# Enabling Design of Low-Volume High-Performance ICs

Submitted in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in

Electrical and Computer Engineering

Mehmet Meric Isgenc

B.S., Electronics Engineering, Sabanci University

M.S, Electrical and Computer Engineering, Carnegie Mellon University

Carnegie Mellon University

Pittsburgh, PA

March, 2019

*Dedicated to my parents who*

*gave me the gift of a happy life*

## Abstract

Integrated circuits (ICs) are ubiquitous, ranging from consumer electronics to custom hardware. While scaling of CMOS feature sizes has enabled faster and smaller ICs, manufacturing smaller transistors and wires has become more challenging, particularly with the sub-wavelength lithography era. In response, foundries, to ensure the manufacturability of ICs at advanced CMOS nodes, have increased the number of design rules. The resultant difficulty in achieving design closure threatens the ability to create custom ICs within required timeframes. As a result, the use of advanced CMOS nodes has become unfavorable particularly for low-volume ICs wherein the design cost is higher than the fabrication cost. To this end, this dissertation explores opportunities to extend the use of advanced CMOS nodes for low-volume ICs by trading some amount of chip area for a reduction in design complexity, but without significantly affecting the performance and power consumption of ICs. To pursue this objective, we propose finding more optimal logic cell heights and wire pitches by selecting values more relaxed than the technology-allowed minimums.

To evaluate the impact of this optimization, we designed multiple digital ICs in a commercial 14/16 nm FinFET process. The silicon measurements indicate that relaxing the wiring pitches can reduce power consumption through coupling capacitance reduction. Moreover, taller logic cells that pose a minimal area increase penalty still perform comparably to shorter logic cells while eliminating routing problems. Lastly, using a layout pattern enumerator developed in-house, cell height increase is shown to mitigate manufacturing risks via layout simplification and re-use. These results suggest that logic cell height and wire pitch optimizations provide an excellent foundation for enabling low-volume customers to fabricate ICs in advanced CMOS nodes.

**Acknowledgments**

First and foremost, I would like to thank my advisor Professor Larry Pileggi. His endless support and unique, clear vision helped me grow immensely as a researcher and a person. Without his technical and philosophical contributions, this work would not be possible. I will always look up to his dedication to his work and students. Next, I would like to thank Dr. Samuel Nascimento Pagliarini for mentoring me over the past four years. The key contributions of this work have been possible owing to his managerial acumen and scholarly ability. I will always look up to his ability to listen and think clearly.

I would like to thank Dr. Kaushik Vaidyanathan who welcomed me on my first day at CMU and guided me until the end. I will always try to achieve the clarity and quality of his work. I am grateful to Professor Andrzej Strojwas for enlightening me about the manufacturing of circuits and I would like to acknowledge Professor Shawn Blanton for taking time out to be on my thesis committee and providing me with valuable feedback.

Several pieces of my layout regularity work have been done in collaboration with Dr. Mayler Gama Alvarenga Martins, an extraordinary problem solver. My first year at CMU would be more challenging if I did not get to know Ekin Sumbul whose mentorship and feedback on my early work have been invaluable. I would also like to thank my fellow group members Renzhi Liu, Joe P. Sweeny, Mohammed Zackriya Vanaikar, and Ioannis Karageorgos for contributing to various aspects of this work.

I would also like to express my gratitude to fellow Electrical Computer Engineering Department members. I thank Professor Ken Mai for all of his technical help, as well as his group members Prashanth Mohan and Etkin Akkaya. I am thankful to Professor Jeffrey Weldon and his group

members Yunus Emre Kesim and Mohammed Omar Darwish for insightful discussions. I am also thankful to Professor Diana Marculescu for enabling a collaboration with her student Ahmet Fatih Inci, a promising fellow researcher. I am also grateful to Judy Bandola, Shelley Phelps, Holly Skovira, and Nathan Snizaski for their administrative support. Regarding my background before graduate school, I would like to acknowledge Professor Yasar Gurbuz of Sabanci University who provided me with opportunities to improve my analog-mixed signal design skills. I would fail my duty if I did not acknowledge Dr. Florin Dartu. His guidance during my internship at Taiwan Semiconductor Manufacturing Company has contributed significantly to my understanding of the bleeding edge problems in the semiconductor industry.

I am eternally thankful for the sense of family that my friends provided. I would like to individually thank Amrit Pandey, Antonis Manousis, Dimitrios Stamoulis, Beril Cakir, Berkin Akin, Ekin Sumbul, Onur Albayrak, Sercan Yildiz, Burcu Karagoz, Olivia and Buddy Cho-Smith, Rise Cohen and Maria Simonic, Onur Kibar, Thomas C. Eley, Baris Otus and Ozgun Elci, Elina Zhang, Volkan Cirik, and Mert Terzi. They transformed Pittsburgh into home for me. I will cherish every happy moment we shared.

Finally, I would like to thank my parents Aynur and Faruk Isgenc who supported me through high and low from thousands of miles away in Turkey. Their vision of a fairer, happier world will guide me forever.

---

This work was sponsored by DARPA contract HR0011-16-C-0038, CRAFT (Circuit Realization At Faster Timescales) program.

Table of Contents

## List of Figures

## List of Tables

# 1 Introduction

Moore's law [1] and Dennard's scaling theory [2] established a foundation for the improvement of IC capability via CMOS miniaturization. State-of-the-art CMOS processes have reached the 5 nm node, in spite of numerous manufacturing challenges [3]–[6]. To ensure manufacturability, foundries significantly increased the number of required and recommended design rules (DRs) in advanced nodes [7]–[9]. The resultant challenges to achieve design closure [10] correspond to longer development cycles and silicon re-spins, which inflates the design cost.

Compared to the cost of design, however, the fabrication cost is astronomical for ICs that are manufactured in high volumes (thousands, perhaps millions). Thus, the main design objective for high-volume ICs is to maximize the density of components to lower the chip area, obtain more chips per wafer, and, ultimately, reduce the fabrication cost per chip. To fulfill the area miniaturization goals, building blocks of digital ICs are tailored for high density, which comes at the cost of design closure [11]–[14] and manufacturability issues [7], [15]. Solving these issues require significantly more design effort – tolerable by major high-volume customers. This "high-density, high-design cost" paradigm, however, has rendered state-of-the-art CMOS technologies unfavorable for low-volume ICs of which cost of design is more dominant than the cost of fabrication. As a result, there is a portion of the semiconductor market that could benefit from the performance and power advantages of FinFETs, but unable to do so, due to daunting impact of design challenges.

To this end, we explore trading some amount of chip area for achieving design closure in shorter timeframes; thus, reducing the cost of design significantly. We propose a holistic design methodology that addresses design closure issues that stem from logic cell and interconnect options that are commercially available today. Particularly, the proposed design methodology

focuses on the BEOL metal stack and the logic cells. By relaxing the density of BEOL wiring, we target timing, power, and printability advantages over maximum density designs. Through extreme layout regularity in logic cells, we tackle routing closure and manufacturability issues. We demonstrate the efficacy of the proposed design methodology in a commercial 14/16 nm FinFET process.

Figure 1.1 maps out the process stack of our technology of choice. Metal layers 1 (M1) to 9 (M9) corresponding vias, and dielectric (shown in gray) form the back-end-of-line (BEOL) of the process stack. BEOL layers are routed horizontally (H) or vertically (V) but M1 is an exception, which can be routed in both directions. Metal layer 0 (M0; horizontal for gate contacts, vertical for diffusion contacts) forms the middle-of-line (MOL) between the BEOL and transistors. The front-end-of-line (FEOL) is comprised of FinFETs in the active region. BEOL relaxation, as discussed and demonstrated in depth in the following chapters, corresponds to increasing the width and spacing of intermediate and semi-global routing layers. Layout regularity corresponds to designing logic cells with simpler and similar M1 polygons. Changes in the BEOL, consequently, propagate to the MEOL and FEOL, of which power and performance impact we will be discussing in depth.

The remainder of this dissertation is organized as the following:

- Chapter 2 provides background on the problems caused by CMOS scaling.
- Chapter 3 elaborates our methodology of relaxing the BEOL metal wiring pitches.
- Chapter 4 focuses on optimizing logic cell IP for layout regularity.
- Chapter 5 delivers a discussion about the impact of reverse scaling on manufacturability.
- Chapter 6 concludes with a discussion of key findings and future directions.

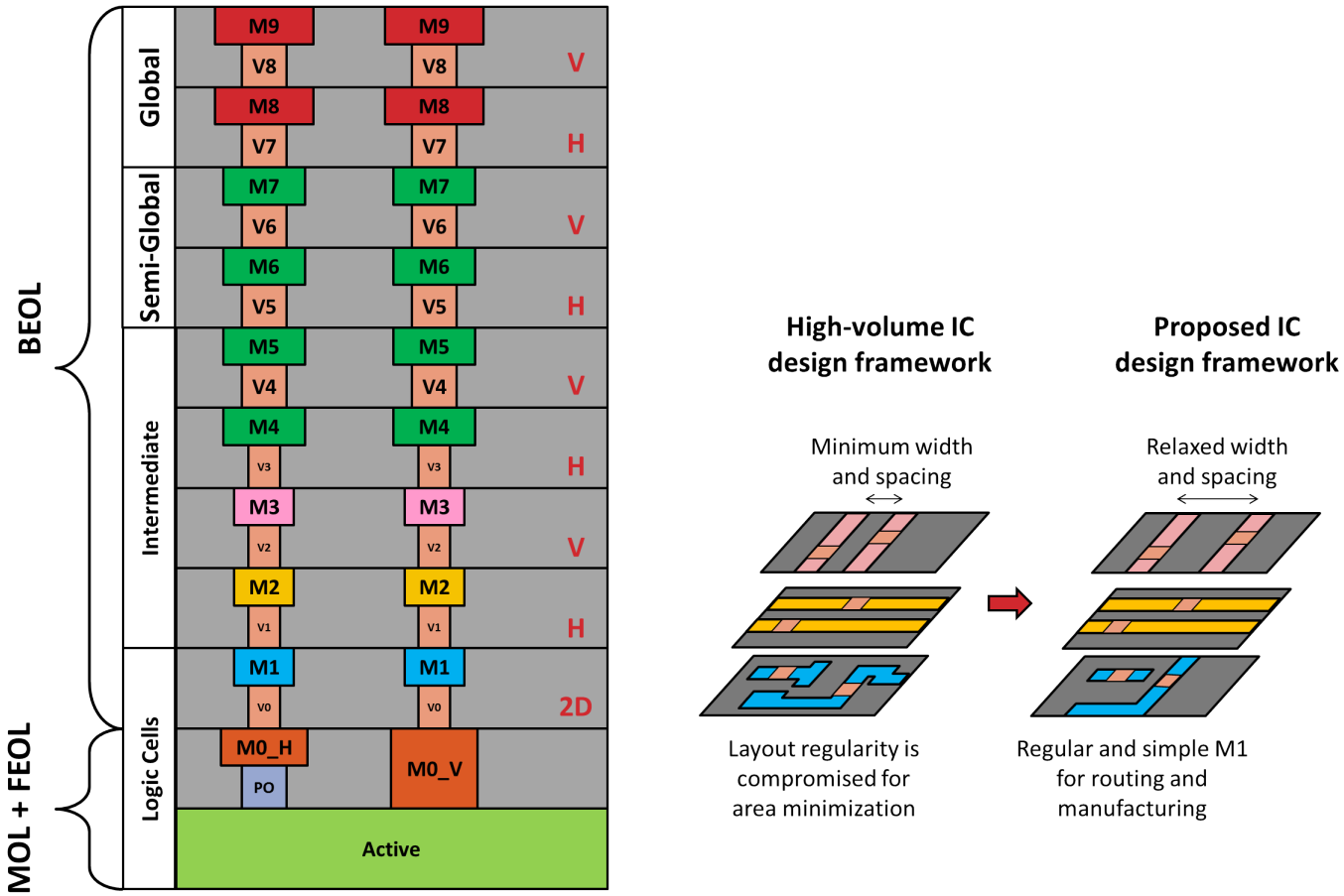


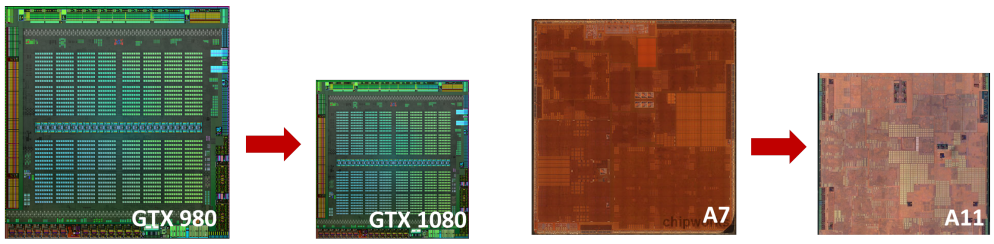

**Figure 1.1 The proposed design methodology shown on a 14/16 nm process stack**

# 2 Background

This chapter details the impact of density scaling on IC development costs and discusses the reasons for design challenges. Furthermore, the shortcomings of the prominent approaches from the literature are detailed.

## 2.1 Moore's Law and Density Scaling

Over four decades ago, Moore's law [1] predicted that the number of components would double every 12 to 18 months for a given chip area. Since its debut in the semiconductor literature, this law has driven the dimensions of merit for CMOS scaling and the enabling technologies. Despite the popularity of the discussions that declare the death of Moore's law, CMOS density scaling has persevered through substantial advancements in photolithography and clever design strategies. Figure 2.1 shows how Moore's law applies to state-of-the-art graphical processing units (GPUs) and mobile processors. Keeping Moore's law alive to these advanced nodes, however, has been at a very high cost to manufacturing and design. Today, these increased costs make advanced CMOS nodes quite unfavorable for almost any companies but the major players in the consumer electronics market.




| Product | NVIDIA GTX 980 | NVIDIA GTX 1080 | Apple A7 | Apple A11 |
|---|---|---|---|---|
| Year | 2015 | 2017 | 2013 | 2017 |
| Process Node | 28 nm | 16 nm | 28 nm | 10 nm |
| Die Area | 601 $mm^2$ | 471 $mm^2$ | 102 $mm^2$ | 88 $mm^2$ |
| Transistor Count | 8 Billion | 12 Billion | 1 Billion | 4.3 Billion |
| Clock Speed | 1 GHz | 1.5 GHz | 1.3 GHz | 2.4 GHz |

**Figure 2.1 Impact of CMOS density scaling on recent GPUs and mobile processors.**

## 2.2 Impact of Scaling on Masks and Fabrication Costs

Miniaturization of CMOS dimensions is enabled through photolithographic and material advancements [16]. In the absence of a reliable and much anticipated extreme ultra-violet (EUV) solution, state-of-the-art CMOS dimensions are well below the illumination wavelength that is 193 nm. As a result, a series of resolution enhancement techniques (RETs) were developed to continue a full pitch scaling that approximates to a factor of 70% [8]. RETs include optical proximity correction (OPC), phase shift mask (PSM), off-axis illumination (OAI), and multiple patterning techniques [8].

RETs require materially and physically more complex masks that become increasingly costlier. In advanced CMOS nodes, the cost of mask sets is millions of dollars, and it approximately doubles for every new technology node [10], [16]. For example, a typical 11-metal process in the 14/16 nm node requires more than 60 masks, and this number is expected to go above 80 in early 7 nm projections [17], [18]. The increase in the number of masks is due to multiple patterning and new layers needed for routing. Some of these routing layers are low-resistance high-level BEOL layers (for performance). Additionally, in FinFET CMOS stacks, the Metal-0 (M0) layer was introduced for very short, local connections within logic cells. This layer is commonly referred to as middle-of-line (MOL) because it interfaces between the FEOL and BEOL. The resultant manufacturing complexity of advanced CMOS stacks correspond to extremely high fabrication costs – foundries can charge millions of dollars depending on the manufacturing volume. However, IC design effort at advanced CMOS nodes is significant, and resultant design costs can be more dominant than fabrication costs at lower volumes of manufacturing [19], [20].

## 2.3 Impact of Scaling on Design Costs

Figure 2.2 shows the evolution of IC development cost factors with scaling. This trend shows that hardware design (including layout and verification) challenges need to be addressed to enable cost-effective use of technology nodes that are below 20 nm. According to [17], to develop a system-on-chip (SoC), the amount of effort needed is 100 engineer years in the 28 nm node, whereas this number doubles to 200 engineer years in the 14/16 nm node. The increasing design cost is associated with shortcomings of state-of-the-art commercial electronic design automation (EDA) tools, increasing verification complexity, issues with re-using IPs, as well as manufacturability. We will be addressing these issues with novel interconnect and IP design methodologies in the following chapters.

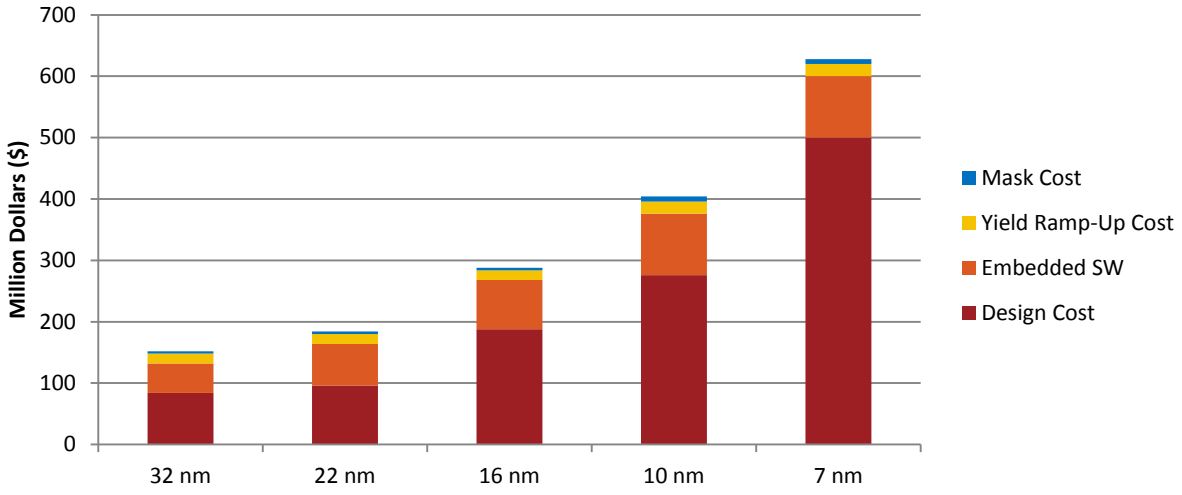



**Figure 2.2 IC design cost becomes higher and more dominant with scaling [19].**

### 2.3.1 Timing Closure Challenges

To be able to route between downscaled transistors in a design, wire dimensions and spacings need to be downscaled as well. The resultant increase in wire resistance has been traditionally compensated by asymmetrical aspect ratio scaling in the sub-micron regime [21], where the wire height is not scaled as aggressively as the width. Tall and skinny wires, when tightly spaced,

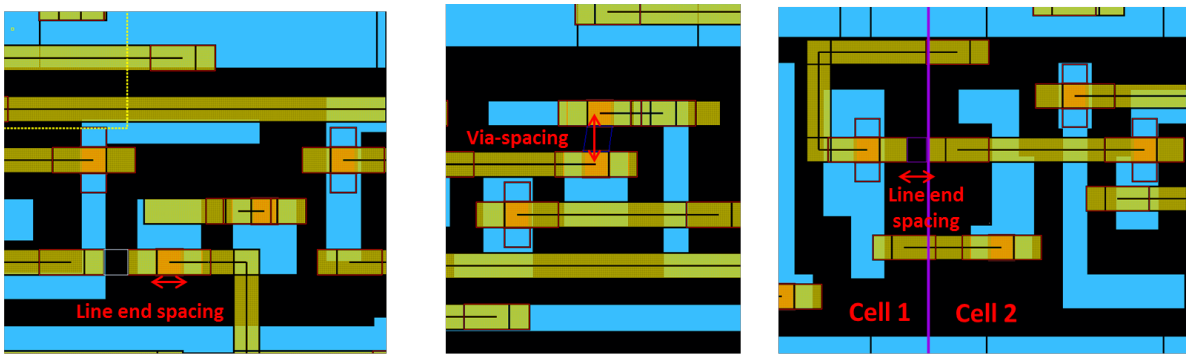
suffer from dominant sidewall coupling capacitance that exacerbates the cross-talk among neighboring wires [21], [22], [23]. As a result, signal integrity and timing closure become difficult to attain, particularly in high-speed, noise-sensitive designs. To reduce cross-talk, placing static power/ground (PG) "shield" wires [24] is a commonly employed technique that consumes additional routing resources. Alternatively, net ordering algorithms [24] and repeater insertion [25] are used in physical design to avoid significant cross-talk. Instead of adding shields, [26] proposes spreading out minimum width wires at an optimal rate. In addition to the signal integrity benefits of wire spreading, it offers improved printability that has become a more concerning issue with scaling. Altering only the wire spacing while keeping the width at a minimum, however, is often not design rule (DR) compliant in advanced CMOS nodes due to process difficulties.

