Performance- and Energy-Aware Optimization of BEOL Interconnect Stack Geometry in Advanced Technology Nodes

(Invited)

Kwangsoo Han[‡], Andrew B. Kahng[†‡], Hyein Lee[‡] and Lutong Wang[‡]
[†]CSE and [‡]ECE Departments, UC San Diego, La Jolla, CA 92093
{kwhan, abk, hyeinlee, luw002}@ucsd.edu

***Abstract*—In advanced technology nodes, BEOL interconnect stack geometry has become a key lever for design enablement. The rapid increase of interconnect RC leads to not only performance loss from interconnect delay increase, but circuit power and area degradation as well. Thus, optimization of BEOL dimensions (i.e., wire width, spacing and thickness subject to a given layers pitch constraint) is crucial to achieve better product performance, power and area. However, it is not obvious how to optimize BEOL dimensions, especially in sub-10nm nodes. In this work, we study BEOL interconnect stack geometry by exploring wire aspect ratio (AR) and wire line-space duty cycle (DC). We perform SPICE-based analyses of timing path delays to find delay- or power-optimal (AR,DC) combinations, and also perform block-level studies with placed and routed designs. Based on our experimental results, we provide various insights on BEOL stack geometry: (i) optimal (AR,DC) for a given wire pitch with respect to power and delay; (ii) sensitivities of optimal (AR,DC) to circuit parameters (e.g., driver strength, input slew, output load, wirelength); (iii) optimal (AR,DC) when multiple interconnect layers are considered; and (iv) potential impacts of BEOL stack optimizations within future design-aware manufacturing and/or manufacturing-aware design methodologies.**

I. Introduction

In advanced technology nodes, power and performance requirements are increasingly stringent even as classical Moore's-Law scaling has slowed down. BEOL interconnect stack geometry has become a key lever for design enablement. Reasons for this include: (i) the resistance of Cu interconnect has increased dramatically in sub-100nm nodes due to grain boundary and trench liner effects [10]; and (ii) the scaling of effective dielectric constant has slowed in recent years, resulting in severely increased interconnect capacitance [20] and diminished performance benefits at new nodes. The resulting rapid increase of interconnect RC leads to not only performance loss from interconnect delay increase, but circuit power and area degradation as well. In this paper, we study the potential value of BEOL interconnect stack geometry optimizations, by exploring wire aspect ratio (AR) and wire line-space duty cycle (DC). Our broad objective is to assess whether new manufacturing-aware design (MAD) [7] and design-aware manufacturing (DAM) methodologies can contribute "equivalent scaling" in the N7/N5 nodes and beyond.

Industry implementations. For interconnect geometry optimization, wire height/width aspect ratio (AR) and wire width/pitch duty cycle (DC) are the two obvious levers for a given metal pitch value and BEOL process (see Figure 1). In the most recent technology nodes, IC companies have deviated from "classic" 2:1 AR and 50% DC for each metal layer, for reasons of performance, energy, reliability and manufacturability. Narasimha et al. [11] achieve 20% reduction in RC delay by optimizing liner resistivity and metal line aspect ratio of 1× layers in IBM's high-performance 45nm SOI technology node. Jan et al. [9] describe two different interconnect geometries that meet different product types, power and performance goals in Intel's 22nm node. Figure 2 depicts the BEOL stacks for high-performance CPU and high-density SoC in Intel's 22nm node [9]. For high-performance CPU, thicker and wider wires with large AR and DC are observed. By contrast, flat wires with moderate DC are used for high-density SoC. Zhu et al. [16] have patented a local optimization to improve SoC performance using two BEOL stacks. The first stack is used for non-critical blocks while the second stack, with larger line width and via width, is used for critical blocks.

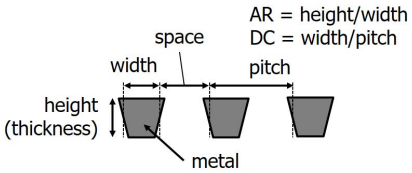

Fig. 1. Illustration of height/width aspect ratio (AR) and width/pitch duty cycle (DC).

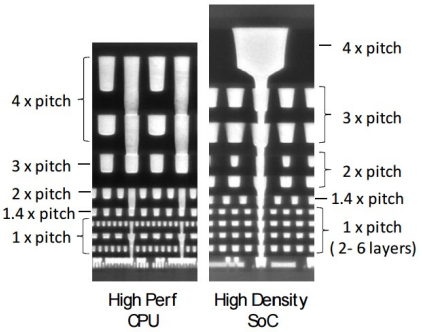

Fig. 2. Interconnect architecture comparison of 22nm CPU and SoC [9].

Current approaches and limitations. Even as sophisticated IC companies have adopted various choices of AR and DC for different design targets in each technology node, to our knowledge it is not obvious how to optimize BEOL dimensions, and there is no general methodology to identify an "optimal" BEOL stack option for a given design. Two levels of optimization exist in previous works: (i) device level, and (ii) block level. For (i), the Elmore delay model [15] provides fast modeling of RC networks. Elfadel et al. [5] describe AQUAIA, which enables fast modeling and simulation of delay, slew and crosstalk. Faruk et al. [6] utilize AQUAIA for variability modeling with different line widths, heights, pitches and dielectric constants. Bakoglu et al. [3], Ismail et al. [8] and Pamanuwa et al. [12] develop delay models considering repeater insertion, inductance and coupling capacitance. These techniques enable fast modeling of delay and energy for given driver, load and interconnect structures. Thus, for instance, "optimal" single-stage (AR,DC) with fixed driver and load can be determined by sweeping (AR,DC) combinations. For (ii), Anand et al. [1][2] develop a framework to optimize a metal stack in a more global sense. The authors conclude with a suggestion of low AR and DC. Takahashi et al. [14] propose a methodology for determining overall interconnect strategy, including repeater insertion, as well as adoption of new metal and dielectric materials. Other works including [17] optimize DC only, for long interconnects. Recently, [13] performs block-level validations of BEOL optimization based on results from single-stage simulation for advanced nodes. The work of [4] suggests that different optimal (AR,DC) combinations may apply when considering driver and load.

