Simplify to Survive,
prescriptive layouts ensure profitable scaling to 32nm and beyond

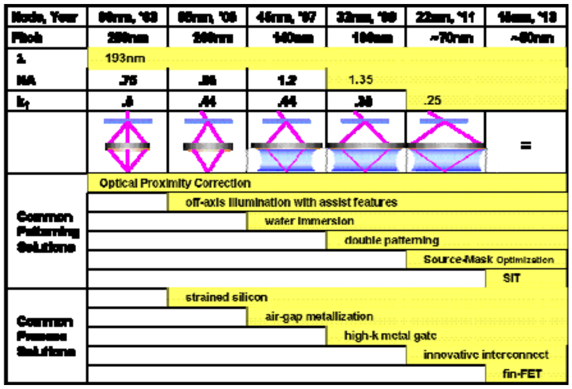
Lars Liebmann[1], Larry Pileggi[2], Jason Hibbeler[1], Vyacheslav Rovner[2],
Tejas Jhaveri[2], Greg Northrop[1]
[1] IBM Microelectronics, 2070 Route 52, Hopewell Junction, NY 12533
[2] PDF Solutions, Pittsburgh 5830 Ellsworth Ave, Suite 304. Pittsburgh PA 15232

ABSTRACT

The time-to-market driven need to maintain concurrent process-design co-development, even in spite of discontinuous patterning, process, and device innovation is reiterated. The escalating design rule complexity resulting from increasing layout sensitivities in physical and electrical yield and the resulting risk to profitable technology scaling is reviewed. Shortcomings in traditional Design for Manufacturability (DfM) solutions are identified and contrasted to the highly successful integrated design-technology co-optimization used for SRAM and other memory arrays. The feasibility of extending memory-style design-technology co-optimization, based on a highly simplified layout environment, to logic chips is demonstrated. Layout density benefits, modeled patterning and electrical yield improvements, as well as substantially improved layout simplicity are quantified in a conventional versus template-based design comparison on a 65nm IBM PowerPC 405 microprocessor core. The adaptability of this highly regularized template-based design solution to different yield concerns and design styles is shown in the extension of this work to 32nm with an increased focus on interconnect redundancy. In closing, the work not covered in this paper, focused on the process side of the integrated process-design co-optimization, is introduced.

Keywords: Design for Manufacturability (DfM), template-based design, design-technology co-optimization (DTCO), predictably composable logic, pdBrix

1. INTRODUCTION



Node, Year	90nm, '03	65nm, '05	45nm, '07	32nm, '09	22nm, '11	16nm, '13
Pitch	280nm	200nm	140nm	100nm	~70nm	~50nm
λ	193nm					
NA	.75	.85	1.2	1.35		
k_1	.5	.44	.44	.35	.25	
						=
Common Patterning Solutions	Optical Proximity Correction					
		off-axis illumination with assist features				
			water immersion			
				double patterning		
					Source-Mask Optimization	
						SIT
Common Process Solutions		strained silicon				
			air-gap metallization			
				high-k metal gate		
					innovative interconnect	
						fin-FET

Figure 1 CMOS scaling is being challenged by radical innovations in patterning, process, device, and interconnect technology, yet the pressure to stay on a 2 year node-to-node cycle remains.

It is well known that profitability in the microelectronics industry is driven by a two year cycle in which transistor density doubles, performance noticeably increases, power consumption drops, and wafer manufacturing cost remains largely constant. With wavelength scaling reaching its limit at 193nm in the 90nm node, patterning resolution improvement has been achieved through a series of discontinuous innovations rather than predictable evolutionary enhancements like wavelength reduction (Fig.1, top). Both computational resolution enhancements, such as off-axis illumination with sub-resolution assist features and optimized model-based optical proximity correction, as well as physical resolution enhancements, such as ultra-high numerical aperture lithography enabled through the use of water immersion, share the common trait of introducing more severe and more complex layout sensitivities in physical and electrical yield detractors.