### 2.3.2 Routing Closure Challenges

Routing closure is honoring a netlist in a fully DR-compliant manner. In that sense, achieving routing closure refers to a design with no opens, shorts, and other physical DR violations. Routing closure has become more challenging in advanced CMOS nodes. In sub-20 nm process nodes, adoption of FinFETs by the industry has increased the manufacturing complexity but also enabled packing more drive strength per unit area – which IP vendors leveraged to offer shorter logic cells for high density. Shorter logic cells, however, are inherently harder to route, especially in the face of harsher DRs. As a result, routers commonly cannot converge for significant amounts of time (perhaps for weeks in million gate designs) due to physical DR violations and shorts. Fixing such violations incur significant manual designer labor.

Figure 2.3 (a) and (b) are samples of Metal-2 and Via-1 spacing violations that occurred while the router was trying to establish a connection to a logic cell in a design block implemented in a

commercial 14/16 nm process. Figure 2.3 (c) is a particularly concerning problem – in terms of DR it is same as (a), but it happens because one of the two neighboring cells exhausts the local routing resources. The router, which runs out of routing track options, establishes a connection that is not DR-compliant. In the literature, this problem is referred as track stealing [12], and it is a very significant problem given that it is neighbor-dependent – all logic cell pairs in a library can be susceptible to track stealing unless they are validated against it.




(a) M2 line-end spacing (b) VIA1 center spacing (c) M2 track stealing

**Figure 2.3 Routing problems among M2 (yellow), VIA1 (orange), M1 (blue) layers.**

To mitigate the routing closure challenges, [12], [13], [14] propose "quantifying" the *pin accessibility* of logic cells using the proximity of pins, their surface area, and the pin features of the neighboring cells in an already placed design. Based on the pin accessibility metric, [12] compares twelve different logic cell libraries and shows that the minimum area is not achieved by the smallest logic cell library, but by the most routing friendly one. Modeling the pin accessibility of every logic cell in a library considering different abutment scenarios, however, is not scalable given the sizes of commercial libraries can be in the order of thousand logic cells.

### 2.3.3 DFM Challenges

For CMOS stacks with FinFETs, to ensure manufacturability, front-end-of-line (FEOL) design rules require grating the transistor gates and channels with a single pitch. In contrast, BEOL design rules are more relaxed at the cost of the metal density. This relaxation is leveraged for

power-performance-area (PPA) optimization in pre-designed commercial logic cell intellectual properties (IPs), and it leads to a high degree of layout customization — i.e., the layout regularity is compromised. While the electrical impact of FEOL layout-dependent effects (LDE) are modeled in device parameters with high confidence, capturing lithographic imperfections in BEOL layers is not as reliable [7]. In that regard, extremely custom layouts of commercial logic IPs introduce a significant number of layout patterns each of which can be *quasi-defects* or actual manufacturing hazards. As a result, foundries provide their customers with a set of required manufacturability checks to attain a certain level of yield. Performing manufacturability checks and fixes further burdens the IC designers in advanced CMOS nodes.

A prominent design-for-manufacturability (DFM) technique is using restrictive layout design guidelines for logic cells [27], [7], [13], [28], [29], and optimizing the CMOS process specifically for the resultant small set of patterns [8], [9], [15], [30]. For DFM, controlling layout patterns is critical because the printability of cells with finer FEOL and BEOL features are more sensitive to their surroundings due to diffraction induced variations [8].

DFM compliance in interconnects is an even more challenging matter since routing complexity—namely the number, length, and density of wires—can vary from one design to another. However, one clear trend is that in an era where wire widths are more than five times smaller than the illumination wavelength that is used for lithography, wire printability degrades with shrinking dimensions [8], [31]. As a result, cumbersome DFM checks for the BEOL interconnect are inevitable unless wires are designed to be more structured and the wire pitch is optimized for printability at the cost of the area.

## 2.4 Impact of Scaling on Design Re-spins

Designing an IC needs increasingly more resources due to the ever-rising demand for design complexity. For multi-million gate SoCs, verification of logical functionality and manufacturability plays a role that is as crucial as design closure. Survey data from the industry [32], [33] show that approximately 50% of the project time was devoted to functional and physical verification by 2014. Data suggests that the verification effort and the cost increase (almost exponentially) with scaling. Despite the substantial engineering time that goes into the verification, more than 40% of ICs require a second design spin [33]. Based on a survey conducted by Synopsys [32], functional errors are the dominant reason for first spin failure. The other reasons are signal integrity, timing closure (paths that violate setup and hold time), and reliability problems. Moreover, with scaling, manufacturing issues worsen, causing unforeseen performance deviations that are hard to debug. The use of a BEOL metal stack optimized for design closure and manufacturability can eliminate some of the re-spin causes cost-effectively; hence, help allocate more resources to functional verification instead of design.

## 2.5 Yield and Relation to Chip Area and Cost

The cost of design for a commercial SoC is on the order of half a billion dollars at the 14/16 nm node [17]. To amortize astronomical costs, the consumer electronics industry is desperate to increase the number of good chips per wafer (GCPW) that is a function of wafer size, chip area, and yield. If there were no relationship between chip area and yield, maximizing GCPW would mean designing a chip in the smallest area possible. However, area and yield are related, and "burning" some extra area for DFM countermeasures can guarantee a higher GCPW by elevating the yield. For example, using redundancy in memory designs or via duplication in logic circuits can increase the area but also boost the yield. Particularly, as the critical dimensions shrink with

CMOS scaling, yield loss mechanisms related to lithography and etch become more dominant, and area gains achieved by aggressive scaling can be overturned by issues that can lower the yield. That being said, for a fixed defect density, increasing chip area beyond a certain level could increase the number of defects on the chip [34]. As a result, achieving a good GCPW requires a co-optimization of the chip area and DFM countermeasures, wherein an optimal balance between the area and yield needs to be explored for a given IC.

## 2.6 Summary

Due to increasing manufacturing complexity, both fabrication and design costs rise steeply with scaling. In the face of harsher DRs and aggressive scaling trends in the FEOL and BEOL, achieving timing and routing closure has become more challenging. Moreover, design methodologies with integrated DFM flows are costly, and failure to comply with manufacturability guidelines can correspond to silicon re-spins.

# 3 Relaxing the Pitch of BEOL Wiring

Transistor scaling is generally beneficial power, performance, and area of digital ICs. In contrast, the downscaling of wire dimensions leads to higher wire resistance and higher capacitance (due to coupling capacitance dominance), hence larger RC delays. With scaling, wires correspond to a more significant portion of on-chip delays and power consumption in advanced CMOS nodes.

Furthermore, printing wires with finer dimensions and at higher densities keep becoming more difficult. Unfortunately, as state-of-the-art EDA tools cannot offer a holistic *push-button* solution to these issues, a substantial amount of designer effort is necessary to minimize the adverse effects of wire scaling in large-scale ICs with millions of signal and power nets. Instead, we propose optimizing the density of the BEOL metal stack for alleviating these problems.

We apply our design methodology to a commercial 14/16 nm FinFET technology. The range and the definition of the layers for this technology are as the following. FEOL is comprised of a mesh of fins and poly lines and corresponding cut layers. MOL layers are vertical M0 (for connecting to diffusion contacts), horizontal M0 (for connecting to gates), and VIA0. We divide BEOL layers into two categories: internal logic cell routing (M1 and M2) and global routing layers (M2 to M10). In the following chapters, r-BEOL corresponds to the relaxation of layers M3 to M7, the majority of intermediate and long distance on-chip interconnects. We do not alter M2 since it is the main detail routing layer for physical pin connections, and we use M8 and M9 at their original state given their width, and spacing values are already very large.

## 3.1 Optimizing Wire Width and Pitch

The unit wire resistance and the total wire length in ICs have been rising with scaling [35]. The resultant increase in the total wire resistance has been a growing concern in terms of timing and

reliability. To alleviate the wire resistance increase, foundries have begun downscaling wire heights with a smaller factor than the width, leading to an asymmetrical aspect ratio of wire cross-sections [36]. This approach has offered some wire resistance benefits but at the cost of capacitance challenges.

For tall and skinny wires, the sidewall coupling capacitance among adjacent wires has become a more dominant portion of the total wire capacitance, increasing the intensity of cross-talk. Although EDA tools can mitigate some of the cross-talk induced glitches, the timing-critical nets may still require custom modifications to eliminate the likelihood of setup or hold timing errors. Furthermore, at the absence of a EUV solution, printing dense and skinny wires has become increasingly challenging [31]. Printing challenges render variations from the actual layout drawings of wires. Since modeling and simulating the lithographic process variations for every wire in a design block is not computationally feasible, commercial technology files contain built-in timing margins for optimistic and pessimistic deviations in wire dimensions and spacings. Using default timing margins, however, can result in sub-optimal IC performance and reduction of power efficiency due to overdesign.

Another critical problem is connecting wires in different metal layers together. As wire dimensions shrink, the surface area of vias becomes smaller, increasing the via resistance significantly. At the face of harsh via spacing and enclosure DRs, using low-resistance via stacks with redundancy has become more difficult. Thus, IR-drops and electromigration risks have become growing sources of concern in advanced CMOS nodes, and IC design houses dedicate more resources to analyzing and fixing them.

Since wire density is at the core of these challenges, increasing the width and spacing of the wires (wire pitch) to improve performance and reliability would be at an apparent cost to density. However, it is conceivable that a methodology and toolset could be formulated to provide such a trade-off as a function of cost vs. performance. We refer to the design methodology that utilizes wire pitches that are greater than the technology-allowed minimums as relaxed BEOL (r-BEOL). We explore the r-BEOL opportunities for a commercial 14/16 nm technology.

### 3.2 Forbidden Pitches

Self-aligned patterning and OPC requirements allow for discrete wire spacing and width values in sub-20 nm nodes. Particularly, for metal layers between 3 and 7, arbitrarily increasing the spacing among minimum width wires is *forbidden* by DR. Instead, wire widths need to be concurrently increased as shown in Figure 3.1. Although some minimum width and large spacing options are DR compliant, they fail to meet the metal density requirements. Thus, spreading out wires alone, as proposed in [26], is not an effective solution for advanced CMOS nodes. Therefore, we need to explore a DR-compliant wiring pitch that fulfills our design goals.

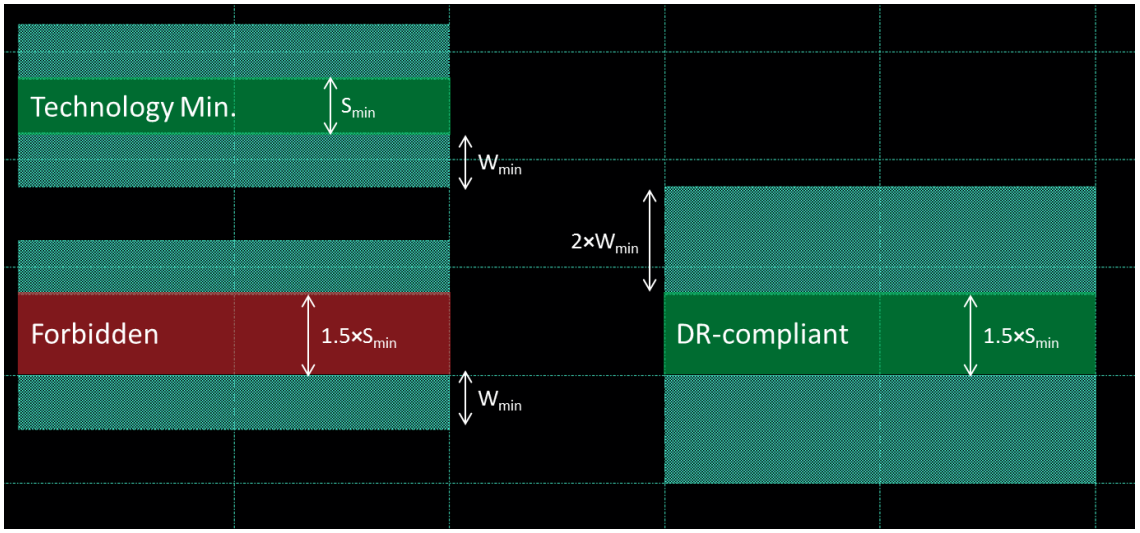



**Figure 3.1 M4 wires shown on the routing grid. Minimum width wires cannot be arbitrarily spaced out.**

### 3.3 Coupling Capacitance

In the 14/16 nm node, for a set of wires routed at the tightest pitch ($P_{min}$), the side-wall coupling capacitance comprises 80% of the total wire capacitance. As the wire length increases, the side-wall coupling capacitance becomes more dominant (Figure 3.2). To observe the impact of r-BEOL on the side-wall, vertical, and fringe wire capacitances, we prepared the experimental setup shown in Figure 3.3. We increased the wiring pitch of adjacent wires in M4, M5, and M6 layers (a total of nine wires). We chose M4, M5, and M6 since the majority of the intermediate length, timing critical paths are routed in these layers. To preserve the relevance of fringe at the wire ends, we elongated the wires as we increased the wiring pitch. Using a commercial 3D field solver, we extracted the wire capacitances of the layout clips at an optimistic (best) and a pessimistic (worst) capacitance corner. The worst capacitance corner assumes an increased side-wall dielectric constant and a reduction in wire spacings due to mask misalignments. The best corner assumes a decreased dielectric constant and an increase in wire spacings.

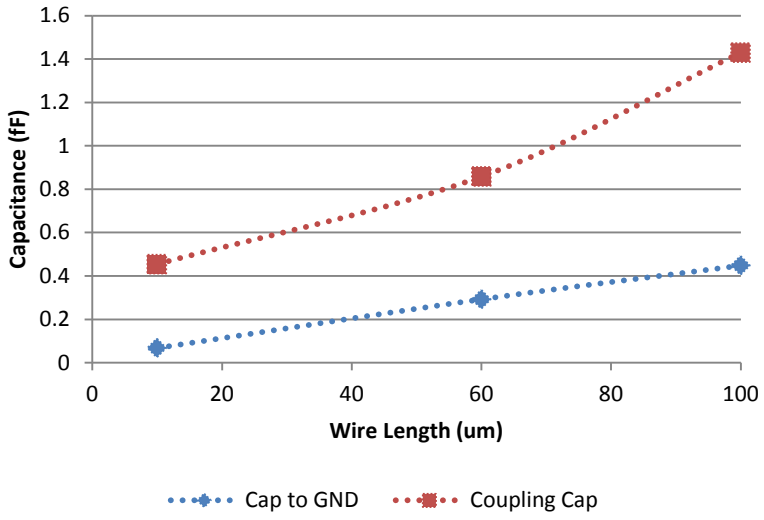



**Figure 3.2 Coupling capacitance of interconnects are dominant in the 14/16 nm node.**

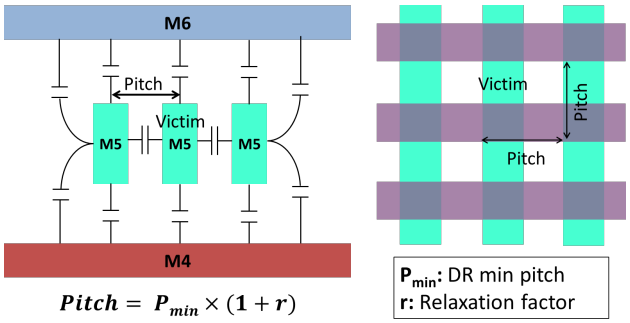



**Figure 3.3. Cross-section and overviews of the experimental metal mesh setup.**

In our results, we use the total capacitance of the *victim* in Figure 3.3 as a proxy, since it experiences the most significant side-wall and vertical coupling. Figure 3.4 illustrates how the unit wire capacitance of the victim changes with the wiring pitch. The discontinuities in the graph correspond to forbidden pitches. For both extraction corners, increasing the wiring pitch from the technology-allowed minimum to the next legal one offers a 20% reduction in the unit capacitance. Relaxation beyond a factor of 75% does not reduce the unit capacitance since the coupling capacitance to the ground becomes dominant. This graph suggests that performance and power benefits may exist for designs that can utilize a BEOL pitch that is relaxed by 75%. Interestingly, the optimal wiring pitch we found in this experiment corresponds to that for the previous technology node.

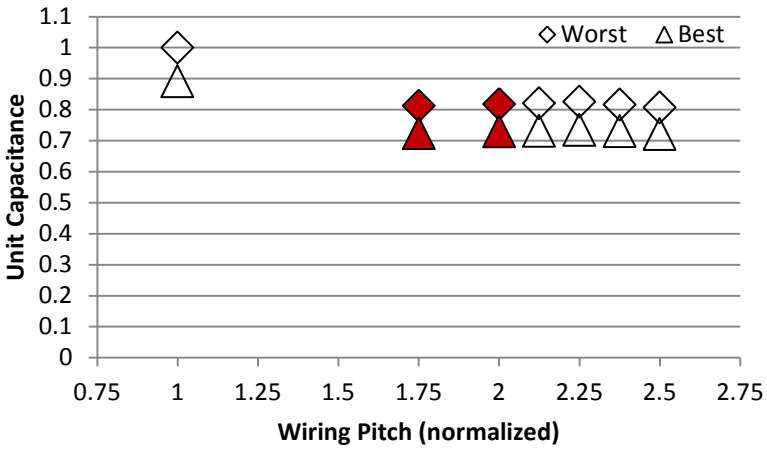



**Figure 3.4 Unit wire capacitance normalized by the highest value achieved in the experiment.**

The capacitance benefit of r-BEOL can be leveraged to improve the worst-case timing of nets. For densely routed wires that switch frequently, the coupling capacitance can effectively double in the worst case, if the drivers switch in opposite polarities. To investigate the impact of cross-talk on timing and demonstrate the efficacy of r-BEOL, we built a circuit-level experimental setup (Figure 3.5). In this setup, we assembled three parallel wires of 2-micron length, with the middle one being the *victim* and the adjacent ones being the *aggressors* (i.e., switching in opposite polarities). For the sake of consistency, we kept the wires at M5, a commonplace routing layer. Since the severity of cross-talk depends on the slew rate of the signals, we loaded the drivers with a fan-out (FO) of 1, 4, and 12 to alter the slew rate. We used the same driver sizing for all load types.

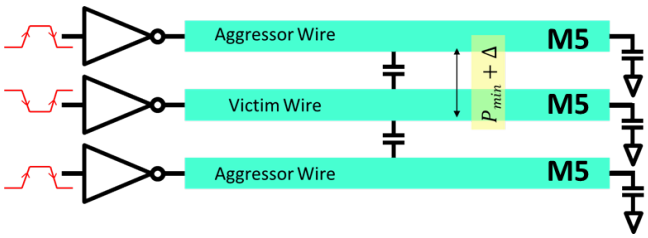



**Figure 3.5 Experimental setup to observe the impact of the metal pitch on timing.**

The entries in Table 3.1 correspond to the propagation delay of the input signal for the standard BEOL and 75% r-BEOL wiring. In the fan-out-of-1 (FO1) loading scenario, where the wire capacitance is more dominant than the load capacitance, r-BEOL can reduce the propagation delay by a factor of two. As the load capacitance becomes more dominant, the benefit of r-BEOL decreases. From the IC design perspective, r-BEOL can be leveraged for reducing the severity of worst-case switching scenarios in terms of timing and power.

| Load scenarios | Normalized Propagation Delay (FO4) | |
|---|---|---|
| | Standard BEOL | 75% r-BEOL |
| **FO1** | 0.68 **(2x)** | 0.34 |
| **FO4** | 1.37 **(1.3x)** | 1.06 |
| **FO12** | 3.60 **(1.1x)** | 3.30 |

**Table 3.1 r-BEOL can significantly reduce the worst-case timing of nets.**

### 3.4 Buffering for Long Routes

Long routes that interconnect IP blocks in ICs can compromise timing since the propagation delay through a wire grows quadratically with the wire length when the wire is unbuffered. Based on the clock and maximum transition time (slew-rate) constraints, timing optimization engines of EDA tools break long wires into shorter segments and insert repeaters [37] to linearize the wire delay. Repeater insertion, however, consumes additional power and reduces the maximum possible speed of signal propagation for that length of wire.

To find the minimum wire length at which repeater insertion becomes profitable for timing, we designed the circuit schematics shown in Figure 3.6. We took the on-resistance ($R_{ON}$) and the gate capacitance ($C_g$) of the repeaters into account. The total wire RC is the product of the unit wire RC and the route length. For the sake of accuracy, we modeled the wire RC with two pi segments. The route in Figure 3.6 (a) has a driver and the same sized load. Dividing the route into two segments, we inserted a repeater between the driver and the load (Figure 3.6 (b)). Transient simulations of the experimental circuits (Figure 3.7) show that the repeater insertion becomes beneficial after 200 microns. Though timing-critical routes longer than couple hundred microns are not a very common net profile, closing timing for such nets can require manual design effort. As a push-button alternative, we modified the route in Figure 3.6 (a) to emulate a 75% r-BEOL using our findings in Figure 3.4. On average, r-BEOL reduced the propagation delay by 20% compared to the standard BEOL without a repeater (Figure 3.7).