Our approach. In this work, we study optimization of BEOL interconnect stack geometry through exploration of wire aspect ratio (AR) and wire width/pitch duty cycle (DC) impacts in sub-10nm nodes. We perform SPICE-based analyses of timing path delays to find delay- or power-optimal (AR,DC) combinations, and also perform block-level studies with placed and routed designs. Based on our experimental results, we provide various insights on BEOL stack geometry: (i) optimal (AR,DC) for a given wire pitch with respect to power and delay; (ii) sensitivities of optimal (AR,DC) to circuit parameters (e.g., driver strength, input slew, output load, wirelength); (iii) block-level optimal (AR,DC) with multiple interconnect layers; and (iv) potential impacts of future design-aware manufacturing (DAM) and/or manufacturing-aware design (MAD) methodologies [7] that co-optimize product designs and BEOL interconnect stacks.

The contributions of this work are summarized as follows.

- We explore various wire dimensions using SPICE simulation and determine "optimal" wire dimensions (AR,DC) for a given metal pitch with respect to power and performance.
- We study the sensitivities of (AR,DC) optimization to several parameters that determine circuit performance and power (e.g., driver strength, input slew, output load, wirelength).
- We show that performance and power results from a standard place-and-route (P&R) flow (including post-route parasitic RC extraction (PEX) and static timing analysis (STA)) are consistent with our SPICE simulation-based results.
- We investigate the potential impacts of future design-aware manufacturing (DAM) and manufacturing-aware design (MAD) methodologies by performing P&R with real block-level designs.

The remainder of this paper is organized as follows. Section II reviews two important related previous works. In Section III, we describe our study based on path-based single-stage SPICE simulation. In Section IV, we explain our block-level validation using optimal AR and DC choices from SPICE simulation. Section V describes our study of DAM and MAD methodologies, including experimental setup, results and analysis. We give conclusions and future research directions in Section VI.

II. Previous Work

In this section, we review two important related works on BEOL interconnect optimization for sub-14nm nodes.

Shah [13] analyzes the impact of local layer interconnect dimensions on the performance of single-stage, single-size inverter circuits for various technology nodes, and identifies optimal AR and DC values with respect to slew-bounded delay and slew-bounded energy-delay product (EDP). Based on SPICE simulation and field-solver analyses for single gate-interconnect stages, it is shown that the combination of low AR and high DC can achieve a better overall performance for advanced nodes. The author provides validations using predictive technology models, and furthermore studies multi-stage impact using a physical design flow for sample benchmark designs and a random path model. While [13] considers effects seen in advanced process technologies, such as new barrier and dielectric materials and the impact of process variations based on the ITRS roadmap [20], several limitations are noted. First, predictive technology models and scaled libraries for advanced nodes are based on generic planar-bulk 32/28nm libraries, with mismatched scaling of parasitics versus the dimensions of devices and interconnects; this may not accurately match current and impending 7nm/5nm FinFET nodes. Second, optimal wire dimensions are determined only for local metal layers, whereas long interconnects on higher layers may play a more important role in BEOL interconnect optimization. Third, the design-level analysis of [13] only compares power and timing performance implications of the suggested optimal AR and DC values to those of ITRS predicted values. Tradeoffs of AR and DC at the design level are not contemplated in [13], as the work focuses only on single-stage analysis for fixed-size inverters.

Ciofi et al. [4] investigate the impact of wire geometry on the resistance, capacitance, and RC delay of Cu/low-k damascene interconnects for fixed line-to-line pitch for the 7nm logic technology node. The resistance is computed by applying a semiempirical resistivity model and the capacitance is simulated by means of a 2D field solver. The authors show that RC delay can be significantly reduced by trading capacitance for resistance with wider and thicker wires. They also show that a given RC delay can be achieved with several

geometries, which provides a useful degree of freedom for system-level optimization. Next, the authors suggest that the optimal point for circuit performance in terms of power and delay may differ from the RC delay and give delay and power contours for different AR and DC combinations for $\times 1$ and $\times 4$ drivers with a wirelength of 300 contacted poly pitches (CPPs). However, this work does not suggest any method to incorporate the existing findings to block-level designs, which include different types of cells, and semi-global/global interconnect layers.

III. Path-Based Simulation

We now describe our methodology to evaluate power and delay impacts of various BEOL interconnect stack geometries, based on path-based SPICE-level simulations. Based on SPICE-level simulation, we study the sensitivities of delay and power to driver strength, wirelength, output load and input slew. We further show that the P&R flow's analysis results (i.e., including PEX and STA) are well-correlated with SPICE-level simulation results.

