Politecnico di Torino

Politecnico di Torino

MSc in Electronic Engineering

Area and performance evaluation in the physical design of advanced integrated circuits

Candidate: Gabriele Adragna

Supervisors: Prof. Guido Masera
Eng. Afonso Moreira

April 2025

Abstract

The continuous scaling of technology nodes in modern VLSI design introduces significant challenges in managing increasing complexity and design density. Optimizing area, power, and performance (PPA) is crucial to meeting design requirements for reliability, functionality, and production costs. This thesis evaluates PPA metrics, with a focus on area shrinkage possibilities, specifically for sub-5nm technology. While all stages of the physical design flow are considered, particular emphasis is placed on power distribution network (PDN) optimization and its critical impact on the final PPA results. The research also includes a comparison with an older technology node, highlighting the evolving challenges and strategies required for advanced scaling. The results, derived from experimental evaluations using state-of-the-art EDA tools and benchmarks, provide insights into effective strategies for achieving PPA goals and offer guidelines for PDN design in increasingly scaled technology nodes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Context

The rapid advancement of semiconductor technologies, particularly at sub-5nm nodes, poses significant challenges to the physical design of integrated circuits (ICs). The achievement of optimal trade-offs between area, performance, and power has become critical in commercial applications, particularly in industries such as mobile communications and complex SoC manufacturing. As a leader in these fields, Qualcomm has been at the forefront of developing cutting-edge solutions to address these challenges. This thesis, developed in collaboration with Qualcomm, focuses on evaluating and improving the area and performance in advanced physical design flows, with particular attention to the constraints and demands of the latest technology nodes.

1.2 Physical design flow overview

[1] Physical design or back-end design is the step in standard design cycles at which the gate-level representation of the IC logic is converted into real geometrical shapes that, provided in a standard format [3] to the manufacturer, can be used in lithography to produce an actual IC. At this stage, the physical parameters of the technology node used are the main driving factor in the design choices and final results. The standard physical design flow can be summarized in six steps:

- **Partitioning**:

  Partitioning step requires to divide the overall design in smaller sub-blocks, this is done in order to facilitate the following steps and reduce the design complexity, it also helps to isolate potential sub-designs that requires particular attention and treatment. In modern advanced designs several hierarchical level (>3) can be needed in order to have sub-blocks with manageable design complexity.

  One of the key aspect of partitioning is the minimization of the number of connections between different design partitions, a wrongly designed partitioning could lead to performance degradation related to inefficient use of routing resources and interconnections, example in figure 1.1.

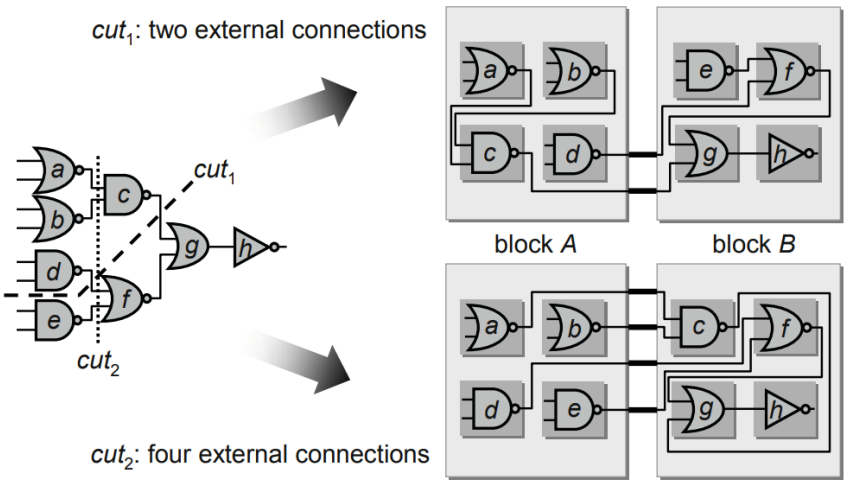

Figure 1.1: Example of two different partitioning strategies [1]

- **Floorplanning**:

  The floorplanning stage determines the shape and dimensions of the module. The aim of the designer is to choose the block area and aspect ratio such that the target metrics regarding performance, power, and area are met. In modern VLSI projects, an area budget is usually split between all the blocks based on their respective needs and empirical data from previous designs. Shrinking the design is commonly an objective when a new technology node is used and RTL complexity has not been increased. Since floorplanning directly affects further stages, it is often iterated with feedback to achieve the desired optimization.

  Several sub-steps are performed at floorplanning , these are needed as a preliminary phase of placement:

  - **Pin assignment**

    During pin assignment, the net terminals of the I/O nets are assigned to individual pins that connect the block to the outer world. The goal at this step is to optimize wire length to limit parasitics and save routing resources. In the context of large designs with high partitioning, the location of pins is strongly driven by the position of the block with respect to adjacent ones. Several figures of merit can be exploited to optimize the design, such as pin size, spacing, and metal layers. Usually, clock pins differ in these parameters from signal pins due to more stringent signal integrity and timing needs.

  - **Macro placement**: Often, designs contain sub-blocks (hard macros), usually memories or IPs, that cannot be modified in shape, being hard-defined in the library files. These macros have to be placed before standard cells since power distribution and routing resources are heavily affected by their presence. Typically, macro placement is carried out by a designer using heuristics. The macro placement problem is strongly multi-factorial and differs in approach from block to block. Automatic placers are available in industry-standard EDA tools, but the results are often not good enough for performance-driven designs. In recent years,

machine learning placers have become increasingly present in the field, highlighting the growing need for automation[5].

– **Power planning**: In general, the power planning phase involves designing a robust power distribution network (PDN) to ensure efficient power delivery across the chip. This includes creating power pads, which act as entry points for power from the external package. Power rings are then designed to surround the core area, distributing power from the pads to the power stripes. These stripes run across the core, connecting to the power rings and distributing power to the power rails trough stacked Vias, which deliver power directly to the standard cells. Additionally, managing IR drop is crucial to ensure minimal voltage drop across the power network, maintaining the chip's performance. Electromigration (EM) management is also important to handle high current densities without causing degradation A schematic view of the PDN is presented in figure 1.2.

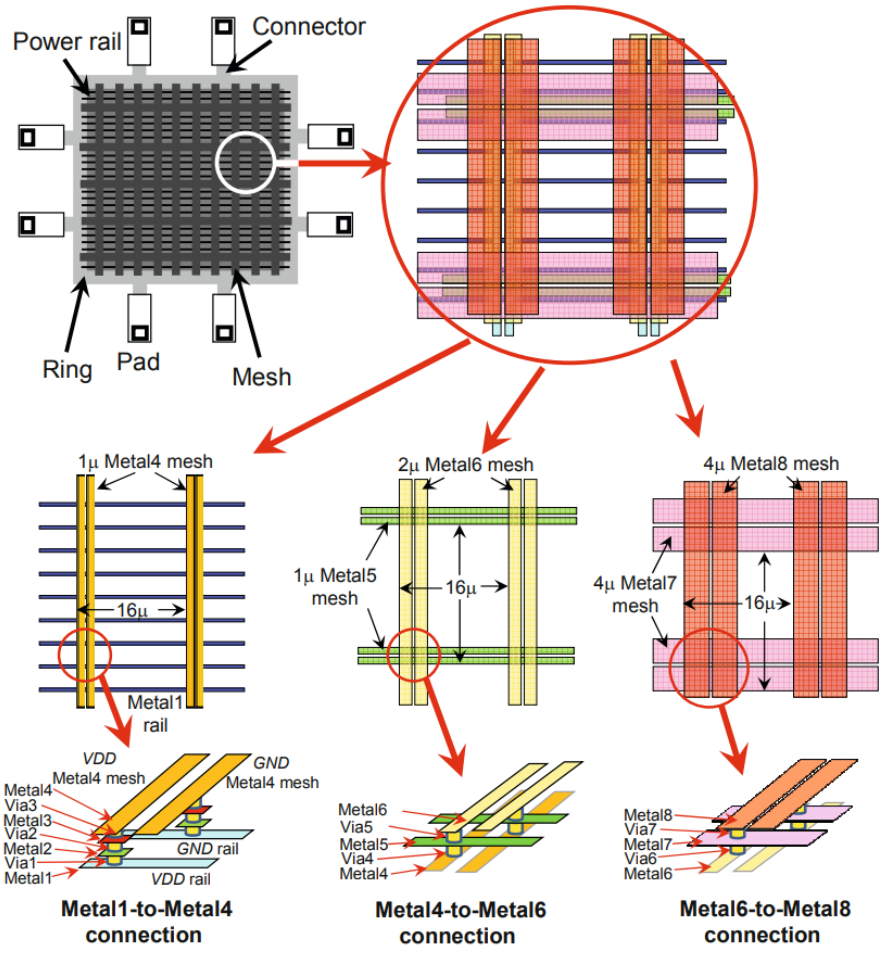

Figure 1.2: Example of power distribution network [1]

The design of the local mesh has a strong impact on the successive phases because it significantly affects the routing resources. Recent designs with high transistor counts and densities can use up to 40% of the routing resources for the power grid. Additional degrees of freedom can be ex-

ploited by using via stubs (or "staples") on M1 instead of full stripes. This usually leads to improved routability since more resources are kept free for signal routing. However, this approach may result in worse IR drop figures because the overall effective resistance of the grid increases. Another important parameter is the mesh density. A denser mesh will provide better performance figures (less IR drop) at the cost of higher usage of resources. In general, the right compromise has to be found. More relaxed power grids lead to better routability and hence more possibilities for area shrinkage. However, this can worsen timing because IR drop affects the timing integrity.

Modern chips have a very hierarchical and partitioned approach, this reflects also on the power delivery network that can adopt more complex schemes and routing strategies. When several voltage domains are present additional cells like power-switches and Global Distributed Head Switches (GDHS) are needed, tap cells are used to prevent latch-up and various types of ESD protection cells can be adopted.