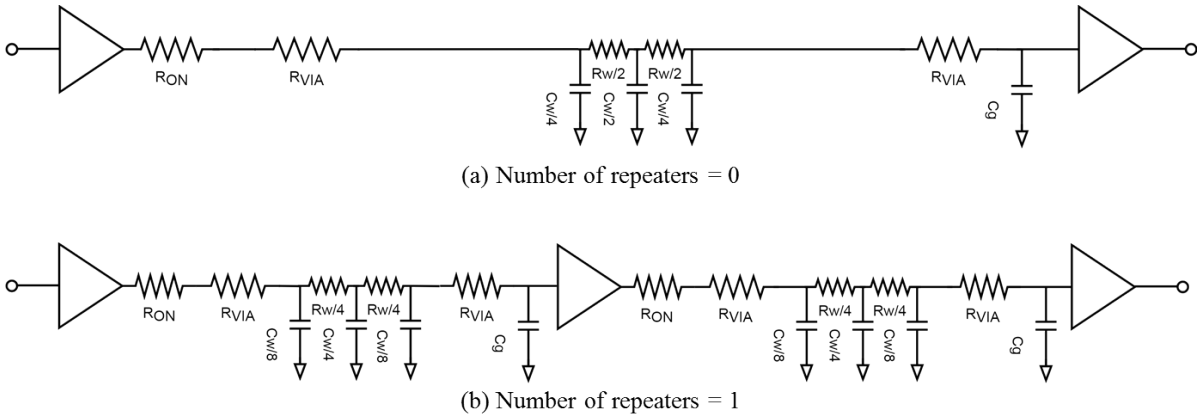



**Figure 3.6 (a) A long route between a driver and a load (b) A long route broken into two segments with a repeater in the middle.**

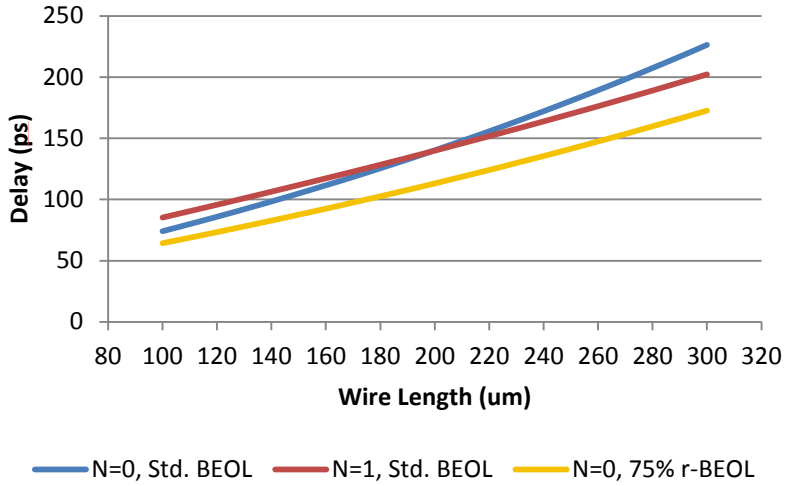



**Figure 3.7 Propagation delay of a long route with a repeater and r-BEOL.**

To observe the larger scale impact of r-BEOL, we implemented an experimental design block with 20 thousand gates in the 14/16 nm node. In this experiment, we used three BEOL variants: standard (tech. min. pitch), 75% and 100% relaxed. For the r-BEOL variants, we modified the minimum wiring pitch values for layers M3 to M7. We did not use M8 and M9 routing, assuming them to be dedicated to the global power network. Although heavily utilized by the router, we did not alter the wiring pitch of M2 to avoid compatibility issues with the logic IP. Figure 3.8 shows the floorplan: two identical encryption modules communicate through a 256-bit wide bus at 1 GHz. We additionally set a maximum transition constraint on every net. Therefore,

the timing optimization engine inserts repeaters for nets of which propagation delay and slew rate exceed the timing constraints.

Increasing the bus length between two modules from 100 microns to 1 mm, we reported the wire RC associated with the 256-bit wide bus, the number of repeater instances, and the power consumption for three BEOL variants. Figure 3.9 and Figure 3.10 show that the standard BEOL implementation has 1.3 times more wire resistance and 1.2 times more wire capacitance than the r-BEOL variants. These reductions in the wire RC translate to a relaxation in the repeater optimization. As seen in Figure 3.11, the number of repeaters is roughly the same for all variants of the design until the bus length reaches 200 microns (the break-even value reported in Figure 3.7). Beyond 200 microns, compared to the r-BEOL variants, the number of repeaters in the standard BEOL variants begins increasing more rapidly with the bus length. Owing to the repeater downsizing enabled by the RC reduction, the r-BEOL variants consume significantly less power than the standard BEOL variant, as shown in Figure 3.12. While enabling such benefits, r-BEOL reduces the number of available routing tracks due to wiring pitch increase. Given our ultimate goal to enable faster design closure, we cannot let routing closure be affected by r-BEOL. Therefore, we next investigate how routing closure changes with r-BEOL.

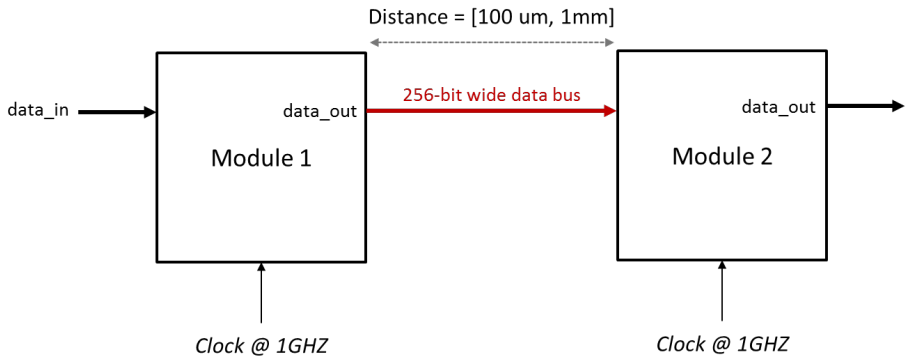



**Figure 3.8 A congested block-to-block routing scenario with a 256-bit wide bus in between.**

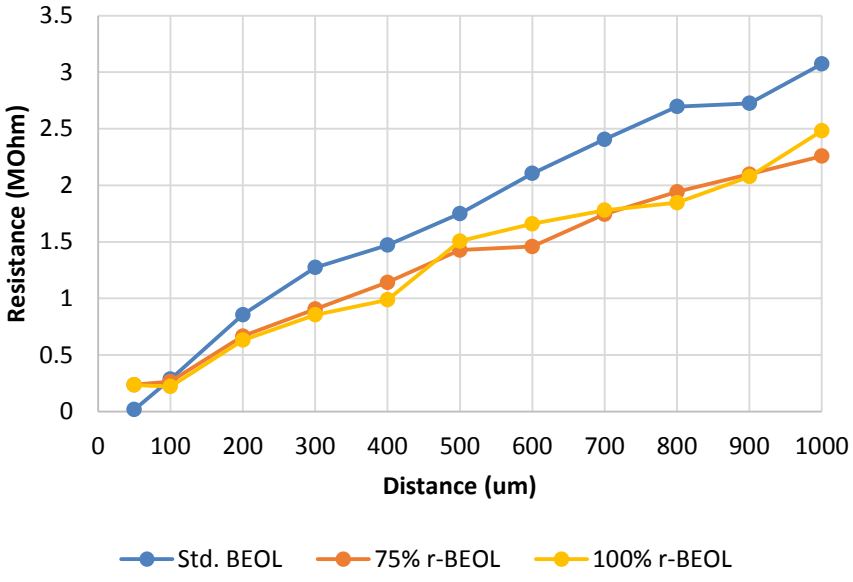



**Figure 3.9 Total resistance of the long routes.**

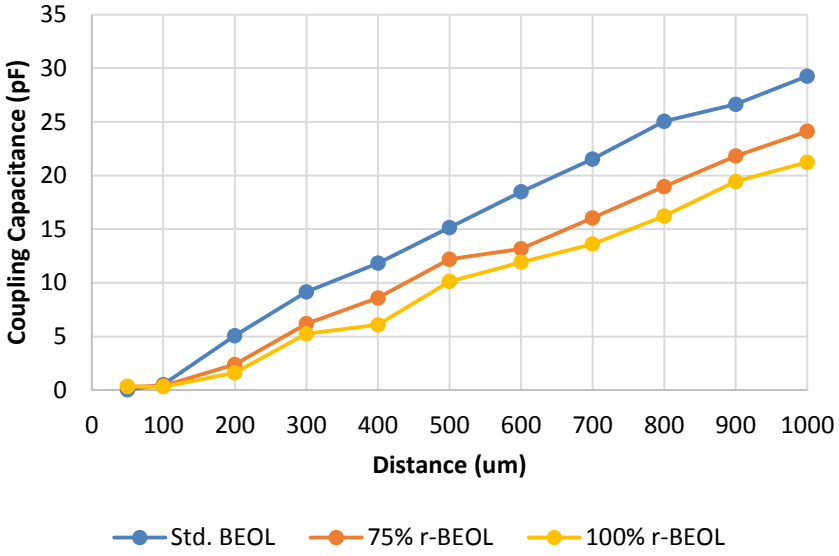



**Figure 3.10 Total coupling capacitance of the long routes.**

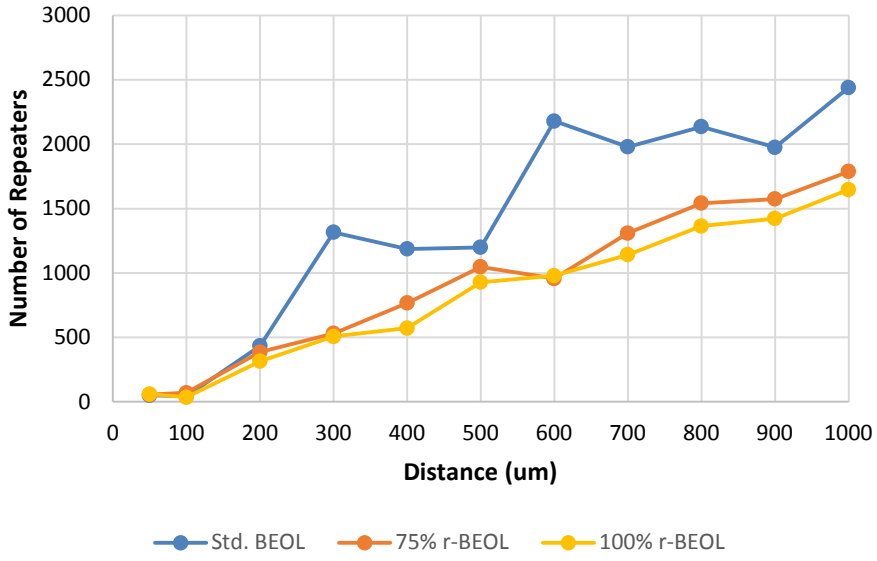



**Figure 3.11 The number of repeater cells between the design blocks.**

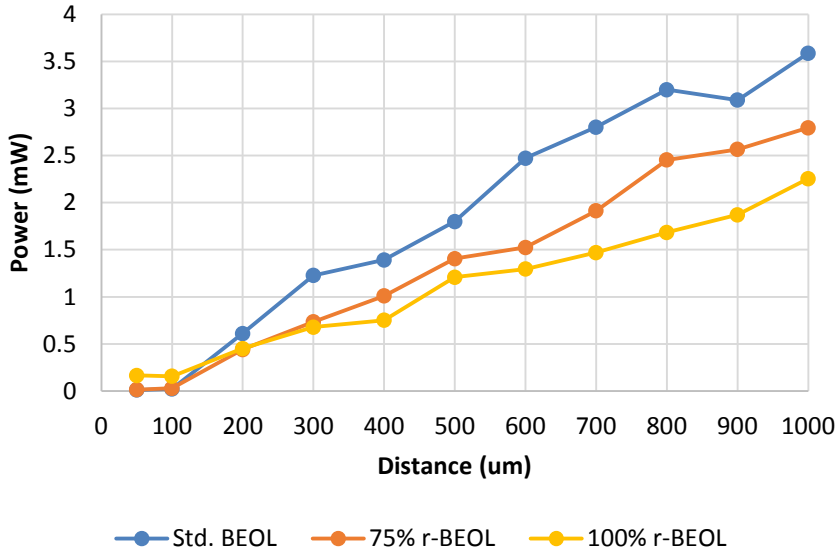



**Figure 3.12 Total power consumed by the repeaters versus route length.**

## 3.5 Routing with r-BEOL

Increasing the BEOL wiring pitch reduces the routing tracks per unit area, and inevitably changes the utilization of BEOL layers in routing. Figure 3.13 shows an increase in the utilization of M6 (orange) and M7 (pink) wires in the layouts of three AES-256 BEOL variants. Figure 3.14 shows that the router decides to push some of the routes from M4 and M5 layers to M6 and M7 layers. Overall, use of r-BEOL incurs only 1% wire length increase compared to the

use of standard BEOL. Figure 3.15 shows that a portion of VIA3 and VIA4 is pushed to VIA5 and VIA6 in the r-BEOL implementations of the design, leading to an insignificant via count difference. Given that vias are increasing sources of reliability issues, it is important that r-BEOL does not introduce additional risk factors. Furthermore, wider wires in the r-BEOL metal stack can be leveraged to upsize the vias, reduce the via resistance, and mitigate reliability issues.

Depending on the routing resources available to a design block, the wiring pitch increase can inflate the routing congestion [38]. To relax this congestion and avoid any adverse impact on the routing closure, we propose trading some amount of the floorplan area. To explore this trade-off, we prepared an experimental setup and, initially, synthesized two variants of AES-256 with the standard BEOL and the 75% r-BEOL with an identical floorplan size. This sizing corresponds to an aggressive utilization factor (UF) of 94%, which is a metric that captures how much of the net design area is covered by standard cells. Since a certain amount of whitespace is dedicated to physical cells (welltaps and endcaps), 100% UF is not feasible. The first two entries in Table 3.2 show the AES-256 synthesized with the standard BEOL and 75% r-BEOL with 94% UF. In this experiment, we used horizontal and vertical congestion overflows and execution time as a proxy to indicate the routing difficulty of designs. As expected, increasing the wiring pitch inflates the congestion, and, consequently, the execution time increases. For the 75% r-BEOL variant of the design, we begin decreasing the UF with 2% steps; enabling more routing tracks by increasing the floorplan sizing. The execution time of r-BEOL matches the standard BEOL when the floorplan sizing is relaxed by 9%.

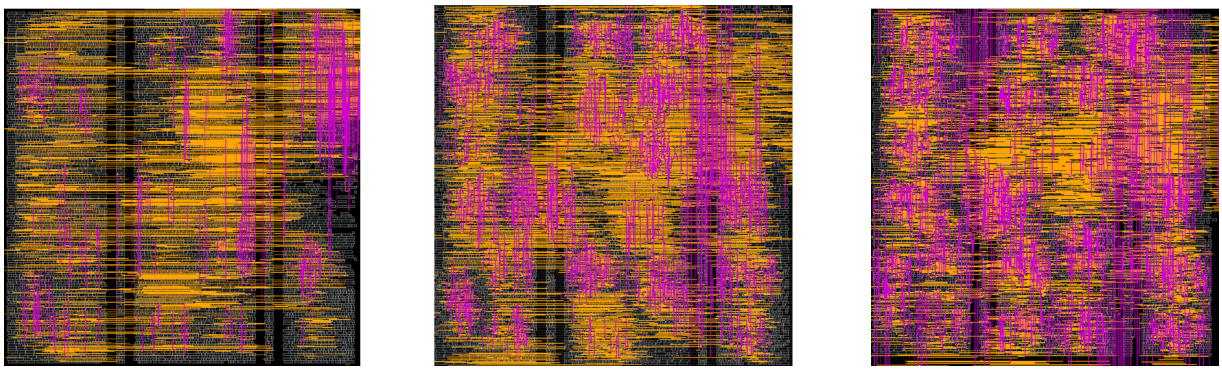



**Figure 3.13 Orange and pink lines stand for M6 and M7 layers respectively. Other layers are disabled.**

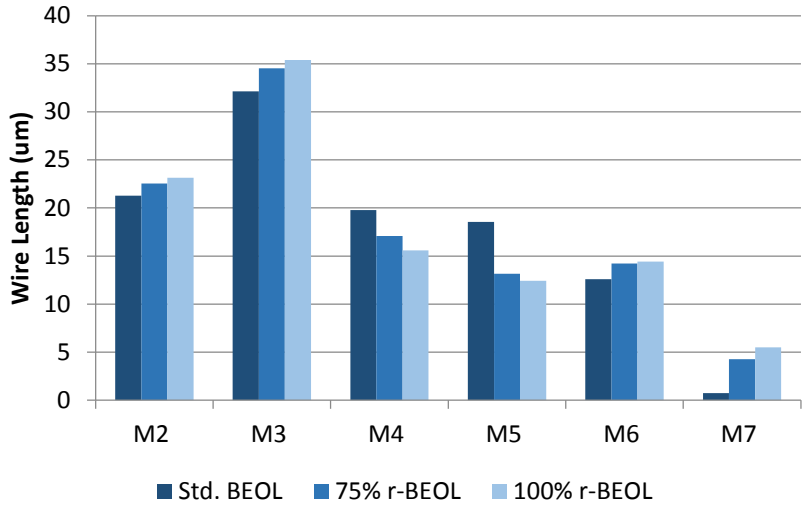



**Figure 3.14 Distribution of wires in AES-256 BEOL variants.**

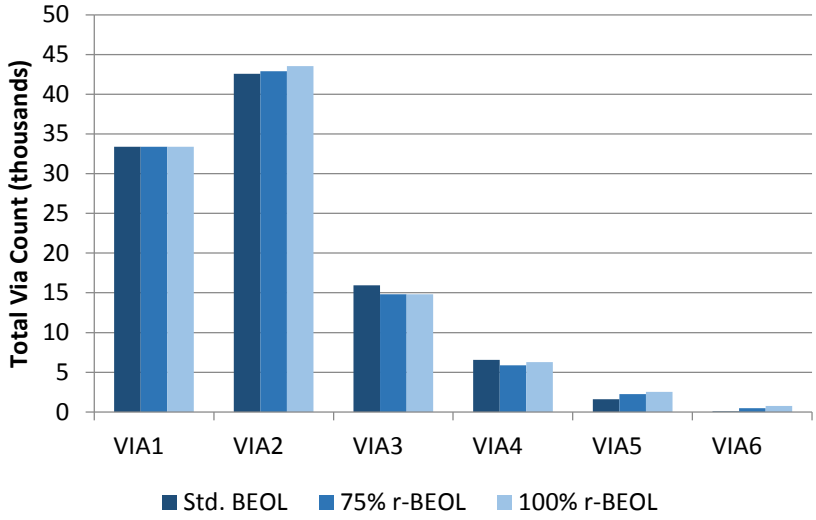



**Figure 3.15 Distribution of vias in AES-256 BEOL variants.**

| BEOL pitch / UF | Hor. Overflow (%) | Ver. Overflow (%) | Exec. Time (s) |
|---|---|---|---|
| **Standard BEOL UF = 0.94** | 0 | 0 | 872 |
| **r-BEOL 75% UF = 0.94** | 0.41 | 0.07 | 1351 |
| **r-BEOL 75% UF = 0.92** | 0.38 | 0.01 | 1208 |
| **r-BEOL 75% UF = 0.90** | 0.36 | 0.02 | 1004 |
| **r-BEOL 75% UF = 0.88** | 0.17 | 0.03 | 967 |
| **r-BEOL 75% UF = 0.86** | 0.19 | 0.01 | 881 |

**Table 3.2 Routing congestion analysis for the r-BEOL variants of the AES-256 circuit.**

### 3.6 Results

We have shown that r-BEOL can reduce the unit coupling capacitance and resistance of wires. For short and medium length wires, this reduction was shown to alleviate the cross-talk and improve the timing of wires. For long routes, r-BEOL was shown to reduce the power needed for compliance with maximum transition constraints. From the routing perspective, wire lengths were shown to increase as expected, which leaves the remaining question: is the reduction in coupling capacitance significant enough to overcome the increase in power coming from the increase in wire lengths?

To answer this question, we synthesized a set of benchmark circuits with sizes ranging from 1K cells to 2M cells, clock frequencies ranging from 200MHz to 2GHz, and of varying complexities. Each circuit was implemented with a standard BEOL and a 75% r-BEOL, wherein wiring pitches from M3 to M7 were relaxed. We used M8 and M9 layers for on-chip power distribution. All circuits were DRC clean and met the timing constraints. At the end of each run, we captured the extracted capacitance and reported the results in Table 3.3.

