Lithography Simulation-Based Full-Chip Design Analyses

Puneet Gupta[a], Andrew B. Kahng[a], Sam Nakagawa[a], Saumil Shah[b] and Puneet Sharma[c]

[a]Blaze DFM, Inc., Sunnyvale, CA;
[b]University of Michigan, Ann Arbor, MI;
[c]University of California San Diego, La Jolla, CA

ABSTRACT

Today's design flows sign-off performance and power prior to application of resolution enhancement techniques (RETs). Together with process variations, RETs can lead to substantial difference between post-layout and on-silicon performance and power. Lithography simulation enables estimation of on-silicon feature sizes at different process conditions. However, current lithography simulation tools are completely shape-based and not connected to the design in any way. This prevents designers from estimating on-silicon performance and power and consequently most chips are designed for pessimistic worst-cases. In this paper we present a novel methodology that uses the result of lithography simulation for estimation of performance and power of a design using standard device- and chip-level analysis tools.

The key challenge addressed by our methodology is to transform shapes generated by lithography simulation to a form that is acceptable by standard analysis tools such that electrical properties are preserved. Our approach is sufficiently fast to be run full-chip on all layers of a large design. We observe that while the difference in power and performance estimates at post-layout and on-silicon is small at ideal process conditions, it increases substantially at non-ideal process conditions. With our RET recipes, linewidths tend to decrease with defocus for most patterns. According to the proposed analyses of layouts litho-simulated at $100nm$ defocus, leakage increases by up to 68%, setup time improves by up to 14%, and dynamic power reduces by up to 2%.

Keywords: Lithography simulation, post-OPC, analysis, verification.

1. INTRODUCTION

RETs are key enablers of the aggressive IC technology scaling that has fast outpaced advancements in lithography hardware solutions. RETs such as optical proximity correction (OPC), phase shift masks (PSM), and off-axis illumination (OAI) dramatically improve resolution and are extremely effective at process variation control. The increased mask and manufacturing costs due to the application of these techniques have been outweighed by the advantages offered, and these techniques are imperative during mask-data preparation. RETs modify the design significantly and there is little similarity left with the design at the post-layout stage at which sign-off is performed. At non-ideal process conditions significant process variations can result even within the process window. Due to RETs and process variations features do not print at their nominal dimensions causing circuit power and performance to be significantly different from sign-off estimates. Today's design flows worst-case process effects and consequently overdesign circuits leaving valuable performance on the table.

Lithography simulation* enables estimation of CD variations at different process points. According to the international technology roadmap for semiconductors (ITRS), a substantial fraction of variations is systematic and can be modeled accurately after layout.[1] So even though random variations cause differences between on-silicon shapes and those predicted by lithography simulation, these difference are relatively small. Consequently, lithography simulation-based design analyses are likely to be significantly more accurate to on-silicon than post-layout analyses. Current lithography simulation tools are completely shape-based and not connected to the design in any way. In this paper we present a novel methodology that uses the results of lithography simulation

Further author information: (Send correspondence to Puneet Gupta)
Puneet Gupta: E-mail: puneet@blaze-dfm.com, Telephone: 1 408 470 4925
Work done by Saumil Shah and Puneet Sharma at Blaze DFM, Inc.

*Residual OPC critical dimension (CD) error or post-OPC edge placement error (EPE) results can be easily used to generate an output similar to that of lithography simulation by adding the errors to the drawn CD.

for estimation of performance and power of a design using standard device- and chip-level analysis tools. The proposed approach can reduce worst-casing and facilitate optimizations that account for process variations.