- **Placement**

  The stage of placement in the physical design flow is a critical step that directly influences the performance, power consumption, and area of the final integrated circuit (IC). This stage involves determining the precise locations of standard cells and other elements within the chip, ensuring that they are optimally positioned to meet design constraints and performance targets.

  Standard cells are the fundamental building blocks of digital ICs, consisting of logic gates, flip-flops, and other basic components. During the placement process, these cells are positioned in predefined regions called "rows." Rows are horizontal strips that span the width of the chip and are separated by channels reserved for routing interconnections.

  Each row is designed to accommodate standard cells of a specific height, ensuring uniformity and alignment. Cells within a row are placed side by side, with their power and ground rails aligned to facilitate efficient power distribution. This arrangement helps in minimizing the wire length and reducing signal delay, as the cells are placed in close proximity to each other. A schematic example is provided in 1.3.

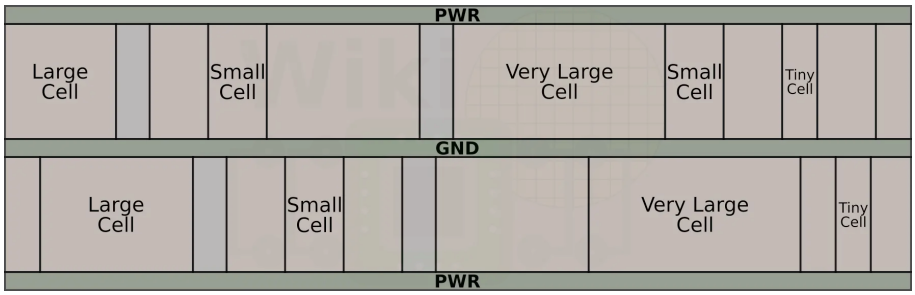

Figure 1.3: Standard cells rows [2]

Placement begins with the global placement phase, where standard cells are roughly positioned to provide an initial solution that meets the overall design

constraints. Techniques such as force-directed placement, simulated annealing, and partitioning-based methods are commonly used. The goal is to distribute the cells evenly across the chip, minimizing congestion and ensuring that the design is scalable. This initial placement sets the stage for more detailed adjustments.

Following global placement, the detailed placement phase fine-tunes the positions of the standard cells to further optimize the design. This step involves adjusting the placement to reduce wire length and improve timing. Algorithms such as branch-and-bound and dynamic programming are often employed to achieve high-quality results. Detailed placement ensures that the cells are positioned with precision, taking into account the intricate requirements of the design.

Several constraints must be considered during the placement process to ensure the design meets all requirements. Design rules, provided by the manufacturing process, dictate the minimum spacing between cells, the width of wires, and other critical parameters. Timing constraints are also crucial, as placement must be performed in a way that minimizes signal delay and meets the setup and hold time constraints of the design. Efficient power distribution is essential to prevent hotspots and ensure reliable operation, and the placement process must consider the power grid design and the distribution of power-hungry cells, clusterizing cells with high switching activity could lead to high IR drop figures compromising timing.

The placement process faces several challenges, particularly with the increasing complexity of modern IC designs. Handling large-scale designs with millions of standard cells requires efficient algorithms and powerful computational resources. Additionally, managing the trade-offs between area, performance, and power is a constant challenge. Process variations and manufacturing constraints add another layer of complexity to the placement process. Designers must navigate these challenges to achieve optimal placement results.

To achieve the best possible placement, various optimization techniques are employed. Heuristic methods, such as simulated annealing and genetic algorithms, are used to explore the design space and find near-optimal solutions. Multi-objective optimization considers multiple objectives simultaneously, such as minimizing wire length while also reducing power consumption and improving timing.

- **Clock Tree Synthesis** The primary objective of CTS is to create a clock distribution network that delivers the clock signal to all sequential elements, such as flip-flops and latches, with minimal skew and latency. At the same time the clock tree should use the minimum amount of routing resources and signal integrity problems like cross-talk need to be minimized.

  The design of the clock tree must be meticulously planned to achieve a balanced and efficient distribution. Various topologies, such as H-tree, X-tree, and balanced binary tree structures, are employed to achieve this balance. To improve the balance the tools perform insertion of clock buffers and inverters, which help drive the clock signal across the chip.

Skew optimization is a critical aspect of CTS. Skew refers to the difference in the arrival times of the clock signal at different sequential elements. Minimizing skew is essential to prevent timing violations and ensure reliable operation. Techniques such as adjusting the positions of buffers, resizing them, or adding delay elements are used to balance the arrival times of the clock signal.

Reducing clock latency is also a key objective of CTS. Clock latency is the delay from the clock source to the sequential elements. Reducing latency helps meet the timing requirements of the design, ensuring that the clock signal is delivered promptly to all parts of the circuit. This can be achieved by minimizing the number of buffers and inverters in the clock path and optimizing their placement.

Power optimization is another important consideration in CTS. The clock network is a significant consumer of power in an IC, and optimizing the clock tree can help reduce overall power consumption. Techniques such as using low-power buffers, optimizing clock gating, and minimizing switching activity in the clock network are employed to achieve this goal.

- **Signal routing** The routing phase in physical design is one of the final and most complex steps in the process of transforming a netlist into a manufacturable layout for an integrated circuit. After the placement stage has fixed the positions of standard cells and macro blocks, routing is responsible for establishing the physical interconnections between the pins of these components according to the netlist. This involves the creation of metal interconnects across multiple routing layers to carry signals, clock, and power throughout the chip.

  The routing process can be broadly categorized into two major stages: global routing and detailed routing, each addressing different aspects of the interconnect problem.

  - **Global Routing:**

    Global routing operates on an abstracted representation of the chip and aims to plan the approximate routes for nets before assigning exact paths. The chip is divided into a grid, and each grid cell corresponds to a region where wiring can be placed. The goal of global routing is to assign each net to a sequence of grid cells (routing regions) while considering congestion, timing constraints, and wirelength minimization.

    Congestion Estimation:

    Global routing evaluates the routing demand for each grid cell. It aims to balance the wire distribution across the chip and avoid congestion hotspots where too many nets attempt to pass through limited routing resources.

    Timing Closure:

    Global routing works to meet timing constraints by minimizing critical path delays and ensuring nets have feasible path lengths that satisfy their timing budgets. It considers RC delays, where wire resistance and capacitance affect signal propagation times.

    Routing Cost Functions:

To optimize wirelength, congestion, and timing, cost functions are defined for each grid cell. The routing algorithm (e.g., maze routing, A* search, or Steiner tree algorithms) seeks to minimize this cost by assigning nets to low-cost regions.

The output of global routing is a rough assignment of nets to routing regions, without specific details about the metal layers or exact wire geometries.

– **Detailed Routing:**

Detailed routing is a finer-grained process that takes the global routing results as input and determines the exact geometric layout of the wires. It assigns specific metal tracks, vias, and layers to each net, adhering to design rules and manufacturing constraints. This stage must satisfy stringent design rules set by the foundry, such as minimum wire width, spacing between wires, via alignment, and other process technology-specific requirements.

Track Assignment:

Detailed routing begins by selecting routing tracks within the grid regions allocated by global routing. It defines the exact path for each wire in terms of horizontal and vertical metal segments, using different metal layers for different routing directions (e.g., horizontal routing on metal layer M1, vertical routing on M2, etc.).

Design Rule Checking (DRC):

At this stage, Design Rule Checking (DRC) is critical to ensure that all wire geometries and via placements comply with manufacturing requirements such as wire widths, spacings, and via enclosures. Violations in DRC can cause yield loss or circuit failure after fabrication.

Minimizing Crosstalk and Parasitics:

Detailed routing must also address crosstalk, which is the unwanted interference between adjacent wires due to capacitive coupling. This is crucial in high-performance designs where signal integrity is a concern. Additionally, the router aims to reduce parasitic effects (resistance, capacitance, and inductance), which degrade signal quality and timing performance.

Routing Algorithms:

Detailed routing uses advanced algorithms like negotiation-based routers, rip-up and reroute, or pattern-based routing to assign metal tracks efficiently while iterating over congestion and DRC fixes. For example, line-probe algorithms or L-shaped and Z-shaped patterns are common in practical routing strategies.

Multi-Layer Routing:

Modern IC designs use several layers of metal, each optimized for different routing tasks. Lower layers (closer to the transistors) are used for shorter, local connections, while upper layers (with larger pitches and lower resistance) are used for long global interconnects. Detailed routing efficiently manages the transitions between layers vias and ensures that power/ground routing adheres to specific metal layers reserved for power distribution. Do to the growing needs of modern chips of routing resources, the trend for the metal stack has been a constant grow

in metal layers number (1.4), the number of metal layers has a strong impact on the manufacturing costs of the DIE, so also for costs reasons the optimization of routing resources becomes crucial.

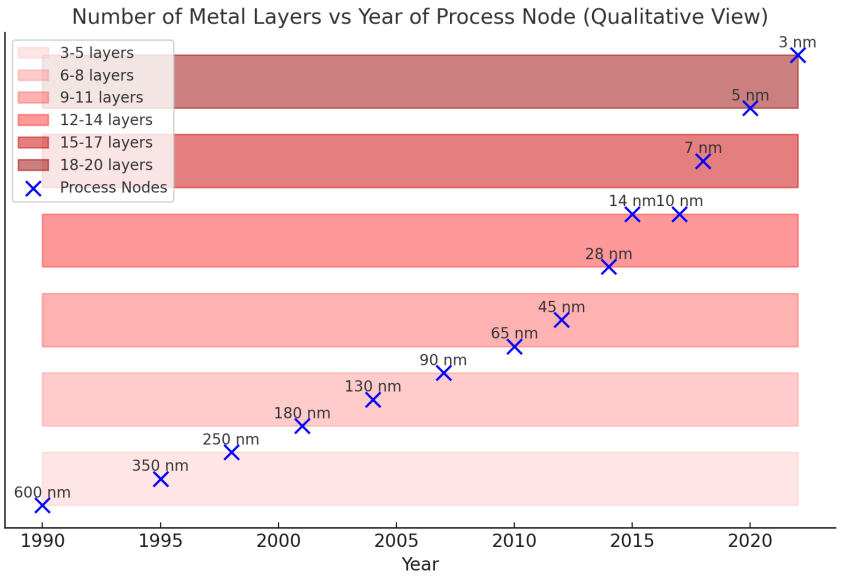

Figure 1.4: Metal stack trends in VLSI

Wirelength and Via Minimization:

Detailed routing attempts to minimize the total wirelength and number of vias used in the layout. Vias introduce additional resistance and can negatively impact reliability, making their reduction critical for performance and manufacturability.