Beyond the aforementioned escalation in patterning complexity, the profitability of the microelectronics industry is further challenged by the fact that dimensional scaling alone is no longer sufficient to achieve the electrical performance targets of the next technology node. Additional device, interconnect, and process innovations, introduced with every new node, further add to the unpredictability of layout sensitive detractors to yield ramp (Fig.1, bottom). Yet, even as process optimization becomes more difficult and layout-specific, the two year node-to-node timetable requires concurrent technology and design co-optimization. It is simply not possible to develop a new process solution and thoroughly characterize all its layout sensitivities before starting the respective node's design work. To have product designs available when the process is scheduled to yield functional chips, design work has to proceed in parallel to the process and device optimization work.

1.1. Design Rules

Traditional design rules had provided an elegantly simple solution that allowed process and design development to occur concurrently rather than sequentially. Design rules, which were predominately derived from competitive scaling targets but also involved a period of intense collaboration, negotiation, simulation, and optimization early in the technology node, established process control targets that set goals for process development and provided rules to initiate layout work. Even as the physical limitations of certain processes started to prevent equally aggressive scaling of all design rules, rules could be published early in the technology node to allow designers to begin their work while process engineers drove the process to the committed control limits. On rare occasion, the committed process control limits could not be achieved, causing highly undesirable churn late in a technology node. In general, however, this approach worked until non-monotonic layout sensitivities drove rapid escalation in design rule complexity (Fig. 2).

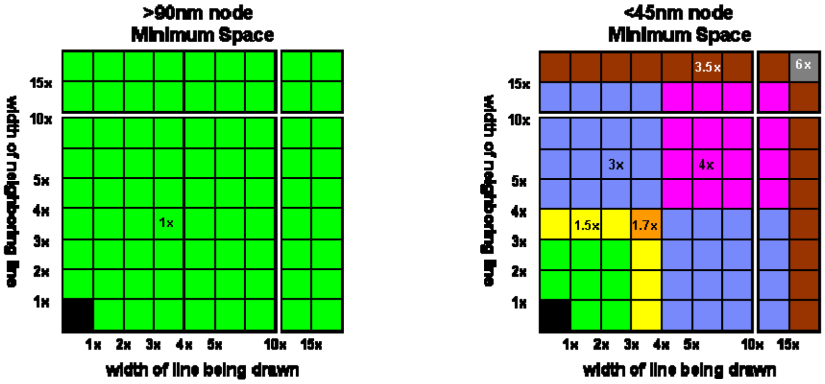


Figure 2 Illustration of the escalation in design rule complexity in response to non-monotonic layout sensitivities. Shown is a typical example of minimum allowed metal spacing as a function of drawn and neighboring widths. What used to be a simple pass-fail limit has turned into a complex multi-body problem.

In the example in Fig. 2, 'forbidden pitches' in the off-axis illumination based patterning solution drive very complex multi-feature width-dependent spacing rules. Unfortunately, even with all this complexity, these rules can not provide absolute assurance that all design-rule-clean layouts will yield or perform adequately. Unanticipated asymmetric width-space combinations or two-dimensional constructs not taken into account in the rule optimization can easily cause yield loss while passing the design rule checking. On the other hand, some layout constructs that may be extremely valuable to a particular design, may fail design rule checking but would be adequately manufacturable. Rules-based design simply has to make compromises between being too conservative and being too complex, which opens the door for yield or performance challenges.

1.2. Traditional DfM

Two traditional Design-for-Manufacturability (DfM) solutions are in direct response to the challenge of design rules becoming both too complex and too conservative; model-based optimization, and restrictive design rules.

Model-based layout optimization seeks to eliminate conservatism in the design by letting designers or design tools interactively and iteratively detect and eliminate layout sensitivities by modeling the anticipated process response of a particular layout construct. Illustrated in Fig.3a is the lithography response to a local wiring (M1) layout showing layout sensitivities in terms of localized line and space narrowing. While in theory the concept of model-based layout legalization may seem like a good idea, it is completely infeasible for all but a few niche applications: Of the large number of process steps that each exhibit unique layout sensitivities, only lithography and chemical mechanical polishing (CMP) have reasonably accurate process models. And even for these models, only the aerial image component of lithography, not the chemical component of resist exposure, is based on predictive fist principle simulations, all other effects are captured in semi-empirical models that require extensive experimental calibration. A partial set of empirical models, calibrated to a rapidly changing process, clearly can not provide an accurate basis for layout legalization early in a technology node. Predictive process modeling challenges aside, integrating model-based layout legalization into a design flow would require substantial reengineering of the entire IP generation, synthesis, placement, and routing flow and would severely err on the side of escalating design complexity in favor of reducing conservatism (1). The noteworthy exception is lithography and CMP modeling based routing optimization late in the design flow (2,3).

