The Past, Present, and Future of Design-Technology Co-Optimization

Greg Yeric[1], Brian Cline[1], Saurabh Sinha[1], David Pietromonaco[2], Vikas Chandra[2], and Rob Aitken[2]
Research and Development, ARM
[1]Austin, TX, USA, [2]San Jose, CA, USA
greg.yeric@arm.com

***Abstract*— Design-Technology Co-Optimization (DTCO) has evolved from early Design-for-Manufacture (DFM) needs into a multi-faceted, multi-lateral co-optimization below 20nm, where multiple patterning and FinFETs add significant complexities. Effective DTCO now involves end product metrics applied to a myriad of design-technology choices. This paper will highlight past and present examples of DTCO in practice for low-power SoC design, and examine a future of even more complexity that will drive a continued evolution in DTCO.**

I. Introduction

In past technology nodes, foundries' continued ability to deliver process node scaling meant everyone except the foundries could remain complacent regarding Moore's Law. As conventional transistor and lithography scaling began to hit significant difficulties, other avenues had to be exploited. Design-Technology Co-Optimization (DTCO) has become an increasingly key component enabling scaling entitlement for advanced process nodes. But like Design-for-Manufacturability (DFM) before it, the term means many things to many people, and in practice it is a moving target. Today's technology choices are increasingly complex, and reaching the optimal result necessarily involves multiple parties in early communication. By examining the evolution from past to present, we can better understand DTCO's positive effects and identify future opportunities.

II. The Past: From Optimization to Co-Optimization

The disaggregation of the semiconductor industry fostered the growth of independent fab, fabless, IP, EDA, and packaging companies. In simpler technology nodes, advance communication between these entities was not required to produce sufficient technology scaling. Fabs produced Process Design Kits (PDKs) including design rules and transistor models, and products scaled accordingly.

Beginning in earnest around the 90nm node, various yield concerns resulted in (DFM) initiatives [1], which attempted to prescribe restrictions to designers that would maximize yield. Communication was primarily unidirectional, from fab to designer, in the form of increasingly restrictive design rule checks (DRC). Designers (of physical IP, for instance), would then optimize their products within the bounds of the PDK. After IP was created, end product development would begin with synthesis, place and route flow.

This past dynamic is depicted in Figure 1. Two to three years before the year of production (YOP), possible technology choices are narrowed and the target process development begins. After the process stabilizes and is characterized, a full PDK is issued, allowing IP to be created and then used to implement early designs in the yield ramp stage. The primary scaling decisions have already occurred.

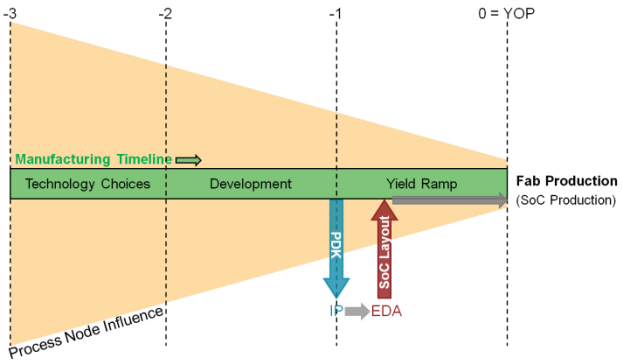

Figure 1: Manufacturing technology development timeline. Year of Production (YOP)=0. Conventional interactions with IP and EDA also noted.

A. Distant past: DFM below 90nm

Restrictive design rules began to emerge in earnest in response to the 65nm node. Physical design engineers continued to rely on historical layout constructs which allowed them to minimize standard cell area, but many of these constructs became increasingly difficult to manufacture at 65nm and below in sub-wavelength lithography. Figure 2 shows layout for a flip-flop typical of the 65nm era, with key area-saving constructs (e.g. "outer channel" poly routes, which provide connectivity between disparate areas of the flop without using metal wires).

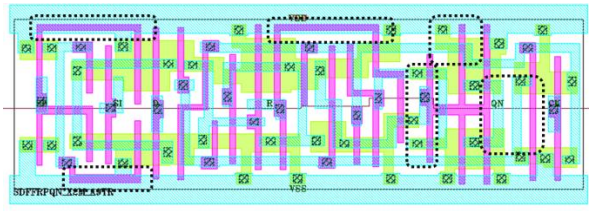

Figure 2: 65nm-era flip-flop. Key area-saving constructs are highlighted.

At 45/40nm and 32/28nm, with no wavelength scaling, increasingly stringent lithography restrictions made many of the area saving layout constructs of the past illegal. A 28nm-era flip-flop layout is shown in Figure 3. Comparing the similar circuit implementations in Figure 2 and Figure 3, the same layout required more area (horizontal gate pitches) to implement, as well as wiring on metal 2. This result can be

978-1-4673-6146-0/13/$31.00 ©2013 IEEE

considered a cost of regularity. A primary issue was dipole printing of 28nm gate pitches, requiring gate poly in only one preferred direction.

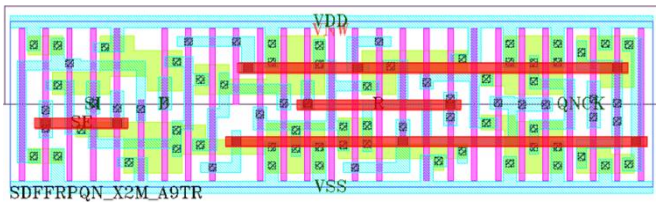

Figure 3: 28nm-era flip-flop. More gate pitches (area) required to implement the same function as compared to 65nm.