A recently proposed work by Yang et al. addressed post-lithography based analysis and optimization. They proposed a timing analysis flow based on residual OPC errors (equivalent to lithography simulation output).[8] In this work lithography simulation was performed only on timing-critical cells and their neighborhood. In modern designs a large fraction of cells are timing-critical and along with their neighborhood can include almost all cells limiting the runtime benefits of the approach. Only setup-time analysis was performed and interconnect variations ignored. Several non-trivial details related to handling non-rectangular gates in SPICE simulations and cell-level hierarchy reconstruction are missing. Recent works have also attempted to capture systematic variations and account for them in analyses and optimizations. Gate-length variability due to proximity effects and across-field lens aberrations was characterized by Orshansky et al. with a set of simple patterns located at different field locations.[7] Systematic variations due to defocus and pitch were captured through lithography simulations of simple test-patterns and used to drive timing and leakage analyses and optimizations.[2, 5] Systematic variability due to lens aberrations was characterized for timing analysis and analytical placement.[6] These works, however, rely on the ability of simple patterns to predict process variations which unfortunately can be quite inaccurate especially as the optical radius of influence (i.e., radius beyond which proximity effects fade) is increasing.

An overview of our approach is as follows. For MOS devices, performance and power is dependent on their gate-length and gate-width. To compute the gate-width, the active region contour is approximated by an equivalent rectangular region.[4] Accurate determination of gate-length is more important due to the heavy dependence of leakage and delay on it. Our gate-length computation depends on the objective to be analyzed (e.g., delay, leakage, dynamic power). We first rectilinearize the simulated gate contour and then approximate it by an electrically-equivalent rectangle. We then use transistor-level modeling (TLM) to estimate the impact of gate-length variations on design metrics. As an alternative to TLM, we propose a cell-level analysis flow that allows standard analyses tools to be used. After lithography simulation cell instances of the same cell differ and cannot be mapped to the same cell in the library for lithography simulation-based analyses. We add variants of each cell in the library; the variants are similar in function and drive strength of the cell but have different gate-lengths assigned to the devices. After rectilinearization and determination of gate-length of all devices in a cell instance, the variant that matches in the electrical behavior of the cell instance is selected and mapped to. The output is generated in the form of a modified Verilog file and can be used by standard analyses tools. The above-mentioned flow generates separate Verilog files for different metrics to be analyzed. We propose a "mixed-mode" flow in which only one Verilog file that is accurate for all metrics is generated. Our interconnect analysis flow modifies the parasitics database to account for variations in wire width and spacing. Interconnects are simplified to polygons and their resistance computed using analytical formulas. For capacitance computation, pairs of interconnects are simultaneously simplified and the change in their coupling capacitance estimated using a pre-created lookup table. The same parasitic extraction approach is used to compute parasitics for corresponding drawn shapes and the *change* in parasitics is computed. The parasitic database is then updated with the change.

The remainder of this paper is organized as follows. In Sections 2 and 3 we respectively describe our gate and interconnect analyses methodologies. Section 4 explains full-chip analyses that uses gate and interconnect analyses. It also explains our mixed-mode analyses flow. In Section 5 we describe our experimental setup and present results. Section 6 concludes with a brief description of ongoing work.

2. DEVICE ANALYSES

The gate-length and gate-width of a MOS device have the most direct impact on its performance and power. While performance and power exhibit complicated dependence on gate-length and gate-width, simpler approximations are as follows: (a) delay is partially determined by the saturation current which increases and decreases linearly with gate-length and gate-width respectively, (b) dynamic power increases linearly with gate-length and gate-width due to the change in gate capacitance, and (c) subthreshold leakage increases linearly with gate-width and exponentially with the squared of gate-length. Our analyses accounts for changes in gate-length and gate-width due to lithography imperfections and those in associated parameters such as source area and parameter, drain area and parameter, and stress parameters. Consequent changes in parasitics, however, are ignored in our analyses.

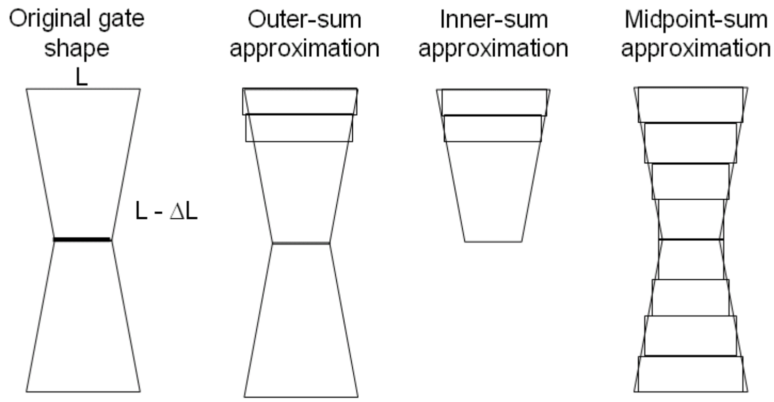

Figure 1. Three possible ways for rectilinearization.