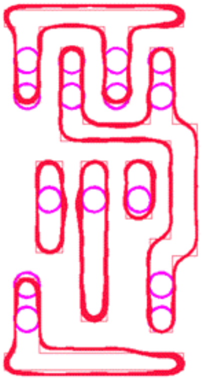


Figure 3a Example of lithography variability bands of a first-metal layout as they would be presented to a designer in interactive model-based DfM.

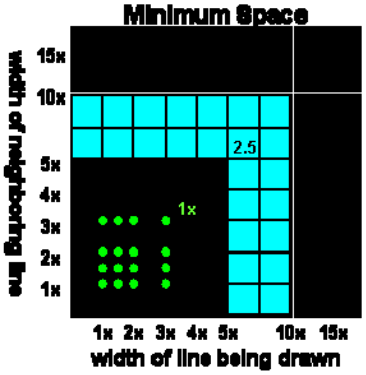


Figure 3b Illustration of an RDR description of the width-dependent-spacing phenomenon captured in Fig 2. Dots represent specific allowed line placements.

In contrast to model-based DfM, Restricted Design Rule (RDR) based DfM focuses on preserving, even enhancing, design efficiency while ensuring yield and performance by eliminating unknown layout sensitivities. Best described as a design approach based on 'prescriptive' design rules rather than the traditional 'prohibitive' design rules, RDR-based design seeks to provide clarity in the design-process handoff by comprehensively defining all allowed feature placements (4). As shown in Fig.3b, RDRs minimize the complexity of width-dependent spacing rules (Fig.2) by defining discrete placement options for narrow lines, followed by a more traditional continuous design space with a single conservative design rule for intermediate width lines, and complete elimination of all lines (with limited design value) of extremely large dimensions. In a process environment where yield and performance no longer improve monotonically above the minimum design rule, RDRs provide clear targets for aggressive process optimization. Key to successful implementation of RDR-based DfM is a close collaboration between the design and process teams in the initial definition of the RDRs to ensure process development for a limited set of design rules that is actually useful to the designers (5). While currently the only feasible design-process co-optimization solution, RDR-based DfM still leaves room for improvement in eliminating design conservatism while ensuring competitive performance and yield.

The common element of current DfM solutions is that, similar to the original design rules, they focus primarily on the 'layout space' after the actual design and before actual process optimization. Therefore, these approaches can only provide sub-optimal solutions.

1.3. SRAM Design-Process Co-optimization

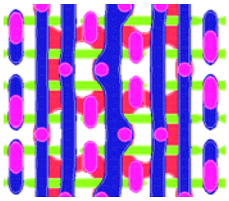


Figure 4 A SRAM layout showing pattern complexity.

A look at how a logic chip's memory arrays are developed for a new technology node provides valuable insights into the next level of design-technology co-optimization. SRAM designs have consistently pushed technology harder than logic, achieving very aggressive transistor densities under very rigorous performance constraints. The aggressive scaling of memory is achieved primarily through deep design-technology co-optimization: technology options, including resolution enhancement techniques but extending to all aspects of process, device, and interconnect development, are fine tuned to the unique requirements of the SRAM layout. As Fig.4 shows, this successful design-technology co-optimization is not enabled by forcing all designs onto simple one-dimensional gratings. Design rules are pushed to the absolute limit for each unique and often complex layout configuration to the point where the design rule is replaced by the actual construct being scaled. Parameterized variants of the SRAM cells are comprehensively qualified and characterized using dedicated test vehicles to ensure a complete understanding of the yield implications of each specific layout construct. Multi-disciplinary engineering teams interactively converge on an optimized layout and matching manufacturing process using rigorous simulation and iterative experimentation. This comprehensive design-technology co-optimization is fundamentally enabled by a simplified design space that allows competitive designs to be synthesized from a small set of predictably composable logic constructs.