This eliminated all of the useful layout constructs shown in Figure 4, including out-bound poly routes, parallel gate connections, offset gate contacts, and non-uniform gate CD and non-uniform pitch poly. Thus, logic scaling from 65nm to 28nm could not scale to the entitlement predicted by the metal pitch and gate pitch scaling.

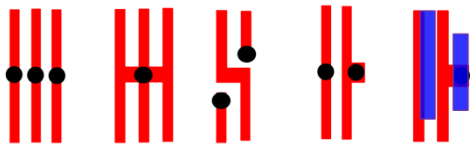

Figure 4: 65nm-era area-savings constructs that were lost to 28nm preferred gate orientation design rules.

B. *Past: Beyond Standard Cell Area*

A problem with the state of DTCO in the 65nm-28nm era was that the final cost of the design restrictions was not known during their definition. IP designers could provide an estimate of flip-flop size increase, but the size of standard cells was not sufficient information, as it had been in the past. For example, each of the gate layer constructs shown in Figure 4 had to be replaced with metal wiring. The increase in metal within the cells could reduce yield according to critical area, but more importantly the pin access (ability of a router to connect to the pins of the standard cell) qualities of the standard cells had to be compromised. In many cases, such as the flip-flop shown in Figure 3, metal 2 wiring had to be employed (red lines) where none was required in previous technology nodes. Metal 2 use inside the cells further restricts routing and decreases implemented block utilization. However, getting to these answers requires complete library construction and well as optimized synthesis, place, and route, and the time line of Figure 1 did not allow for that.

Another example impetus in the evolving DTCO conversation came from the increased strain required to compensate for lack of physical gate length scaling below 90nm. A point was reached that was counter-intuitive at first glance: Larger standard cells produced smaller circuits. With strain, using a gate pitch larger than the minimum resulted in faster standard cells. This performance increase often more than compensated for the area penalty when implemented into logic blocks.

An example from 28nm test chip data is shown in Figure 5. Gate pitches P1-P5 represent minimum to continuously larger pitch. In this case, the intermediate pitches P3 and P4 resulted in performance gains high enough to offset the standard cell size increase in many implementation cases. Higher drive transistors allow downsizing of gates and reduced repeater insertion, resulting in smaller implemented blocks.

However, given the constraints of implementing deep sub-wavelength lithography, decisions involving the gate pitch had been made much earlier, and in many cases could not be reversed. This issue helped underscore the value of using early learning from the implementation of logic blocks back to inform early technology development. In order to optimize the gate pitch, fabs must model and confirm in hardware the transistor-level characteristics (Miller capacitance, mobility) but without understanding the effect of technology choices through the product design flow may not result in the best end result.

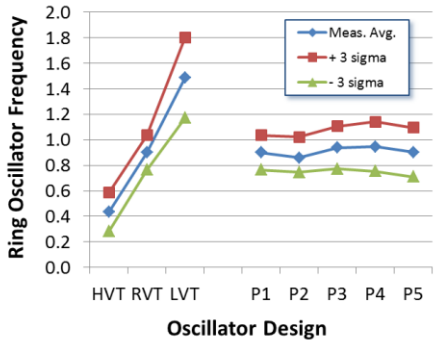

Figure 5: 28nm-era ring oscillator frequency as a function of gate pitch (right-side). Delay with varying VT type is shown in the left for comparison.

C. *Past: Deep sub-λ Lithography and IP*

As the industry continued to scale pitch, the constant lithography wavelength (193nm) forced manufacturers to become much more aggressive with Optical-Proximity Correction (OPC) techniques, including Off-Axis Illumination (OAI) and sub-resolution assist features (SRAFs). This resulted in highly non-linear effects in layout, including significant influence of shapes beyond immediate proximity, which created significant challenges in the creation of standard design rules that would enable robust designs. This dynamic gave rise to the use of lithography simulators as additional design rule checks (Figure 6). By enabling such simulations of logic layout, a feedback loop with designers allowed inspection of local topologies and OPC interactions enabled tuning of OPC to provide higher density of common standard cell constructs.

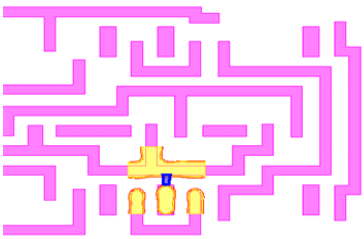

Figure 6: 28nm-era M1 layout and lithographic contour analysis example

Of course, no one can comprehend all possible end topologies and their printability. Thus, it became prudent to test printability schemes using end-user (IP design) layout styles much earlier in the technology development timeline.

Figure 7 shows a successful use case in test chip silicon, showing the identification of an OPC escape identified in implemented standard cell layout [2].

This placed further value on earlier two-way communication between fab and IP designer, before technology choices were set. This two-way collaboration, pushed into the technology development phase, is depicted in Figure 8 (as contrasted to Figure 1).

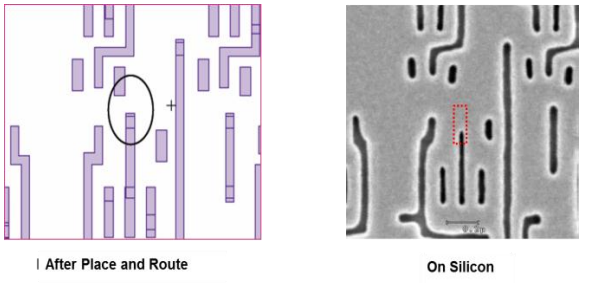

Figure 7: Example of standard-cell like design-OPC interaction failure identified on a printability test wafer.

