners. Figure 9 shows that substantial power or area overheads can result from improper choice of aging signoff corner. Additional AVS-awareness is likely to reap benefits when separately applied to clock vs. datapath circuits in signoff.

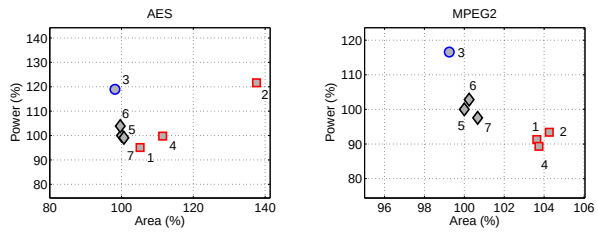


Figure 9: Tradeoff of average power (over 10-year lifetime) versus area, among circuit implementations signed off at different BTI aging corners, assuming DC BTI stress and AVS [1].

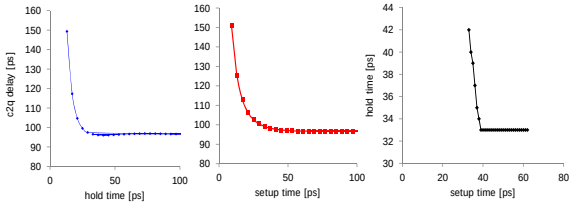


Figure 10: Left to right: (i) c2q delay vs. setup time; (ii) c2q delay vs. hold time; and (iii) setup time vs. hold time.

3.4 Improved Clock Analyses

Conventional STA signoff relies on worst-case assumptions, e.g., late arrival of data, early arrival of clock signal at capture flop, etc. to ensure safe delivery of data to flip-flops. At the same time, excessive pessimism can be mitigated by introduction of *flexible margins*, particularly in clock-related analyses. For example, conventional flip-flop timing models have fixed values of setup and hold times and clock-to-q (c2q) delay; these are characterized via such methodologies as a *pushout criterion* (limiting c2q delay degradation to 10%). However, interdependencies of hold time, setup time and c2q delay on each other are not captured in conventional timing signoff and closure flows. Figure 10 shows (i) c2q delay vs. setup time, (ii) c2q delay vs. hold time, and (iii) setup time vs. hold time from SPICE simulation of a DFQDX flip-flop from a 65nm foundry library. The c2q delay rapidly increases when the setup or hold time is decreased. In the conventional timing analysis enablement, this region is disregarded as a result of the fixed 10% pushout criterion.

Several works propose exploitation of interdependent setup-hold or setup-hold-c2q timing models, e.g., [28] proposes an improved STA that considers variation through use of interdependent setup-hold times. Chen et al. [7] suggest iterative timing analysis based on nonlinear and analytical interdependent flip-flop modeling. Commercial timing analysis tools can also comprehend interdependent setup-hold times to reduce analysis pessimism (cf. "setup-hold pessimism reduction", or SHPR). The recent work of [23] exploits the three-way tradeoff among setup time, hold time and c2q delay to recover "free" margin, essentially by giving flexibility at timing path boundaries. A sequential linear programming optimization across multiple timing corners reduces pessimism in the analysis of setup- and hold-critical paths, and increases worst timing slack by up to 130ps in a 65nm foundry library. Another opportunity to recover clock-related margin is with respect to the jitter margin. As noted above, the clock jitter margin is applied as a flat margin, which is pessimistic consecutive short clock pulses are less likely during circuit operation. Hence, a cycle-to-cycle clock jitter margin can be used to reduce pessimism in future analysis and closure methodology.

4. FUTURES AND CONCLUSIONS

Modern timing closure connects many disciplines and activities: margin definition; model-hardware correlation; variation modeling and testchip/DOE definition[14]; signoff constraint definition; low-power design; EDA tool innovation; design and deployment of (critical path-mimicking) process/aging monitor circuits; awareness of 'new effects' and new device/process/model implications; and coordination of the overall SOC design closure process. Three comments follow.