1.4. Maintaining Profitable Scaling

To assess whether an SRAM-like design-technology co-optimization is feasible and beneficial for logic designs in advanced technology nodes, three key questions need to be answered:

- Can a competitive logic design be generated from a highly constrained set of logic constructs (templates)?
- Are these logic constructs really predictably composable?
- Does the constrained and predictable layout environment provide process optimization and characterization benefits that result in improved yield and performance?

The work reported in this paper focuses primarily on the design'ability aspect of template-based design, though other work, focused on demonstrating the broader design-process co-optimization benefits, is briefly introduced.

2. 65nm PowerPC 405s CORE DESIGN

2.1. Experimental Setup

To assess the feasibility of generating a competitive product design from a simplified set of layout constructs, IBM and PDF collaborated in the redesign of a PowerPC 405s core in the 65nm technology node using PDF's pdBrix design methodology (6,7). Starting with the original netlist, the redesign was constrained to maintain the same:

- macro footprint and area (1111 mm x 1817.5 mm, including reused memory blocks)
- performance (3.5ns Clock Period, ~287 MHz)
- power (RVT / HVT transistor breakdown = 90% / 10%)

2.2. Fabric and Templates

To ensure predictable composability of the resulting logic constructs, the pdBrix design flow begins with the definition of an ultra-regular cell image, referred to as the 'fabric', onto which all logic is mapped. Conceptually, the regularity of a sample fabric is illustrated in Fig.5a, showing the extremely constrained design grid available for poly and metal shapes. Fig.5b is a simple cartoon of a possible logic topology mapped onto this coarse design grid to illustrate that, while all features have to fall onto the prescribed grid, not all grid lines have to be occupied by layout features. The combination of this coarse layout grid and a handful of simple design rules yields a fabric that results in:

- limited diffusion corners
- fixed-pitch, unidirectional Poly (vertical)
- unidirectional metal (first metal, M1, horizontal; second metal, M2, vertical)
- relaxed pitch for M1 (17%) and M2 (25%)
- contacts and vias on grid

The relaxed metal pitch and the predictable nature of the gridded layout enabled the use of tighter rules for metal extension past contacts and vias.

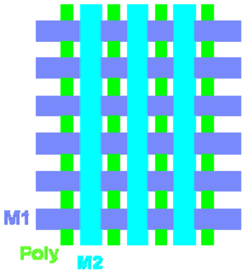


Figure 5a Illustration of the Fabric showing uni-directional fixed pitch layout

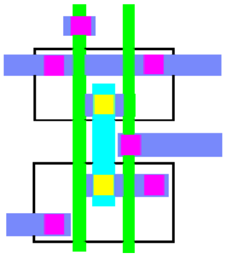


Figure 5b Simple logic mapped onto the Fabric, showing selective track use.

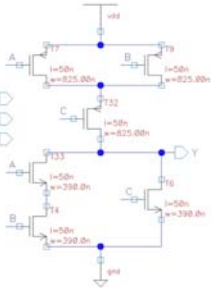


Figure 5c Schematic of a sample logic primitive, !(AB+C), representing a Template.

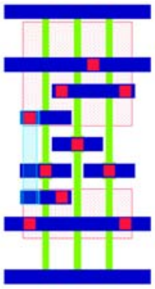


Figure 5d Physical layout of the Template defined in Fig. 5c.

A specific example of the basic logic constructs used in this design is shown in Fig.s 5c and 5d. The schematic for a !(AB+C) function is mapped onto the aforementioned fabric to yield the template layout shown in Fig. 5d.

2.3. Hotspot Reduction

The lithography benefits of ultra regular layouts are quite obvious. Not only does the fixed-pitch, single-orientation layout simplify the frequency space that has to be imaged, it completely eliminates all complex two-dimensional constructs that could lead to yield concerns. To confirm this theory, both the conventional and the pdBrix layouts were scaled to 45nm dimensions and run through process-window lithography simulations. The scaling to 45nm was done to provide a more challenging resolution environment typical of advanced technology nodes and to leverage more accurate through-process lithography models available in 45nm. It provided the additional benefit of assessing the layout migrate'ability differences in the two design styles. Fig.s 6a and 6b show the simulated lithographic process window bands for the pdBrix and conventional layout, while Fig.6c shows examples of typical layout sensitivities discovered in the conventional layout. The hotspot count in Table 1 shows the improved patterning robustness of the pdBrix design.