The circuits used in the experiment have very distinct sizes and cover a wide range of profiles that lowers a possible bias that can favor r-BEOL. The MM variants shown in Table 3.3 are

*matrix multiplication* modules with different degrees of pipelining and redundancy. The ALU circuit is a simple single-cycle 32-bit ALU with four operations. DES and AES [39] cores implement the encryption standards they are named after. The last circuit in the table is a correlation subsystem for a GPS application. Nearly every circuit, once implemented with r-BEOL, produced a small reduction in the extracted capacitance values. The observed capacitance reduction is in the order of 1 to 5%. The only exception to this trend is the MM_v4 circuit, which showed an increase of 0.5% in capacitance when designed with a target clock frequency of 200MHz.

| Circuit and Target Clock Frequency | Size (Kgates) | Standard BEOL (pF) | r-BEOL 75% (pF) |
|---|---|---|---|
| MM_v1 // 1.6 GHz | 0.3 | 1.29 | 1.23 |
| MM_v3 // 1.6 GHz | 1.2 | 2.46 | 2.41 |
| 32-bit ALU // 800 MHz | 1.1 | 5.44 | 5.38 |
| DES // 1 GHz | 1.5 | 6.52 | 6.38 |
| MM_v2 // 1.4 GHz | 6.1 | 28.9 | 27.7 |
| AES // 2 GHz | 10.7 | 42.2 | 41.6 |
| DES // 1.5 GHz | 13.4 | 46.6 | 45.9 |
| MM_v4 // 200 MHz | 62.8 | 386.5 | 388.1 |
| GPS corr. subsystem // 1 GHz | 2,059 | 6,049.4 | 5,962.2 |

**Table 3.3 Extracted capacitance values for various circuits under standard and relaxed BEOLs.**

The accuracy of static power estimations of design blocks depend on the modeling accuracy of logic cells and interconnects. Although IP vendors and foundries provide models that are highly accurate, they are not guaranteed to match the characteristics of transistors and wires of a specific die. Therefore, to validate the efficacy of r-BEOL, actual silicon data is needed. To this end, we designed a set of benchmark circuits in two separate chips, one using standard BEOL and one using 75% r-BEOL. The chips were fabricated in a commercial 14/16 nm technology on

the same wafer. Both chips have identical floorplans, power networks, and timing constraints to avoid skewing power results. The functionality of the benchmark circuits was validated at-speed.

Figure 3.16 [40] shows the current measurements for the 32-bit ALU circuits that are designed for a maximum clock frequency of 800 MHz. At lower frequencies, the r-BEOL variant of the circuit draws more current, whereas there is an inversion beyond 300 MHz. At 800 MHz, the r-BEOL variant consumes approximately 2% less power. Figure 3.17 shows the pipelined DES circuit results. Similar to the 32-bit ALU, the r-BEOL variant catches up with the standard BEOL variant. Since this design can run at a higher clock frequency of 1.5 GHz, r-BEOL enables a 4.4% power reduction.

At lower clock frequencies, standard and r-BEOL variants consume approximately the same amount of power. As the clock frequency increases and nets begin switching faster, the wire capacitance becomes more dominant in power consumption. In that regard, r-BEOL's capability to "shield" nets from cross-talk mitigates the severity of worst-case switching scenarios and lowers the power consumption.

As may be expected, process variation can have an impact on the r-BEOL silicon results presented. To discern r-BEOL power savings from sheer process variation, we first eliminate lot-to-lot variation since all chips were fabricated on the same wafer/lot. To ensure that the observed power differences are not significantly skewed from die-to-die variation, we evaluated multiple dies to obtain a sense of the statistical distribution. This result is shown in Figure 3.18, for which 30 measurements were performed: 15 on standard BEOL chips and another 15 on r-BEOL chips. The circuit under test is a DES crypto engine working at 1.2 GHz on a set of entirely random inputs (i.e., both the plaintext being encrypted and the key are randomized with an on-chip input

vector generator). The test setup was identical for all tested chips (same PCB, same discrete components, and same power sources). The average current measured for the standard BEOL variant was 39.85 mA with a standard deviation of 0.36 mA. The r-BEOL variant, on the other hand, registered an average of 39.07 mA with a standard deviation of 0.37 mA.

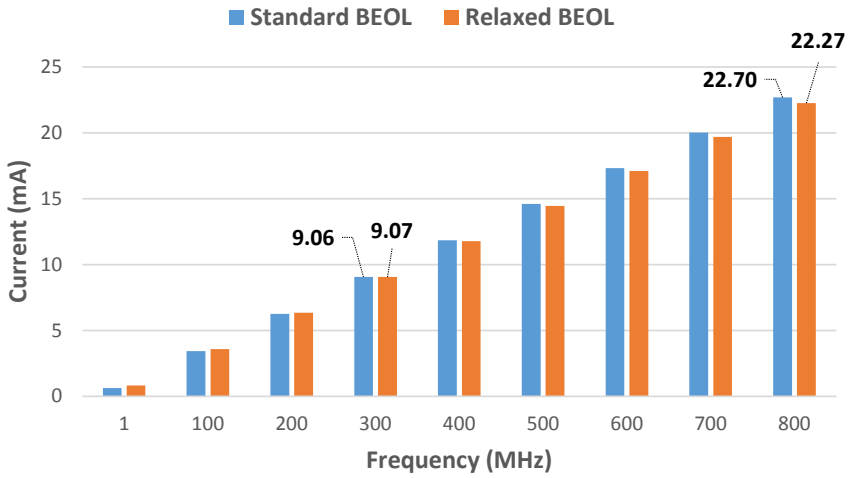



**Figure 3.16 Current measurements for a 32-bit ALU circuit fabricated with standard and relaxed BEOL variants.**

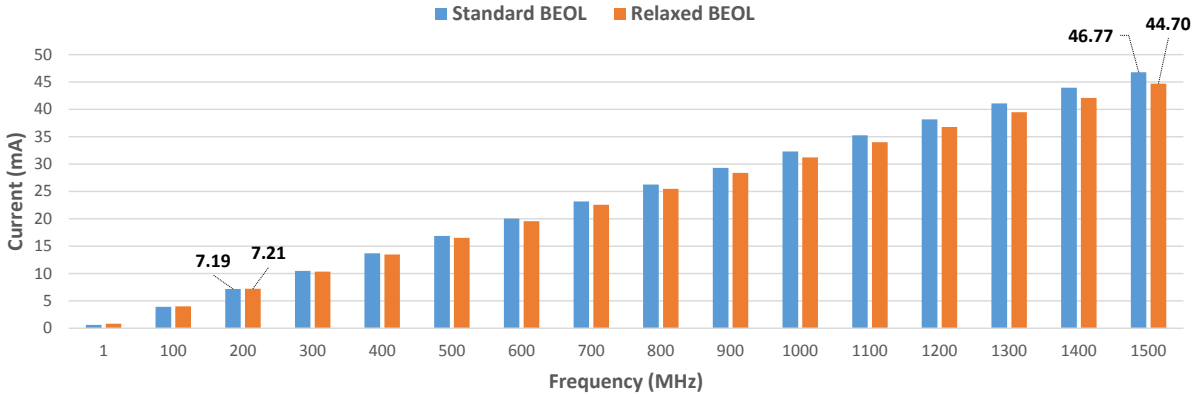



**Figure 3.17 Current measurements for a DES circuit fabricated with standard and relaxed BEOL variants.**

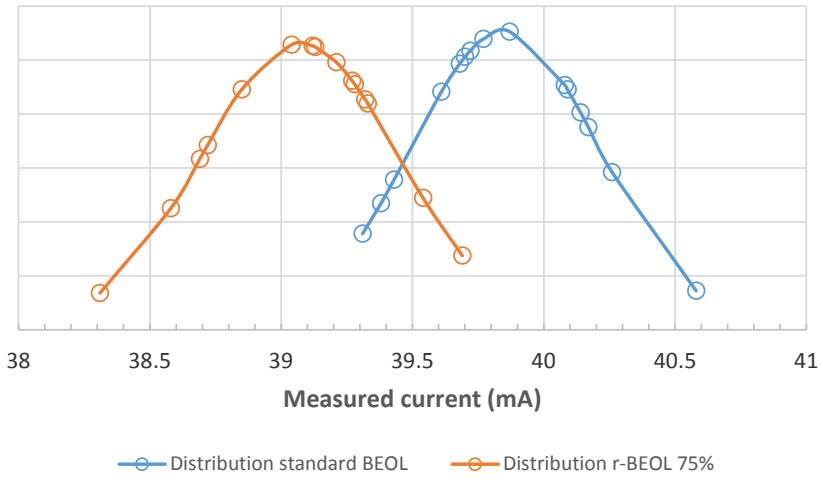



**Figure 3.18 Current measurements for a DES crypto engine running at 1.2GHz, multiple chips.**

While the power savings shown here are welcomed, power reduction is only one of the many aspects of a relaxed BEOL pitch. More importantly, we have shown how r-BEOL can counter the adverse effects of wiring density on SI and timing. Despite not being detailed in our results, the 75% r-BEOL wiring pitch is so relaxed that double patterning is no longer required for certain layers. As a result, r-BEOL can further alleviate DFM concerns and possibly lower fabrication costs without additional designer effort. In that regard, we see r-BEOL as an excellent, cost-effective methodology to enable the use of advanced CMOS nodes. For a holistic, cost-effective solution to the design problems in advanced CMOS nodes, we will next optimize the logic cells that are fundamental components of digital ICs.

## 3.7 Summary

In this chapter, we detailed the issues caused by the BEOL wiring density in advanced CMOS nodes and introduced a methodology to distribute the wiring more homogeneously across the BEOL metal stack by relaxing the wiring pitch. Figure 3.19 summarizes the methodology we presented in this chapter. In a commercial 14/16 nm, using circuits of varying sizes, we explored optimal r-BEOL pitches and modified the technology files for physical design. Thereon, we

demonstrated a reduction in the BEOL wire parasitics that resulted in timing and power benefits. We validated the capabilities of r-BEOL with multiple benchmark chips fabricated in the same technology. We have shown that, for designs with sufficient routing resources, there is no direct area overhead of r-BEOL; however, we have shown that some area increase might be needed to reduce the routing congestion. For congested designs, a designer can decide if a metal stack is feasible based on the area constraints dictated by the fabrication volume. Moreover, if collaboration with a foundry is a possibility, pitch relaxation can be leveraged to simplify some of the metal masks, lowering the manufacturing complexity and cost.

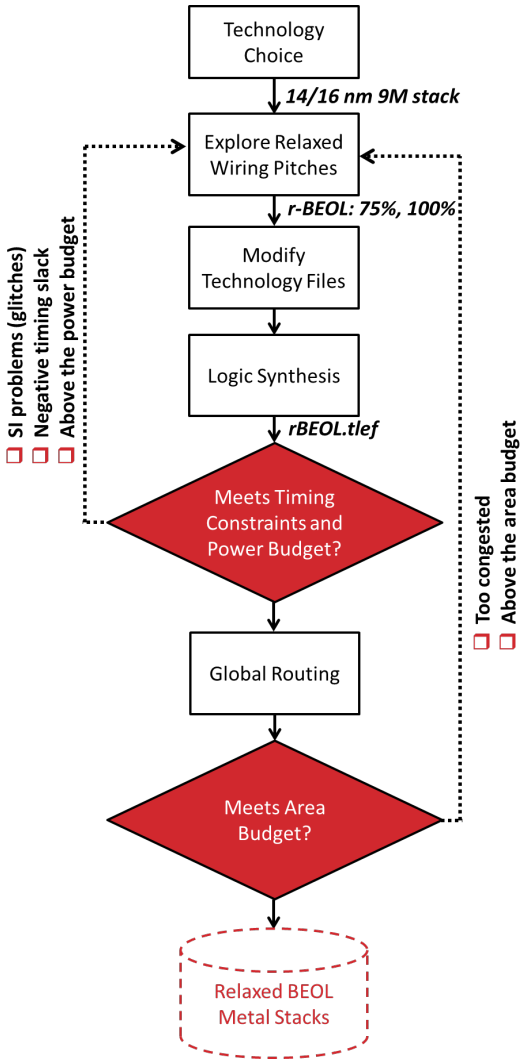



**Figure 3.19 Design methodology for r-BEOL.**

# 4 Optimizing Logic Cell IP

Before the 20 nm node, the traditional technique to scale logic cells was to keep the layout topology identical to previous nodes while meeting DRs in the newer node [30]. Availability of FinFETs, local connectors in MOL, and the need for multiple patterning have changed the logic cell layouts substantially. Owing to their 3-D structure, FinFETs can drive more current than planar FETs. Foundries and IP vendors leveraged FinFETs to offer shorter logic cells for high-density applications. Although appealing, the aggressive scaling of transistor dimensions could not be reproduced for the BEOL metal stack due to process difficulties. The resultant gap between the scaling factors of the logic cell heights and the wires created routing closure and DFM issues that increased the design cost dramatically in the 14/16 nm node and below. In this chapter, we propose a methodology to design logic IPs with a more structured layout to address these issues. We detail six steps that form the backbone of our proposed approach:

1) Select a design rule compliant cell height that is taller than the technology minimum
2) Use layout design guidelines to maximize pin access and layout simplicity
3) Design complex cells re-using simple *constructs* that have good pin accessibility
4) Analyze layout dependent effects induced by neighboring cells, and fix or eliminate cells that degrade the library performance, if any
5) Tune the fin counts in the logic cells to guarantee sufficiently diverse drive strengths with minimal impact on power
6) Determine the library composition in terms of Boolean functions and their drive strength flavors

## 4.1 Cell Heights

In more mature CMOS nodes (28 nm and above), IP vendors typically offer logic cells that are 12 and 9 M2 tracks (T) tall. Usually, the 12T cells are used for high-performance designs, and the 9T cells are used for low power applications. Owing to stronger current drive capabilities of FinFETs, 7.5T cells became more prevalent for high-density IC designs in sub-20 nm nodes. Recent studies show that cell heights can be downscaled to 6T in the 5 nm node [41]. Due to process difficulties, however, BEOL wire pitches do not follow the same downscaling trend as transistors.

In the 14/16 nm node, due to electromigration and IR drop requirements, foundries recommend dedicating one and a half M2 tracks to power rails. As a result, a 7.5T tall cell is left with six M2 tracks for signal routing. Usually, half of the remaining tracks are dedicated to the internal signal routing; therefore, a 7.5T cell at most has three M2 tracks for accessing to its input pins. Due to metal spacing and via enclosure rules, complex logic cells with high input counts commonly employ pins that are single or two M2 tracks tall. Such cells, when placed closely on a row, lead to routing problems, such as track stealing [12], [42].

### 4.1.1 Reverse Scaling

There are two possible directions to explore the trade-off between the cell area and pin accessibility: extending the cell horizontally or vertically. Horizontal extension, while keeping the cell height at 7.5T, could help spread the pins of a cell, therefore, reducing the likelihood of routing hotspots induced by pin proximity [14] inside the cells. However, the horizontal extension does not increase the number of M2 tracks that intersect with the M1 pins and does not guarantee a solution to the track stealing problem. Instead, we explore increasing the cell height to elongate and improve the accessibility of pins in the advanced CMOS nodes. We refer to this

cell architecture choice as reverse scaling (RS), since commercial logic cells display the opposite trend.

For the commercial 14/16 nm technology used in this work, the DR-compliant cell heights after 7.5T are 9T and 10.5T. A clear trend is visible in Figure 4.1 that shows the layouts of an AOI22 (and-or-invert with four inputs) cell in 7.5T, 9T, and 10.5T: M1 polygons become simpler as cell height increases. The 7.5T variant is a projection from the commercial library whereas the 9T and 10.5T variants are from RS libraries. In the following subsections, we detail our layout design decisions for simplicity and re-use.

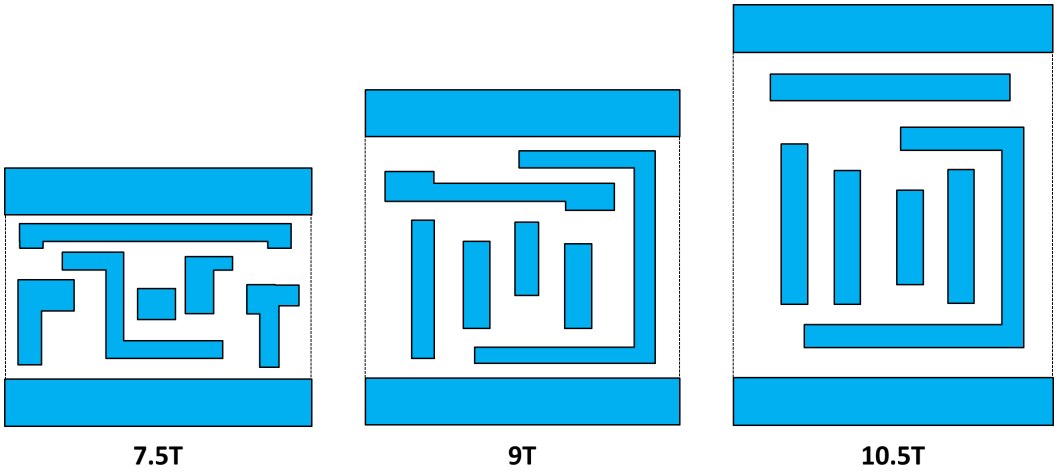



**Figure 4.1. Layouts of AOI22 cells become simpler with as cell height increases.**

### 4.1.2 Relation of Reverse Scaling to r-BEOL

In Chapter 3, we demonstrated r-BEOL's efficacy at mitigating SI, timing, power, and reliability issues induced by wire density. Although r-BEOL is compatible with the 7.5T commercial logic IP, increasing the wiring pitch would further limit the pin accessibility of short logic cells, creating more routing problems. In that regard, RS logic IPs complement r-BEOL more effectively. Therefore, availability of RS logic IPs is invaluable for a low-cost IC design flow

that holistically addresses design closure and manufacturability issues at every level of the BEOL stack in sub-20 nm CMOS nodes.

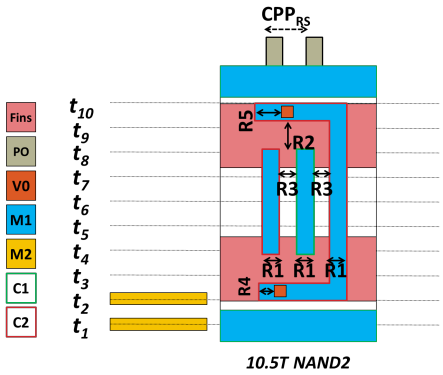
### 4.2 Pin Shapes and Internal Cell Routing

Pin accessibility, particularly in sub-20 nm nodes, boils down to congestion between the pin layers and the first routing layer [43]. Specifically for the 14/16 node, pin access means DR-compliant connections among M1 pins and M2 routing with VIA1 in between. Since pin density increases as cells become shorter, VIA1 spacing problems begin arising. To space out the vias in commercial logic cells, small horizontal M1 stubs are added to pins as shown in the 7.5T cell projection in Figure 4.1. This strategy compromises the layout regularity but does not guarantee DR-compliant pin access for all cells.

To further reduce the pin density, a contacted poly pitch ($CPP_{7.5T}$) that is approximately 10% bigger than the technology-minimum $CPP_{min}$ is used in 7.5T cells. In our RS cell architecture, we take advantage of the pin density relaxation enabled by height increase and use a $CPP_{RS}$ that is equal to $CPP_{min}$. Using CPP at minimum helps compensate for a portion of the area increase induced by cell height increase. On the flip side, the FEOL and MOL features (vertical and horizontal M0) need to be spaced 10% tighter, rendering a trade-off between the cell width and internal coupling capacitances.

Figure 4.2 marks the layout choices used in both RS libraries on the 10.5T NAND2 cell. For the sake of brevity, the MOL layers (horizontal and vertical M0 connectors) are not shown. The legend, on the left of the figure, shows the layers: fins (active region), poly (PO), V0, M1, and M2. Outlines, C1 (green) and C2 (red), represent the M1 coloring masks. For the sake of brevity,

the foundry DRs corresponding to our layout choices are abbreviated as Rule (R) 1, R2, R3, R4, and R5.




| Layout Choice | Definition |
|---|---|
| $CPP_{RS} = CPP_{min}$ | $CPP_{RS} < CPP_{7.5T}$ |
| R1 | > Minimum width for tip-to-side spacing |
| R2 | Min tip-to-side spacing, same color |
| R3 | Side-to-side spacing, different color |
| R1 + R3 = $CPP_{RS}$ | Pin pitch is equal to the contacted poly pitch |
| R4 | > Minimum VIA-0 enclosure,<br>= Recommended DFM VIA-0 enclosure |
| R5 | > R4 for pin symmetry |

**Figure 4.2. The layout simplification choices are shown on the 10.5T NAND2 cell.**

R1, the width of vertical only M1 pins, is wider than the technology-minimum width to enable technology-minimum R2 and R3 spacing among M1 shapes of same (C1-C1 and C2-C2) and different colors (C1-C2) respectively. Thus, vertical input pins that are orthogonal to other M1 polygons can be elongated as much as possible. Input pins are placed at a uniform pitch 'R1+R3' that is equal to $CPP_{RS}$. This pin architecture forces the shapes in the MOL layers, which interface between M1 and FEOL, to be distributed regularly as well. The output pin is allowed to have only two 90-degree turns, and it has a width (for both directions) of R1 as well. The smallest M1 enclosure of VIA0 is R4 that is a DFM recommendation (larger than the technology-allowed minimum). The other enclosure choice, R5, is larger to create a pin shape that is perfectly symmetrical to the cell's equator, if possible. Symmetry makes a cell's BEOL indifferent to its row orientation. Logic cells such as NAND-NOR and AND-OR can be designed with identical BEOL shapes.