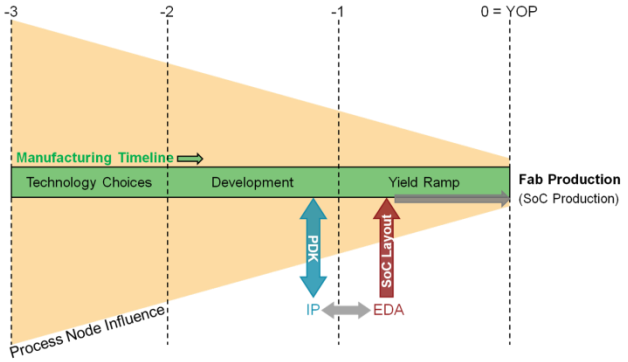

Figure 8: Evolution of DTCO to include 2-way feedback during the technology development phase.

III. The Present: Early, Multi-Lateral DTCO

The technology below 28nm represents an inflection point in the application and benefit of DTCO. 20nm introduced double patterning of the metal layers, and the 16/14 nodes introduced FinFETs. Each of these issues added new levels of complexity into design/technology co-development which required early evaluation of higher level design metrics in order to achieve desired product scaling.

A. FinFETs

The transition to FinFETs below 20nm provided a key paradigm shift in DTCO. In previous technologies with relatively un-quantized device widths, a standard cell designer could receive M2 pitch and device characterization from a PDK and then independently determine which cell height(s) (in number of M2 tracks) represented the optimal end result for standard cell library(ies). The additional constraint of fitting a discrete number of fins within a cell changed this.

The existing paradigm can remain if the fin pitch equals the M2 pitch. However, in low-power standard cells there was significant pressure to reduce the fin pitch to below the M2 pitch. To understand this impetus, consider that not all fin tracks are available for active transistors. Power rail connections at the top and bottom of the cell force the removal of 1 fin each, and typically 2 additional fin tracks must be removed in the center of the cell to accommodate gate input connections, which (as of today) are not allowed over the active diffusion regions. Because low power standard cell libraries have historically been 7-9 M2 tracks tall, if fin pitch equaled M2 pitch, there would be too few active fins available. This is illustrated in Figure 9. The drawing on the left shows the example of an 8-track standard cell, where only 4 active fins would remain. That results in only 2 fins per FET (PMOS and NMOS), which would not be acceptable for performance, but perhaps more importantly creates unacceptable quantization of device strength (only one device tuning option would remain, a -50% option). Figure 10 shows a circuit tuning example for a conventional, planar technology. Continuous (non-discrete) tuning of device widths allows the devices to achieve optimum power/performance. Discrete tuning between 1 and 2 fins only would render moot any real device tuning and result in higher power implementations.

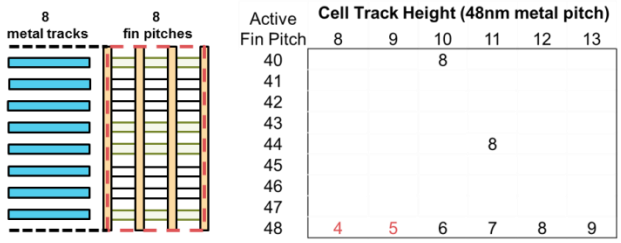

Figure 9: With multi-gate devices, an integer number of fin pitches must fit within an integer number of metal tracks, defining the standard cell height (left). The table on the right lists the total number of active fins (total fins minus 4) in a standard cell for various fin pitches, if a solution exists.

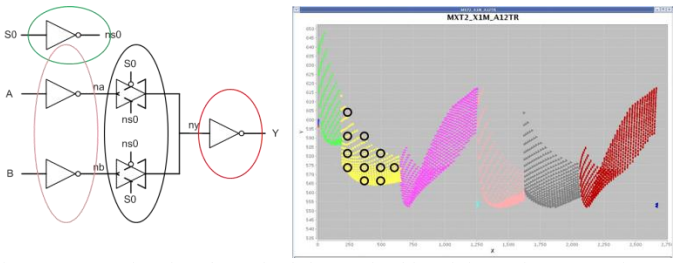

Figure 10: Devices in schematic (left) are sized in minimum increments in order to find global circuit optimization. The right graph records cell delay as a function of tuning experiment, for the case of planar (non-multigate) technology.

As shown in the table on the right of Figure 9, there are limited solutions to the number of fins that can fit in a standard cell height, (only 2 integer solutions exist when fin pitch does not equal the metal pitch). One could populate the table with more solutions by allowing slight adjustments to specific fin locations. But this fine tuning is necessarily tightly linked to standard cell layout evaluation, including power rail construction, transistor contacting, etc., in order to determine exactly which fin configuration results in the best marriage to contacts and wire design rules for most cells. Answers to sufficient accuracy must evaluate the aggregate effect on hundreds of key cells, many of which are quite complicated to construct (various And-Or-Invert, Flip-Flops, etc.). The fin quantization era means the lead time required to create fairly complete cell libraries must be factored in while the fin patterning image is being determined, which is much earlier in the technology development cycle than before.

As an added complexity, multiple standard cell heights are typical for a technology. Generally speaking, there is a minimum size standard cell that is rational within a technology, and a larger cell size that represents the maximum performance available. SoC implementations can mix and match these minimum area and maximum performance cell libraries in various logic blocks to optimize power and performance. Thus, complex evaluations of IP become part of the discussion represented by Figure 8.

Fin patterning below 20nm does not escape problems with deep sub-wavelength lithography. For example, standard cell designers prefer to taper devices in order to optimize the power and performance. To support device taper below 20nm may involve additional mask layers (cuts). Thus, the situation in Figure 11 must be addressed: Increase standard cell area when taper is required (due to additional spaces between active regions of varying width) or increase wafer cost. Issues such as this bring the end customer more tightly into the discussion, because the right answer is a combination of two suppliers' metrics: Wafer cost and IP area.