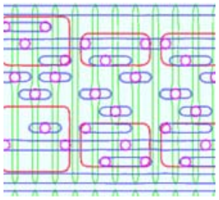


Figure 6a pdBrix layout

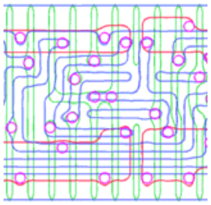


Figure 6b standard layout

Lithography simulations for: diffusion (red) poly (green), 1st metal (blue), and contact (purple) including dose, focus, and mask error

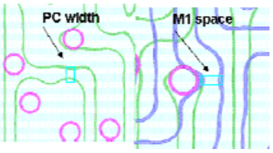


Figure 6c Litho layout sensitivities: pinching and bridging hotspots occurring around 2-d constructs.

Litho Hotspots after scaling to 45nm					
Poly Width		M1 Width		M1 Space	
pdBrix	stndrd	pdBrix	stndrd	pdBrix	stndrd
0	34	0	23	0	1634

Table 1 count of lithography hotspots that introduce a potential yield risk for the standard and pdBrix layout.

2.4. Variability Improvement

In addition to process limited yield (PLY) effects, advanced technology nodes are suffering from loss of circuit limited yield (CLY); i.e. chips that don't have any hard fails but don't meet timing or performance requirements. A primary factor in controlling CLY is the reduction of electrical variability. Fig.7 shows the improvement in transistor variability achieved with the pdBrix layout. The lithography variability-bands for two layouts of the same logic function are shown in Fig.7a. The variability-bands are comprised of individual contours that capture deviations in exposure dose, defocus, and mask size. Fig.7b shows the electrically equivalent channel length extracted from these layouts based on the range of process contours. Electrical channel length variability is reduced from 4.1nm to 3.5nm, which is equivalent to as much as 10-15mV reduction in threshold voltage variation. This translates into a very significant improvement in CLY.

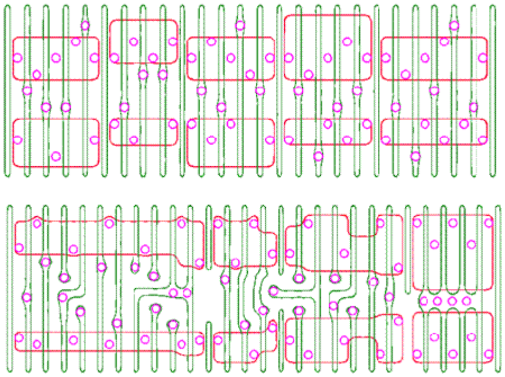


Figure 7a Diffusion, poly, and contact lithography variability-bands for pdBrix layout (top) and conventional layout (bottom)

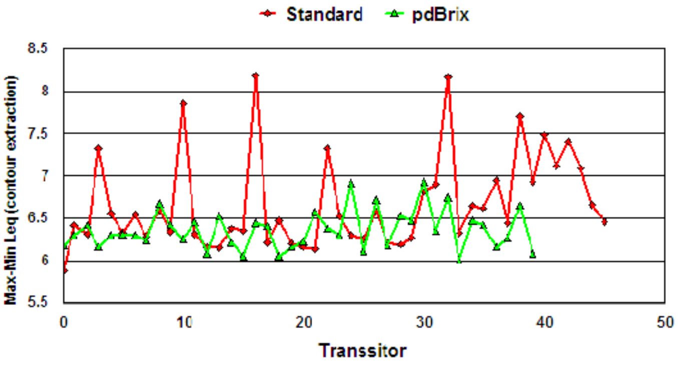


Figure 7b Range of electrical channel length extracted for each transistor in the two layouts of Fig 7a. Increased electrical variability as a response to lithography variation due to dose, focus, and mask size is seen in the standard layout.