2.1. Device Gate-Length and Gate-Width Computation

The gate is formed where the poly and active regions overlap. The non-rectangular shapes of poly and active regions makes the gate-width computation non-trivial. We use the flow previously proposed to find the *equivalent* active region and compute gate-width (W_{Avg}).[4]

Delay and power are heavily dependent on the small variations in gate-length introduced by lithography. Therefore, it is important to accurately access the delay and power impact due to variations in gate-length. After a sequence of simplification steps described below, a gate contour is reduced to a rectangle and the *average* gate-length computed. We differentiate between the different analyses metrics - setup time, hold time, leakage, and dynamic power - for simplifications to preserve electrical equivalence as much as possible. The first step is to rectilinearize the gate contour generated by lithography simulation. Three possible ways for rectilinearization are: (1) interior-point, (2) exterior-point, and (3) mid-point and are illustrated in Figure 1. For setup time as the objective, exterior point rectilinearization may be performed. This is because gate delay, transition time, and capacitance increase with gate-length; exterior-point rectilinearization yields an upper-bound on the gate-length and consequently setup time. For similar reasons interior-point rectilinearization may be used for hold time and leakage. Dynamic power depends on the gate capacitance which is determined by the gate area. Therefore, an area-preserving rectilinearization is desirable for dynamic power; mid-point rectilinearization may be used as a less computation-intensive approximation. While the described objective-specific methods of rectilinearization ensure pessimistic estimates, we use only mid-point rectilinearization in our experiments to reduce memory requirements.

Since standard circuit simulation tools such as Synopsys HSPICE can only simulate rectangular gates, the next simplification step reduces a rectilinear polygon to a rectangle. L_{Avg} denotes the gate-length of the rectangle and is computed differently for the various analyses objectives. We allow two modes for L_{Avg} computation:

1. *Lookup table mode.* In this mode we use a flow similar to one proposed previously.[4] Lookup tables for device on- and off-currents are created for different gate-lengths and gate-widths by SPICE simulations. The rectilinear polygons are sliced into rectangles and the on-currents (for setup and hold delays) and off-currents (for leakage) of all slices are summed up. L_{Avg} is the gate-length of the rectangle of the same gate-width and that yields the same on- or off-current. For dynamic power objective, rectangle that preserves the area is used.

2. *Expression mode.* If device SPICE models are not available, the lookup tables cannot be generated. In this case analytical expressions may be used to compute L_{Avg} for different objectives. For example for dynamic power objective an area-preserving rectangle would have $L_{\text{Avg}} = \sum_i W_i L_i / W_{\text{Avg}}$, where W_i and L_i are the width and length of the i^{th} slice.

2.2. Cell-Level Analyses

While device-level analyses tools are more accurate, they are not sufficiently fast to be run full-chip. Standard full-chip analyses tools are cell-level; they use characterized libraries that are created by SPICE simulations to perform design analyses. We propose two flows to use standard cell-level analyses tools for lithography simulation-based analyses:

1. *Cell library-based.* Cell-level performance and power analyses tools require each cell *instance* in the design to refer to a cell *master* in the library. All drawn cells that refer to one cell master are alike. However, after lithography simulation they may all differ. Unfortunately, it is not feasible to create different cells masters for each instance and the hierarchy needs to be reconstructed. We create a library that contains multiple variants of each cell master. Two types of library variants may be characterized: (1) *cell-level* in which gate-lengths of devices in a variant are equal but differ from those of devices in another variant, and (2) *transistor-level* in which devices in a variant may have different gate-lengths. To reduce library size, we do not alter device gate-widths in the variants because the percentage variation in gate-width after lithography simulation is very small. For each cell instance, a different cell variant may be chosen for different objectives. For setup time, the variants in which gate-length of each device is larger than the L_{avg} of the corresponding device are selected. Similarly for hold time and leakage objectives, the variants in which gate-length of each device is smaller than the L_{Avg} of the corresponding device get selected. For dynamic power, the variants in which the gate-area of each device is larger than the gate-area of the corresponding litho-simulated device are selected. When multiple variants meet the selection criteria, the one that minimizes the error ($= |L^j_{\text{Avg}} - L^j_{\text{Var}}|$, where L^j_{Avg} is the L_{Avg} of device j and L^j_{Var} is the gate-length of device j in the variant) is chosen.