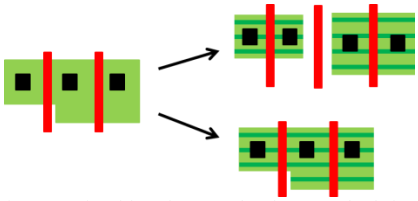

Figure 11: Device taper in older planar technology on the left, and choices for sub-20nm FinFET on the right.

Effects related to process, voltage and temperature (PVT) variability can also be significantly different with FinFETs. Fin width variation will affect circuits in new ways, because it should be more locally correlated than random dopant fluctuations and LER, but also because it directly affects drain-induced barrier lowering (DIBL) and sub-threshold slope [3] and should not be modeled by simple V_T shift. Voltage scaling, above and below the nominal supply point, is an important design lever, and FinFETs can provide extra benefit in both directions. Additionally, MOSFET inverted temperature dependence (ITD) [4], which has been increasingly penal in recent nodes, may be ameliorated with proper FinFET construction.

B. *SRAM and Design assist*

SRAM bitcells have historically been optimized with specific transistor size ratios that do not fit the integer world of FinFETs. The minimum area bitcell, with one fin each in the pull down, pass gate, and pull up devices in a bitcell (often referred to as a "111" bitcell), is inherently less stable than a traditionally-sized bitcell.

At the memory instance level, this means that design assist techniques which were optional in the past can be required with FinFETs below 20nm [5]. While FinFETs can provide better device matching than planar devices, owing to lower channel doping when properly constructed, Pelgrom's law makes life more difficult at each successive node [6].

Multiple bitcell types are usually offered (minimum area vs. higher performance options, for example). This is depicted in Figure 12, showing a "111" bitcell next to a "221" bitcell. Additional larger cells can be more stable, but an additional complexity arises for the memory designer. Due to the discrete transistor sizing, some cell types will benefit from certain types of design assist but others could be degraded by the same assist method. Careful matching of bit cell with assist methodology is increasingly a key part of memory compiler design.

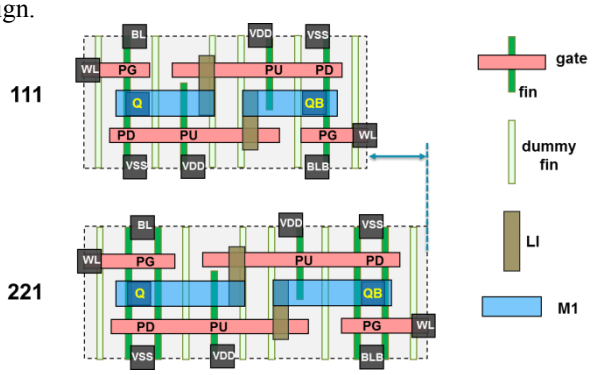

Figure 12: Minimum "111" fin bitcell compared to larger potentially more stable 221 bitcell.

C. *Multiple Patterning*

Beyond 28nm, the required pitches forced the use of multiple patterning [7]. While a line/space grating can achieve the "peak" density offered by double patterning lithography (DPL), the average density in implemented standard cells ranges from this peak value toward the single-mask density. Figure 13 shows example layout of a cell with a high pin density and the set of shapes requiring resolution of a two-color conflict (where the two masks in double patterning are referred to as being different "colors"). In many cases such as this, there is no solution except to increase cell size, and the full entitlement of the pitch scaling is not achievable [8].

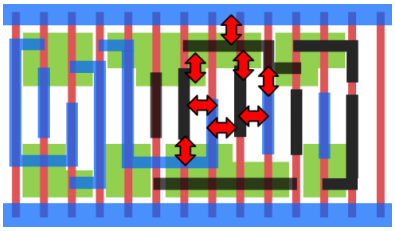

Figure 13: Example of complex DPL coloring conflict in standard cell.

Consider the simplified standard cell areas of Figure 14(a), in a two color layout. Shapes within a cell must not conflict with the power rail at the top and shapes from arbitrary neighboring cells on the left and right. That would result in a two-color loop conflict, often called an "odd cycle" [9]. However, maximally dense standard cell layout requires exactly this. One straightforward fix would be to add white space both vertically and horizontally as in Figure 14(b), but this is typically illegal due to contacting rules unless the cell width is extended by an additional poly pitch at every point where horizontal space needs to be added. Another option is to maintain the horizontal density but then require placement restrictions, as shown in Figure 14(c) [10]. This latter option may result in the best tradeoff, but that result requires comprehension of placer capabilities during technology definition, thus this scenario highlights the need for early EDA

involvement in the technology definition, especially with regard to multiple patterning optimization.

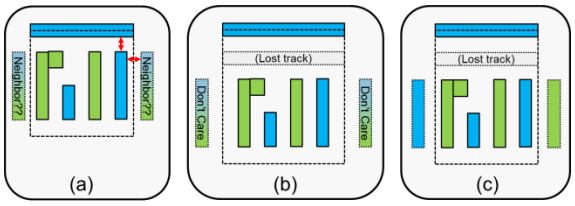

Figure 14: Standard cell DPL color conflict (a) and potential fixes. Fix (b) sacrifices area at left/right cell edges to make placement of abutting cells color-insensitive. Fix (c) does not increase cell size horizontally but requires restrictions on cell abutment combinations.