A. Single-Stage SPICE Simulation

For SPICE-level simulation, we evaluate the power and delay impacts of various wire dimensions using single-stage circuits. We first extract RC values per unit length (μm) with different AR and DC values for three metal layer types[1] and use these values to construct single-stage circuits with various configurations (i.e., sizes of buffers, wirelength, output load and input slew).

RC extraction for various wire dimensions. We perform parasitic RC extraction using Cadence QRC [19] with QRC techfiles and LEF files [21] to obtain per-unit length (μm) R and C values for various wire dimensions. To model different wire thicknesses, we generate multiple QRC techfiles with ICT [19] files that are modified with various thickness values for each metal layer, using Cadence QRC Techgen [19]. To sweep metal width values, we modify the "WIDTH" and "SPACING" fields in LEF files.[2] To obtain wire RC with fine-grained width and thickness values, we perform linear interpolation. Figures 3(a) and (b) show the contour maps of per-unit length (μm) $\bar{r}$ and $\bar{c}$ for metal pitch $32nm$[3], respectively. Both $\bar{r}$ and $\bar{c}$ increase with larger width and thickness values, as expected. Our observations are consistent with those reported in [4].

Circuit structure for SPICE simulation. Figure 4 shows the circuit structure that we use for SPICE simulation. With the extracted per-unit length (μm) values $\bar{r}$ and $\bar{c}$, we compute wire resistance R_{wire} and capacitance C_{wire} for a given wirelength. We then construct an RC circuit using the $\Pi 3$ model for wire segments.

Sensitivity of power and delay to input configurations. Using SPICE simulation with the circuit structure described above, we study the sensitivities of performance and power

[1]We consider $1\times$, $1.5\times$ and $2.5\times$ layers (i.e., pitch = 32, 48 and $80nm$).

[2]The default ICT and LEF files that we use are provided by our collaborators at a leading technology consortium. In our study, we do not investigate patterning or manufacturing issues that may pertain to different AR,DC combinations.

[3]As in [4], we focus on $1\times$ metal layer in N7/N5 nodes.

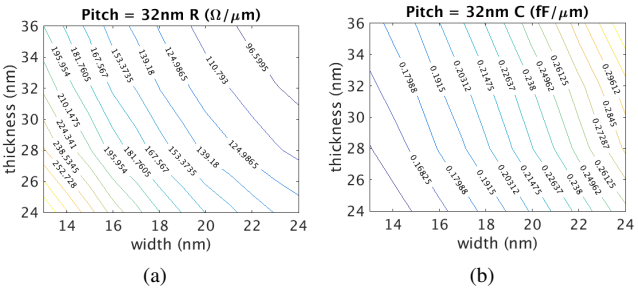

Fig. 3. Contour maps of (a) resistance and (b) capacitance per unit length (μm) for metal pitch $32nm$.

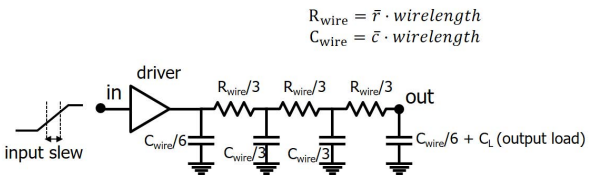

Fig. 4. Circuit structure for SPICE simulation.

to several parameters with $1\times$ metal layer (pitch = $32nm$). We vary driver strength, wirelength, output load and input slew, as follows. The values in bold font are defaults.

- driver strength = {X1, **X4**, X8, X16}
- wirelength = {$5\mu m$, **10**μm, $15\mu m$, $20\mu m$}
- output load = {$2fF$, $3fF$, **5**fF, $10fF$}
- input slew = {**50**ps, $100ps$}

Figures 5(a), (b), (c) and (d) show the power and delay contour maps for various driver sizes, i.e., BUF_X1, BUF_X2, BUF_X8 and BUF_X16, respectively. We observe that (i) with BUF_X1, smaller width and thickness values are always better for both power and delay (no tradeoff between power and delay is observed), and (ii) delay-optimal wire dimension changes according to the driver strength. The reason for (i) is that the effective resistance of the BUF_X1 is relatively larger than the resistance of the wire, which results in a larger impact of wire capacitance (compared to that of wire resistance).

Figures 6(a), (b), (c) and (d) show power and delay contour maps with different wirelength values, i.e., $5\mu m$, $10\mu m$, $15\mu m$ and $20\mu m$, respectively. The delay contours move toward the right and upward as the wirelength increases, as expected.

Figures 7(a), (b), (c) and (d) show power and delay contour maps with output load values $2fF$, $3fF$, $5fF$ and $10fF$, respectively. We observe that larger width values are preferred as the load cap increases. This might be because as the load capacitance increases, the stage delay dependence on wire capacitance lessens, and the relative sensitivity to wire resistance increases.

Figures 8(a) and (b) give power and delay contours showing the sensitivity to input slew. Although the absolute delay and power values are different, the relative delay and power values do not change significantly, suggesting that input slew is not a critical factor in determining power- and/or delay-optimal wire dimensions.