   After a cell variant is chosen for a cell instance for an analyses objective, the Verilog file is modified to reflect the binding of the cell instance to the cell variant. The modified Verilog can be used by an off-the-shelf analyses tool. The accuracy of the analyses increases with the number of variants in the library at the cost of library characterization time.

2. *Transistor-level modeling.* If library characterization is not feasible, transistor-level modeling (TLM) may be used to estimate the variations in performance and power due to process variations. At the core of our delay modeling routine is an RC delay model. The delay is recomputed for every target output state. Currently the model does not distinguish between different input transitions leading to the same output state. A set of channel connected devices is referred to as a stage, and forms the basic unit of analysis. The next step is to identify devices that contribute significantly to a particular transition. These dominant devices are all devices that are part of a stack that connects the stage output to power or ground. Performing series-parallel reduction on the dominant devices, each stage is reduced to an RC pair. Both gate and junction capacitances are considered. The delay of each stage is expressed as the product of the equivalent resistance and the load capacitance of that stage. The resistance of each device is a function of its gate-width and length. TLM obtains these values from pre-created lookup tables generated by SPICE simulations. The computed delay overheads are used to modify the timing arc delays during static timing analysis to facilitate full-chip timing analyses.

   To estimate the leakage of the cell, we identify leakage dominant devices for each input state. It is known that stacked devices have very low leakage. Only devices that are in the off state and are not series connected to any other off devices are labeled dominant. The total cell leakage corresponding to a state is simply the sum of the off-currents of the dominant devices. Similar to resistance values, the off-currents are obtained from a lookup table generated by SPICE simulations. The average leakage of the cell is the average over all input states. We note that the absolute delay and leakage values are not of interest here and only the relative overheads due to gate-length variation are required to be accurate. Table 1 shows the delay and leakage overhead accuracies for our transistor level modeling method for various cells, for a particular length assignment. The accuracy suffers due to layout dependent effects such as well proximity, stress, etc. as well as the “lumped” nature of the delay model we use in the current implementation. However, the level of accuracy provided by TLM is sufficient for the modeling purpose described in this paper.

Cell	Delay Overhead (%)		Leakage Overhead (%)	
	SPICE	TLM	SPICE	TLM
INV	-4.8	-8.1	42.68	49.37
NAND	-7.3	-11.2	53.93	60.69
AND	-6.8	-7.8	50.02	56.56
AOI	-6.3	-6.5	51.82	57.49
MUX	-5.8	-4.7	50.94	56.77

Table 1. Transistor-level modeling matching accuracy.

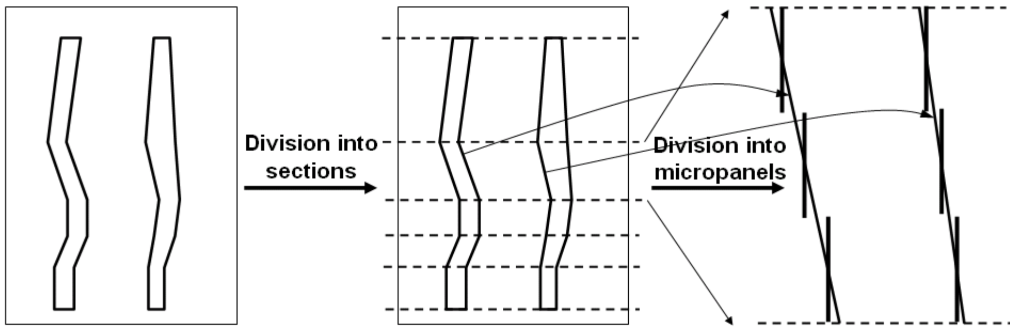

Figure 2. Steps involved in shape simplification for capacitance computation.