An additional issue at the 20nm node is the introduction of local interconnect (LI) layers between the transistors and M1. Contact layers became rectangular as dipole lithography was required to meet contacted gate pitch requirements. The rectangles provided a benefit in recapturing diffusion tabs (convex corners), one of the lost constructs shown in comparisons of Figure 2 and Figure 3, allowing transistor source/drain regions to be connected to the outside metal rails. In Figure 3 you see the diffusion tabs of Figure 2 replaced with metal tabs, which then blocked other routes from using those outer tracks. Additionally, by adding a second orthogonal local interconnect patterning layer, the process was able to support some of the key gate constructs without using M1, as depicted in Figure 15 [11].

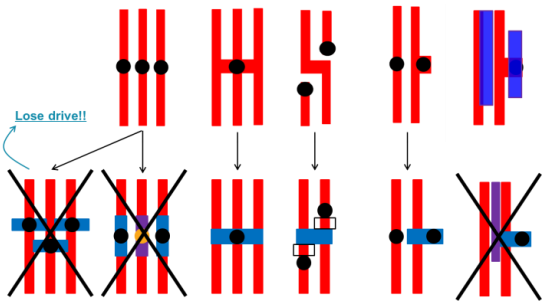

Figure 15: Key standard cell constructs, in 65nm-era 2D poly style (top) and in 20nm DLP/LI style (bottom).

The cost/benefit analysis of specific local interconnect and double patterning options requires understanding beyond simple pitch scaling, and even beyond the examination of a few standard cells. Consider the question of whether or not to change a layer from single to double patterning. If that were to add 2-3% to the wafer cost, then the resulting effect on final block area scaling must be determined to at least this accuracy, in order to provide meaningful feedback. Because these patterning details have varied effect on cell route-ability, one must build a fairly complete set of standard cell in order to be able to accurately calculate the aggregate result. Commensurately, the effects of these technology choices could not be evaluated on physical IP without a detailed understanding of EDA tool efficiency in these technologies. Parasitic variation due to misalignment of wires on different masks must also be taken into account [12]. These double-patterning issues further drove DTCO to the current paradigm depicted in Figure 16, where the development timeline includes multiple learning cycles involving fab, IP, and EDA.

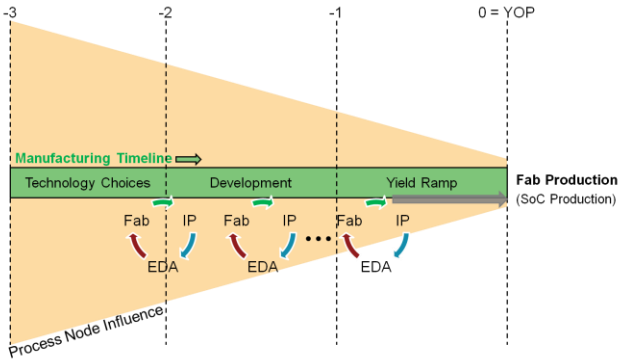

Figure 16: Contemporary DTCO, with multiple learning cycles involving close collaboration between fab, IP, and EDA.

DTCO of this nature, which begins more than 2 years prior to production, is currently targeted to the 10nm node. Given the lack of production-ready EUV at 10nm, at least in the development period, DPL cannot provide adequate pitch scaling, and therefore triple patterning lithography (TPL) may be required. Triple patterning coloring conflicts at 10nm cannot be contained/moderated by the solutions depicted in Figure 14, and furthermore TPL DRC is NP complete [13]. Thus, block level coloring (also known as decomposition) and TPL-aware placement and routing is a primary co-optimization concern. DRC, cell layouts, and decomposition algorithms must all be developed concurrently. To further complicate matters, multiple patterning must now be extended into the local routing layers, adding router algorithms to the above set of concerns. 10nm patterning, without EUV, promises to be significantly more complex than 20nm.

IV. The Future: An Expanded DTCO Ecosystem

The increasingly complex technology choices discussed in the previous section represent merely the beginning of an inflection point in the semiconductor industry. As Figure 17 attempts to show qualitatively, FinFETs and DPL/TPL were just the beginning of intensified technological change. Continued attainment of scaling entitlement will require accelerated change and heterogeneity to the technology development landscape.

A prime example is the transistors themselves. The planar MOSFET was able to last 4 decades, but the silicon FinFET may last only 2 technology nodes. Active development for replacement devices includes mobility enhanced devices such as Quantum Well FETs (QWFETs), both horizontal and vertical nanowires (HNW, VNW), a host of 2D semiconductors, carbon nanotubes (CNT) and even non-field-effect devices. It is likely that multiple foundries will evaluate (different) multiple devices across multiple nodes in the near future.

The patterning roadmap portends a future of mixed-lithography choices. Triple patterning lithography (also known as LELELE) may coexist with self-aligned double patterning (SADP), and EUV, and then beyond the 10nm even multiple

patterning versions of EUV. Augmentations at specific layers may come from direct write e-beam (DWEB) and/or directed self-assembly (DSA). All of these technologies must be evaluated in specific use cases in order to accurately understand the tradeoffs. Lithographic restrictions in the M1 layer will affect the efficiency of the place and route, which itself will be considering tradeoffs between gridding restrictions and runtime [14].

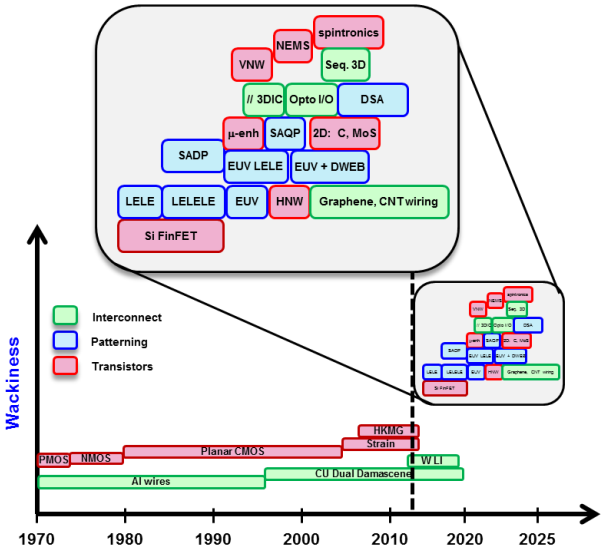

Figure 17: Qualitative assessment of past and future device complexity.