Comment 1. EDA tool innovation in the timing closure space has been impressive. Designers now have a choice of physically-aware ECO tools (e.g., Dorado Tweaker [33], Synopsys DMSA [35], Cadence EDI) that are congestion- and legal location-aware, and scale well onto hundreds of threads. There is a

[14] [41] notes such open issues as design of testchips that are targeted to model-hardware correlation; minimized DOEs for global and local variation modeling in the BEOL stack; and FEOL testchip design and test methodologies that dramatically increase the number of accessible transistors, or testable DUTs per wafer.

choice of improved variational and statistical modeling and analysis tools (e.g., from Solido [37], or FXM from CLK DA [31]). Signoff STA tools offer improved support of voltage scaling (interpolation across lib groups) and comprehension of dynamic IR effects ('-dynamic' analysis options).

Comment 2. Process and device innovation will continue to challenge timing closure. Oncoming worries include metal fill effects, as density constraints continue to tighten and the freedom to define fill exclude windows (e.g., around clock routes) decreases. How to comprehend "actual" foundry-specific fill early in the design closure process is an open issue that will soon become critical. Process enhancements such as air gaps may help mute impacts of BEOL RC and noise scaling, with associated PD and timing optimizations yet to be developed. FinFET current densities bring self-heating and reliability concerns into performance analyses; higher drive strengths in smaller footprints may cause further placement-optimization interferences, e.g., with fractional-track (7.5T, 8.25T) libraries.

Comment 3. SOC design closure complexity requires around-the-clock effort from globally distributed engineering teams, brutal work schedules, and huge investments in EDA tooling and compute resources. Beyond this, strategies and methodology for timing budgeting, constraints evolution, and coordination of top- vs. block-level effort (and, flat vs. ETM-based/hierarchical analysis and optimization) all affect design schedule and QOR. The ability to handle even a few additional functional ECOs or constraints changes within a 60-day tapeout march can be the difference between market success and failure. Above and beyond this, there can be huge impact from better methodologies and optimizations long before the PD team ever embarks on its tapeout march.[15]

Last, futures might include the following. (1) *General observations.* (i) As margin becomes scarcer, analysis accuracy and model-hardware correlation gain importance. (ii) Model-hardware correlation is progressively weakening, and the traditional model - design kit - P&R flow is inapplicable during early (unstable) stages of a new technology node. This demands fundamentally faster techniques for modeling, characterization and P&R [41]. (iii) Recovery of margin from setup-hold-c2q flexibility, improved signoff corner definition, etc. will have increased value as fewer such "mitigations" remain on the table. (2) *Rise of BEOL and MOL.* (i) BEOL and MOL will become "first-class citizens", with increased mindshare in variation modeling and signoff corner definition (even, in variation-aware path-based STA). (ii) Improved library, placement and routing strategies for restricted (SADP/SAQP) BEOL patterning in FinFET nodes will be needed. (3) *Variation modeling and analysis.* (i) Statistical SPEF or similar will be revived (cf. "BEOL as first-class citizen"). (ii) LVF or similar will replace 'relative margin'-based OCV formats; non-Gaussian variance models will enter standard use. (iii) Hopefully, progress toward a unified model of PVT variation (FEOL, BEOL, voltage, temperature) will be made, with unification of process variation and voltage variation being the first step. (4) *Signoff.* (i) AVS (and/or, PVS-like [2]) process adaptivity will be widely adopted, along with typical-plus-flat-margin strategies for closing setup with reduced pessimism. (ii) Design-specific tightened corner methodologies for both BEOL and FEOL can improve PPA as well as schedule. (iii) Cross-corners (FSG, SFG), already required for clock network analysis, will further permeate the timing closure process. (iv) Improved methods for reducing the number of timing libraries or library variants will be needed. (5) *3D integration.* New 3DIC-specific timing closure challenges will include (i) (partitioning, clocking interface design methodology to avoid) variation-aware analysis across multiple die; (ii) closure of power integrity and thermal loops with timing analysis; and (iii) variability-mitigating optimizations.