2.5. Area Improvement

While the yield and performance benefits of regularized layouts may be well accepted, the biggest barrier to broader implementation of regularized layout styles is the perceived impact on layout density. The pdBrix design of IBM's PowerPC microprocessor showed that it was possible to contain the extremely regularized layout in the same footprint as the original layout. Further, the total area occupied by sequential logic was identical in the two layout styles and the area occupied by combinatorial logic actually decreased by 25%. Further analysis of the specific contributors to this 25% area reduction indicated that:

- 25% of the reduction was achieved through Template and Fabric co-optimization; i.e. elimination of layout conservatism by implementing construct-specific design rules.
- 5% of the reduction was achieved through the use of design specific complex gates, referred to as Brix, synthesized from the primitive logic functions rendered in Templates.
- 70% of the reduction was attributed to optimal construction of application-specific logic cell functions; i.e. eliminating conservatism in layout variants like power levels by synthesizing an application specific library from a technology node specific set of Templates rather than using generic set of standard cells designed to cover all possible power/performance needs.

2.6. Design Simplicity Improvement

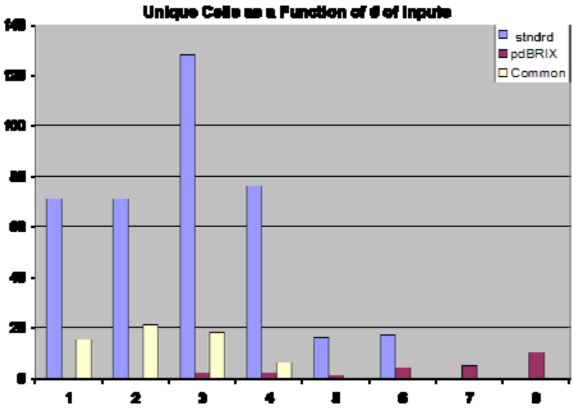


Figure 8 Substantial simplification of the design space as measured by the number of unique logic cells used in the standard versus the pdBrix design.

Reduction of lithography hotspots and improvement in variability were expected benefits of highly regularized layouts. Very effective use of the physical and electrical predictability of the highly regularized layout style to eliminate layout density penalties was necessary to make this design approach possible. But the core value proposition hinges on demonstrating the feasibility of generating a competitive design from a substantially simplified set of layout primitives. Fig.8 compares the number of unique logic cells used in the standard microprocessor core design to the number of cells used in the pdBrix version of that design. The cell count is reported as a function of the number of logical inputs to the operation, showing that the pdBrix methodology shifted the design towards more complex and compact logic functions. The small number of logic cells common to both design styles further emphasizes that the reduction in cell count is fundamentally enabled by the optimal use of application specific logic cells, not merely a constrained physical synthesis.

3. Fabric Adaptability

The 65nm PowerPC design exercise:

a) confirmed the inherent patterning benefits of regularized layouts;

b) demonstrated that an optimally integrated design flow can achieve competitive layout density with a highly regularized layout style; and

c) showed that it is possible to generate a competitive design from a substantially simplified set of logic elements.

Based on these positive results, it was decided to continue this work in the 32nm node, driving the project closer to the leading edge of the technology node to be better aligned with new product design starts. In porting the 65nm fabric to 32nm, three opportunities to align the Fabric more with product specific internal design objectives were identified:

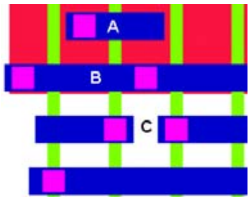


Figure 9 Illustration of 3 specific layout considerations in the original pattern-count optimized Fabric: contact redundancy, narrow power wires, tight tip-to-tip space.

- Running first metal perpendicular to the poly gates and forcing it to be completely uni-directional eliminates all possibility of multiple diffusion contacts, as shown in Fig.9A. It is well known that redundant diffusion contacts are preferred in some designs for device performance and yield benefits. Further, the strictly unidirectional metal adds additional vias for all ‘wrong’ way connections, which is not ideal for processes with via yield challenges.

- Forcing all metal to be equal width (Fig.9B) eliminates the possibility of using wide wires for power distribution, preferred in some designs to improve reliability on these high current carrying constructs.

- Linking the contacted device pitch (i.e. minimum poly pitch) to the minimum contacted tip-to-tip spacing of first metal (Fig.9C) creates a scaling problem. While pitch scaling is driven to 30% reduction per node, tip-to-tip spacing has been scaling at roughly 20% per node, forcing a decoupling of these two constructs in the overall density scaling.