Cost of ownership is a primary concern regarding EUV, and that is driven by throughput, which is limited by source power. But simple wafer cost is not the complete view. The EUV source power issue is in a sense multiplicative, because Line Edge Roughness (LER) is inversely proportional to source power. If source power remains low, LER will remain higher, increasing variability that must be dealt with in design at the expense of circuit power, performance, and/or area.

EUV may present additional value at the design level by supporting fewer lithography-related design restrictions, possibly turning back the clock on some of the trends discussed above. A potentially significant example relates to M1 routing. As lithography progressed into deeper sub-wavelength nodes, at some point routers had to simply ignore M1. This is because the rules governing M1 became too complex, and routers are evaluated in large part by run time. In not allowing M1 route, block sizes can increase by up to 20%, or a significant fraction of a full technology node. If EUV were to allow M1 design rules to be simplified to the point where routers could once again utilize that layer, this would add to the EUV value statement.

If LER can be reduced, EUV could possibly be used to enable more complex active and poly shapes, which may allow for a reversal of the trend depicted in Figure 2 and Figure 3 and ease the pressure on the metal layers.

A possible design-technology limitation is EUV flare, which may need to be addressed via expanded, and possibly variable [15], transition regions between areas of differing pattern density. This will increasingly affect the density of memory arrays, as parasitic limitations continue to push memory toward smaller instances, and smaller bitcell sub-array blocks within memory instances, as seen below:

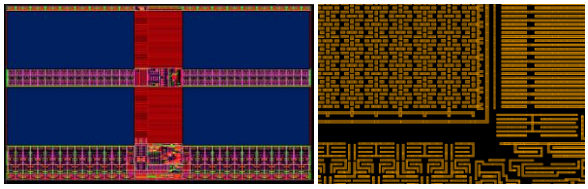

Figure 18: Example large SRAM array with bitcells broken into sub-arrays and surrounded by up to 50% periphery (left). At right, array-periphery transition regions in M1.

In the following sections, we consider the inflection point of increasing technology complexity shown in Figure 17. We expect the following trends in DTCO to promote continued scaling.

A. DTCO ecosystem expands along conventional axis

The need to quantify end product metrics earlier and earlier in the process has added communication links that were previously non-existent. Research consortia and universities working on fundamental device and process R&D now regularly partner with fabless companies in order to quantify and validate fundamental technology choices. Technology evaluation through to the design level has also extended to equipment developers, as the processing challenges have become more intertwined with equipment characteristics. An example is given in Figure 19, showing ARM standard cell layout image tests on ASML EUV lithography tools.

Figure 19: Example of 10nm-node standard-cell like image on EUV printability test wafer. From [1]

End product cost is no longer neatly compartmentalized in fab and design buckets. The fab may consider many possible multiple patterning solutions, all with different costs, and different effects on IP scaling. Wafer cost differences might be rather straightforward to quantify, but their results on implemented designs are not. The choice between bulk and SOI wafers is one example. Improved device isolation and/or performance may compensate for an increase in wafer cost. Because specific patterning options have different effects on the scaling of specific cells, the design cost (in terms of total implemented chip area) is a statistical function of the micro-architecture, the standard cell library, and the performance target. This then necessitates end customer needs to be considered in order to accurately quantify options and choose the best technology path.

As an example, consider two design implementations of the same architecture and IP libraries, but with two different performance targets. These two implementations can arrive at very different conclusions regarding the optimum design-technology tradeoffs. Figure 20 illustrates this dynamic, using

the example of varying transistor V_T option. For lower frequency targets, cells with different performance (here via V_T option), will converge to a minimum implementation area. As the performance target is increased, the synthesis, place and route flow will include larger drive strength cells (via transistor folding), and/or more buffer insertions. It will need to do this more aggressively with the lower performance cells (higher V_T). Thus, customer with a higher performance target can see a much larger benefit to any technology choice that increases the cell drive strength. This same differentiation would arise from reduced device size to accommodate staggered gate contacts as depicted in the lower left case of Figure 15. As shown in Figure 20, the difference in block area can be much greater than one or two additional mask steps, but is variable.

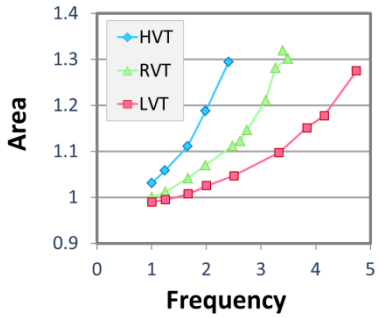

Figure 20: Implemented area as a function of target frequency and V_T option.

In the future, results varying by product target may put pressure to bifurcate process choices in order to better meet criteria of different end customers. This is not a new paradigm, in the sense that older technologies provided distinctly different "HP" and "LP" transistor integrations. The difference for future technologies would be that this concept would extend to patterning choices and other fundamental process concepts.