Congestion Management:

High routing congestion can lead to an infeasible layout, requiring iterative rip-up and reroute strategies to reduce congestion while maintaining timing and DRC compliance.A common practice at this stage is to graphically evaluate congestion across the chip, this is done with a congestion map, the latter can be reported by the tool and it is a simple but effective way to understand how far the design is from routing closure. An example is provided in 1.5, it is basically an heat-map, for this example a warmer color is associated with a more congested area. Another useful map is the cell density map, a colour map that identifies the cell density across the design.

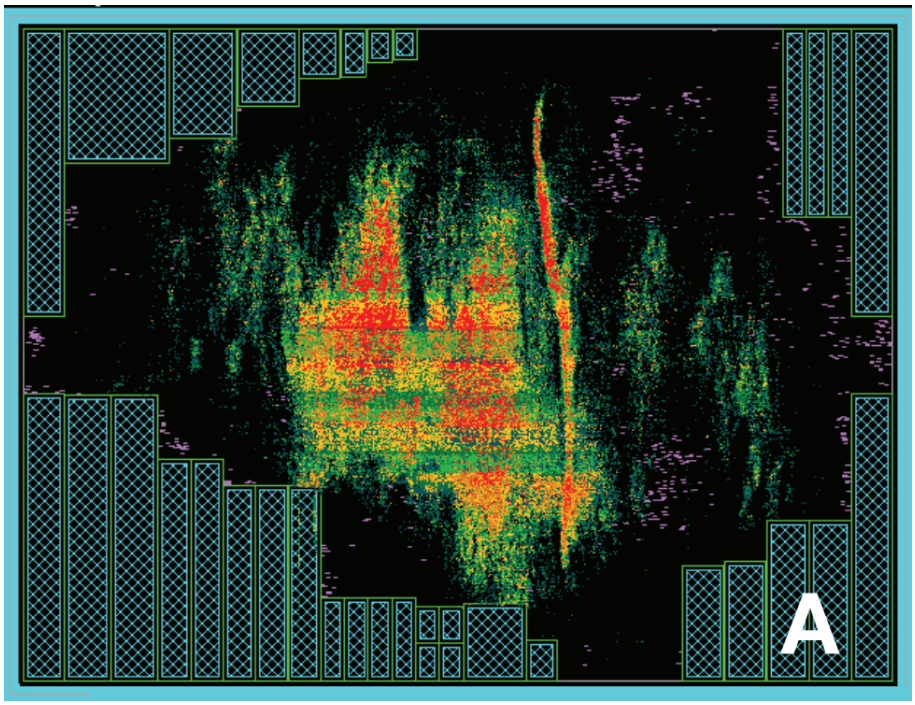

Figure 1.5: Congestion Map example

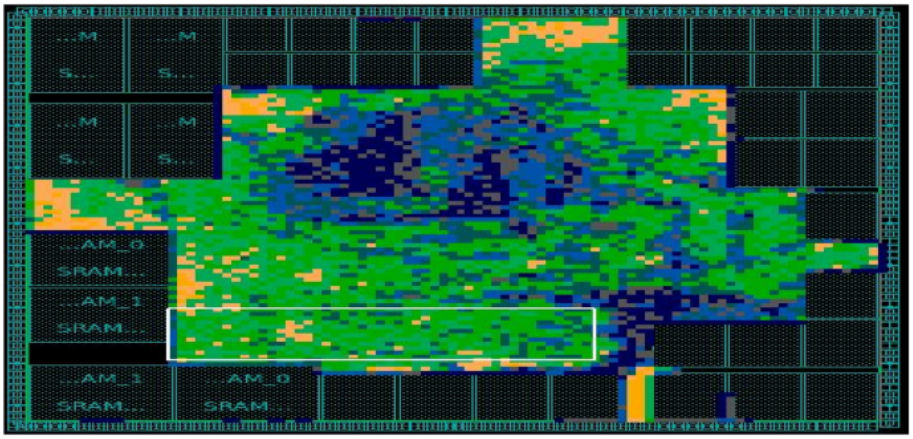

Figure 1.6: Cell density Map example

- **Timing closure** Timing closure occurs after the routing phase and involves a detailed process of analyzing and resolving any violations in timing requirements, such as setup and hold times, clock skews, and overall delay margins. Performing timing closure is essential to ensure that the chip functions correctly at its intended clock speed and operating conditions.

  After the routing phase, the design is placed and interconnected with the physical routing of the wires. At this stage, the design is close to its final form, but routing introduces parasitics such as resistance and capacitance to the interconnects. These parasitics can significantly affect signal propagation delays, causing timing violations that were not present in previous stages of the design. Therefore, the design undergoes an iterative process of timing analysis and optimization.

The first step in timing closure involves performing a comprehensive static timing analysis (STA) to identify paths that do not meet the required timing constraints. The STA tool calculates delays across all critical paths, considering the parasitics extracted from the routed design. Paths with violations are flagged for optimization.

For setup time violations, which occur when a signal fails to arrive at a flip-flop in time for the next clock edge, designers might adjust the clock tree or resize cells to improve propagation delay. Increasing the drive strength of the logic cells or reducing the fanout can also help reduce delays. If the violation is related to interconnect delays, designers may reroute the affected nets to shorter or less congested paths, or apply buffering to minimize delays introduced by parasitic effects.

For hold time violations, which occur when a signal arrives too early and interferes with the current data, the focus is on adding delays to the path. This can be achieved by inserting delay buffers or adjusting routing to increase the wire length, thereby introducing additional delay.

Another critical aspect of timing closure is clock tree optimization. Clock skew, the difference in clock arrival times at different flip-flops, can significantly impact timing. Techniques such as clock tree balancing or adding skew buffers are used to minimize skew and ensure synchronized clock signals across the design.

Signal integrity issues, such as crosstalk, also come into play during timing closure. Crosstalk can cause unwanted coupling between neighboring signals, leading to timing violations. Designers address this by increasing spacing between sensitive nets, shielding critical signals with ground wires, or using alternative routing layers.

Power considerations are also intertwined with timing closure. Aggressive timing fixes, such as resizing cells or adding buffers, can increase dynamic and leakage power. Designers must strike a balance between meeting timing constraints and adhering to power budgets. Also IR drop is usually considered in order to have a more precise estimate of the path delays.

Finally, after all adjustments and optimizations, the design is re-analyzed using STA to ensure all timing violations have been resolved. This process is iterative and may require multiple cycles of analysis and correction until the design meets all timing constraints under various operating conditions, including worst-case and best-case scenarios.

In summary, timing closure is a highly detailed and iterative process that ensures the design meets timing requirements after routing. It involves identifying and resolving timing violations through cell resizing, rerouting, buffering, clock tree optimization, and managing parasitics and crosstalk. Achieving timing closure is a complex balancing act that ensures the chip functions correctly while maintaining power and performance targets. A diagram of the complete flow is shown in 1.7

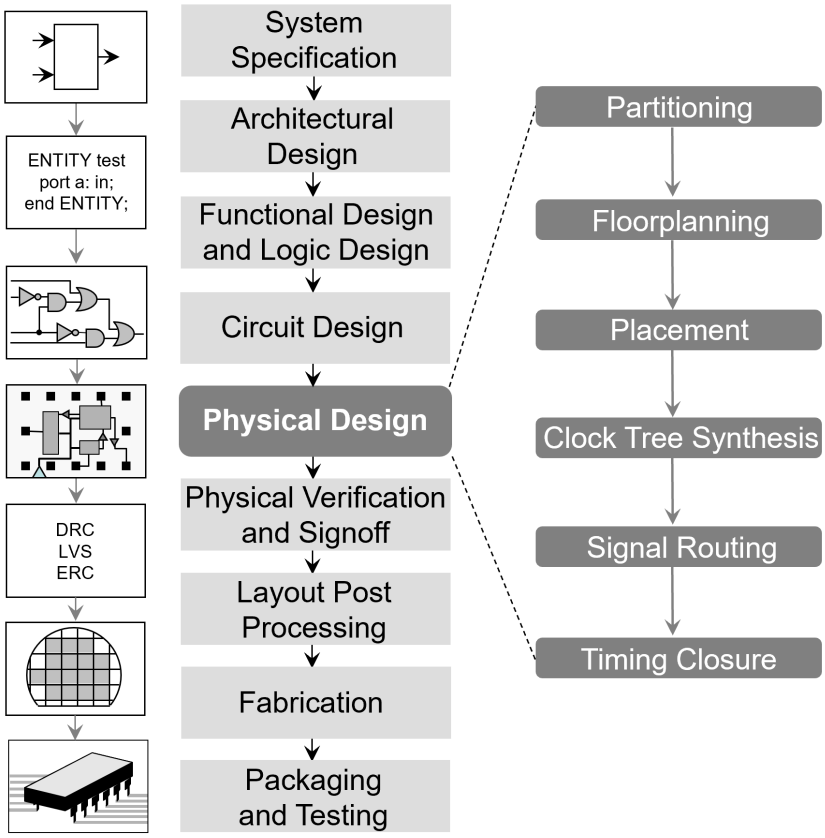

Figure 1.7: Overview of the physical design flow [3]

1.3 Problem Statement

In the physical design phase of advanced integrated circuits, the design of the Power Delivery Network (PDN) plays a critical role in determining the final Power, Performance, and Area (PPA) results. Extremely scaled technologies, such as the one being analyzed in this thesis, pose additional complexities that demand in-depth investigations and empirical results that can be generalized and extended across the entire design phase [6].