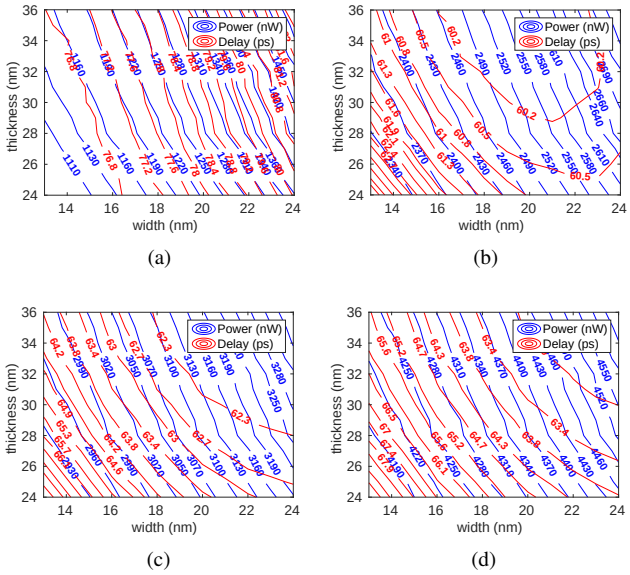

Fig. 5. Sensitivity of power and delay to driver strength: (a) BUF_X1 (b) BUF_X2, (c) BUF_X8 and (d) BUF_X16.

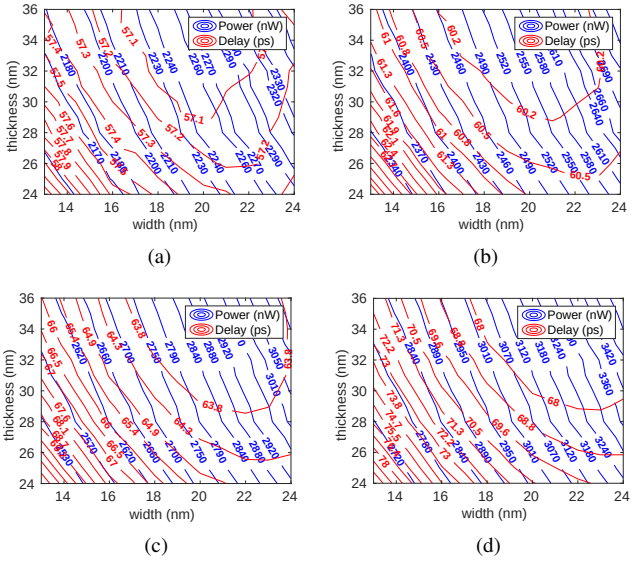

Fig. 6. Sensitivity of power and delay to wirelength: (a) $5\mu m$, (b) $10\mu m$, (c) $15\mu m$ and (d) $20\mu m$.

B. Validation on single-stage paths.

We have confirmed consistency between our SPICE-based results and the block-level analysis flow within a commercial P&R tool [18]. As shown in Figure 9, for each type of layer, we generate eight bits[4] of signal wires, and compare (i) the stage delay of the middle signal as reported by the P&R tool's static timing analysis (STA) capability to (ii) the delay reported from SPICE simulation. To avoid effects of via parasitics, the drivers are located at each wire input port with modified driver output pins on the metal layer of the wire. For each type of layer, we sweep DC (i.e., 0.4, 0.45, 0.5, 0.55, 0.6, 0.65 and 0.7) and AR (i.e., 1.5, 1.75, 2.0). We use three wirelength values (i.e., $100\mu m$, $200\mu m$ and $300\mu m$) and three output load values (i.e., $5fF$, $10fF$ and $15fF$). Figure 10 plots the delays reported by the P&R tool and SPICE simulation, suggesting strong correlation of STA in P&R with SPICE simulation.

[4]Our background study indicates that for each signal wire, more than one neighboring signal wire contributes to its capacitance. Out of eight parallel wires, we consider the fourth and fifth to be "middle" wires.

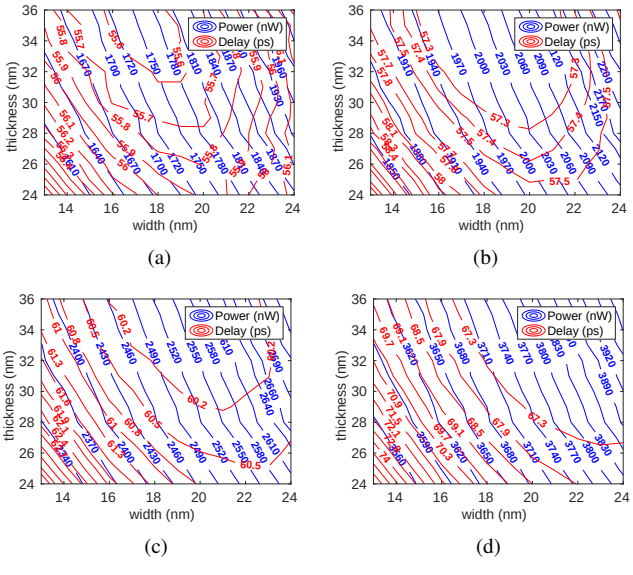

Fig. 7. Sensitivity of power and delay to output load: (a) $2fF$, (b) $3fF$, (c) $5fF$ and (d) $10fF$.

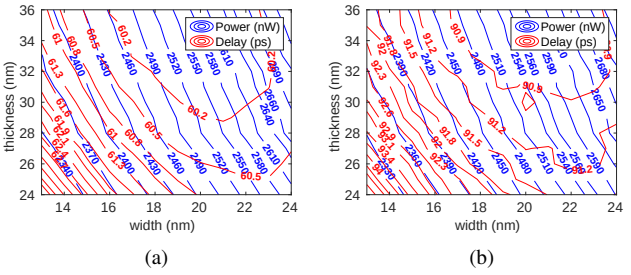

Fig. 8. Sensitivity of power and delay to input slew: (a) $50ps$ and (b) $100ps$.