B. "More than Moore" brings more ecosystem expansion

Packaging, especially 3D IC technologies are expected to be a key enabler of future scaling progress. Opportunities lie in re-structuring memory and logic topology and in mixing different process technologies within the same final "IC". The value of the latter concept will be of increasing value, as supporting specific I/O requirements, including ESD, has become less and less attractive in advanced technology nodes. This might have occupied 10% of chip area in the past but can easily occupy 20% now, which will be compounded by the anticipated poor scaling of wafer cost into the future. Adding a lower-cost layer (that might allow better I/O performance anyway) is a tradeoff that will clearly be considered as cost-effective 3DIC comes online. Therefore, packaging entities will be an increasingly important part of future DTCO. Added device options, including opto-electronic I/O and embedded non-volatile memory could also be considered.

C. Reliability as added DTCO dimension

As continued technology scaling faces increased difficulty in achieving desired progress, the DTCO ecosystem will need to continue to uncover additional areas of scaling opportunity. Reliability awareness should become an integral part of the complete technology offering to meet the power, performance and area requirements. Design for reliability is a perfect example of DTCO, where close collaboration of innovation in devices, process, materials and circuit design can improve overall scaling as compared to past implementations (which generally simply guard-band around the unknown).

Reliability mechanisms which pose challenges to future scaling include Bias Temperature Instability (BTI), Time-dependent Dielectric Breakdown (TDDB), Hot Carrier Injection (HCI), Electromigration (EM), Soft errors and Random Telegraph Noise (RTN) [17]. Figure 21 depicts the increased pressure on reliability with continued technology scaling. A key enabler will be providing accurate reliability modeling in the early stages of technology assessment.

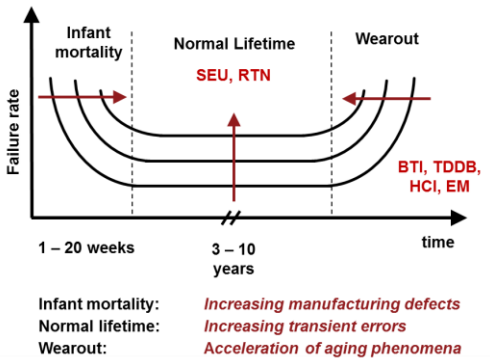

Figure 21: Reliability mechanisms becoming worse with technology scaling

Bias temperature instability, which relates to formation of traps in the gate oxide, is fairly well understood today, but may fundamentally change as entirely new channel materials are employed. Unlike other aging effects, BTI can be partially offset by "healing" – when the device is oppositely biased, some of the traps collapse. Designers can thus mitigate the effects of BTI by balancing bias states.

Time-dependent dielectric breakdown (TDDB), also known as soft oxide breakdown, is also trap related. FinFET devices can have improved TDDB due to vertical field reduction and increased barrier to tunneling [18], but 3D features must be carefully engineered and TDDB is still a critical reliability mechanism. With continued dimension scaling, in combination with pressure to employ lower dielectric constant films, increasing electric fields in the interconnect has extended TDDB concerns to the wiring [19].

Electromigration (EM) has moved from a problem of mild effect around high drive buffers to something that must be carefully checked throughout a design. Figure 22 shows that the maximum DC current allowed through local metal has not been keeping pace with device current scaling. This trend, which should intensify into the FinFET era, necessitates more robust power delivery network (PDN) design, wider metal in critical nets, and even in some cases strapping outputs in M2. All of these remedies necessarily increase block area.

While the soft error rate (SER) per bit cell has stabilized in recent technologies, and with FinFETs may actually improve as compared to planar, the rate per area has been increasing [20]. SER has conventionally been most important in SRAM arrays and dealt with via error correction. With continued scaling, the flip-flop SER has become comparable to that of

SRAMs [21]. However, protecting (hardening) flip-flops against soft error is challenging due to the fact that they are spatially distributed [22].

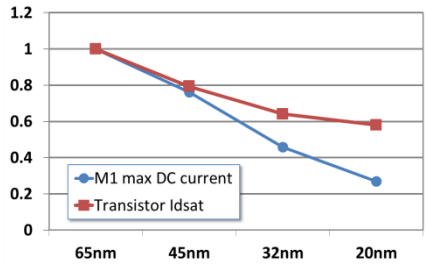

Figure 22: Relative scaling of metal maximum DC current versus I_{DSAT} of inverter-sized transistor.

Random telegraph noise (RTN), which causes time-varying threshold voltage [23], is a future reliability concern. While this effect has historically not been significant in digital design, the trend for RTN variation, $\sim (L \times W)^{-1}$, is steeper than that for random variability, $\sim (L \times W)^{-0.5}$.

Mitigation of reliability effects extends past physical IP because many of the mechanisms are activity dependent and state dependent, which is complicated enough even before considering effects such as BTI healing. This activity dependence means consideration of workloads can result in substantial changes to product reliability prediction [24]. This will be yet another reason to add more understanding of end product into the DTCO evaluation.

D. Continued importance of hardware optimization

As discussed above, design-technology choices can depend highly on power and performance targets. But beyond that, the choices for optimal logic design will vary as well. An example of this is heterogeneous multicore designs [25] where two or more entirely different, but software-compatible, classes of processor implementation are combined to provide hardware optimization for power and performance. Another example is the use of specialized graphics hardware (GPU) to provide higher performance for certain parallel compute tasks. Increasingly, system programmers and developers will have to make intelligent choices as to what specialized hardware to use for what tasks. Solving this problem, while managing software costs, will be key to improving hardware that has been optimized down to the process level for specific tasks.

V. CONCLUSIONS

The semiconductor ecosystem has evolved from fab-centric process scaling to multi-lateral design-technology co-optimization as the scaling challenges have intensified. All evidence suggests that this trend will continue and intensify in the future.

REFERENCES

[1] V. Pitchumani, "A hitchhiker's guide to the DFM universe", IEEE Asia Pacific Conf. Circuits and Systems (APCCAS) , 2006, pp. 1103-1106.