In a demonstration of the adaptability of the pdBrix design approach, a new fabric with different optimization priorities was defined, as shown in Fig.10. The new fabric continues to enforce: limited diffusion corners, fixed-pitch unidirectional poly (vertical), preferred orientation metal (M1 now vertical, M2 now horizontal), contacts and vias on grid, and relaxed pitch for M1 (25%) and M2 (5%), but also accommodates limited wrong-way metal (on-grid) and wider power rails. The improvement in via and contact redundancy for different layout options is shown in Fig 11. The initial improvement is achieved by exploiting opportunities to insert redundant contacts and vias into the original image (labeled ‘plus redundancy’ in Fig.11), followed by allowing limited use of wrong-way metal in the original image, and finally by switching to a new cell image. Fig.11 compares the ‘number of non-redundant connections’ for all four cases. Reporting interconnect improvement in this fashion highlights the fact that both ‘eliminating connections’ and ‘adding redundancy to connections’ fulfills the stated design intent.

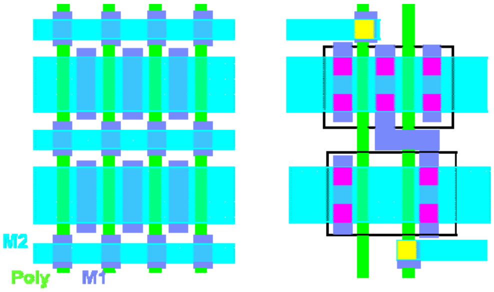


Figure 10 Fabric chosen to optimize redundancy, reduce necessary connections, and allow wide power wire.

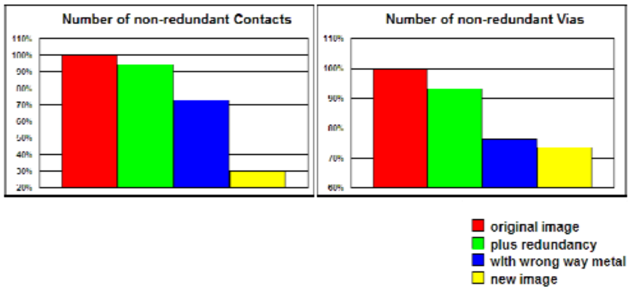


Figure 11 Reduction of non redundant connections, relative to the original PowerPC design, based on different fabric trade-offs.

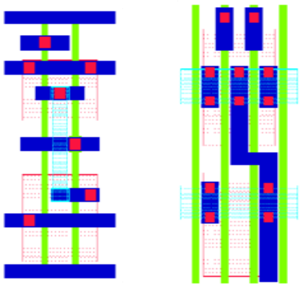


Figure 12 A nand2 rendered in the ‘low pattern count’ (left) and ‘high redundancy’ (right) Fabric

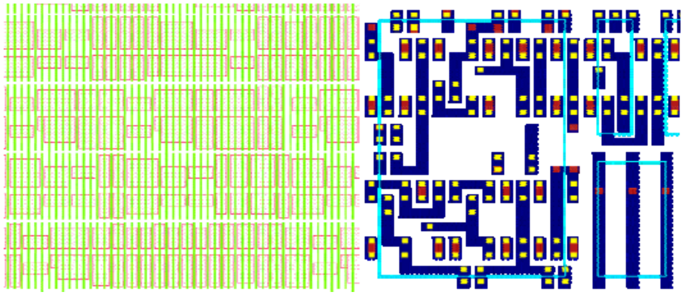


Figure 13 Long range parametric layout sensitivities on diffusion and poly are minimized through macro regularity (left), local yield hotspots on first metal are prevented through control of boundary conditions.