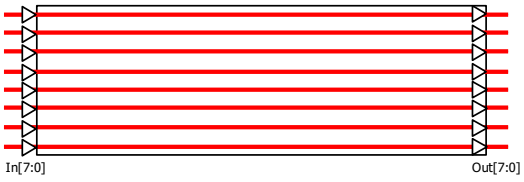

Fig. 9. The artificial testcase with eight bits of single-stage paths.

IV. Block-Level Validation

For block-level validation on real designs, we enable a commercial P&R tool flow with the following steps. (i) We first group cells by their driver strength, and place-and-route using Cadence Innovus [18] within each group. From SPICE-based simulation studies, driver strength is the key factor in determining optimal wire dimensions. Thus, by limiting the range of driver strength of cells, we can find better correlation to SPICE-based results. In our implementation, we partition

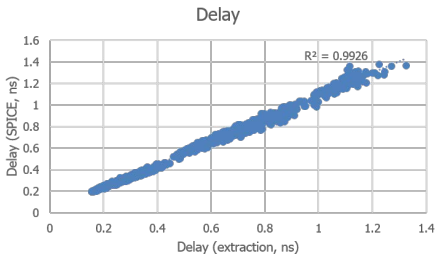

Fig. 10. Correlation between timing reports from P&R timing analysis and SPICE simulation.

the cells into two groups, with driver strengths of $\times 1$ and $\times 4$, respectively. (ii) To avoid the tool's noise, we use one fixed post-routed layout for a design with default wire dimensions (i.e., AR[5] = 2.0, DC = 0.5) for all metal layers. Then we extract parasitics by using different BEOL stack by varying (AR,DC) combinations, and report design metrics (timing and power). For DC, we sweep from 0.5 to 0.7 with $1nm$ step in wire width for $1\times$ and $1.5\times$ layers, and with $2nm$ step for the $2.5\times$ layer. For AR, we use 1.50, 1.75, 2.00 and 2.25 for selection. In our experiment, we run our extraction flow for all (AR,DC) combinations for each layer type, while fixing the other layer types with the default configuration.

Figures 11 and 12 show the contour maps of delay (power) from both SPICE simulation and place-and-route runs. For SPICE simulation, we assume a wirelength of $10\mu m$ and FO3 capacitance load.[6] For block-level validation, we use a low-density parity-check (LDPC) decoder block [22] as our reference design. We report the contour maps of delay and power, with varying metal width and thickness. From Figures 11(a) – (c), we can see that for the $\times 1$ cell group, there is no tradeoff between power and delay. Therefore, optimal power and delay are always achieved with smaller width and thickness, which is verified in Figures 11(d) – (f). From Figures 11(a) – (c), we see that there is a tradeoff for the $\times 4$ cell group. For better delay, medium (resp. small) DC is preferred with higher AR for $1\times$ (resp. $1.5\times$) metal layers, while smaller (AR,DC) is preferred for $2.5\times$ metal layers, which is also verified from our block-level design.

For both cell groups, smaller (AR,DC) is always preferable in terms of power, due to smaller capacitance; this can be seen in Figures 13(d) – (f). To create a simplified real-world configuration for high-performance blocks, we rerun the P&R flow enabling both $\times 1$ and $\times 4$ cells with the tightest clock period achievable[7] and we plot the contour maps in Figure 13. We also show the wirelength distribution per layer type, labeled by the cell group of the driver. As shown in Table I, since $\times 4$ cells drive more than double the wirelength on every layer, the contour plot is more similar to that of $\times 4$'s.

Overall, if the designs with $\times 1$ cells can be seen as low frequency and low power, and designs with $\times 4$ cells can be seen as high frequency and high performance, our observations show that those designs prefer distinct BEOL stacks, as shown in Figure 2. For a simplified real-world high-performance configuration with cells of multiple driver strengths, the above preference of BEOL stack from high-drive cells still holds as larger cells drive at least $2\times$ wirelength on each metal layer, suggesting that optimization of (AR,DC) towards high-drive cells may be beneficial for every layer.

TABLE I
WIRELENGTH DISTRIBUTION PER LAYER TYPE (NORMALIZED) GROUPED BY DRIVER CELLS.

Layer		1×		1.5×		2.5×	
Driver		×1	×4	×1	×4	×1	×4
Design	AES	0.15	0.41	0.06	0.31	0.02	0.05
	LDPC	0.11	0.21	0.03	0.28	0.02	0.36

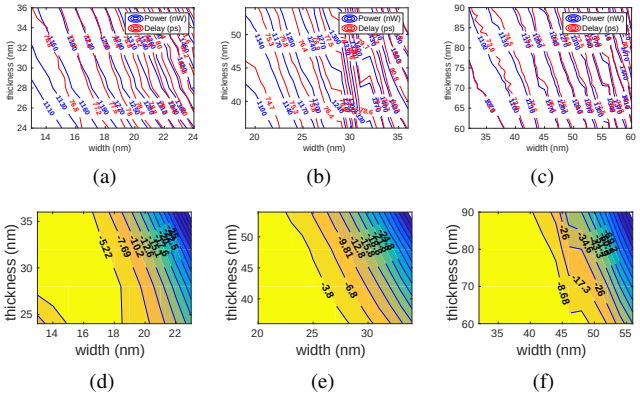

Fig. 11. Block-level validation to single-stage SPICE simulation: (a) – (c) contour maps of power and delay (BUF_X1) for $1\times$, $1.5\times$ and $2.5\times$ metal layers, respectively, assuming wirelength of $10\mu m$ and load of $2fF$; (d) – (f) contour maps of TNS (total negative slack) when varying (AR,DC) for $1\times$, $1.5\times$ and $2.5\times$ layers, respectively.