For internal signal routing, we use M1 shapes in a restricted bidirectional fashion that duplicates the layout choices for the input and output pins. Although unidirectional routing has been shown to lead to extreme layout regularity [8], [15], [28], [29], it pushes the layout complexity and pin

access problems to the upper routing metal layers. For 10, 7, and 5 nm process nodes; however, uni-directional routing is a requirement of self-aligned patterning; thus, M0 is used perpendicular to poly and M1 is parallel to poly. Owing to restrictions of turns we apply on M1 shapes; it is possible to map horizontal M1 shapes to M0 shapes in more advanced process nodes.

The utilization profile of horizontal M1 routing tracks inside the cells varies across the library. For simpler cells like inverters, NANDs, NORs, two horizontal M1 tracks are dedicated to internal routing. For cells like AOI and OAI (or-and-invert) three horizontal M1 tracks are utilized. M2 internal routing is used in a horizontal-only fashion inside complex combinational cells (adders, XORs) and sequential cells for clock routing. We routed internal M2 wires at technology minimum pitch and used four M2 tracks at most (in a DFF with asynchronous reset input). For efficient use of routing resources, we allow M2 signal routing over our cells.

Table 4.1 shows how much the capacitance values of an input pin, an intermediate node, and the output pin differ for taller cells when compared the equivalent parts of a 7.5T NAND2 cell. The capacitance values were extracted for the most pessimistic corner and normalized by the input capacitance of the 7.5T variant of the cell. The input capacitance for the 9T variant is lower because we transformed the irregular pin shapes into vertical ones. The input capacitance of the 10.5T variant is larger because of the surface area of pin shapes. Using a tighter CPP increases the internal parasitics of cells. Since the MOL features are spaced closer in the 9T and 10.5T cells, the coupling is stronger; therefore, the capacitance of intermediate and outputs nodes is more significant.

| Cell Height | Input Cap | Intermediate Cap | Output Cap |
|---|---|---|---|
| 9T | -1.6% | +6.3% | +7.0% |
| 10.5T | +3.0% | +5.9% | +7.6% |

**Table 4.1 Normalized pin capacitances for NAND2 cells from three libraries.**

## 4.3 Construct-based Design

All surrounding neighbors can block access to a pin that is close to a cell's boundary. For a commercial library with thousands of cells, which differ in their pin layouts substantially, guaranteeing pin-access regardless of the neighboring cells is almost impossible. To quantify and show the scale of this problem, Figure 4.3 shows four cells on two consecutive rows. A cross marks an input pin of the center cell $C_i$ that is close to the boundary. The accessibility of the marked pin depends on the use of vertical and horizontal routing tracks by its horizontal ($N_H$), vertical ($N_V$), and diagonal ($N_D$) neighbors. For one center and three neighbor cells, assuming mirroring around the y-axis is possible, there can be $2^4$ different placement orientations. Moreover, cells can shift in delta steps that are multiples of the CPP (contacted poly pitch) and still influence the accessibility of the pin. Thus, for a 1000-cell commercial library, there can be more than 16 ($1000^3$ by 16) billions of such abutment scenarios (Equation 1).

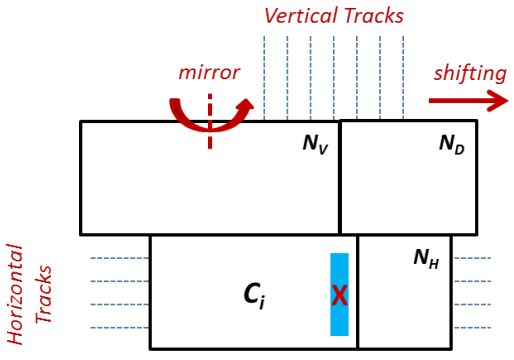



**Figure 4.3. The accessibility of a pin depends on all surrounding neighbors.**

$$\#of\ abutments = \sum_{\Delta=PP} \sum_{C_1}^{C_{1000}} \sum_{N_{v,1}}^{N_{v,1000}} \sum_{N_{d,1}}^{N_{d,1000}} \sum_{N_{h,1}}^{N_{h,1000}} 2^4 \quad (1)$$

There are two ways to reduce the number of different abutments. The first direction is to reduce the number of cells in the library. The second approach is to adopt a restrictive layout design approach that ensures geometric similarity among pin shapes. In this work, we tune the library composition and restrict the layout design to achieve our goal to improve pin access. To this end, we re-use the layouts of simple logic cells to implement more complex cells. These simple layouts are referred to as constructs, and the approach is named as *construct-based* library design. If the pin access of constructs can be proven to be good, construct-based complex cells have, thus, good pin access.

Cells that are good construct candidates should have low pin count density (equal to or less than four consecutive pins), simple pin shapes (minimum no. of turns, no indentations or stubs), and high preference rate by logic synthesis (good PPA). To choose a set of constructs, we surveyed a set of IWLS 2005 benchmark circuits [39]. We then performed logic synthesis on these benchmarks using a commercial tool and a 1000-cell commercial library. In the resultant design blocks, we ranked the cells by how often they appear (i.e., what percentage of the block area is coming from a given cell). This metric indicates that a cell is frequently picked by the synthesis tool and it is large enough to abut many different cells. Figure 4.4 shows that DFFs relatively cover the 50% of floorplan area. NAND2, NOR2, OR-AND-INV (OAI21, OAI22) and AND-OR-INV (AOI21, AOI22) cells are the next group of frequently used cells. Our interpretation of this data is that if the layouts of AOI22 (OAI22), NAND (NOR), and INV cells are used as baseline constructs to design more complex cells, such as DFF, XOR, and adders, the library will achieve more layout regularity. Making sure that the baseline constructs provide reasonable pin

access would, therefore, propagate the same level of pin access to the whole library. Table 4.2 shows the construct composition for a selection of complex Boolean functions. We explored the design space for the ordering of constructs and the internal routing among them for the optimal PPA, minimum M2 use, and the least modification in FEOL, MOL and BEOL layers inside the cells. The M1 utilization of more complex cells follow their core constructs. For instance, latches in DFFs take up three horizontal M1 tracks and an additional track for latch-to-latch connections.

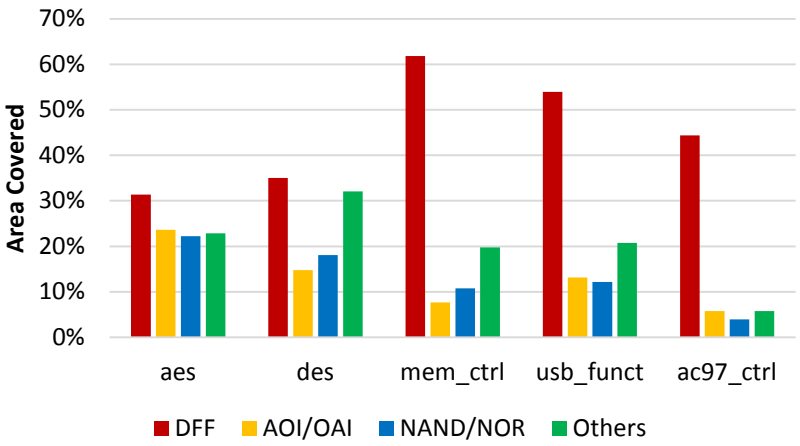


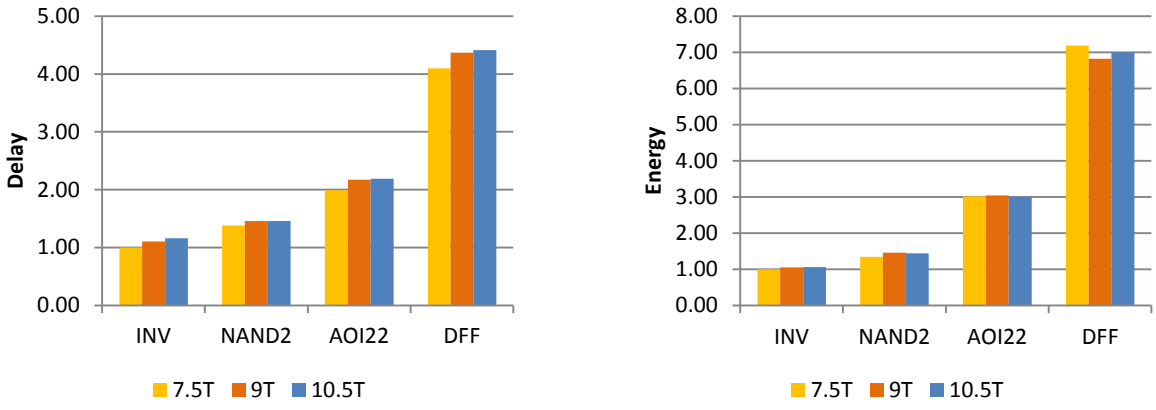

**Figure 4.4. DFF and AOI/OAI account for the majority of design areas.**

| Cell | Function | Constructs |
|---|---|---|
| MUX2 | IN0`.SE+IN1.SE` | AOI22+2×INV |
| XOR2 | A`B+AB` | OAI21+NAND2 |
| XOR3 | A⊕B⊕C | 2×NOR2+2×AOI21 |
| Latch | D.CK + Q.CK` | AOI22+2×INV |
| DFF | Posedge DFF | 2×AOI22+2×INV |
| DFFR | Posedge Reset DFF | 2×AOI22+2×INV+2×NOR2 |

**Table 4.2 Construct composition of complex logic cells.**

Figure 4.5 compare the performance (a) and energy consumption (b) of a set of simple (INV, NAND2) and more complex cells (AOI22, DFF) of different heights. The values in charts are normalized by the 7.5T inverter's delay and energy consumption values. To collect these values,

we performed SPICE simulations on pessimistic RC netlists of the cells at slow-slow, low-supply voltage (90%), high temperature (125C) process corner. Each cell has a fan-out of 4 (FO4) load to consider the impact of input pin capacitance differences among cells of different heights as demonstrated in Table 4.1. All cells used in the experiment have identical fin counts and drive strengths. Figure 4.5 suggests that the performance of simpler cells tends to be closer whereas the construct-based complex cells are relatively slower. Traditionally, taller cells are leveraged for packing wider devices for high-performance applications but construct-based design style allows for less layout customization such as Euler path, diffusion overlaps, poly cuts etc. Due to the parasitic capacitance differences shown in Table I, the 9T and 10.5T cells consume more energy. In overall, 9T cells are more energy efficient than their 10.5T counterparts. DFF is one of the few cells construct-based design style relaxes the internal routing and the corresponding capacitance, leading to more energy-efficient but 10% slower DFFs.




**(a)** **(b)**

**Figure 4.5 FO4 delay and energy consumption of cells of different heights.**

To assess the trade-off between the layout regularity and performance at the library level, we characterized the timing of RS libraries and the equivalent subset of the 7.5T (7.5T SS) cells using an identical extraction, simulation, and modeling flow. Then, we synthesized a set of

benchmark circuits of different sizes, complexities, and functionalities. We used an AES-256 [39], a matrix multiplication (MM) module, two 32-bit open source RISC processors (OR1200 [42] and an ARM core [43]), and an all-digital spiking neural network (SNN). We set a hard-to-achieve timing constraint on these designs to capture the maximum achievable clock frequency. Figure 4.6 shows that all libraries achieve almost similar maximum clock frequencies for all designs. Depending on the design, the construct-based libraries can enable slight performance improvements or a cause a marginal slowdown. The SNN design is the one that the 9T and 10.5T libraries performed the poorest compared to the 7.5T SS; they lag by 3.5% and 4.5% respectively. In contrast, for the ARM core design both construct-based libraries performed 3% better than their 7.5T counterpart. These performance differences are attributable to the cell choices made by the logic synthesis tool.

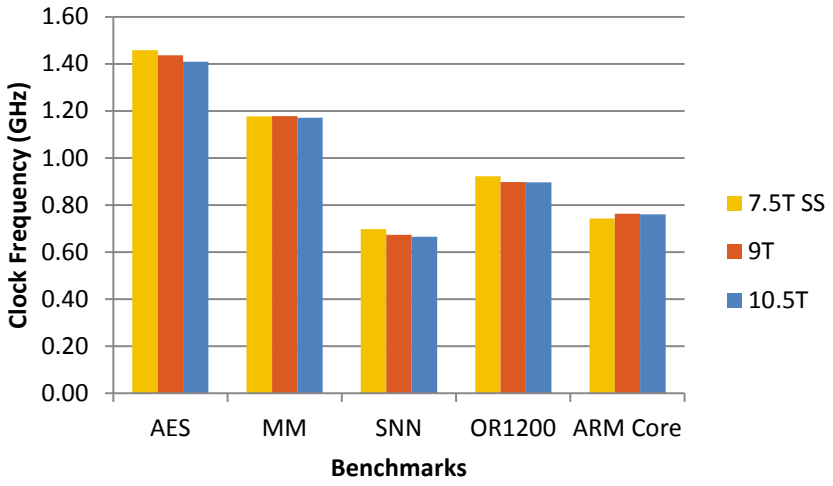



**Figure 4.6. Performance impact of construct-based logic IP design.**

## 4.4 Layout Dependent Effects

In advanced CMOS nodes, systematic variations caused by lithographic imperfections are modeled in post-layout (extracted) netlists of cells. Spacing to oxide edge (SOD), cut poly effect (CPE), well proximity effect (WPE) are major LDEs (layout dependent effects) that alter a

transistor's carrier mobility and threshold voltage depending on its layout. The post-layout extraction engine determines the LDE parameters of a transistor based on the surrounding layout shapes within a window that is several poly pitches wide and several fin pitches tall. The extraction window is comparable to the size of a smaller logic cell, and the "context" [46] that surrounds a cell plays an important role in its power and timing characteristics.

Design blocks with a high number of cells and significantly different layouts have too many contexts to be characterized for timing and power – some cell combinations can have better power and performance, while some have worse. As a result, static timing analysis is performed in an LDE aware fashion for timing critical paths in an IC with overly conservative margins. Alternatively, limiting the types of layout arrangements can translate to complete knowledge of timing for all possible contexts in construct-based logic IPs. This approach can improve the accuracy of timing estimations and eliminate the need for conservative/pessimistic timing margins that might lead to sub-optimal IC performance.

Without having access to known failure mechanisms and patterns from the foundry, we seek detractor cells in our in-house logic cell libraries that cause a significant deviation in the timing of a target inverter. For this experiment, every cell with a different layout has been abutted to an inverter. In both 9T and 10.5T logic IPs, slightly more than 40 cells differ in their layouts. For the 7.5T library, equivalent counterparts of the abutment cells have been used. Following the abutment, RC extraction is performed at the worst corner, assuming higher dielectric constant and pessimistic misalignment between neighboring polygons. Thereon, the target inverter's propagation delay is calculated via transient SPICE simulations for a FO4 load.

Figure 4.7 shows that the target inverter's delay varies for a set of different neighbors in all libraries. Note that due to the height difference, 9T and 10.5T inverters start with a slightly higher delay. The difference between the best and worst cases is less than 2% for all libraries. Although this is not a significant difference, it should be noted that the samples used in this experiment correspond to 100% of all single neighbor abutment scenarios with unique layouts for both 9T and 10.5T libraries, whereas they cover less than 5% of all cases for the 7.5T library. The core conclusion of this experiment is that a simple and construct-based layout design can relax the pessimistic timing margins that emulate "the worst contexts" because all possible cell abutment scenarios can be feasibly characterized, owing to layout regularity.

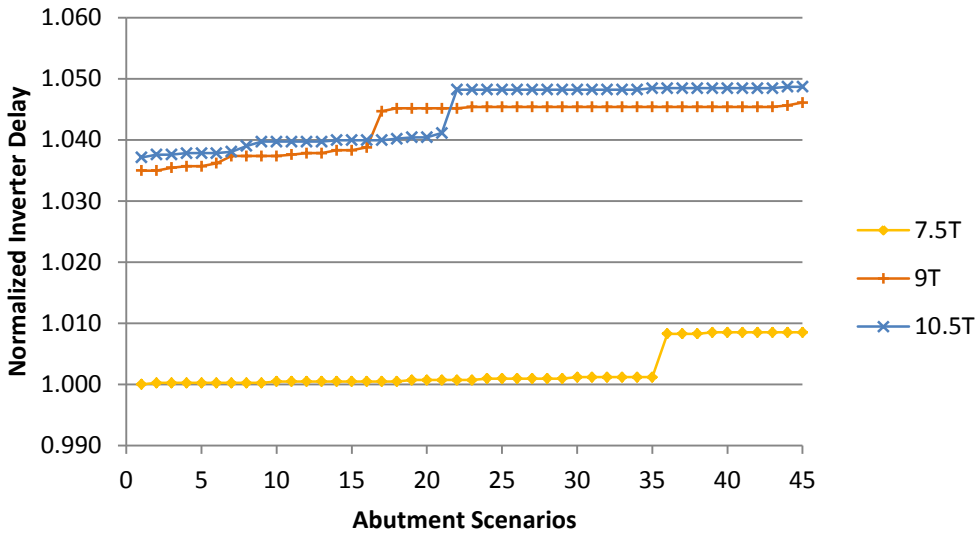



**Figure 4.7. Normalized inverter delay for a set of abutment scenarios.**

## 4.5 Fin Efficiency

For legacy technology nodes with planar FETs, speed and drive strength of a gate could be tuned via transistor widths, which was a continuous variable. Since the number of fins is discrete in FinFETs, there is a much restrictive design space for a given logic cell. In the 14/16 nm node, 7.5T, 9T, and 10.5T cells can contain six, eight, and ten fins (half PFET, half NFET). Because of

the highly capacitive nature of FinFETs, fin count increase may not yield the typical power-performance trade-offs seen in previous planar MOSFET nodes.

To explore the power-performance impact of the fin count choice, we used fin efficiency (FE), the ratio of the active fin area to a cell's area [15], as a proxy. On a 10.5T inverter, we tried three FE flavors: low (43%), medium (57%), and maximum (71%, not 100%, since some area is dedicated to pins). Figure 4.8 shows the FO4 delay and dynamic power consumption of 10.5T inverters with different FE values. This graph suggests that, going from low to medium FE, there is a 10% reduction in propagation delay at the cost of a 24% increase in dynamic power consumption. This increase is expected since the total number of fins goes from six (three in PFETs and three in NFETs) to eight. However, the delay improvement is reversed at maximum FE due to capacitance increase. This trend suggests that cocktailing low and medium FE cells in a library can enable a richer selection to logic synthesis for power and performance optimization.

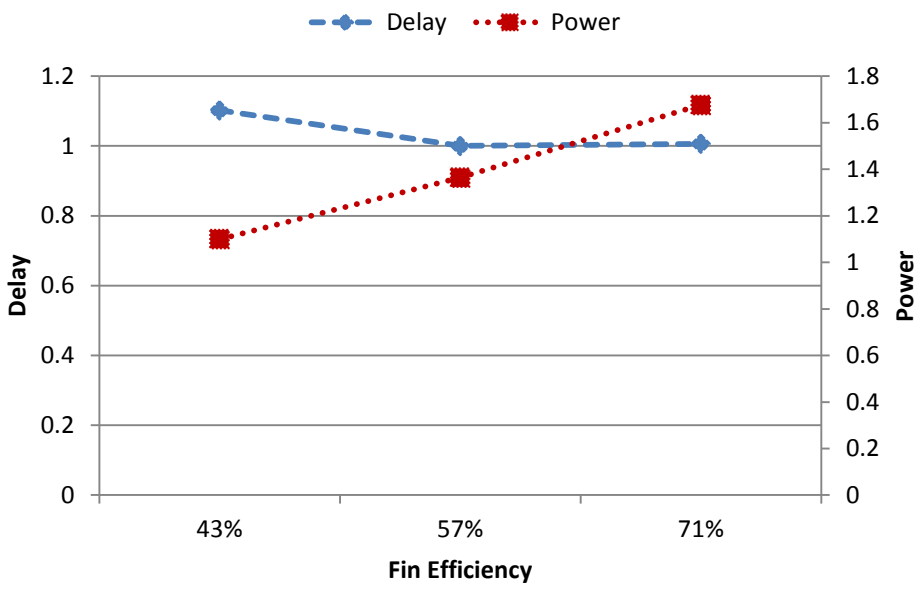



**Figure 4.8. FE impact on delay and power of an INV.**

## 4.6 Library Composition

The diversity of Boolean functions and their drive strengths enriches the selection of cells for logic synthesis optimization. As a result, commercial logic cell libraries usually contain thousands of cells to improve the PPA [42]. In this work, we limited the library composition to a set of essential combinational and sequential cells and their various drive strengths. As a result, cell layouts can be limited to a small set to mitigate routing, timing, and DFM issues.