A well-designed power delivery network must ensure that the target metrics for Power and Performance are achieved. However, PDNs significantly impact area metrics, consuming up to 20-40% of routing resources in advanced technology nodes [7]. This constraint not only limits the available space for signal interconnects but also exacerbates routing congestion, leading to potential performance bottlenecks and increased design complexity. Furthermore, PDN-induced area overheads can contribute to as much as 10-15% of the total layout area, making optimization a critical requirement for achieving efficient designs.

Additional challenges arise with advanced technology nodes (< 5nm), where the demand for higher transistor density, tighter voltage margins (e.g., IR-drop budgets as low as 10-20mV), and reduced routing layers intensify the need for innovative PDN configurations. These nodes introduce significant trade-offs between ensuring robust power delivery and minimizing routing congestion.

This study aims to identify the optimal PDN configuration that strikes a balance among these metrics, enabling maximum area shrinkage while maintaining power and performance within acceptable thresholds. By addressing these challenges through empirical analysis and innovative methodologies, this work seeks to contribute practical insights and solutions for PDN optimization in cutting-edge technology nodes.

1.4 Scope and Limitations

This research focuses on an advanced technology nodes, specifically < 5nm, utilizing cutting-edge processes and design methodologies. Due to the sensitive nature of the technology and the confidentiality agreement under which this research was conducted, certain specific details, such as the exact node name and proprietary methodologies, cannot be disclosed. However, the analyses and results presented in this thesis remain highly relevant and valuable, offering insights and empirical data that can serve as a reference for future work in cutting-edge nodes and similar advanced technologies.

This study leverages industry-standard Electronic Design Automation (EDA) tools widely adopted in the semiconductor industry. While specific tool names can not be disclosed due to confidentiality agreement, the results provided are accurate and aligned with industry practices.

1.5 Thesis Organization

Thesis Organization

This section outlines the organization of the thesis, summarizing the content of each chapter to provide the reader with a clear understanding of the research flow.

- **Chapter 1: Introduction**
  - Provides the background and context for the research topic.
  - Offers an overview of the physical design flow in integrated circuits.
  - Clearly articulates the problem statement and the motivation for this study.
  - Defines the scope and limitations of the research, including considerations of confidentiality.
  - Summarizes the structure of the thesis to guide the reader.
- **Chapter 2: Literature Review**
  - Reviews existing research and methodologies relevant to Power Delivery Networks (PDNs).
  - Identifies gaps and challenges in PDN design for advanced technology nodes.
- **Chapter 3: Methodology**
  - Details the research design and approach adopted in this study.
  - Describes the data collection techniques employed.
  - Explains the analysis methods applied to evaluate PDN configurations and PPA metrics.
- **Chapter 4: Design and Implementation**
  - Presents the design process and implementation of the Power Delivery Network configurations.
- **Chapter 5: Results**
  - Provides a detailed presentation of the results obtained from the study.
  - Includes quantitative analyses and visualizations such as graphs and tables to support findings.
  - Interprets the results and their implications for PDN design in advanced nodes.
  - Discusses trade-offs between routing resources, area, and PPA metrics.
- **Chapter 6: Conclusion**
  - Summarizes the main findings and contributions of the research.
  - Suggests directions for future research.

Chapter 2

Literature Review

Recent publishing have been posing a stronger accent on the importance of the PDN implementation for highly optimized designs using extremely scaled tech nodes.

2.1 Congestion and Area Trade-offs:

"Benchmarking Power Delivery Network Designs at the 5-nm Technology Node"[8] Lanzillo et al. provided a comprehensive evaluation of 96 PDN configurations for a 5-nm FinFET CMOS technology, focusing on the impact of PDN density on routing congestion and resource availability:

- **Dense PDNs**:
  - Prioritize power integrity but significantly constrain routing resources.
  - Result in a **17% reduction** in available signal track density compared to sparse configurations, leading to increased congestion.
  - Ideal for high-performance designs where congestion can be mitigated by other means.
- **Sparse PDNs**:
  - Provide greater routing flexibility, with an increase of in signal track density compared to dense configurations.
  - Improve routing availability but at the expense of power integrity.
  - Suitable for low-power designs with relaxed IR-drop budgets.

Lanzillo et al. also explored advanced PDN features that address congestion and area challenges:

- **Power Staples**:
  - Replace continuous power rails, reducing routing blockage.
  - Dense configurations perform comparably to continuous rails while offering enhanced routing flexibility.
- **Backside Power Delivery**:

  - Relocating power rails to the wafer's backside improves signal track density by **10–30%**.
  - Dense PDNs benefit the most, with a **25–30% increase**, while sparse PDNs show around **10% improvement**.

Implications for Area and Routing Resource Optimization:

- The study demonstrates the importance of balancing power grid density and routing availability.
- Findings provide a robust framework for evaluating PDN design trade-offs, particularly regarding area and congestion.
- This research builds on these insights, further exploring how advanced PDN features, such as power staples, influence area shrinkage and routing resource utilization.

Lanzillo et al. highlighted that their benchmarking methodology does not require a full Place-and-Route (PnR) process, allowing metrics such as signal track density and routing resource utilization to be evaluated at an early design stage. While this approach enables rapid preliminary analysis, it does not capture fully the post-PnR effects that influence final routing congestion and area usage.

In contrast, this research focuses on retrieving metrics after the PnR phase, providing a detailed and accurate understanding of the impact of PDN designs on area shrinkage and routing congestion.

"Full Chip Impact Study of Power Delivery Network Designs in Gate-Level Monolithic 3-D ICs" [9]

Samal et al. analyzed the impact of Power Delivery Networks (PDNs) on routing congestion and area utilization across different technology nodes. The study highlights several key findings:

- **Impact on Routing Congestion and Signal Wirelength**:
  - PDNs occupy significant routing resources, particularly in top and intermediate metal layers, leading to signal detours and increased wirelength.
  - At advanced nodes, PDNs increase total power consumption by up to **19% at 7 nm**, primarily due to longer signal wirelength and parasitics.
- **Metal Layer Utilization**:
  - Up to **40% of the top metal layer** and **20% of intermediate layers** are dedicated to PDN in typical designs.
  - The remaining routing capacity often becomes insufficient for signal nets, worsening congestion and area efficiency.
- **Scaling Challenges at Advanced Nodes**:
  - Increased resistivity of copper wires at advanced nodes exacerbates the impact of PDNs, further constraining routing flexibility.

The study underscores the critical trade-offs between routing resource availability and power delivery efficiency in PDN designs. These findings are relevant to this work, which focuses on evaluating PDN-induced congestion in advanced technology nodes. The focus of the study is on 3-D ICs, but most of the conclusions can be extended to 2-D architectures as the one being examined in this thesis.

2.2 Optimization Techniques

"Machine Learning-Driven Optimization of Metal Stack and PDN" [7]

Shin et al. introduced an application-driven optimization framework for metal stack and Power Delivery Network (PDN) designs using a machine learning-based tool, Synopsys DSO.ai. The study addressed key challenges in advanced technology nodes, particularly in balancing PPA (Power, Performance, and Area) metrics during Design-Technology Co-Optimization (DTCO).

- **Trade-offs in BEOL Scaling**:
  - Increasing metal width and spacing reduces wire delay per unit distance, improving performance.
  - However, this approach negatively impacts routing resources, reducing wire track availability per unit area.
  - Strengthening the PDN improves IR-drop performance but consumes additional BEOL resources, which can degrade overall PPA metrics.
- **Machine Learning Framework for PDN Optimization**:
  - Synopsys DSO.ai explored parameters such as layer sheet count, pitch, spacing, and PDN horizontal/vertical pitches.
  - The optimization was constrained to maximize achieved frequency while maintaining IR-drop targets.
- **Results and Comparison with Human Experts**:
  - Machine learning achieved a **+2.2% frequency improvement** with a **5.5% worse IR-drop**.
  - This outperformed human experts, who achieved a **+1.4% frequency improvement** with a **2.5% worse IR-drop**.

The study demonstrated that machine learning frameworks can effectively optimize PDN and metal stack configurations, achieving superior results compared to traditional human-driven DTCO methods. These findings are particularly relevant for advanced nodes, where PPA trade-offs are critical, and routing resource constraints are significant.

2.2.1 Material and design challenges

"Power Delivery Design, Signal Routing, and Performance of On-Chip Cobalt Interconnects in Advanced Technology Nodes" [10]

Lanzillo et al. investigated the trade-offs in using cobalt and copper interconnects for power delivery and signal routing in advanced technology nodes. Key findings include:

- **Routing Resource Utilization**:
  - Wider cobalt-based power lines result in up to **50% higher IR drop** compared to copper lines.
  - Shared routing layers for signal and power exacerbate congestion, particularly at advanced nodes with a **24-nm minimum pitch**.
- **Area Penalties in Power Line Design**:
  - Increasing power line width from 3× to 5× reduces IR drop but results in a **30% increase in standard cell height**.
  - Reducing power tap spacing improves IR drop but incurs a **25% area penalty** due to restricted pin access.
- **Design Trade-offs**:
  - Wider lines mitigate IR drop but reduce routing flexibility and increase congestion in high-density designs.
  - Routing signals higher in the BEOL stack alleviates delays but increases via-related resistance.

This study highlights the critical balance between routing resource availability and PDN efficiency, offering insights into the area penalties associated with wider power lines and denser PDNs.

Emerging Architectures

"A Holistic Evaluation of Buried Power Rails and Back-Side Power for Sub-5-nm Technology Nodes" [11]

Panth et al. investigated innovative PDN architectures, including buried power rails (BPRs) and backside power delivery (BSP), aimed at addressing routing congestion and power integrity challenges in sub-5-nm technology nodes. While these approaches provide significant benefits, such as **30% lower off-chip voltage droop** and **85% lower on-chip IR drop**, they require deep structural changes to the chip design, such as integrating power rails into the silicon substrate or relocating power grids to the wafer's backside.