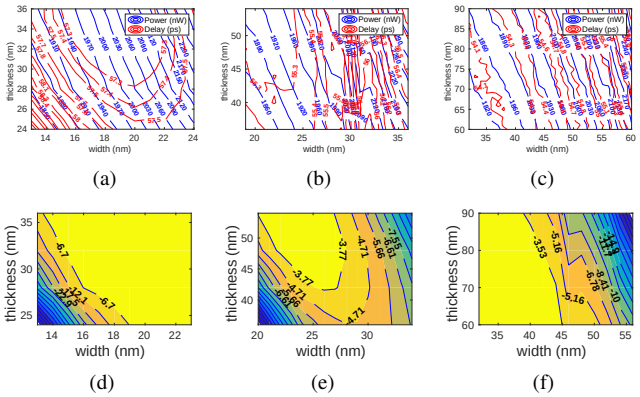

Fig. 12. Block-level validation to single-stage SPICE simulation: (a) – (c) contour maps of power and delay (BUF_X4) for $1\times$, $1.5\times$ and $2.5\times$ metal layers, respectively, assuming wirelength of $10\mu m$ and load of $3fF$; (d) – (f) contour maps of TNS (total negative slack) when varying (AR,DC) for $1\times$, $1.5\times$ and $2.5\times$ layers, respectively.

[5]For consistency over all metal widths, AR is henceforth defined as metal thickness divided by metal half-pitch.

[6]In our background study, we place-and-route seven designs from OpenCores [22], observing average net length of $2\mu m$ to $8\mu m$, and fanout of approximately three for each design. A similar configuration is used in [4].

[7]A timing target is considered to have been achieved if setup worst negative slack (WNS) $>$ -50ps.

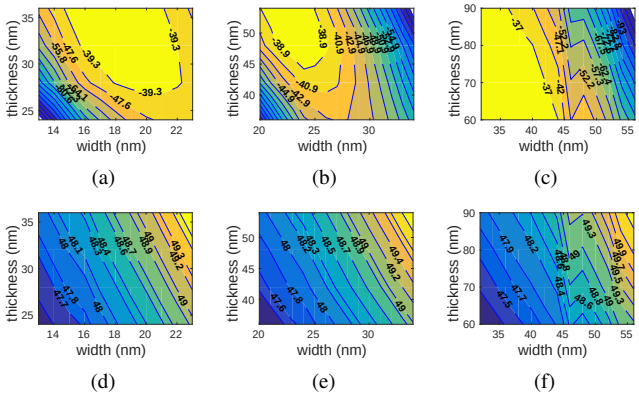

Fig. 13. Contour maps of (a) – (c) TNS (total negative slack) and (d) – (f) power when varying (AR,DC) for 1×, 1.5× and 2.5× layers, respectively.

V. Potential Benefits of Design-aware Manufacturing and Manufacturing-aware Design Methodologies

In this section, we explore the potential future benefits of design-aware manufacturing (DAM) and manufacturing-aware design (MAD) [7] methodologies. Design-aware manufacturing refers to the optimization of manufacturing to maximize the quality of a given product design. Here, DAM means that (AR,DC) is tuned according to the characteristics of each design. For example, in a DAM flow, we may propose specific BEOL stack for a given P&R solution. Manufacturing-aware design refers to the optimization during physical implementation, where a downstream, post-P&R optimization of (AR,DC) is assumed in the P&R flow, and thus is applied in both P&R and manufacturing.

A. Experimental Setup

We investigate the design freedoms of DAM / MAD, and their impacts, for a total of 108 BEOL stacks covering wide (AR,DC) ranges as follows. (i) We perform P&R using the 108 BEOL stacks. By covering a variety of BEOL stacks, we effectively explore MAD. (ii) For each implementation with a different BEOL stack, we perform PEX and STA using all BEOL stacks of the DAM study. This step shows DAM for each stack.

After P&R and PEX, we report the total negative slack (TNS) and the total power from STA for 108×108 data points (the total number of pairs of a MAD stack and a DAM stack). In the following discussion, we use the naming convention *(P{stack number}, R{stack number})* to represent the pair of a MAD stack and a DAM stack.

In our BEOL stack, we have two 1× layers, two 1.5× layers, and four 2.5× layers. For each type of layer, we choose from three DC combinations (i.e., 0.5, 0.6, 0.7). For all the layers, we apply a uniform AR from four combinations (i.e., 1.5, 1.75, 2.0, 2.25) to reduce the number of combinations. Thus, we have a total of $3\times3\times3\times4 = 108$ BEOL stacks. Table II summarizes a noteworthy subset of the BEOL stack configurations[8]. All experiments are performed at the typical-typical (TT) process corner, for the LDPC testcase with clock period of $0.8ns$, and with availability of both ×1 and ×4 cells.

TABLE II
A noteworthy subset of BEOL stack configurations.