3. INTERCONNECT ANALYSIS

Our interconnect analysis flow computes the change in parasitics caused due to mismatch between drawn and litho-simulated interconnect shapes. We update the standard parasitic extended format (SPEF) database with the changed parasitics and then an off-the-shelf timing analysis tool can be run. If the output of lithography simulation are contours, we simplify the contours to polygons by a piecewise linear approximation. Resistance of an interconnect is computed by integration over the length. Since interconnects are polygons this reduces to addition of resistances of trapezoids (from top view) and can be done very efficiently by analytical formulas.

We iterate over coupling capacitances found in the SPEF database and analyze the two interconnects between which the coupling capacitance is computed simultaneously. Since SPEF may have long interconnects fractured during parasitic extraction, we use node coordinates, that can be optionally specified in SPEF, to establish a mapping between fractured interconnects and routing segments in the design. Figure 2 shows the steps involved in our shape simplification flow for coupling capacitance computation. Without loss of generality we assume the pair of interconnects to be vertical. Horizontal lines are drawn through all vertices of the two interconnect polygons to cut the polygons into sections. Two adjacent horizontal lines contain a section pair, one from each of the two polygons, between them. To compute the coupling capacitance between a section pair, the sections are split into horizontally-aligned micropanels. Values of capacitances for a pair of horizontally-aligned micropanels are obtained from a lookup table. The lookup table lists values of parallel plate capacitance and also includes fringing components of the capacitance values, which are impacted by the presence of metal wires above and below pair of polygon shaped metal wires. Capacitances for different micropanels are summed up to find the capacitance between sections and summed up for all sections to find capacitance between the interconnect pair. With the same method, we also find the capacitance for corresponding drawn shapes and compute the change in capacitance to update the SPEF database. The capacitance lookup table is created using 3D field solver simulations for template geometries generated for a technology. Capacitance values are obtained after interpolation using the following parameters: (1) widths of the interconnect pair, (2) spacing, (3) layer, and (4) densities of above and below layers. The runtime increases with the complexity of the interconnect polygons and the number of micropanels created.

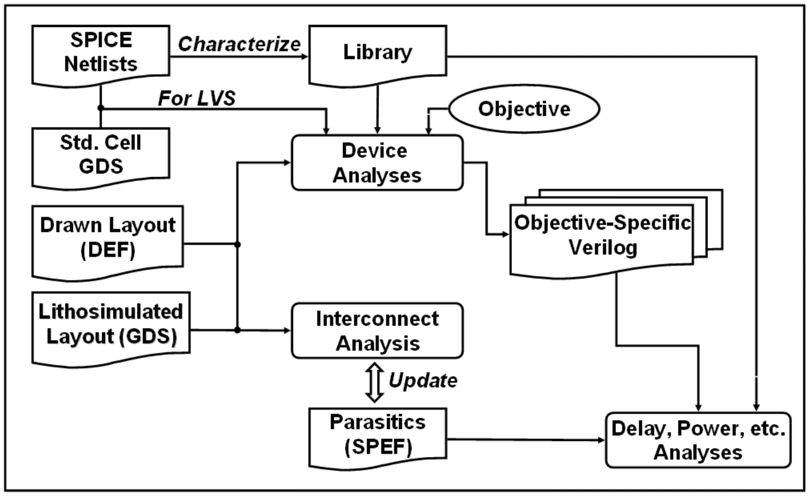

Figure 3. Lithography simulation-based design analyses flow.

4. FULL-CHIP ANALYSES

Figure 3 illustrates the complete analyses flow. The DEF file, that contains layout information for cells and interconnects along with connectivity, is used to correlate the shapes in litho-simulated GDS with cells and interconnects. Within a cell, device locations are correlated with device names as a byproduct of layout versus schematic (LVS) between cell GDS's and SPICE netlists. Our full-chip analyses flow iterates over all cell instances and invokes device analyses. To improve the runtime, full-chip analyses optionally takes the optical radius (i.e., radius beyond which proximity effects fade) and for cells that do not have other cells within the optical radius cached results are used instead of device analyses.