Both 9T and 10.5T libraries are comprised of multiple drive strength versions of adders, falling and rising edge DFFs, scan DFFs (SDFFs), clock gates, asynchronous input (reset and set) versions of all DFFs, delay cells, combinational logic cells up to four inputs (e.g. AND2, AND3, AND4), AOI and OAI cells, buffers, and inverters. Additionally, physical (logically non-functional) cells such as well-taps, termination cells, antenna cells, and decoupling capacitors are designed. The number of cells is 100 for both libraries. For the sake of fairness, the PPA comparisons are performed for an equivalent 100-cell subset (7.5T subset) of the 1000-cell commercial library, as well as for the full library (7.5T full).

The cell level simulations presented in Figure 4.8 suggested that combining LP and HP cells could yield to better power and performance tuning in logic synthesis. To assess the impact of FE diversity on power and performance, we designed three AES-256 [39] blocks with LP-only, HP-only, and LP-HP mixed libraries. No clock gating was used to assess the power-performance trade-off based on the combinational and sequential cells. Table 4.3 shows that the HP-only library improves the performance by 10% and incurs a 20% power overhead compared to the LP-only design. It should also be noted that fins are ‘power hungry’ even when they are not switching. Therefore, there is a substantial increase in leakage. The last row shows that when LP and HP cells are combined, the logic synthesis tool manages to achieve the same performance as

it did in the HP-only case, but at much less dynamic and leakage power. This observation validates the initial hypothesis that enabling a richer drive strength selection within the library can improve power-performance optimization. Another implication of this observation is for the layout regularity of cells. Since there is more room to vary the fin counts inside taller cells, more drive strength flavors can be achieved without altering the BEOL layouts. As a result, in RS libraries, a richer timing and power design space were achieved with fewer layout patterns. In the rest of this dissertation, only the LP+HP versions of the RS libraries are used in the experiments.

| | Leakage (uW) | Dynamic (uW) | Total (uW) | CK Frequency (GHz) |
|---|---|---|---|---|
| **LP only** | 0.18 | 12.42 | 12.60 | 1.90 |
| **HP only** | 0.30 (1.7x) | 15.00 (1.2x) | 15.30 (1.2x) | 2.13 (1.1x) |
| **LP+HP** | 0.24 (1.3x) | 13.60 (1.1x) | 13.84 (1.1x) | 2.12 (1.1x) |

**Table 4.3 Library composition's impact on power and performance**

## 4.7 Results

In this next section, we analyze the effects of all RS techniques on the power and performance of multiple digital designs in a fabricated test chip. Further, we analyze the impact of RS libraries on the ease of routing and demonstrate the area trade-off associated with the use of taller cells.

### 4.7.1 Power and Performance

To validate the functionality of our in-house logic IPs and to assess the impact of our design choices, we taped-out a chip in a commercial 14/16 nm FinFET process. This die contains four AES-256 encryption block designs implemented with the 7.5T full commercial library, a subset of the same 7.5T commercial library, and our 9T and 10.5T libraries. For the sake of simplicity, we use the following names for each library in the remainder of the text: 7.5T Full, 7.5T Subset, 9T, and 10.5T. Figure 4.9 (a) shows a microscopic image of a group of the die and the next image highlights the AES-256 design blocks on a simplified overview of the chip.

The design blocks were synthesized with an identical clock speed constraint of 1 GHz. All designs were able to meet the constraint with a positive slack of 10ps for hold and setup checks. Clock gating was allowed during synthesis to minimize power consumption. The power measurements for every block were performed when the other blocks received no clock signal. We used a linear forward shift register (LFSR) to generate pseudo-random encryption input data and keys.

There are two test modes in the chip: functional and power. In the functional test mode, the blocks encrypt and decrypt a 256-bit input data coming from the LFSR. The output data is compared to the input data, and a valid flag is raised if they match. For the power test mode, all blocks run continuously to allow for external power measurement. When in the power test mode, the output pads are logically disconnected to eliminate the impact of IO pad power consumption in the measured results. The power testing was performed at room temperature with a 0.8V core supply voltage from ten chips to account for die-to-die variation. Figure 4.10 shows the average root-mean-square (RMS) power consumption of four variants of the AES-256 design. According to the measurement data, the 7.5T Subset and 9T variants consume only 1% more power than the 7.5T variant. The 10.5T variant, on the other hand, consumes increasingly more power which peaks around 5% at 1 GHz. This difference is attributable to the higher input capacitances and high fin count HP cells in the library.

These measurements show that the improvements we achieved in the BEOL through RS and construct-based design have a marginal impact on the power and performance of design blocks. The 9T and 10.5T cells, owing to their FEOL layout, can operate in GHz regime while only consuming slightly more power than their 7.5T counterparts. Next, we investigate our logic IP's impact on routing closure and the associated area trade-off.

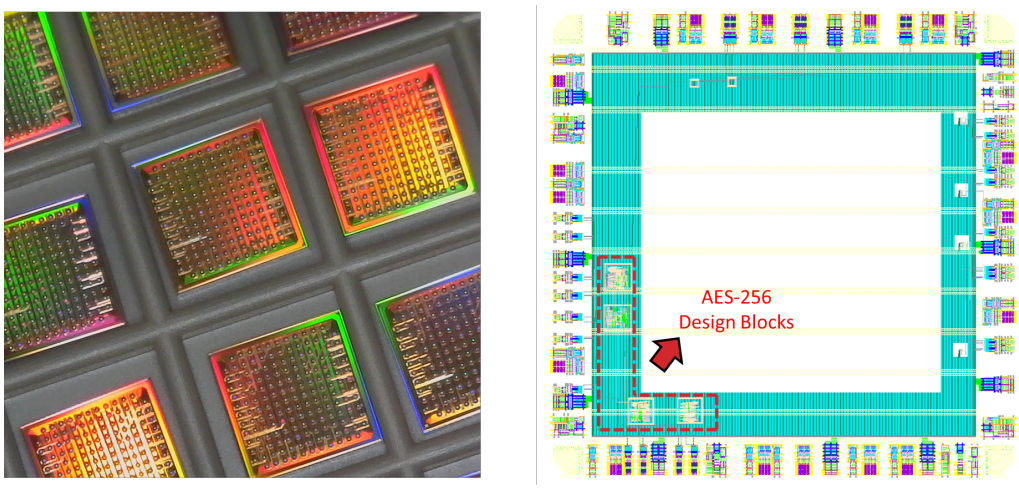



**Figure 4.9. (a) Microscopic image of the batch of dies in the tray; bumps and routing in upper metal layers (M10, M9) are visible (b) The GDS view of the chip with a simplified power mesh.**

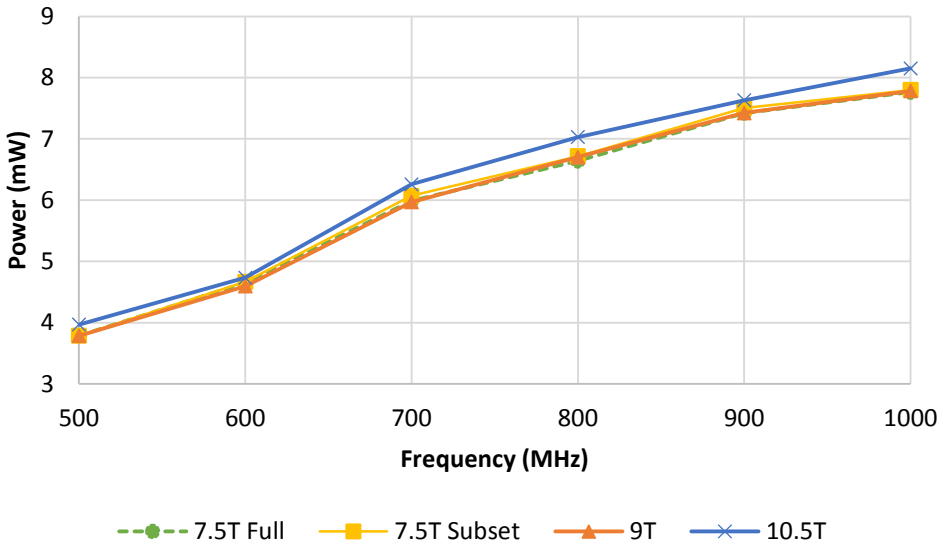



**Figure 4.10. The total power consumption of AES-256 blocks implemented with different libraries at 1 GHz clock speed.**

### 4.7.2 Routing and Area

Area of logic cells, placement density, and timing specifications affect the area of a design block. The literature shows that the smallest design block area is not always achieved with the smallest library but by the most routable one [12]. Increasing the cell height imposes a cell-by-cell area increase by default; however, it also enables more pin accessibility. In this section, we show that

the pin access improvement can overcome the area penalty due to the use of taller cells.

Figure 4.11 shows the normalized area of a selection of combinational, buffer, and sequential cells. Typically, the area overhead compared to 7.5T cells is around 15% for 9T cells and approximately 30% for 10.5T cells. This overhead is proportionally less than the cell height increase factor (20% for 9T and 40% for 10.5T) due to the reduction in CPP enabled by vertical expansion. AOI22_X1 is highlighted because increasing cell height from 9T to 10.5T allows for further horizontal compactness and the area overhead comes to 13% for both versions of the cell. Due to construct-based DFF layout design, the area overhead of sequential cells is more significant. The total areas of all 1000 cells in the 9T and 10.5T libraries are 18% and 34% bigger than the 7.5T subset, respectively.

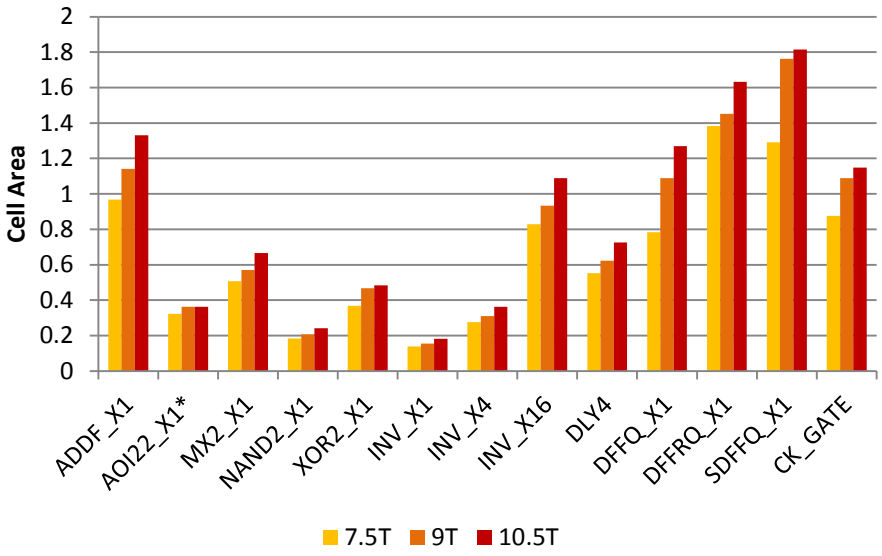



**Figure 4.11. Normalized area of the combinational, buffer, and sequential cells.**

To observe the impact of cell height on the routing and area, we used an AES-256 encryption module with 30K gates and OR1200 [44] RISC core with 300K gates. In logic synthesis, we set a clock target of 1 GHz for both designs. In the power network of both design blocks, we used wide M3 stripes for power routing. Additionally, following the IP vendor's recommendations to

prevent IR-drops in larger designs, horizontal M2 power rails were used inside the OR1200 blocks.

In this experiment, we implemented both designs in a commercial physical design tool using 7.5T Subset, 9T, and 10.5T libraries. While increasing the UF (placement density of logic cells) in small steps, we demonstrated at what area DRC violations reach to an unfixable number for each variant of the designs. The number of DRC violations reported hereby is the smallest that the router can reach after reasonable attempts at finishing detail routing. To emulate congestion, we assumed that the metal layers above M5 are dedicated to global signal routing, and we limited the number of routing layers to four in the AES-256 blocks. In the OR1200 blocks, given the higher complexity of the design, we allowed signal routing up to M6.

Figure 4.12 shows the number of DRC violations versus UF for 7.5T Subset, 9T, and 10.5T library variants of AES-256. The routing of the 7.5T variant shows an early onset of DRC violations after 60% UF. The violations statistics are as the following: 62% of the total violations occur in M1, M2, M3, and the corresponding via layers. More than a quarter of all violations are VIA1 and M2 spacing problems that are the primary interface for logic cell pin access. While trying to solve these problems, the detail router eventually creates more problems in the upper metal layers. In contrast, the 9T and 10.5T variants of AES-256 are free of routing violations until 67% and 77% UF respectively. Beyond these points, the 9T and 10.5T variants show similar types of DRC violations as the 7.5T variant.

The areas of the 9T and 10.5T variants of AES-256 are only 4.9% and 5.3% bigger than the 7.5T variant's area (Table 4.4). These numbers are much smaller than the cell-by-cell area overheads presented in Figure 4.11 because improving the pin access enables higher placement density with

less routing failures. This finding validates the core objective of this work: reverse scaling the BEOL of logic IP can enable substantial routing closure advantages at the cost of minimal area overhead for congested designs.

Figure 4.13 shows that after 50% UF, the number of DRC violations increases exponentially in the 7.5T variant of OR1200. On the other hand, at 70% UF, 9T and 10.5T variants of OR-1200 have less than ten violations that are four orders of magnitude fewer than their 7.5T counterpart. No routing was performed beyond 70% UF because well-taps, termination cells, and other required physical cells consume a considerable amount of area. The second row of Table 4.4 presents the area overheads for OR1200 blocks. For this area comparison, we have accepted ten or fewer DRC violations as being fixable with a minimal designer intervention (perhaps another iteration of detail routing would be sufficient). We highlight that the area of the 9T variant of OR1200 fits into a smaller area than the 7.5T variant. Most likely, some of the DRC violations could have been eliminated by pinpointing the problematic cells in the 7.5T variant of OR1200 and applying non-default rules for their placement and routing. This approach would most likely be able to reduce the area of the 7.5T variant below the 9T; however, at the cost of manual designer effort. We also highlight that 10.5T cells, which are on average 30% larger than their 7.5T counterparts, incurs only 13% area overhead for this design.

It is important to note that the block designs used in this experiment are much smaller compared to complex SoCs. In a much larger design with millions of gates, the number of violations would be exponentially higher (for all libraries), and the router would take substantially longer time to converge to a solution (even with millions of violations). Therefore, the routing results exhibited in the routing experiments imply that taller cells can be utilized to eliminate major routing-induced DRC issues, thereby lowering the NRE costs in IC design. We note that we do not

present measured execution times for routing, although they tend to increase substantially with the increase in DRCs.

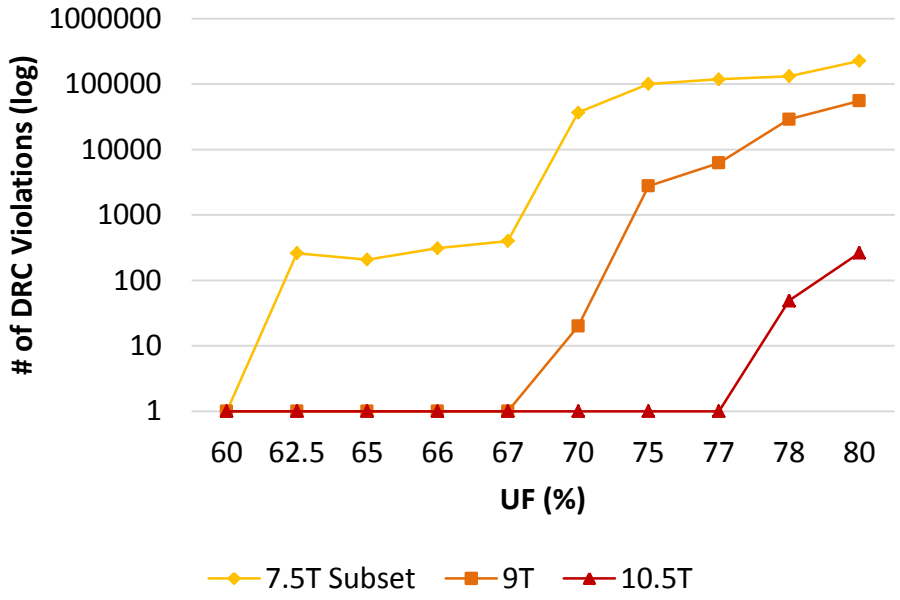



**Figure 4.12. Number of routing violations in the 7.5T, 9T, and 10.5T variants of AES-256.**

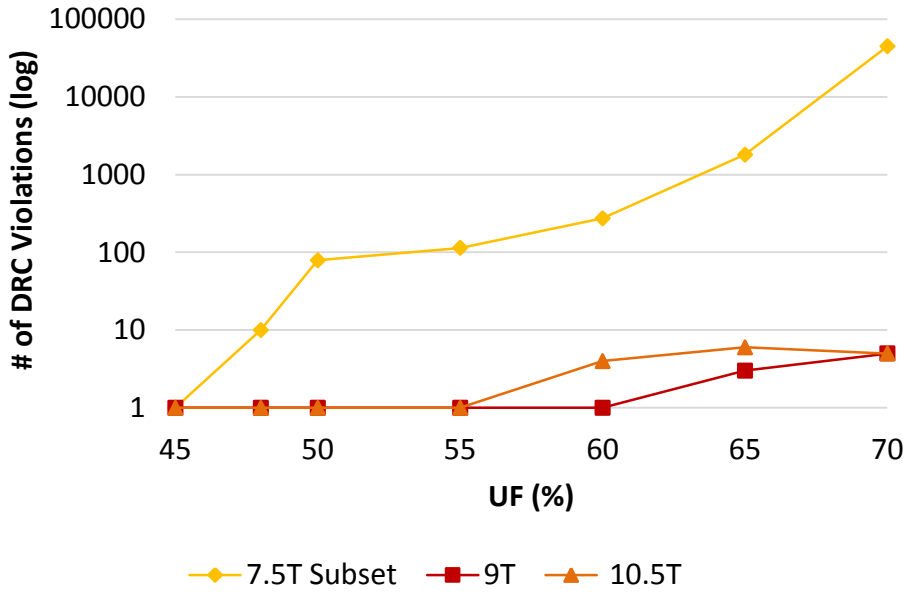



**Figure 4.13. Number of routing violations in the 7.5T, 9T, and 10.5T variants of OR1200.**

| Design | Metric | 7.5T Subset | 9T | 10.5T |
|---|---|---|---|---|
| AES-256 | Max UF (%) | 60 | 67 | 77 |
| | Area Overhead | - | 4.9% | 5.3% |
| OR1200 | Max UF (%) | 48 | 70 | 70 |
| | Area Overhead | - | -11% | 13% |

**Table 4.4 Maximum UF and area overheads in design blocks with reverse scaled libraries.**

## 4.8 Summary

In this chapter, we presented a methodology to design a logic IP with better pin access, as summarized in Figure 4.14. We designed constructs with simple layout features and re-used these layout constructs to design two libraries in a commercial 14/16 nm technology. Owing to their pin access qualities, taller cells were able to endure higher placement densities that rendered a minimal increase in floorplan sizes. For a RISC processor, 9T cells were even able to offer area reduction benefits on the first DRC clean routing pass. Our evaluations showed no signs of adverse timing impact of LDEs in our libraries; however, a set of design blocks implemented with construct-based libraries were on average 2% slower than their commercial library counterparts. This difference is attributable to the changes in the cell parasitic profile caused by BEOL, MOL, and FEOL layout choices that are less customized. Similarly, a test chip designed and fabricated in a commercial 14/16 nm technology showed that 10.5T cells can cause a 5% increase in power whereas 9T cells can almost match commercial 7.5T cells. In summary, we have shown that combining state-of-the-art transistors with minimum channel lengths with relaxed, regular BEOL features can enable high-performance ICs with minimal power and area impact while significantly lowering the effort needed for physical design closure.