Stack index	1×		1.5×		2.5×	
	AR	DC	AR	DC	AR	DC
1	1.50	0.5	1.50	0.5	1.50	0.5
4	2.25	0.5	2.25	0.5	2.25	0.5
18	1.75	0.5	1.75	0.6	1.75	0.6
25	1.50	0.5	1.50	0.7	1.50	0.5
55	2.00	0.6	2.00	0.6	2.00	0.6
108	2.25	0.7	2.25	0.7	2.25	0.7

B. Experimental Results

Figure 14 shows the results of the DAM and MAD studies. In the figure, The x-axis gives 108 different BEOL stacks[9] for P&R, and the y-axis gives both TNS and total power. For each P&R stack, we give the TNS after PEX and STA with the default BEOL stack in red dots and the respective power is represented in black bars. red columns show TNS after PEX and STA with all 108 stacks, while the orange columns show the power, also for all 108 stacks. We sort the P&R stack indices along x-axis, according to TNS based on the default BEOL stack. In this way, the impact of DAM is presented horizontally (by comparing between red dots and black bars), and the impact of MAD is presented vertically (shown in red and orange bars) for each P&R stack.

The red dots indicate that different physical implementations may result in up to 40% difference in TNS. For example, the TNS difference between the leftmost BEOL stack (P55,R55) and the rightmost BEOL stack (P25,R25) in red dots is $2.59ns$ (40% of TNS with (P55,R55)). This means that TNS can be improved by up to 40% by exploiting the DAM. Regarding the MAD, we observe that for the P&R implementation with the default BEOL stack P108, the TNS can vary from -9.15ns to -5.28ns. This means that we can improve the TNS by up to 49%. By combining the DAM and MAD exploration, the maximum improvement of the TNS is from -9.15ns ((P108,R1)) to -3.66ns ((P18,R4)), which is a 60% improvement, albeit dependent on the given timing specification.

The black bars in Figure 14 indicate that different physical implementations may result in up to 7% difference in power. Also, we can observe that the red dots and black bars stay steady for each red and orange column, suggesting that given a routed design, a particular BEOL stack may be preferred regardless of the BEOL stack used for P&R. Even though we sort the P&R stack by TNS, we cannot find a monotonic trend for power, suggesting a weak correlation of timing and power for each design.

VI. Conclusions

In this paper, we have investigated the potential impact of design-aware manufacturing (DAM) and manufacturing-aware design (MAD) methodologies to optimize BEOL dimensions in sub-10nm nodes. We study BEOL interconnect

[8]Index=36·i+12·j+4·k+l, where i, j and k are the indices of DC for 1×, 1.5× and 2.5× layers, respectively, and l is the index for AR. See the examples in Table II.

[9]Due to the limited space, we do not show the names of all the 108 BEOL stack options in the chart. Wire dimension information of noteworthy BEOL stack options is shown in Table II.

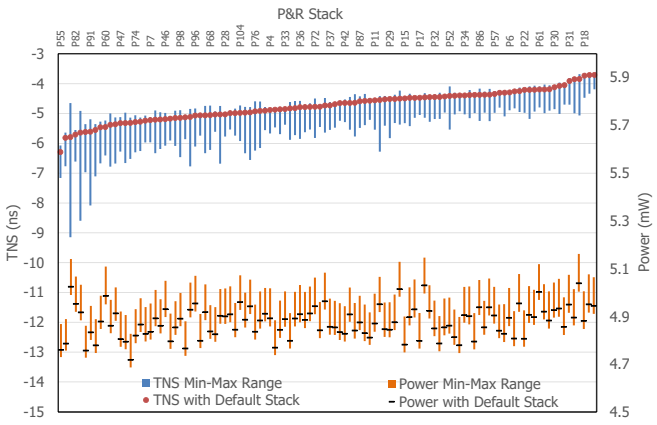

Fig. 14. Study of design-aware manufacturing (DAM) and manufacturing-aware design (MAD).

stack geometry by exploring the wire aspect ratio (AR) and duty cycle (DC). We perform SPICE-based analyses of timing path delays and correlate these with analyses in the P&R tool, using a single-stage artificial netlist construction. We also perform block-level studies with placed and routed designs. Based on our studies, we find the optimal (AR,DC) for a given wire pitch with respect to power and delay; we also show the sensitivities of BEOL stack geometry to circuit parameters and validate our SPICE analyses with real block-level designs. We further perform studies on design-aware manufacturing and manufacturing-aware design to explore the design freedoms and potential benefits of DAM and MAD. Large differences in design metrics exist across DAM and MAD. By proper utilization of DAM and MAD, we can save up to 60% in TNS and 7% in power for a particular LDPC testcase. Furthermore, based on our experiments, we conjecture that an optimal MAD and DAM BEOL stack exists for any given design.

Our future works include (i) the co-optimization of the front-end (i.e., gate sizing / buffer insertion, etc.) with the back-end (BEOL stack geometry); (ii) study on the impact of airgap layer and airgap-aware BEOL stack optimization. More specifically, we hope to study "chicken-and-egg" loop for MAD and DAM, where P&R is guided by the input BEOL stack option and the netlist changes accordingly, while the optimal BEOL stack option changes according to the design (netlist) information, such as driver strength, wirelength and slack distribution. Figure 15 shows an example flow for co-optimization of the design implementation front-end with the manufacturing technology back-end. In this flow, we suggest a big-loop optimization that considers the interactions between P&R and optimal BEOL stack options (including airgap layers).