4.1. Mixed-mode analyses

As described in Section 2, Verilog files differ for the various analyses objectives. Maintaining objective-specific Verilog files can be cumbersome and is non-standard. We propose a hybrid approach that assigns different objectives to individual cells, based on the sensitivity of an objective to each cell, to generate a single Verilog file that allows accurate analyses for all objectives. An objective for each cell is selected in the following order:

1. Hold time. We first perform hold-time analysis with objective set as hold. We note that hold-time analysis yields the most accurate hold estimates while setup-time analysis yields the most inaccurate hold estimates. So we perform hold-time analysis again but with setup as the objective. For each cell we find the difference between hold-time slack at the two objectives and if that difference is larger than a user-configurable fraction of the hold-time (for hold objective), then the cell is flagged as hold critical. Hold objective is assigned to the hold-critical cells. The user-configurable fraction allows near-critical paths and not just the most critical path to be flagged as timing critical.

2. Setup time. In a way similar to hold time, setup-critical cells are identified and the ones that are not hold-critical are assigned the setup objective. If many cells are found to be simultaneously hold- and setup-critical, mixed-mode analyses should not be performed.

3. Capacitance. To accurately load the hold- and setup-critical cells, capacitance objective is assigned to the cells that: (1) load cells with hold or setup objective assigned, and (2) have not been assigned an objective.

4. Leakage or dynamic power. Only one of leakage or dynamic power can be performed at a time. We assign the leakage or dynamic power objective to all cells that do not have an objective assigned.

Table 2. Testcases used in our experiments.

Circuit	Source	Cells	Nets	IO Pads
s1423	ISCAS'89	1406	708	23
c5315	ISCAS'85	1520	1698	301
AES	opencores.org	25824	26083	388
OpenRisc	opencores.org	58999	59374	374

Table 3. Delay, leakage, and dynamic power estimates after layout, and after lithography simulation at $0nm$ and $100nm$ defocus using the proposed flow.

Circuit	Post-layout			Litho-sim at 0nm defocus			Litho-sim at 100nm defocus			CPU
	Delay	Leakage	Dynamic	Delay	Leakage	Dynamic	Delay	Leakage	Dynamic	
	(ns)	(μW)	(mW)	(ns)	(μW)	(mW)	(ns)	(μW)	(mW)	(s)
s1423	1.221	29.112	0.222	1.229	29.122	0.222	1.064	44.723	0.218	60
c5315	0.639	95.220	1.369	0.647	95.235	1.372	0.550	160.805	1.337	151
AES	2.155	582.707	2.858	2.157	584.917	2.864	2.135	904.541	2.797	2221
OpenRisc	0.700	3415.424	16.920	0.704	3411.665	16.962	-	-	-	5022

5. EXPERIMENTS AND RESULTS

In this section we present our experimental setup and results. We show that delay and power estimates after layout and from our flow differ considerably. The proposed flow is fast enough to be run on large testcases in practical runtimes.

5.1. Experimental Setup

Testcases used in our experiments are summarized in Table 2. We use SPICE models and cell netlists from a leading foundry and commercial tools for cell characterization, and testcase synthesis, layout and extraction. We use Mentor Calibre v.v9.3_5.9 for OPC and lithography simulation.

5.2. Results

Table 3 presents circuit delay, leakage, and dynamic power of our four testcases analyzed: (1) after layout, (2) using the proposed flow with lithography simulation performed with zero defocus, and (3) using the proposed flow with lithography simulation performed with $100nm$ defocus. We observe that circuit performance and power is close to post-layout estimates at zero defocus. Unfortunately, RETs are not as effective at non-ideal process conditions and we observe significant change in performance and power are $100nm$ defocus. With our OPC recipes, linewidths tend to decrease with increasing defocus for most patterns. Therefore, leakage increases tremendously, circuit time improves and hold-time violations become likely.

Table 4 presents the accuracy of mixed-mode analyses with respect to objective-specific analyses. As discussed in Section 4, mixed-mode analyses generates a single design that is accurate for all objectives (i.e., matches for a particular objective with the design generated by analyses specific for that objective). We observe that setup slack, hold slack, and leakage estimates from mixed-mode analysis match reasonably well with analyses with setup, hold, and leakage as objectives respectively.