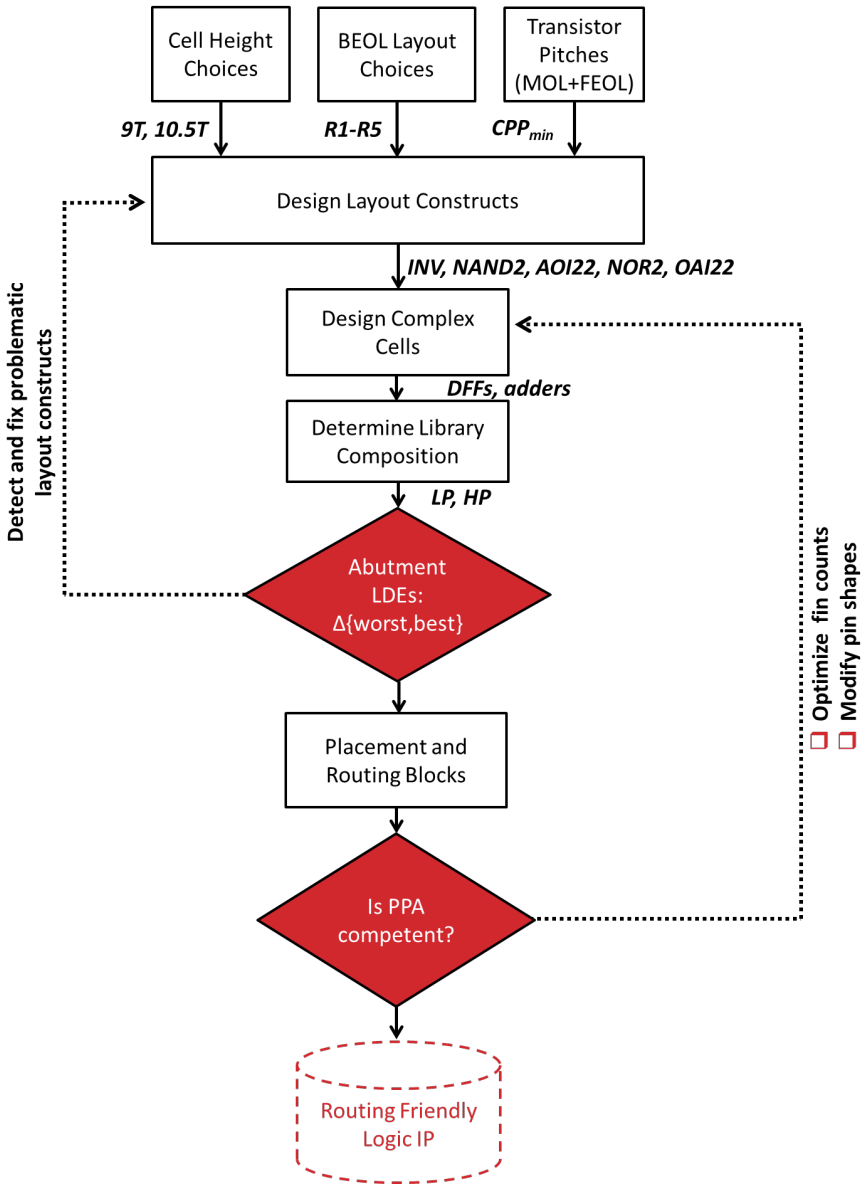



**Figure 4.14 Logic IP design methodology.**

# 5 Reducing the Likelihood of Re-Spins

One of the major impediments to a faster turnaround time in sub-20 nm nodes is the manufacturability concerns. Although the designer's responsibility and effort in ensuring the manufacturability of an IC has grown tremendously in advanced nodes, the risk of severe systematic variations and corresponding re-spins is still non-zero. Instead of addressing DFM issues during physical design, we take an alternative approach and leverage the layout regularity of the logic IPs designed in Chapter 4 for improved manufacturability. To this end, we use layout pattern counts as a proxy to evaluate the manufacturability of logic IPs, and provide fixes to potentially problematic cells as needed.

## 5.1 DFM in Sub-20 nm

With increasing demand to push CMOS scaling to its limits, the gap between the lithographic wavelength and the actual dimensions of the printed shapes has been growing. The manufacturing complexity needed for deeply scaled sub-wavelength CMOS nodes has changed the design scene dramatically. As FinFETs became the device of preference to continue the scaling roadmap, new challenges have emerged in the DFM landscape.

In FinFETs, FEOL is comprised of extremely regular layout polygons with uniform pitches and widths. Poly lines, fins, and MOL connectors are strictly unidirectional and cut layers define their line-ends. On the other hand, BEOL layers are relatively flexible for routing inside and among logic cells; multiple pitches and bidirectional routing are still allowed, but DR restrict them to a small set of options. Despite such limitations on the layout, DRC compliance is not sufficient to avoid defects and foundry imposed DFM checks are commonplace in physical design flows [47], [48] at sub-20 nm nodes.

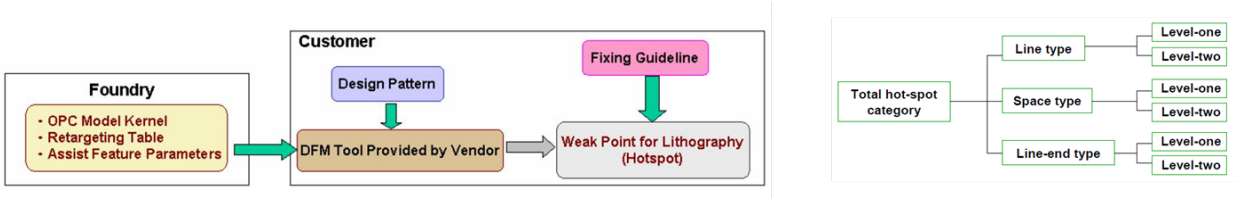



**Figure 5.1. DFM flow proposed by Chang et al.** [48]**.**

Figure 5.1 shows the DFM flow for deeply scaled CMOS nodes proposed by a foundry [48]. In this flow, DFM checks are performed after placement and routing is complete [49]. Based on the process-specific models, the DFM tool searches for lithographic weak points (hotspots) in the design [48]. To this end, the DFM tool decomposes layout polygons into small windows called layout patterns (LP) [50] as shown in Figure 5.2. Thereon, the DFM tool evaluates the likelihood of manufacturing risks of LPs by geometrically comparing them to known-to-be-problematic LPs. This process is called LP matching [50]. Either the DFM tool fixes the LPs that are found to be highly critical, or the designer is given guidelines to repair the problem. Although efficient, this DFM flow can incur additional NRE costs in more significant and more complex IC designs with a high number of distinct LPs. To speed up the LP matching, [51] proposes extracting only the LPs within the proximity of a "critical feature" that is usually not shared with low-volume customers of foundries. For single-layer LP analyses, line-endings are common critical features [48], [52] due to rounding, necking, and bridging problems. For multi-layer LPs, vias (and contacts) serve as critical features due to overlay (misalignment) problems [47]. Multi-layer LPs consist of metal-via-metal triplets.

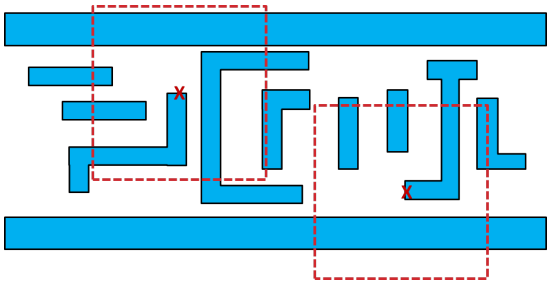


**Figure 5.2. Two layout pattern windows in a dense Metal-1 region.**

LP size, indeed, is an abstraction of non-idealities in lithography, etch, and overlay of which severity depend on the distance from a critical feature [47]. For 193 nm immersion lithography, shapes within a radius between 500 nm to 1 micron [25] tend to influence each other's printing due to wave interference. As a result, in the literature, LP dimensions can be as large as '2 microns by 2 microns' [47], [50], whereas '500 nm by 500 nm' is commonplace [45] for 193 nm immersion lithography. We note that in the 14/16 nm node, 500 nm is approximately equal to the width and the height of a NAND2 cell. Therefore, the printing quality of a cell strongly depends on the surrounding neighboring cells and wires, which is referred to as *'context'* [46]. Given the layout irregularities and the heights of commercial logic IPs discussed in Chapter 4, the number of contexts can be daunting. As an alternative, we propose leveraging the construct-based design approach introduced in Chapter 4 to minimize the number of distinct LPs. Thus, all contexts that could occur in a logic IP can be pre-characterized in silicon and validated against hotspots. To this end, we introduce an assisting tool that performs LP enumeration for all possible contexts of a given logic IP.

## 5.2 Virtual Library Characterization

In a design block, a cell that is not at the fringes of the floorplan is most likely to be surrounded by some other cells with different layouts. Hence, a cell's *context* can vary based on the number, type, position, and the orientation of the neighboring cells. In order to minimize, if not eliminate, the mainstream DFM flow in sub-20 nm CMOS nodes, we approach the LP concept differently. We use the number and types of LPs to characterize the manufacturability of a logic cell library virtually. The tool we use in this flow, LP enumerator (LPE) [54], is able to exhaust all possible cell placement combinations in a given logic cell library and count the number of unique and re-

occurring LPs around user-defined critical features. Such a tool would, in general, allow IP designers to choose between discarding or fixing the problematic cells that add too many LPs.

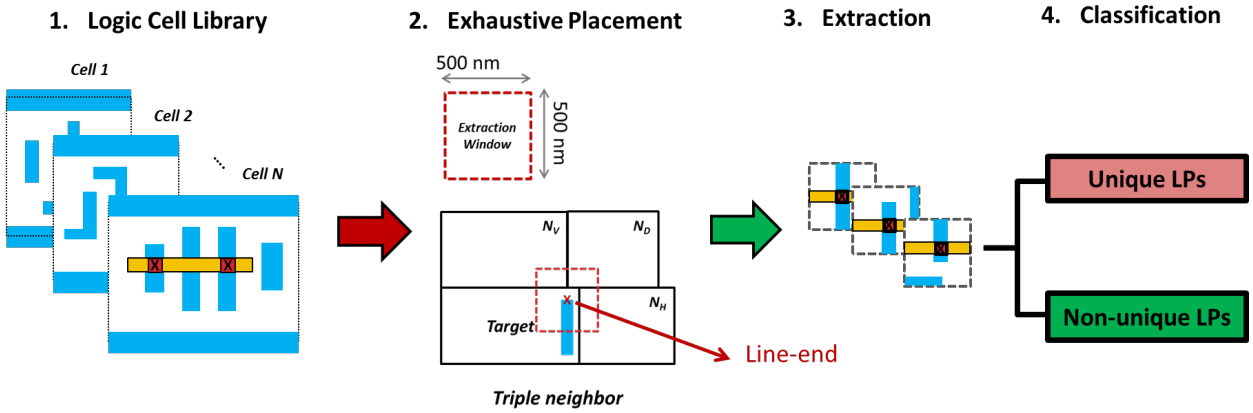



**Figure 5.3 VLC flow with user-defined critical features, target layers, and clip size.**

Figure 5.3 shows the virtual library characterization flow. The user provides a geometric definition (.LEF or .GDS) of the logic cell library, determines the window size for LP extraction, and chooses the target layer(s). In the single-layer enumeration mode, all line endings in the target cell are marked to become the center of LP extraction windows. In multi-layer enumeration mode, extraction windows are centered on vias or contacts. The crosses in Figure 5.4 mark the critical features for single and multi-layer enumeration modes, which are respectively line-endings (a) and via weight centers (b). The first analysis mode is intended for layers that are mainly used for routing inside the cells, while the second is a better fit for input pins and clock/control signal routing within cells.

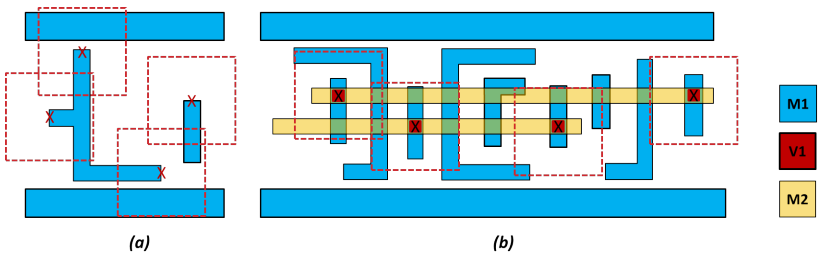



**Figure 5.4 (a) M1 line-endings marked for single-layer (b) V1 centers marked for multi-layer LP extractions.**

To account for all LPs that could occur in an IC, we exhaust every DR-compliant *context* for each cell for LPE. The realism and the computational load of the contextual analysis are controlled by the number of neighboring cells (NNC). Figure 5.5 depicts the type of neighbor contexts we considered for LPE. For *NNC=1*, the target cell is abutted by either a vertical or a horizontal neighbor. For *NNC=2*, the vertical and horizontal neighbors are concurrently present. Lastly, for *NNC=3*, a diagonal neighbor is considered as well. For each context, we take all of the valid neighbor cell orientations and positions in the library into account. The LP extraction engine shifts the neighbor cells in steps that are integer multiple of the CPP until they exit the LP extraction window that is marked by the red dashes. Optionally, the analysis can be performed without any neighbors (*NNC=0*) to reveal the inherent layout regularity attributes of cells.

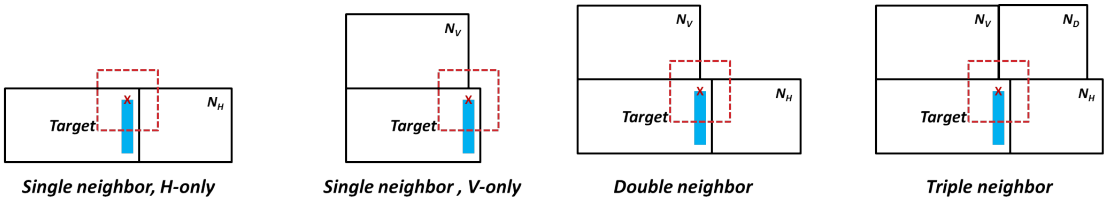



**Figure 5.5 Neighbor context options available for LPE.**

Once the extraction is over, the tool begins enumerating the number of different LPs and creates databases for each cell. A geometric XOR [50], [52] operation determines if two LPs are identical or not. If an LP is geometrically unique, it only occurs in one specific context. If an LP re-occurs in different contexts, it is non-unique. This classification enables an evaluation at cell basis by revealing which polygons are likely to create new, distinct patterns. Although lowering the total LP count is the goal of our logic IP optimization approach, there are different DFM implications of unique and non-unique LPs. A library that is predominantly comprised of non-unique LPs is preferable since repeating patterns enable better design technology co-optimization (DTCO) [55] and cheaper DFM validation [52]. On the contrary, unique LPs are risky since they

occur in rare contexts. Therefore, DFM checks may not even recognize their criticality, and they can cause a manufacturing problem that was not faced before, thereby requiring a silicon re-spin.

Some critical features can be more likely to form unique patterns, depending on their position within the cell. For example, the extraction window for an M1 line ending is very likely to contain polygons from neighboring cells, if the line is too close to the cell boundary. We refer to such windows as *external,* since the LPs extracted from the corresponding window would vary based on the neighboring cells. In contrast, critical features that are well inside a cell are most likely to create an LP extraction window that is inherent to the cell and agnostic to the neighboring cells. We refer to these as *internal windows*. Figure 5.6 shows the difference between internal and external windows. Regardless of the neighboring cells, the window marked in yellow will only produce the same LP, whereas the red one will cover different polygons when abutted by different cells. In our logic IP optimization approach, construct-based logic IP layouts are inherently more regular, creating more non-unique LPs than unique ones.

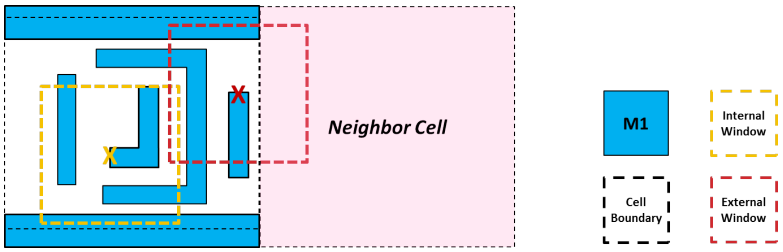



**Figure 5.6 Internal and external LP extraction windows.**

Ultimately, virtual library characterization serves to detect problematic cells early in the IP development for revision. Figure 5.7 shows how we incorporated virtual library characterization into our logic IP design flow for DFM optimization. When analyzed at a cell-by-cell basis, single and multi-layer LPs of a library can reveal the source problematic layout polygons, which may be stemming from cell architecture and primary layout choices. Depending on the severity of the number of LPs, problematic cells can be repaired or discarded from the library.

Figure 5.8 shows the impact of an iteration of layout repair on the number of unique LPs. After repairing the "irregularities" in the layouts of ADDF and DFFQ, the number of unique LPs has decreased by 20% for the whole library. We emphasize that after repair, the number of unique LPs became zero for DFFNQ (the falling edge variant of the DFFQ), since the M1 masks of these flip-flops became identical. This observation shows the interrelation between the LP databases of cells due to the context-dependent nature of the analysis. Therefore, repairing only a few problematic cell layouts can help substantially reduce the size of the LP network.

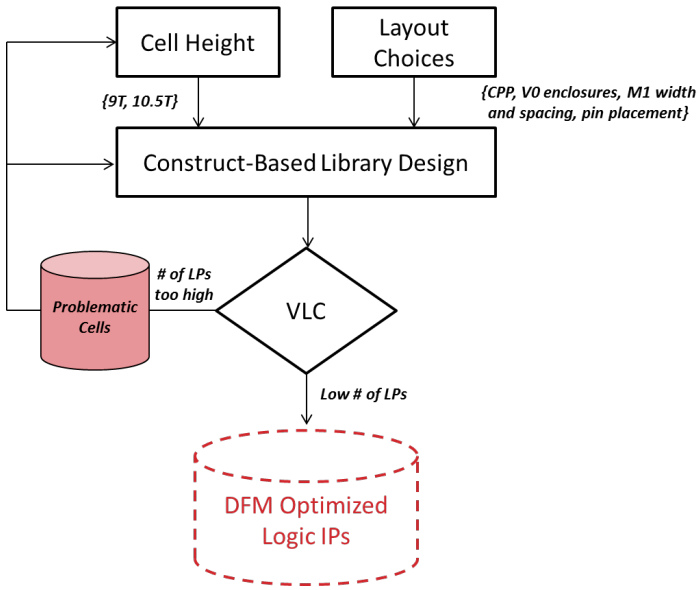



**Figure 5.7 LPE assisted IP design flow.**

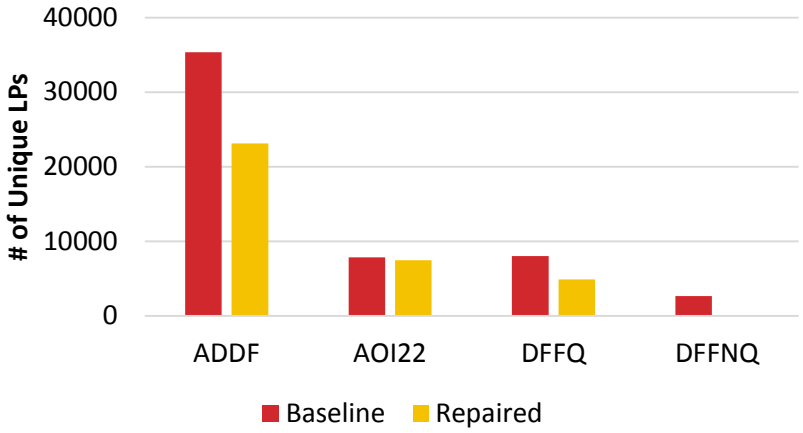



**Figure 5.8 LPE-assisted cell layout repair iterations for M1 polygons.**

In a commercial logic IP with highly customized cell layouts, selecting the right candidates for repair may not be straightforward. In contrast, for the minimization of unique LPs, the proposed construct-based layout design approach presented in Chapter 4 has the upper hand; alterations to one construct can improve the layout regularity for the rest of the library. To demonstrate the library level impact of the proposed design approach, we performed virtual library characterization on our 9T and 10.5T cells and their commercial 7.5T equivalent.

## 5.3 Results

As we demonstrated in Chapter 4, RS of cell heights allows for layout simplification and a more homogenous distribution of polygons. Additionally, construct-based design approach enables layout similarity among the cells in a library. To assess the impact of NNC on the number of LPs, we performed M1 LPE for target libraries 7.5T Subset, 9T, 10.5T. The LP extraction window was sized 500 nm by 500 nm for this experiment. Figure 5.9 shows that all libraries have significantly fewer LPs for NNC=0, wherein LPE extracts internal windows only. When the NNC increases, the impact of layout regularity on LPs becomes more evident. The number of LPs increases more rapidly for the 7.5T Subset compared to the other libraries. At NNC=3, 7.5T Subset has almost 10 and 30 times more patterns than 9T and 10.5T. Note that the analysis covers less than one-tenth of the 7.5T library. If we were to include all cells in the analysis, the number of LPs would have been exponentially higher.