Relevance to This Work: This research focuses on optimizing conventional PDN networks in advanced technology nodes rather than exploring architectural changes like BPRs or BSP. While emerging techniques like these offer exciting opportunities for future designs, they are beyond the scope of this study, which concentrates on evaluating optimizations for existing PDN configurations.

Chapter 3

Methodology

3.1 Research Design

This study focuses on optimizing Power Delivery Network (PDN) configurations in advanced technology nodes (< 5 nm) by performing various analysis on a design block. The primary objective is to evaluate the trade-offs between area utilization, routing congestion, and power integrity metrics to identify optimal PDN schemes.

An automation framework was developed to:

- Perform rapid area sweeps and assign PG schemes automatically.
- Execute the entire flow from floorplanning to Place-and-Route (PnR) with minimal manual intervention.
- Collect and analyze metrics at each stage of the PnR flow, e.g. congestion, wirelength, and overflow.
- Identify routable designs based on predefined thresholds and maintain a ranking table for comparison.

To validate the design choices, in-depth power analyses were conducted on the top-ranked designs, focusing on static and dynamic IR drop as well as grid resistance.

When the data-set was completed, comparisons with an older tech-node were performed.

3.1.1 automation tool

This kind of in-depth analysis required a huge amount of different trials in order to build a sufficiently big dataset. In order to facilitate the study, some tools have been developed to improve the automation factor. Implementation details of such tools goes out of the scope of this research, hence a schematic high-level picture will be described.

The requirements for the automation tools were:

- Allow for fast area sweep across the complete flow, from floor-planning to PnR
- Implement shape-adaptive pin placement
- Implement user-friendly PDN scheme selection from a pre-defined database

This tooling was successfully developed and effectively improved execution times by a good margin.

A block scheme of the automated flow is shown in picture 3.1

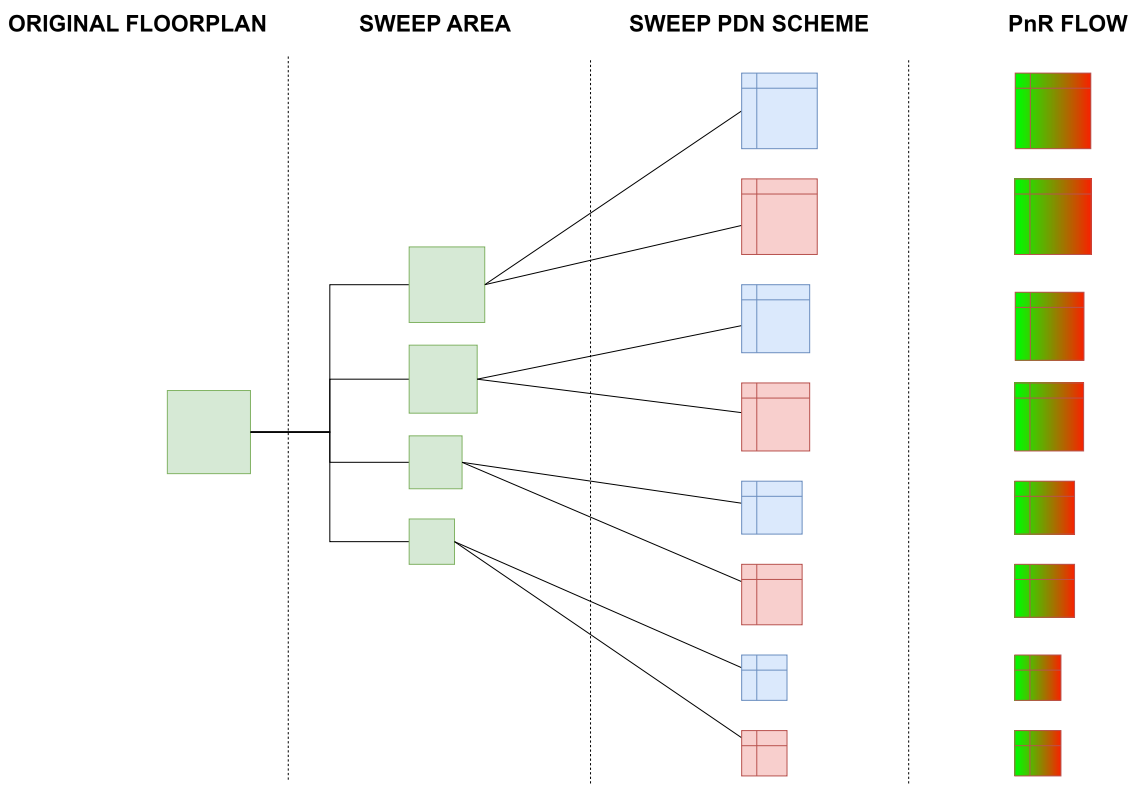

Figure 3.1: Diagram of the automated flow

The tool allows the user to preset a flow sweep between two parameters: Area and power grid scheme.

3.2 Analysis setup

Since the research aimed at investigating routing results at different area values and PDN grids, the chosen block was a standard cell - only block. With this assumption the macro presence could be factored out, also the problem of managing macros while changing the floorplan boundary is avoided.

A total of 9 different PDN schemes have been tested, each of them across different area values. Aspect ratio is kept constant, hence $\Longrightarrow$

For a given shrink percentage ω:

$$\text{new\_horizontal\_edge\_i} = \text{horizontal\_edge\_i} \cdot \text{x\_factor}$$

$$\text{new\_vertical\_edge\_i} = \text{vertical\_edge\_i} \cdot \text{y\_factor}$$

where,

$$\text{x\_factor} = \text{y\_factor} = 0.5 \cdot \left(1 - \frac{\omega}{100}\right)$$

For each different trial the most relevant figures of merit have been considered, focusing on the routability parameters e.g. (overflow, congestion maps, DRCs).

Each PDN configuration is in general characterized by different parameters:

- **Pitch**: The spacing between each rail of the grid. For some schemes a double spacing is possible, in that case, a pitch */beta* will be used for M1 and lower metals, while on upper metals both */alpha* and */beta* are used. This is visually explained in figure /refsinglevsdouble. This value is usually expressed in CPP (Contact poly to poly) units, in this analysis normalized values will used due to confidentiality.

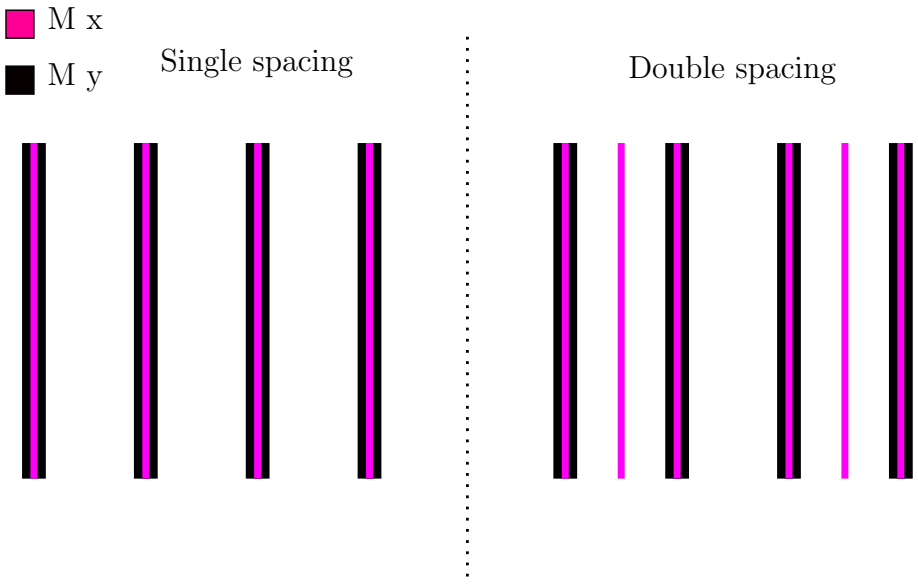

Figure 3.2: Single and double spacing

- **Stubs or Stripes**: Each PDN scheme can be characterized by the use of Stripes or Stubs (Or staples) at the M1 layer. In this analysis the keywords **STR/STU** will indicate the presence of either one ore the other. In figures /reffig:M1 stubs and /reffig:M1 stripes a schematic view is presented. Stubs

are short pieces of M1 metal that are connected to the M0 power rails directly with vias, while Stripes are continuous and interconnected in a grid fashion. Stubs are introduced in order to save routing resources at M1, hence potentially improving congestion. This of course , due to the lack of parallel connections, makes the grid more resistive. The expectation for such configurations is to have improved routability at the cost of worse IR drop.