[15] Should the methodology include deskewing buffers? Hysteresis flops? On-chip regulators? How should maxcap, max fanout, and maxtrans constraints evolve as the design progresses from physical synthesis through post-route optimization? Etc. With regard to optimization, future ability to achieve timing closure will demand such innovations as (i) optimization of the top-level clock plan [10] or useful skew [6]; (ii) improved layout-dependent effect-aware placement and timing-driven routing; (iii) explicitly process variation-aware optimization; or (iv) late-stage optimization that can be driven effectively by path-based timing analysis.

Acknowledgments

I thank Rob Aitken for the invitation to write this paper, and Christian Lutkemeyer, Isadore Katz, Sorin Dobre, Tuck-Boon Chan, Kwangok Jeong, Nancy MacDonald and John Redmond for helpful discussions and inputs, a number of which have been incorporated here. At the UCSD VLSI CAD Laboratory, Hyein Lee and Jiajia Li, along with Mulong Luo, Yaping Sun and Wei-Ting Jonas Chan, provided invaluable help with pulling everything together in the usual compressed time frame.

5. REFERENCES

[1] T.-B. Chan, W.-T. J. Chan and A. B. Kahng, "On Aging-Aware Signoff for Circuits with Adaptive Voltage Scaling", *IEEE Trans. on CAS-I* 61(10) (2014), pp. 2920-2930.

[2] T.-B. Chan, S. Dobre and A. B. Kahng, "Improved Signoff Methodology with Tightened BEOL Corners", *Proc. ICCD*, 2014, pp. 311-316.

[3] T.-B. Chan, P. Gupta, A. B. Kahng and L. Lai, "Synthesis and Analysis of Design-Dependent Ring Oscillator (DDRO) Performance Monitors", *IEEE Trans. on VLSI Systems* 22(10) (2013), pp. 2117-2130.

[4] T.-B. Chan, A. B. Kahng, J. Li, S. Nath and B. Park, "Optimization of Overdrive Signoff in High-Performance and Low-Power ICs", *IEEE Trans. on VLSI Systems* (2014).

[5] T.-B. Chan and A. B. Kahng, "Tunable Sensors for Process-Aware Voltage Scaling", *Proc. ICCAD*, 2012, pp. 7-14.

[6] T.-B. Chan, A. B. Kahng and J. Li, "NOLO: A No-Loop, Predictive Useful Skew Methodology for Improved Timing in IC Implementation", *Proc. ISQED*, 2014, pp. 504-509.

[7] N. Chen, B. Li and U. Schlichtmann, "Iterative Timing Analysis Based on Nonlinear and Interdependent Flipflop Modelling", *IET Circuits, Devices & Systems* 6(5) (2012), pp. 330-337.

[8] R. Ginosar, "Fourteen Ways to Fool Your Synchronizer", *Proc. Async*, 2003, pp. 89-96.

[9] M. Gupta, K. Jeong and A. B. Kahng, "Timing Yield-Aware Color Reassignment and Detailed Placement Perturbation for Bimodal CD Distribution in Double Patterning Lithography", *IEEE Trans on CAD* 29(8) (2010), pp. 1229-1242.

[10] K. Han, A. B. Kahng, J. Lee, J. Li and S. Nath, "A Global-Local Optimization Framework for Simultaneous Multi-Mode Multi-Corner Skew Variation Reduction", *Proc. DAC*, 2015.

[11] K. Han, A. B. Kahng and H. Lee, "Evaluation of BEOL Design Rule Impacts Using an Optimal ILP-Based Detailed Router", *Proc. DAC*, 2015.

[12] *International Technology Roadmap for Semiconductors*, Design Chapter, 2013. http://www.itrs.net/

[13] K. Jeong, "Variability Assessment and Mitigation in Advanced VLSI Manufacturing Through Design-Manufacturing Co-optimization", *Ph.D. Thesis*, UCSD ECE Dept., 2011.

[14] 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.

[15] K. Jeong, A. B. Kahng and K. Samadi, "Impacts of Guardband Reduction on Design Process Outcomes: A Quantitative Approach", *IEEE Trans. on Semiconductor Manufacturing* 22(4) (2009), pp. 552-565.