The differences in the two fabric styles, one optimized to reduce overall pattern count, the other optimized for redundancy and wide power rails, can be seen in Fig.12. Shown are two templates representing the same logic gate (a nand2) mapped onto the two fabrics. Defining the fabric with selective wrong-way metal does not affect the layout density or logic simplicity as demonstrated above, but stresses the need to actively manage predictable composability of the logic elements. To ensure predictable circuit performance regardless of placement, layout sensitivities in the active device parameters have to be minimized. In addition to local proximity effects like poly line-width variation and diffusion corner rounding, device performance is affected by long range layout sensitivities in processes such as etch, stress, or rapid thermal anneal. These long range effects are best controlled through macro-regularity; i.e., global pattern uniformity on diffusion and poly shapes as shown in Fig.13(left). The first metal level, being used for local interconnect only, is not a major contributor to parametric variability, making local patterning hotspots the primary yield concern. The small number of logic constructs allows the safe use of optimally complex layout configurations as long as proper boundary conditions are enforced to ensure that no new complex layout configurations can be formed at cell boundaries. As illustrated in the right layout of Fig.13, in this fabric the more complicated metal patterns (i.e. staircases or dense T-shaped line-ends) are constrained to the center of the cell, as are 'belt buckle' constructs that require tight dimensional control since they can be used for connectivity to other layers. Layout constructs within the boundary region of the cell (i.e. outside of blue outlines in Fig.13) are simpler and more conservative (e.g. line-end connectivity in boundary region belt buckles is not allowed). Predictable composability is thereby maintained for two layout sensitivities of different length scales through two different mechanisms: macro regularity to address long range effects and boundary conditions to address local effects.

4. Integrated Design-Process Co-Optimization

While the work reported here focused on the design and layout aspects of the pdBrix solution, these design studies are complemented by integrated testsite runs under way in both 32nm SOI and bulk technologies. Experiments have been designed to demonstrate:

- variability and design-margin reduction afforded by the predictable and regularized layout environment
- yield and manufacturability improvement facilitated by the simplified design environment
- predictable template composability achieved through layout regularity and control of boundary conditions
- performance and power benefits from the use of complex logic gates
- placement agnostic delay achieved through macroscopic layout regularity
- model-to-hardware correlation improvements through pre-characterized templates

In addition, the power/performance/area benefits, route'ability, and design flow compatibility of the pdBrix solution is being confirmed in another design experiment targeting high performance macros from IBM's Cell processor design.

5. CONCLUSION

It was shown that profitable CMOS scaling under the time pressures of the established two year node-to-node cycle in an environment of continuous disruptive technology innovation requires deep design-technology co-optimization. To overcome the shortcomings of DfM solutions that operate primarily in the physical layout space and do not address fundamental design optimization or design-aware process optimization, the pdBrix design methodology was evaluated. The goal of this collaborative work is to establish a comprehensive design to silicon solution that facilitates rigorous design-technology co-optimization.

The key requirements for an optimal design to silicon solution have been establishes as:

- late binding of logic to physical layout to preserve design-creativity and -performance
- optimally simplified layout to allow construct-driven process development
- targeted characterization and qualification test vehicles to drive yield ramp

The work reported here demonstrated feasibility of design with a simplified set of logic, preserving layout density and improving patterning yield and variability while achieving the stated power/performance targets with significantly fewer layout constructs.

ACKNOWLEDGMENTS

The joint IBM and PDF team of authors thanks their many colleagues that were directly and indirectly involved in this work and acknowledges the specific contributions of Henning Haffner (Infineon) for his work on 32nm design rule optimization and Dureseti Chidambarrao (IBM) for his work on contour-based extraction.

REFERENCES

[1] DFM lessons learned from altPSM design, L. Liebmann, Z. Baum, I. Graur, D. Samuels, Proc. SPIE 6925, 69250C (2008)
[2] Convergent automated chip-level lithography checking and fixing at 45 nm, Valerio Perez, et al. , Proc. SPIE 7275, (2009)
[3] Hotspot detection and design recommendation using silicon-calibrated CMP model, Colin Hui, et.al., Proc. SPIE 7275, (2009)
[4] Layout impact of resolution enhancement techniques: impediment or opportunity?, L. Liebmann, ISPD'03, 2003, Monterey
[5] Intel design for manufacturing and evolution of design rules, Clair Webb, Proc. SPIE 6925, 692503 (2008)
[6] Regular Fabrics for Nano-Scaled CMOS Technologies, L. Pileggi and A.J. Strojwas, ISSCC, (2006).
[7] Maximization of Layout Printability/Manufacturability by Extreme Layout Regularity, Tejas Jhaveri, et al.,, J. Micro/Nanolithography, MEMS, and MOEMS, Vol 6 (03), 2007.