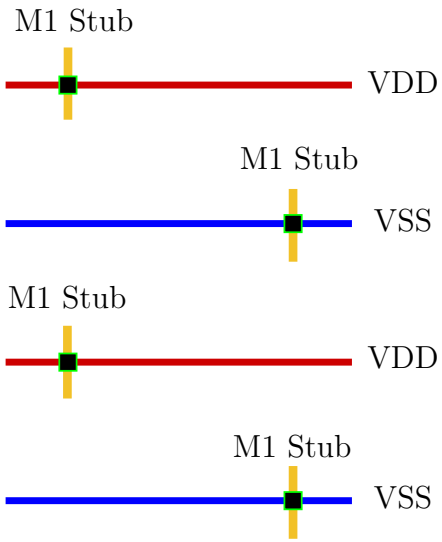

Figure 3.3: M1 stubs configuration

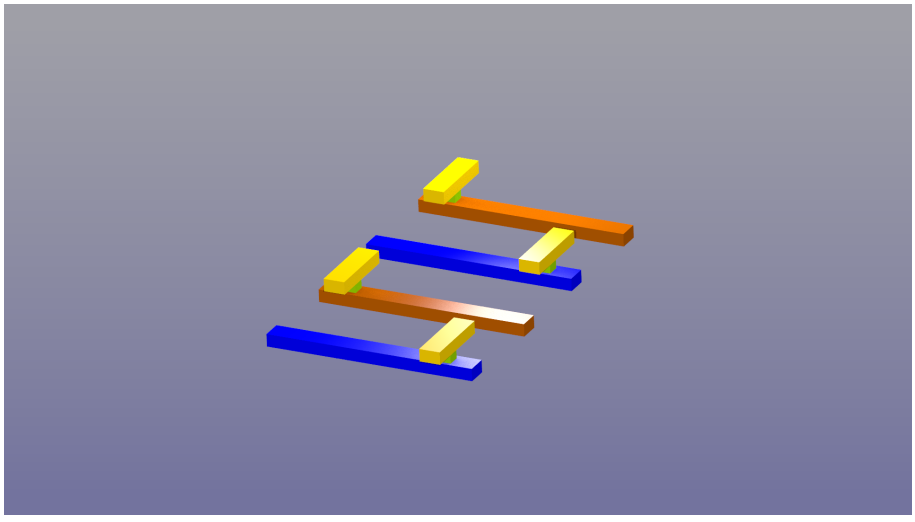

Figure 3.4: M1 stubs configuration - 3D rendering

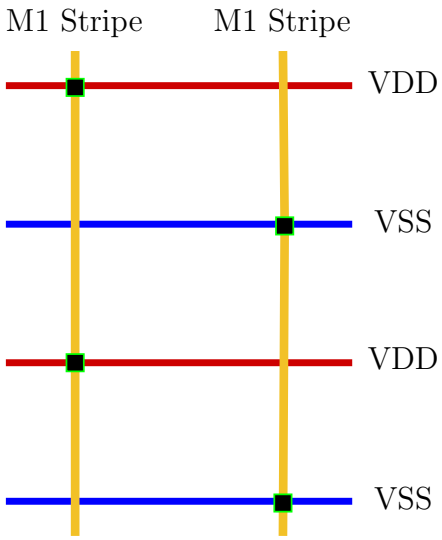

Figure 3.5: M1 stripes configuration

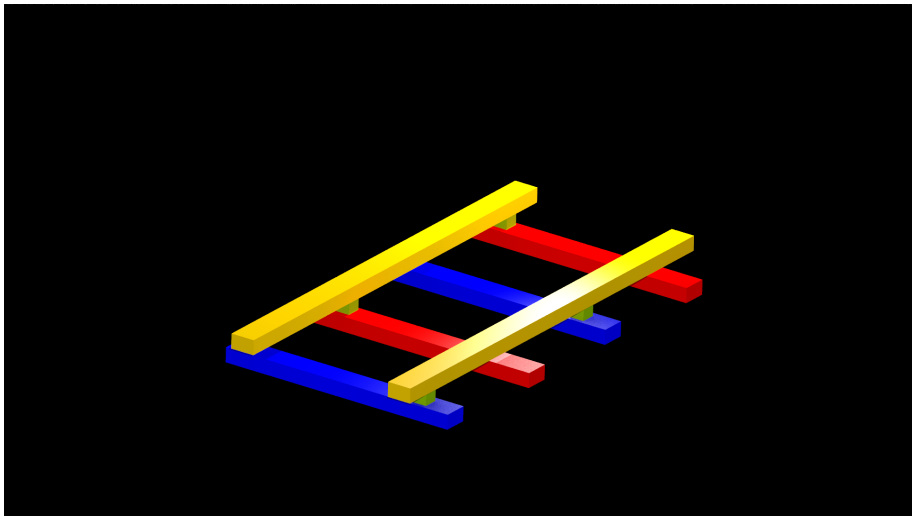

Figure 3.6: M1 stripes configuration - 3D rendering

- **Via stack frequency**: Defines the occurrence rate of vias stacks that routes power from the top metal layers to M1. Since only two configurations are possible, they will be discriminated by the suffixes **A/B**. Option B has an occurrence $\approx 50\%$ lower with respect to A. In figure 3.7 a qualitative scheme is shown, here again a trade-off between resource saving and power integrity is central. Configuration B is expected to improve routability , retaining the same downsides mentioned for the previous case.

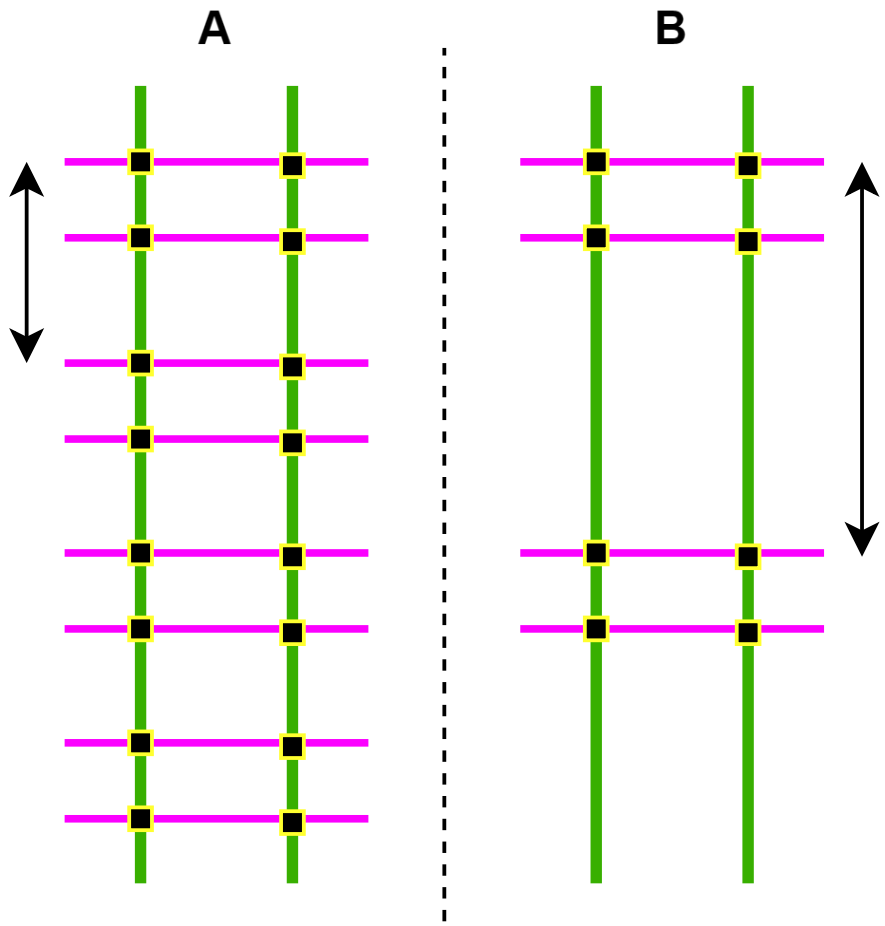

Figure 3.7: A vs B via stack density

Hence, generalizing a convention each PDN scheme can by referred with the TAG $\alpha\_\beta\_STR/STU\_A/B$ for double spacing or $\alpha/\beta\_STR/STU\_A/B$ for single.

The values of α and β in **normalized pitch units** are listed in table **??**.

Normalized pitch values		
alpha	100	75
beta	50	

Table 3.1: Available options for normalized pitch values

The overall summary of the PDN schemes taken into consideration is listed in table 3.2

PDN SCHEME	Lower metals pitch		Higher metals pitch		M1 stripes or stubs		Vias occurrence	
	75	50	100	75	Stubs	Stripes	A	B
75_STU_A	x			x	x		x	
75_STU_B	x			x	x			x
75_STR_B	x			x		x		x
75_STR_A	x			x		x	x	
100_50_STR_A		x	x			x	x	
100_50_STU_A		x	x		x		x	
100_50_STU_B		x	x		x			x
75_50_STR_A		x		x		x	x	
75_50_STU_A		x		x	x		x	

Table 3.2: PDN schemes summary

Considering all the degrees of freedom on the considered PDN schemes, a qualitative classification can be done, discriminating how "stringent" or "relaxed" a scheme is in terms of grid density.

- Relaxed: 75_STU_B, 100_50_STU_B
- Balanced: 75_STU_A, 75_STR_B, 100_50_STU_A, 75_50_STU_A
- Stringent: 75_STR_A, 100_50_STR_A, 75_50_STR_A

This simple classification do not takes into account the difference in terms of pitch. Hence major differences within these groups would not come as a surprise.

Chapter 4

Results

4.1 Congestion

In order to analyze the impact of the PDN scheme on a given design, the overflow is a key metric. Overflow takes into account the supply and demand of routing resources to quantitatively estimate the level of congestion. By analyzing the overflow, designers can identify areas where there is an excessive routing demand. Overflow can be defined as:

$$\text{Overflow} = \text{Max}(0, \text{Demand} - \text{Supply})$$

where:

- Demand: Number of routing tracks required by nets passing trough a region.
- Supply: Number of routing tracks physically available in that region.

Other metrics associated to overflow are:

- Horizontal overflow: Sum of overflows across vertical metal lines.
- Vertical overflow: Sum of overflows across horizontal metal lines
- Total overflow: Sum of overflows across all the regions.

4.1.1 Overflow results

In the following section results regarding overflow will be presented. The data-set spans across 9 different PDN configurations, for area values sweeping from -18% up to +6%. For some PDN schemes not all the area values are present. This is because either a specific combo between PDN scheme and area value did not converge in post-route or because that portion of solution space was already covered by other results.

In figure 4.1 an heatmap of all the available overflow results is presented. The metric considered is the **Total overflow**, hence considering both vertical and horizontal metal lines, averaged across all the design area.