[2] S. Idgunji, V. Chandra, C. Pietrzyk, I. Iqbal, R. Aitken and G. Yeric, "An embedded process monitor test chip architecture", IEEE Intl. Conf. Microelectronic Test Structures, 2010, pp. 122-127.

[3] S. Sinha, B. Cline, G. Yeric, V. Chandra and Y. Cao, "Exploring sub-20nm FinFET design with predictive technology models", IEEE Design Automation Conference (DAC), 2012, pp. 283-288.

[4] A. Calimera, R. Bahar, E. Macii and M. Poncino, "Ensuring temperature-insensitivity of dual-Vt designs through ITD-aware synthesis", Intl. Workshop on Thermal Investigation of ICs and Systems (THERMINIC), 2008, pp. 31-36.

[5] V. Chandra, C. Pietrzyk and R. Aitken, "On the efficacy of write-assist techniques in low voltage nanoscale SRAMs", Design, Automation & Test in Europe (DATE), 2010, pp. 345-350.

[6] T. Matsukawa, Y. Liu, S.-I. O'uchi, K. Endo, et al., "Decomposition of on-current variability of nMOS FinFETs for prediction beyond 20 nm", IEEE. Trans. Electron Devices, v59 n8, Aug. 2012, pp. 2003-2010.

[7] J. Chen, W. Staud, and B. Arnold, "DFM challenges for 32nm node with double dipole lithography (DDL) and double patterning technology (DPT)", IEEE Symp. Semiconductor Manufacturing (ISSM), Sept. 2006, pp. 479-482.

[8] L. Liebmann, D. Pietromonaco, and M. Graf, "Decomposition-aware standard cell design flows to enable double-patterning technology", Proc. SPIE 7974, Apr. 2011, 79740K-1-12.

[9] J. Kye, Y. Ma, L. Yuan, Y. Deng and H. Levinson, "Lithography and Design Interaction – new paradigm for the technology architecture development", IEEE Custom Integrated Circuits Conference (CICC), Sept. 2012, pp. 1-4.

[10] R. Ghaida, K. Agarwal, S. Nassif, X. Yuan, L. Liebmann and P. Gupta, "Layout decomposition and legalization for double-patterning technology", IEEE. Trans. Computer-Aided Design of Integrated Circuits and Systems, v32 n2, Feb. 2013, pp. 202-215.

[11] G. Northrop, "Design technology co-optimization in technology definition for 22nm and beyond", IEEE Symp. on VLSI Technology (VLSI-T), June 2011, pp. 112-113.

[12] K. Jeong, A. Kahng, and R. Topaloglu, "Assessing Chip-Level Impact of Double Patterning Lithography", IEEE Intl. Symp. Quality Electronic Design (ISQED), March 2010, pp. 122-130.

[13] B. Yu, K. Yuan, B. Zhang, D. Ding, and D. Pan, "Layout decomposition for triple patterning lithography", IEEE Intl. Conf. Computer-Aided Design (ICCAD), Nov 2011, pp. 1-8.

[14] C.-T. Lin and Y.-L. Li, "Double patterning lithography aware gridless detailed routing with innovative conflict graph", IEEE Design Automation Conference (DAC), 2010, pp. 398-403.

[15] Fang, S.-Y., and Chang, Y-W., "Simultaneous flare level and flare variation minimization with dummification in EUVL", IEEE Design Automation Conference (DAC), 2012, pp. 1175-1180.

[16] M. van den Brink, "Continuing to shrink: Next-generation lithography - progress and prospects", IEEE Intl. Solid-State Circuits Confernece (ISSCC), Feb. 2013, paper 1.1

[17] R. Aitken, "Reliability Evaluation at the Device Level and its Impact on Design", IEEE Design Automation and Test in Europe (DATE), 2013.

[18] S. Ramey, A. Ashutosh, C. Auth, J. Clifford, *et al.*, "Intrinsic Transistor Reliability Improvements from 22nm Tri-Gate Technology," IEEE Intl. Reliability Physics Symposium (IRPS), 2013, paper 4C.5.

[19] R. Achanta, P. McLaughlin and F. Chen, "Failure rates for interconnect dielectric breakdown: Trends determining technology reliability scaling limits", IEEE. Trans. Device and Materials Reliability, v11 n2, June 2011, pp. 273-277.

[20] N. Seifert, B. Gill, S. Jahinuzzaman, J. Basile, *et al.*, "Soft error suceptibilities of 22 nm tri-gate devices", IEEE Trans. Nuclear Science, v59 n6, Dec. 2012, pp. 2666-2673.

[21] A. Oates, "Reliability challenges for the continued scaling of IC technologies", IEEE Custom Int. Circuits Conf. (CICC), 2012, pp. 1-4.

[22] S. Devarapalli, P. Zarkesh-Ha, and S. Suddarth, "SEU-hardened dual data rate flip-flop using C-elements", IEEE Intl. Symp. on Defect and Fault Tolerance in VLSI Systems (DFT), 2010, pp. 167-171.

[23] K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, *et al.*, "Direct observation of RTN-induced SRAM failure by accelerated testing and its application to product reliability assessment," IEEE Symposium on VLSI Technology, 2010.

[24] E. Mintarno, V. Chandra, D. Pietromonaco, R. Aitken, and R. Dutton, "Workload Dependent NBTI and PBTI Analysis for a sub-45nm Commercial Microprocessor," IEEE Intl. Reliability Physics Symposium (IRPS), 2013, paper 3A.1.

[25] L. Lugini V. Petrucci and D. Mosse, "Online thread assignment for heterogenous multicore systems", Intl. Conf. Parallel Processing Workshops (ICPPW), 2012, pp. 538-544.