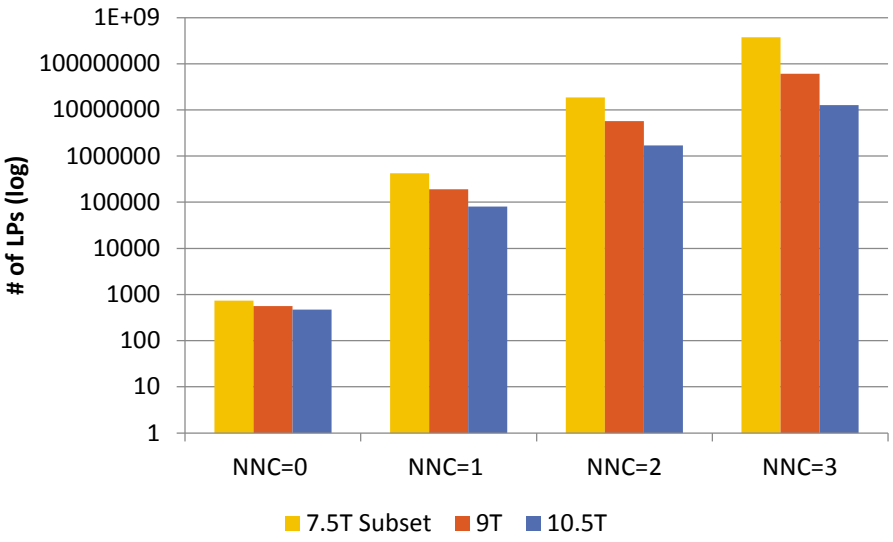



**Figure 5.9 Number of M1 LPs versus NNC.**

The ratio of unique LPs to total LPs reveals an intriguing trend. This metric has significance because it is a projection of layout re-use in the libraries. Figure 5.10 shows the percentage of unique LPs approaches zero for 9T and 10.5T libraries with an increasing number of neighbors. From the library standpoint, this means that cells have identical or similar M1 layout polygons inside them. From the DFM standpoint, only a tiny portion of the M1 LPs on an IC will be geometrically distinct than others, whereas the rest will be repeating. Compared to the construct-based libraries, the 7.5T Subset shows a concerning trend. At NNC=3, more than one-third of the patterns are still unique. This result suggests the commercial 7.5T library would require significantly more effort to verify the manufacturability of all M1 patterns that can occur in an IC.

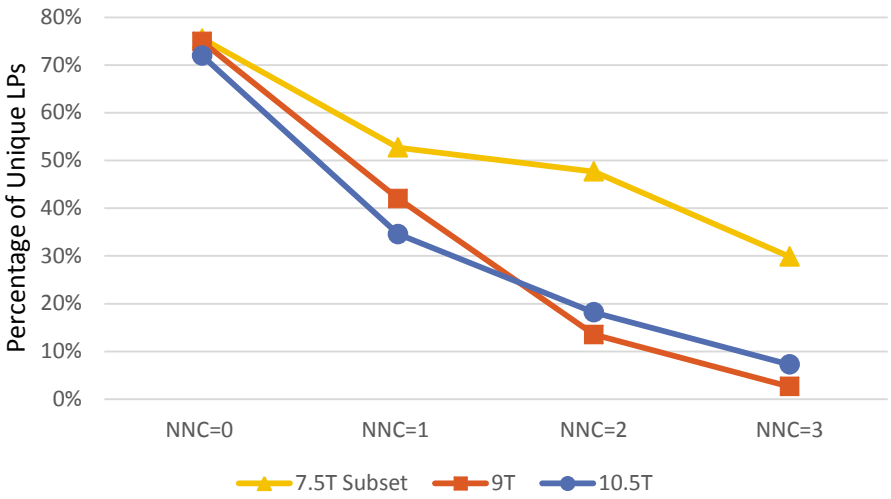



**Figure 5.10 The percentage of unique LPs decrease with increasing NNC.**

M1 regularity achieved through the layout decisions presented in Figure 4.2 propagates to MOL layers. To quantify this effect, we conducted a set of multi-layer LPE analyses. Due to the computational complexity of multi-layer LP extraction, we created a 10-cell subset of all libraries that contain a mix of complex cells (DFF, ADD, XOR2) and simpler cells (NAND, OR, AOI22). Table 5.1 Number of multi-layer LPs compared to NNC=2shows the number of LPs for M0-VIA0-M1 and M1-VIA1-M2 triplets at NNC=2. Here, the first triplet with M0 corresponds to input pin polygons, while the latter corresponds to internal signal routing within complex cells. The results suggest that regularity of input pins in 9T and 10.5T libraries causes a significant reduction in the M0-VIA0-M1 count. On the other hand, the reduction of M1-VIA1-M2 is much smaller because M2 shapes are single width and unidirectional by default.

Ultimately, the objective of extreme layout simplicity and regularity is to be able to design ICs that are inherently free of manufacturability risks so that DFM checks are not needed. This objective requires validation of all kinds of abutment LPs for the logic cells. Such validation, however, is not feasible for the commercial logic IP used herein since one-tenth of it has almost 400 million M1 LPs alone. When mapped to the corresponding cell abutments, the raw area of

M1 LPs in the 7.5T commercial logic IP is approximately 700 $mm^2$, whereas this number is around 30 $mm^2$ for the 9T and 10.5T logic IPs.

| Layer Triplets | **7.5T Subset** | **9T** | **10.5T** |
|---|---|---|---|
| **# of M0-VIA0-M1 LPs** | 6.1M | 2.6M | 2.5M |
| **# of M1-VIA1-M2 LPs** | 6M | 5.7M | 5.2M |

**Table 5.1 Number of multi-layer LPs compared to NNC=2.**

## 5.4 Summary

LPE analyses presented in this chapter demonstrated that reverse scaling of logic cell heights and construct-based design could simplify logic cell layouts and promote re-use. The resultant reduction in the BEOL LPs for the logic cells is multiple orders of magnitude, and it enables the foundations of an alternative DFM flow wherein the manufacturability of all possible cell abutments the IP can be validated. Ultimately, the proposed DFM optimization for logic cells alleviates the existing burden on the designers, hence reduces the design cost. The ability to design a high-performance IC without the risk of re-spins can make the advanced CMOS nodes affordable to low-volume customers.

# 6 Conclusion and Future Work

Manufacturing complexity of advanced CMOS nodes has inflated IC design costs significantly. Since existing design techniques fail to address design closure and manufacturability issues in a cost-effective manner, there is an immediate need for a solution that decreases design closure time and reduces manufacturing risks. Hereby, we created the framework summarized in Figure 6.1 and used this framework to design multiple test chips in a commercial 14/16 nm FinFET process. Ultimately, the framework enables designers to explore and find the most optimal metal stack and logic IP options based on the volume and performance specifications for an IC. Specifically, this work has converged on the following recommendations for low-volume, high performance digital ICs:

- Use state-of-the-art transistors in FEOL
- Use MOL (M0) and immediate BEOL (M1 and M2) layers at technology-minimum pitches
- Relax the BEOL metal stack for layers M3 to M7
- Use higher level metals (M8 and above) as in a state-of-the-art process
- Use slightly taller logic cells for fast routing convergence

As CMOS scaling continues to the end of its roadmap, we believe the solutions that mix the performance benefits of a high-density FEOL and a design-closure friendly BEOL will become more relevant for the future of low-cost IC design. For exploring these solutions more efficiently, our conclusions open up to a set of future research directions:

- Exploring different rates of wiring pitch relaxation for different metals in the BEOL stack can further increase the signal integrity, power, and timing benefits of r-BEOL. Given

that there can be up to 12 metals in a process stack, there is a significant design space to be explored.

- Automating the layout design for construct-based logic cells can enable faster IP development. In particular, with increasing design rule complexity in 10, 7, and 5 nm process nodes, such automation can help co-optimize layout regularity and pin access more efficiently.
- The virtual library characterization flow can be modified to account for future manufacturing issues in more advanced CMOS nodes. If the logic cell downsizing continues with 193 nm immersion lithography, the number of adjacent cells that can impact the printing qualities of a cell will increase, thereby creating a more complex layout pattern extraction and enumeration problem. If EUV becomes more prevalent for FEOL and MOL, the printing non-idealities need to be remodeled.

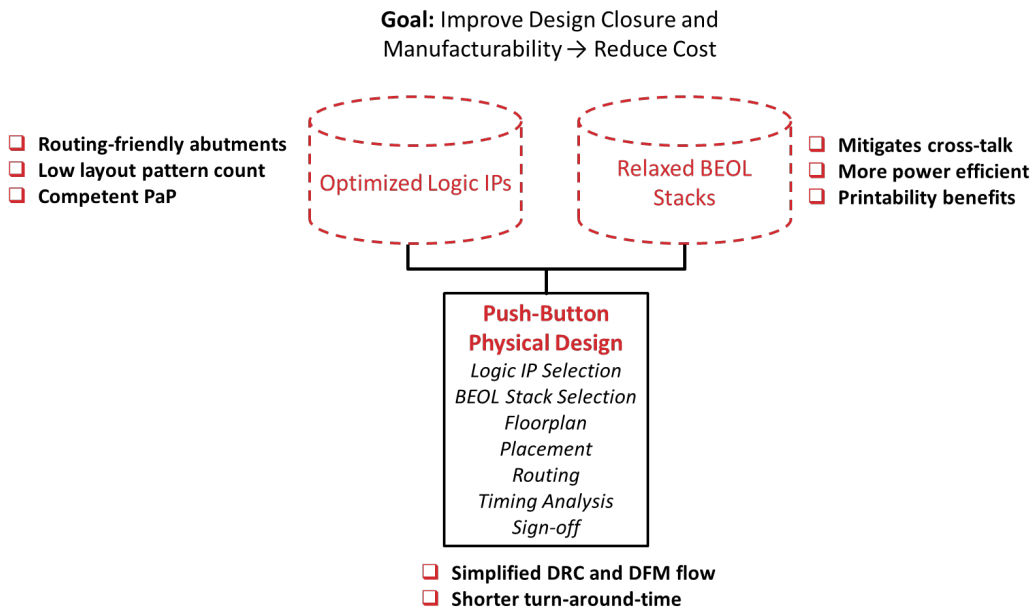



**Figure 6.1 Holistic design flow for high-performance, low-volume ICs**

## References


[1] G. E. Moore, “Progress in digital integrated electronics [Technical literaiture, Copyright 1975 IEEE. Reprinted, with permission. Technical Digest. International Electron Devices Meeting, IEEE, 1975, pp. 11-13.],” *IEEE Solid-State Circuits Soc. Newsl.*, vol. 11, no. 3, pp. 36–37, Sep. 2006.

[2] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, “Design of ion-implanted MOSFET’s with very small physical dimensions,” *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974.

[3] A. B. Kahng, “The ITRS design technology and system drivers roadmap,” in *Proceedings of the 50th Annual Design Automation Conference on - DAC ’13*, 2013, p. 1.

[4] International Technology Roadmap for Semiconductors (ITRS), “More Moore,” *ITRS*, pp. 1–52, 2015.

[5] H. Stork, “Economies of CMOS Scaling,” 2005. [Online]. Available: https://www.nist.gov/sites/default/files/documents/pml/div683/conference/Stork_2005.pdf . [Accessed: 07-Aug-2018].

[6] Semico Research Corporation, “SoC Silicon and Software Design Cost Analysis: Costs for Higher Complexity Continue to Rise,” 2013. [Online]. Available: http://www.semico.com/content/soc-silicon-and-software-design-cost-analysis-costs-higher-complexity-continue-rise. [Accessed: 05-Oct-2018].

[7] L. Liebmann, L. Pileggi, J. Hibbeler, V. Rovner, T. Jhaveri, and G. Northrop, “Simplify to survive: prescriptive layouts ensure profitable scaling to 32nm and beyond,” in *SPIE Advanced Lithography Volume 7275*, 2009, p. 72750A.

[8] L. W. Liebmann, L. Pileggi, and K. Vaidyanathan, *Design Technology Co-Optimization in the Era of Sub-Resolution IC Scaling*. SPIE, 2016.

[9] V. V. Rovner, T. Jhaveri, D. Morris, A. Strojwas, and L. Pileggi, “Performance and manufacturability trade-offs of pattern minimization for sub-22nm technology nodes,” in *SPIE Advanced Lithography*, 2011, vol. 7974, p. 79740I.

[10] D. Kochpatcharin, “OIP Era,” 2013. [Online]. Available: https://www.youtube.com/watch?v=OXy_jCtq1ZI. [Accessed: 01-Sep-2018].

[11] A. B. Kahng, “New Game , New Goal Posts : A Recent History of Timing Closure Invited,” *Proc. 52nd Annu. Des. Autom. Conf. - DAC ’15*, no. 2, 2015.

[12] M. Hsu, N. Katta, H. Y. Lin, K. T. Lin, K. H. Tam, and K. C. Wang, “Design and Manufacturing Process Co-optimization in,” in *ICCAD*, 2014, pp. 574–581.

[13] W. Ye, B. Yu, D. Z. O. Pan, Y.-C. Ban, and L. Liebmann, “Standard Cell Layout Regularity and Pin Access Optimization Considering Middle-of-Line,” in *Proceedings of the 25th edition on Great Lakes Symposium on VLSI - GLSVLSI ’15*, 2015, pp. 289–294.

[14] T. Taghavi, Z. Li, C. Alpert, G.-J. Nam, A. Huber, and S. Ramji, “New placement prediction and mitigation techniques for local routing congestion,” in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2010, pp. 621–624.

[15] K. Vaidyanathan, “Exploiting Challenges of Sub-20 nm CMOS for Affordable Technology Scaling,” Carnegie Mellon University, 2015.

[16] C. M. Weber, C. N. Berglund, and P. Gabella, “Mask cost and profitability in photomask manufacturing: An empirical analysis,” *IEEE Trans. Semicond. Manuf.*, vol. 19, no. 4, pp. 465–474, 2006.

[17] B. Kasprowicz, “Euv Mask Technology and Economics: Impact of Mask Costs on

Patterning Strategy," *EUVL Work.*, p. P33, 2017.
[18] J. Hruska, "Nvidia deeply unhappy with TSMC, claims 20nm essentially worthless," 2012. [Online]. Available: http://www.extremetech.com/computing/123529-nvidia-deeply-unhappy-with-tsmc-claims-22nm-essentially-worthless. [Accessed: 07-Feb-2018].
[19] S. Wang, "2014 Quarterly Report on Semiconductors," 2014. [Online]. Available: https://www.euvlitho.com/2017/P33.pdf.
[20] J. (IBS) Handel, "Factors for Success in System IC Business and Impact on Business Model." 2012.
[21] H. O. Ron, K. W. Mai, and A. Horowitz, Mark, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, 2001.
[22] K. L. Shepard, V. Narayanan, and R. Rose, "Harmony: static noise analysis of deep submicron digital integrated circuits," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 18, no. 8, pp. 1132–1150, 1999.
[23] P. Chen, D. A. Kirkpatrick, and K. Keutzer, "Miller factor for gate-level coupling delay calculation," *IEEE/ACM Int. Conf. Comput. Des. Dig. Tech. Pap. ICCAD*, vol. 2000–Janua, pp. 68–74, 2000.
[24] J. D. Z. Ma and L. He, "Formulae and applications of interconnect estimation considering shield insertion and net ordering," *IEEE/ACM Int. Conf. Comput. Des. Dig. Tech. Pap.*, pp. 327–332, 2001.
[25] H. Kaul, D. Sylvester, and D. Blaauw, "Active shields: a new approach to shielding global wires," *Proc. 12th ACM Gt. Lakes Symp. VLSI*, pp. 112–117, 2002.
[26] R. Arunachalam, E. Acar, and S. R. Nassif, "Optimal shielding/spacing metrics for low power design," *Proc. IEEE Comput. Soc. Annu. Symp. VLSI, ISVLSI*, vol. 2003–Janua, pp. 167–172, 2003.
[27] M. Palusinski, A. Strojwas, and W. Maly, "Regularity in Physical Design," in *GSRC Workshop*, 2001, pp. 17–18.
[28] D. Morris, V. Rovner, L. Pileggi, A. Strojwas, and K. Vaidyanathan, "Enabling application-specific integrated circuits on limited pattern constructs," in *2010 Symposium on VLSI Technology*, 2010, pp. 139–140.
[29] D. Morris, K. Vaidyanathan, N. Lafferty, K. Lai, L. Liebmann, and L. Pileggi, "Design of Embedded Memory and Logic Based On Pattern Constructs," in *Symposium on VLSI Technology*, 2011, vol. 1, pp. 104–105.
[30] K. Vaidyanathan, L. Liebmann, A. Strojwas, and L. Pileggi, "Sub-20 nm design technology co-optimization for standard cell logic," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2014, vol. 2015–Janua, no. January, pp. 124–131.
[31] K. Cao, J. Hu, and M. Cheng, "Wire sizing and spacing for lithographic printability and timing optimization," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 15, no. 12, pp. 1323–1340, 2007.
[32] M. Bartley, "Lies , Damned Lies and Hardware Verification," *Snug 2008*, 2008. [Online]. Available: https://www.testandverification.com/wp-content/uploads/2012/05/Mike_Bartley_SNUG_europe_2008.pdf.
[33] H. Foster, "Smaller designs face greater risk of respins," 2015. [Online]. Available: http://www.techdesignforums.com/practice/technique/wilson-research-2014-functional-verification-study/.
[34] H. Melzner, "Smaller is better? Maximization of good chips per wafer by co-optimization

of yield and chip area," *ASMC (Advanced Semicond. Manuf. Conf. Proc.*, vol. 2006, pp. 372–379, 2006.
[35] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron II," *Proc. 1999 Int. Symp. Phys. Des. - ISPD '99*, pp. 193–200, 1999.
[36] J. H. C. Chen, N. LiCausi, E. T. Ryan, T. E. Standaert, and G. Bonilla, "Interconnect performance and scaling strategy at the 5 nm Node," *2016 IEEE Int. Interconnect Technol. Conf. / Adv. Met. Conf. IITC/AMC 2016*, vol. 7139, no. 2, pp. 12–14, 2016.
[37] H. B. Bakoglu and J. D. Meindl, "Optimal Interconnection Circuits for VLSI," *IEEE Trans. Electron Devices*, vol. 32, no. 5, pp. 903–909, 1985.
[38] M. M. Isgenc, S. Pagliarini, R. Liu, and L. Pileggi, "Evaluating the benefits of relaxed BEOL pitch for deeply scaled ICs," in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, 2017, pp. 180–185.
[39] C. Albrecht, "IWLS 2005 Benchmarks." 2005.
[40] S. N. Pagliarini, M. M. Isgenc, M. G. A. Martins, and L. Pileggi, "Application and Product-Volume-Specific Customization of BEOL Metal Pitch," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 26, no. 9, pp. 1627–1636, Sep. 2018.
[41] P. Debacker *et al.*, "Low track height standard-cells enable high-placement density and low-BEOL cost (Conference Presentation)," in *Design-Process-Technology Co-optimization for Manufacturability XI*, 2017, p. 2.
[42] X. Xu, N. Shah, A. Evans, S. Sinha, B. Cline, and G. Yeric, "Standard cell library design and optimization methodology for ASAP7 PDK: (Invited paper)," in *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2017, vol. 2017–Novem, pp. 999–1004.
[43] X. Xu, B. Cline, G. Yeric, B. Yu, and D. Z. Pan, "Self-aligned double patterning aware pin access and standard cell layout co-optimization," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 34, no. 5, pp. 699–712, 2015.
[44] D. Lampret, "OR1200." [Online]. Available: https://openrisc.io/or1k.html.
[45] C. Santifort, "ARM Core," 2011. [Online]. Available: https://opencores.org/projects/amber.
[46] P. Hurat *et al.*, "Timing variability analysis for layout-dependent-effects in 28nm custom and standard cell-based designs," in *SPIE Advanced Lithography*, 2011, vol. 7974, no. April 2011, p. 797412.
[47] R. O. Topaloglu, "Design with FinFETs: Design rules, patterns, and variability," *IEEE/ACM Int. Conf. Comput. Des. Dig. Tech. Pap. ICCAD*, pp. 569–571, 2013.
[48] C. C. Chang *et al.*, "Layout patterning check for DFM," *Proc. SPIE*, vol. 6925, no. March 2008, p. 69251R–69251R–7, 2008.
[49] K. Krishnamoorthy, "In-Design and Signoff Pattern Detection and Fixing Flows for Accelerated DFM Convergence," 2016. [Online]. Available: https://www.globalfoundries.com/sites/default/files/articles/in-design-and-signoff-pattern-detection-and-fixing-flows-for-accelerated-dfm-convergence.pdf. [Accessed: 05-Oct-2018].
[50] R. O. Topaloglu, "ICCAD-2016 CAD Contest in Pattern Classification for Integrated Circuit Design Space Analysis and Benchmark Suite," 2016.
[51] D. Ding, X. Wu, J. Ghosh, and D. Z. Pan, "Machine learning based lithographic hotspot detection with critical-feature extraction and classification," *2009 IEEE Int. Conf. Integr. Circuit Des. Technol. ICICDT 2009*, pp. 219–222, 2009.

[52] S. Pagliarini, M. Martins, and L. Pileggi, “Virtual characterization for exhaustive DFM evaluation of logic cell libraries,” in *2017 18th International Symposium on Quality Electronic Design (ISQED)*, 2017, no. Vcv, pp. 93–98.
[53] M. Orshansky, S. R. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design*. Boston, MA: Springer US, 2008.
[54] M. G. A. Martins, S. Pagliarini, M. Isgenc, and L. Pileggi, “From Virtual Characterization to Test-Chips: DFM Analysis through Pattern Enumeration,” *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, pp. 1–1, 2018.
[55] T. Jhaveri, V. Rovner, L. Liebmann, L. Pileggi, A. J. Strojwas, and J. D. Hibbeler, “Co-optimization of circuits, layout and lithography for predictive technology scaling beyond gratings,” *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 29, no. 4, pp. 509–527, 2010.