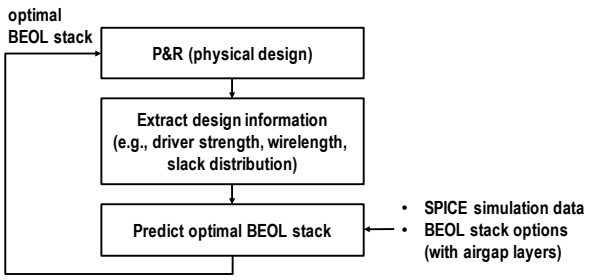

Fig. 15. Co-optimization of SoC physical implementation (design process) with BEOL stack optimization (manufacturing process). BEOL stack options include airgap layers.

Acknowledgments

We thank Praveen Raghavan and Peter Debacker of IMEC for providing key enablements used in our studies.

References

[1] M. B. Anand, H. Shibata and M. Kakumu, "Multiobjective Optimization of VLSI Interconnect Parameters", *IEEE Trans. on CAD* 17(12) (1998), pp. 1252-1261.

[2] M. B. Anand, H. Shibata and M. Kakumu, "Optimization Study of VLSI Interconnect Parameters", *IEEE Trans. on Electron Devices* 47(1) (2000), pp. 178-186.

[3] H. B. Bakoglu and J. D. Meindl, "Optimal Interconnection Circuits for VLSI", *IEEE Trans. on Electron Devices* 32(5) (1985), pp. 903-909.

[4] I. Ciofi, A. Contino, P. J. Roussel, R. Baert, V. H. Vega-Gonzalez, K. Croes, M. Badaroglu, C. J. Wilson, P. Raghavan, A. Mercha and D. Verkest, "Impact of Wire Geometry on Interconnect RC and Circuit Delay", *IEEE Trans. on Electron Devices* 63(6) (2016), pp. 2488-2496.

[5] I. M. Elfadel, M. B. Anand, A. Deutsch, O. Adekanmbi, M. Angyal, H. Smith, B. Rubin and G. Kopcsay, "AQUAIA: A CAD Tool for On-Chip Interconnect Modeling, Analysis, and Optimization", *Proc. Electrical Performance of Electronic Packaging*, 2002, pp. 337-340.

[6] M. G. Faruk, M. S. Angyal, O. Ogunsola, D. K. Watts and R. Wilkins, "Variability Modeling and Process Optimization for the 32 nm BEOL Using In-Line Scatterometry Data", *IEEE Trans. on Semiconductor Manufacturing* 27(2) (2014), pp. 260-268.

[7] P. Gupta and A. B. Kahng, "Manufacturing-Aware Physical Design", *Proc. ICCAD*, 2003, pp. 681-687.

[8] Y. I. Ismail, E. G. Friedman and J. L. Neves, "Equivalent Elmore Delay for RLC Trees", *IEEE Trans. on CAD* 19(1) (2000), pp. 83-97.

[9] C.-H. Jan, B. Uddalak, R. Brain, S.-J. Choi, G. Curello, G. Gupta and W. Hafez, "A 22nm SoC Platform Technology Featuring 3-D Tri-gate and High-k/Metal Gate, Optimized for Ultra Low Power, High Performance and High Density SoC Applications", *Proc. IEDM*, 2012, pp. 311-314.

[10] Q.-T. Jiang, M.-H. Tsai and R. H. Havemann, "Line Width Dependence of Copper Resistivity", *Proc. Interconnect Technology Conference*, 2001, pp. 227-229.

[11] S. Narasimha, K. Onishi, H. M. Nayfeh, A. Waite, M. Weybright, J. Johnson, C. Fonseca et al., "High Performance 45-nm SOI Technology with Enhanced Strain, Porous Low-k BEOL, and Immersion Lithography", *Proc. IEDM*, 2006, pp. 1-4.

[12] D. Pamunuwa and H. Tenhunen, "Repeater Insertion to Minimise Delay in Coupled Interconnects", *Proc. VLSI Design*, 2001, pp. 513-517.

[13] P. P. Shah, "Optimization of the BEOL Interconnect Stack for Advanced Semiconductor Technology Nodes", *M.S. Thesis*, UC San Diego ECE Department, 2015.

[14] S. Takahashi, M. Edahiro and Y. Hayashi, "Interconnect Design Strategy: Structures, Repeaters and Materials with Strategic System Performance Analysis Model", *IEEE Trans. on Electron Devices* 48(2) (2001), pp. 239-251.

[15] B. Tutuianu, F. Dartu and L. Pileggi, "An Explicit RC-Circuit Delay Approximation Based on the First Three Moments of the Impulse Response", *Proc. DAC*, 1996, pp. 611-616.

[16] J. J. Zhu, P. Chidambaram, G. Nallapati and C. Yeap, "Back End of Line (BEOL) Local Optimization to Improve Product Performance", *U.S. Patent Application 20150303145 A1*, 2014.

[17] Z. Zhu, D. Wan and Y. Yang, "An Interconnect-Line-Size Optimization Model Considering Scattering Effect", *IEEE Electron Device Letters* 31(7) (2010), pp. 641-643.

[18] Cadence Innovus User Guide, http://www.cadence.com

[19] Cadence Quantus QRC User Guide, http://www.cadence.com

[20] International Technology Roadmap for Semiconductors, www.itrs2.net

[21] LEF DEF reference 5.7. http://www.si2.org/openeda.si2.org/projects/lefdef

[22] OpenCores: Open Source IP-Cores, http://www.opencores.org

[23] Synopsys HSPICE User Guide, http://www.synopsys.com