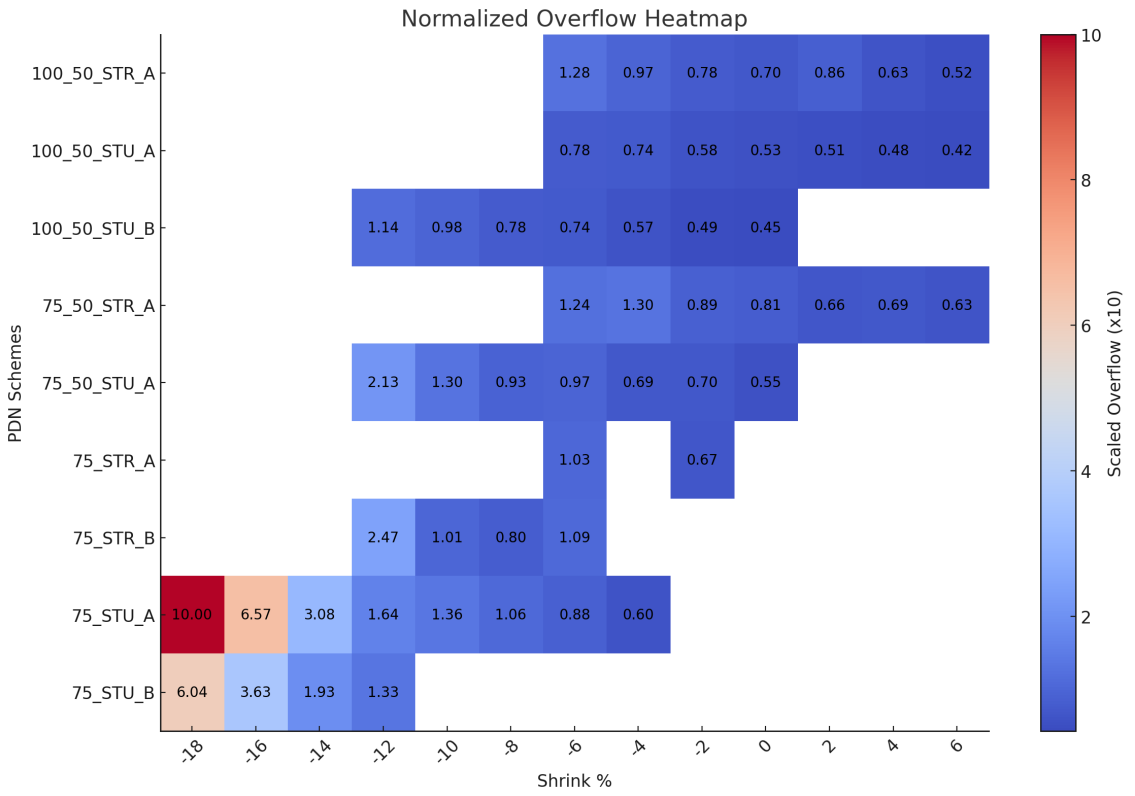

Figure 4.1: Overflow results for all the PDN configurations - Area values from -18% to +6%

As expected overflow values degrade significantly with shrink%, in general has been noted that designs with overflow values $>\approx 1$ have an high DRCs (Design rule checks) count. The DRCs number together with the short-circuit number has been used trough all the evaluation phase to grade the designs with more routability potential.

In general, near the design breaking point (high DRC count) there is not a clear relation between overflow values and the DRCs number itself, but is a very effective metric to evaluate the routability across a large span of area values.

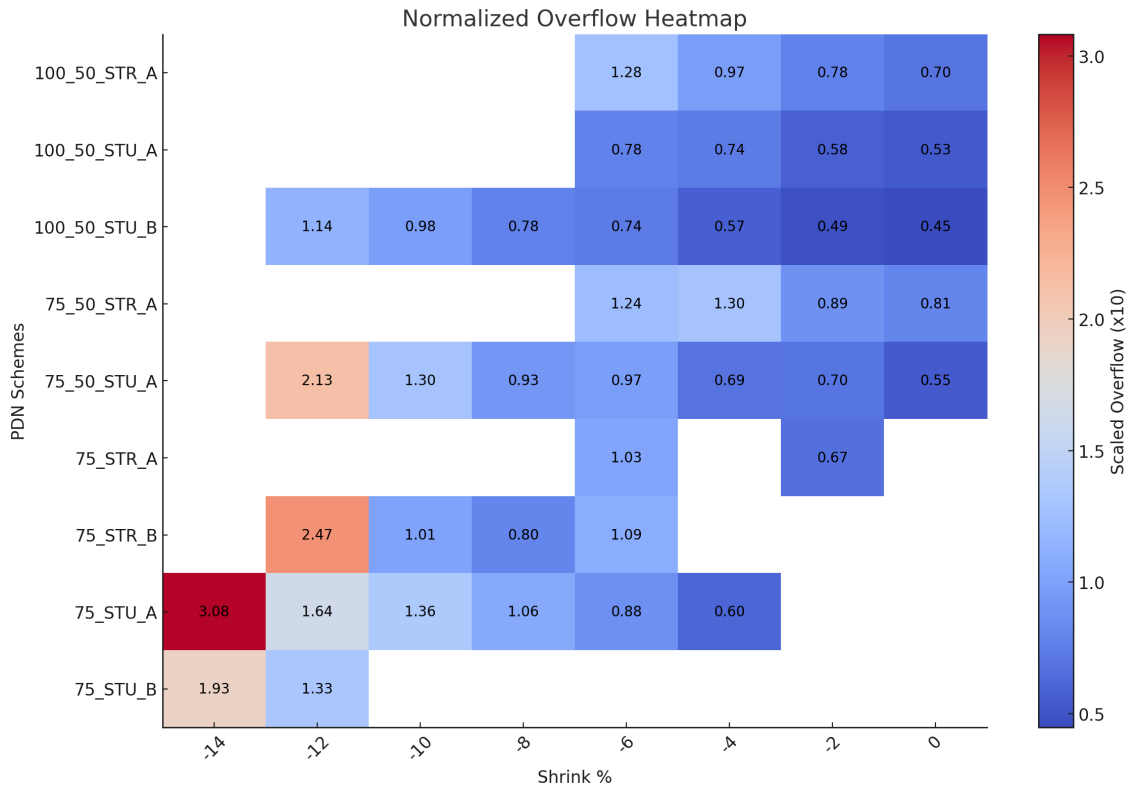

Figure 4.2: Overflow results for all the PDN configurations - Area values from -14% to +0%

The figure 4.2 presents an heatmap with values covering the -14% to 0% span. The only two configurations that converged at -14% are **75_STU_A** and **75_STU_B**. Those two schemes adopt a single pitch approach, having a normalized pitch of 75 units for both lower and upper metals. Considering the fact that double pitch configurations like the 100_50* have a more relaxed (100) pitch at higher metals, and a more stringent one (50) at lower metals, the relative impact of lower metal pitch is overall stronger.

Going more in depth, in figure 4.3 overflow comparisons are made between PDN schemes that have identical parameters apart from via occurrence. In this way the effect of the via frequency can be isolated to evaluate the specific impact of this parameter. N.B case B has 50% lower vias connections at M1 The first case compares **75_STU_A vs 75_STU_B**

Case B has:

- 20% lower overflow at -12% shrink
- 45% lower overflow at -14% shrink

The second case compares **75_STR_A vs 75_STR_B**

Case B has 24% lower overflow at -6%.

So in general the impact of Vias occurrence at M1 is strong on overflow, hence congestion. Of course the penalty in terms of power integrity should be evaluated, this will be done in the following sections.

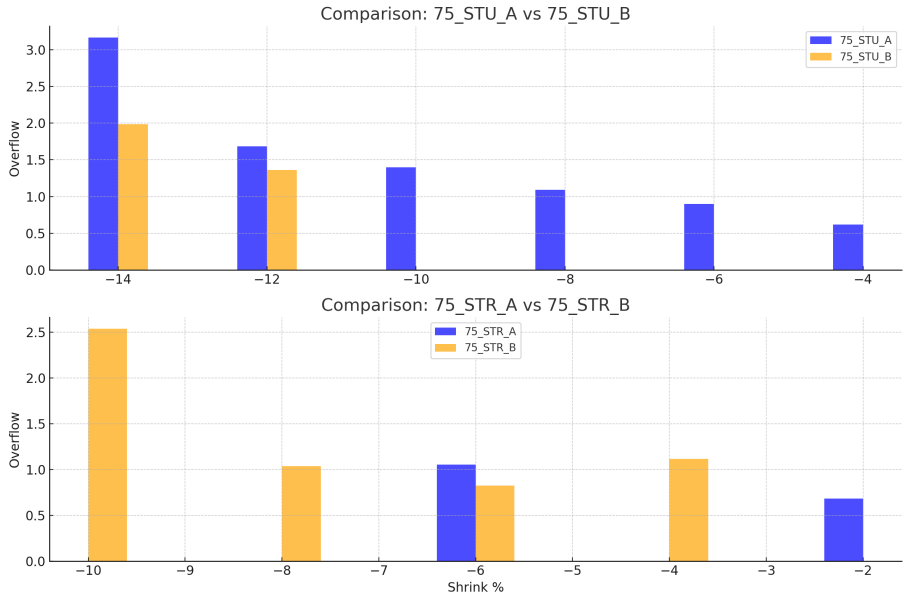

Figure 4.3: Overflow comparison - A vs B Vias occurrence

In figure 4.4 a chart comparing Stripes and Stubs configuration is presented. As done for the previous analysis, just equal schemes are compared.

First case, **75_STR_A vs 75_STU_A**:

- Stubs has $\approx 15\%$ lower overflow at -6%.

- Stubs has lower overflow (0.62) at -4% with respect to Stripes (0.68) at -2%.

Second case, **75_STR_B vs 75_STU_B**:

In this case there is no data for overlapping area values, still is quite evident that overflow figures of the Stripes configuration are worse; $\approx 25\%$ lower overflow at -16% stubs with respect to -10% stripes.

Third case, **100_50_STR_A vs 100_50_STU_A**:

In this case there is an average overflow difference of $\approx 30\%$, favoring the Stubs case.

Third case, **75_50_STR_A vs 75_50_STU_A**:

In this case there is an average overflow difference of $\approx 27\%$, favoring the Stubs case.