6. CONCLUSIONS

Power and performance estimates after layout can be substantially different from on-silicon performance. Lithography simulation predicts on-silicon geometries for given process settings. In this paper we proposed a flow to use the lithography simulation results to predict on-silicon power and performance. For device analyses, we perform steps to simplify gate contours from lithography simulation to regular rectangular gates. To facilitate cell-level power and performance analyses, we proposed a methodology to map printed cells to cell variants (cells of similar functionality and drive strength but with non-nominal gate-lengths) in the library. We also proposed

Table 4. Accuracy of mixed-mode analyses with respect to individual objective-specific analyses. Circuit c5315 is combinational so hold-time analysis is not applicable

Setup timing slack (ns)				
Circuit	Objective			
	setup	hold	leakage	mixed-mode
s1423	**-0.0043**	0.0000	0.0000	**-0.0039**
c5315	**0.0000**	0.0016	0.0016	**-0.0001**
AES	**0.1000**	0.1017	0.1017	**0.10173**

Hold timing slack (ns)				
Circuit	Objective			
	setup	hold	leakage	mixed-mode
s1423	0.0931	**0.0927**	0.0927	**0.0927**
c5315	NA	NA	NA	NA
AES	0.0002	**0.0002**	0.0002	**0.0002**

Leakage (μW)				
Circuit	Objective			
	setup	hold	leakage	mixed-mode
s1423	43.112	44.741	**44.723**	**44.616**
c5315	158.807	160.851	**160.805**	**160.771**
AES	895.552	904.751	**904.541**	**904.541**

an alternative transistor-level modeling-based analyses flow. Imperfections in printing of the interconnects alter their parasitics and consequently performance. Our interconnect analyses flow simplifies the contours from lithography simulation and uses a pre-created lookup table to estimate the changes in parasitics. The parasitic database is then updated to enable lithography simulation-based timing analyses.

In addition to speed and accuracy improvements of the various modules, we are specifically working in the following directions:

- Currently for L_{Avg} computation we do not consider the impact of the slice locations. However, due to narrow-width effects, slices near the gate edge affect device performance and power differently than those at the center.[3] We plan to consider these effects for more accurate L_{Avg} computation.
- For cell-level analyses we plan to develop a hybrid methodology that uses transistor-level modeling to interpolate between characterized cell variants to improve the analyses quality without the need for large number of variants.

REFERENCES

1. "International Technology Roadmap for Semiconductors," http://public.itrs.net
2. P. Gupta and F.-L. Heng, "Toward a Systematic-Variation Aware Timing Methodology," in *ACM/IEEE Design Automation Conference*, pp. 321–326, 2004.
3. P. Gupta, A. B. Kahng, Y. Kim, S. Shah and D. Sylvester, "Modeling of Non-Uniform Device Geometries for Post-Lithography Circuit Analysis," in *SPIE Microlithography Conference*, to appear, 2006.
4. F.-L. Heng, J.-F. Lee and P. Gupta, "Toward Through-Process Layout Quality Metrics," in *SPIE Microlithography Conference*, pp. 161–167, 2005.
5. A. B. Kahng, S. Muddu and P. Sharma, "Defocus-Aware Leakage Estimation and Control," in *International Symposium on Low Power Electronics and Design*, pp. 263–268, 2005.
6. A. B. Kahng, C.-H. Park, P. Sharma and Q. Wang, "Lens Aberration-Aware Placement for Across Field Line-Width Control," in *Proc. Design Automation and Testing in Europe*, 2006, to appear.
7. M. Orshansky, L. Milor, P. Chen, K. Keutzer and C. Hu, "Impact of Spatial Intrachip Gate Length Variability on the Performance of high-Speed Digital Circuits," in *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, pp. 544-553, 2002.
8. J. Yang, L. Capodieci and D. Sylvester, "Advanced timing analysis based on post-opc extraction of critical dimensions," in *ACM/IEEE Design Automation Conference*, pp. 359–364, 2005.