[16] A. B. Kahng, "The Road Ahead: Shared Red Bricks", *IEEE Design and Test*, March-April 2002, pp. 70-71.

[17] A. B. Kahng, "Opportunities in Future Physical Implementation and Manufacturing Handoff Flows", *Proc. ISOCC*, 2007, pp. 46-50.

[18] A. B. Kahng, "The Road Ahead: The Future of Signoff", *IEEE Design and Test*, May-June 2011, pp. 86-88.

[19] A. B. Kahng, "The Road Ahead: Roadmapping Power", *IEEE Design and Test*, Sept.-Oct. 2011, pp. 104-106.

[20] A. B. Kahng, "DfX and Signoff: Challenges and Opportunities", presentation (*ISVLSI*, 2012), http://vlsicad.ucsd.edu/Presentations/talk/ISVLSI-2012-Kahng-final-distributed.pdf

[21] A. B. Kahng, "Toward Holistic Modeling, Margining and Tolerance of IC Variability", *Proc. ISVLSI*, 2014, pp. 284-289.

[22] A. B. Kahng, S. Kang, J. Li, and J. Pineda de Gyvez, "An Improved Methodology for Resilient Design Implementation", *ACM TODAES* (2015).

[23] A. B. Kahng and H. Lee, "Margin Recovery with Flexible Flip-Flop Timing", *Proc. ISQED*, 2014, pp. 496-503.

[24] A. B. Kahng and H. Lee, "Minimum Implant Area-Aware Gate Sizing and Placement", *Proc. GLSVLSI*, 2014, pp. 57-62.

[25] C. Lutkemeyer and P. Ghanta, "Modeling Slew Dependent Constraint Arc Variation in Static Timing Analysis", *Proc. TAU*, 2014.

[26] C. Lutkemeyer, "A Practical Model to Reduce Margin Pessimism for Multi-Input Switching in Static Timing Analysis of Digital CMOS Circuits", *Proc. TAU*, 2015.

[27] R. Rithe, J. Gu, A. Wang, S. Datla, G. Gammie, D. Buss and A. Chandrakasan, "Non-linear Operating Point Statistical Analysis for Local Variations in Logic Timing at Low Voltage", *Proc. DATE*, 2010, pp. 965-968.

[28] E. Salman and E. G. Friedman, "Utilizing Interdependent Timing Constraints to Enhance Robustness in Synchronous Circuits", *Microelectronics Journal* 43(2) (2012), pp. 119-127.

[29] Y.-S. Su, W.-K. Hon, C.-C. Yang, S.-C. Chang and Y.-J. Chang, "Clock Skew Minimization in Multi-voltage Mode Designs Using Adjustable Delay Buffers", *IEEE Trans. on CAD*, 29(12) (2010), pp. 1921-1930.

[30] N. D. MacDonald, "Timing Closure in Deep Submicron Designs", *DAC.com Knowledge Center Article*, March 2010. http://vlsicad.ucsd.edu/DAC15/MACDONALD_TIMINGCLOSURE.pdf

[31] "CLK Design Automation." http://www.clkda.com/

[32] CLK Design Automation, "A Brief Introduction to Liberty Variance Format – LVF", *white paper*, http://www.clkda.com/, April 2015.

[33] "Dorado Design Automation." http://www.dorado-da.com/

[34] "Synopsys IC Compiler User Guide."

[35] "Synopsys PrimeTime User's Manual."

[36] "Synopsys HSPICE User Guide."

[37] "Solido Design Automation." http://www.solidodesign.com/

[38] "Open Source Liberty." http://opensourceliberty.org/

[39] T.-B. Chan, *personal communication*, March 2015.

[40] S. Dobre, *personal communication*, March 2015.

[41] K. Jeong, *personal communication*, April 2015.

[42] I. Katz, *personal communication*, April 2015.

[43] C. Lutkemeyer, *personal communication*, March 2015.

[44] N. MacDonald, *personal communication*, April 2015.

[45] J. Redmond, *personal communication*, March 2015.