In general choosing Stubs instead of Stripes leads to important gains in terms of congestion, in terms of overflow the difference is found to be $\approx 20 - 30\%$ for the same area value. Of course the same considerations done in the A vs B Vias case held here, the penalties in terms of power integrity should be evaluated.

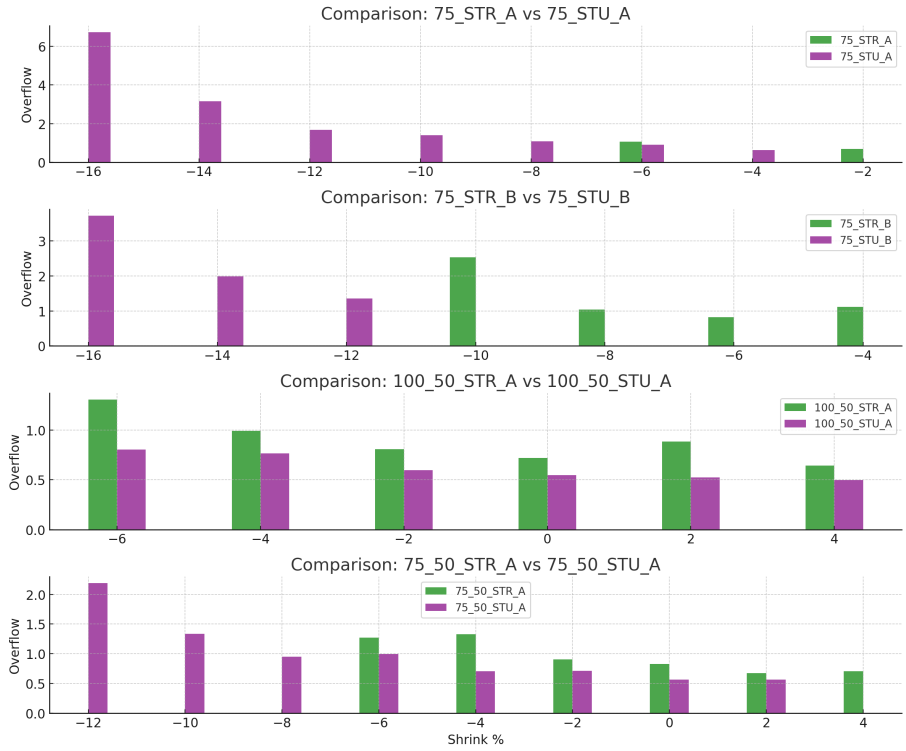

Figure 4.4: Overflow comparison - Stripes vs Stubs

The figure 4.5 presents a chart with the average Overflow/Shrink values. Despite not being a standard figure of merit, it can be useful to discriminate which PDN schemes are more efficient in this context. Consistently with the previous analysis the schemes with better results are characterized by the use of Stubs and a more relaxed Vias presence (option B). The difference between the best and the worst ratio is ≈ 3, meaning that while retaining the same overflow, the area can be reduced by 3 times in relative terms.

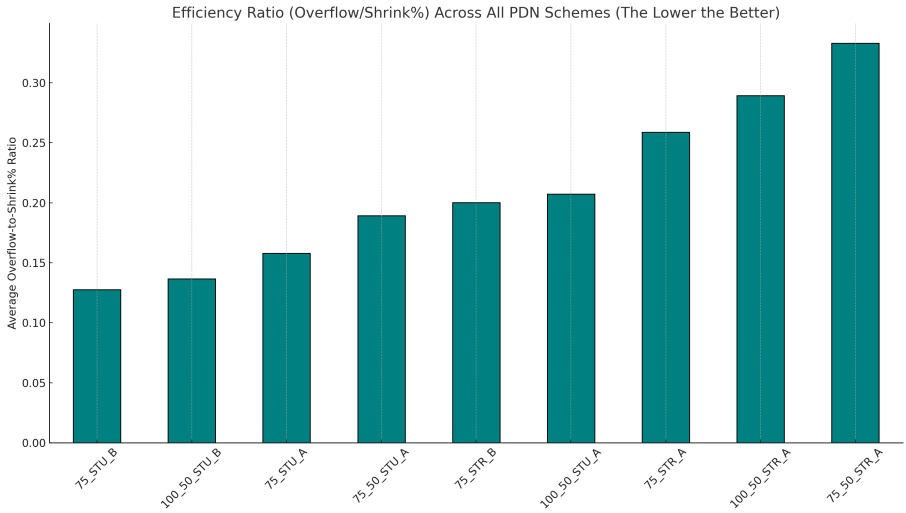

Figure 4.5: Overflow/shrink ratio

4.1.2 Metal congestion

A further analysis can be done on the congestion distribution on the metal stack. The graph in figure 4.6 shows how the congestion distribution changes for the Lower, Mid and Upper metal layers:

- Lower Metals going from 0% to 26% of the metal stack
- Mid Metals going from 26% to 60% of the metal stack
- Upper Metals going from 60% to 100% of the metal stack

The congestion levels are divided in bins going from 0.0 to 1.2, at steps of 0.1. Where 0.0 is no congestion and 1.2 is High congestion. The Y axis represents the average (across the portion of metal stack) fraction of instances falling on that bin.

For this test case the data set is composed of all the PDN schemes with an area reduction of -6% with respect to the reference. This values has been asserted to be the threshold at which most of the schemes shows congested behavior.

- Lower metals: For all the PDN schemes moderate to high congestion is present, with the most frequent bin being 1.1 (high congestion). Stubs schemes have a "flatter" distribution with respect to Stripes. The latter present a peak at $\approx 40\%$ for the 1.1 bin, while Stubs schemes at $\approx 25 - 30\%$.
- Mid metals: For Mid metals the situation is similar, in this case the 0.9 bin is the most frequent, so in general there is less congestion with respect to the lower metals. Here the difference between Stripes and Stubs is less prominent but still present.
- Upper metals: for the last portion of the metal stack the congestion distribution appears more homogeneous, with higher presence of mildly congested areas.

In general, all the samples at -6% area shrink suffer from high congestion at the lower levels. Thus implementing alternative schemes at lower metals could significantly improve congestion.

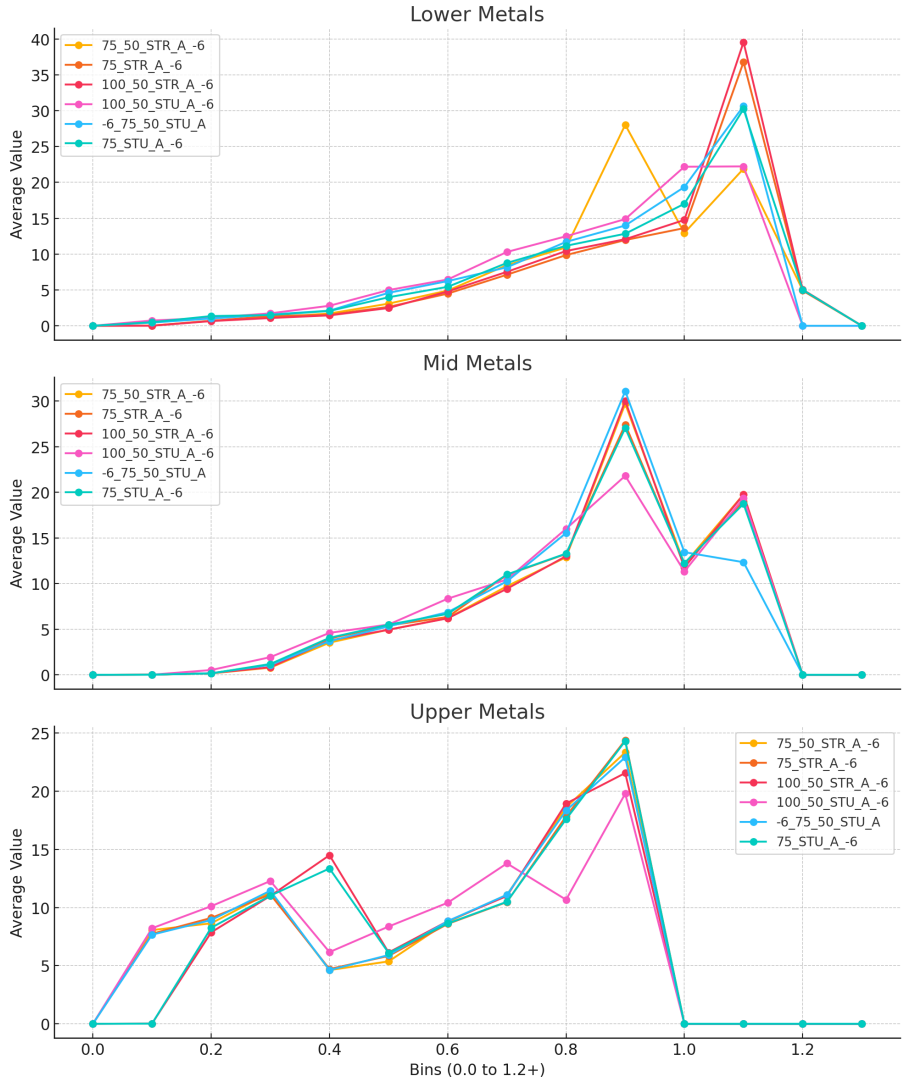

Figure 4.6: Congestion across all metal stack

To emphasize the effect of lower metal congestion, in figure 4.7 congestion distribution at M1 is shown. The data set is composed of an area sweep from +2% to -12% for **75_50_STU_A**. The distribution does not change shape, but the peak at 1.1 bins goes higher as area diminishes. For the case +2% $\approx$ 62% of M1 has high congestion (1.1), goes up to $\approx$ 75% for the case -12%.

An interesting internal comparison can be done on the **75_STU_B** case, that among all is the one with the most relaxed specifications in terms of grid density. In figure 4.8 a comparison among the -14%, -16% and -18% is presented. These area values represents the breaking point of the design. In this case the difference in congestion is mostly present in the Mid metals.

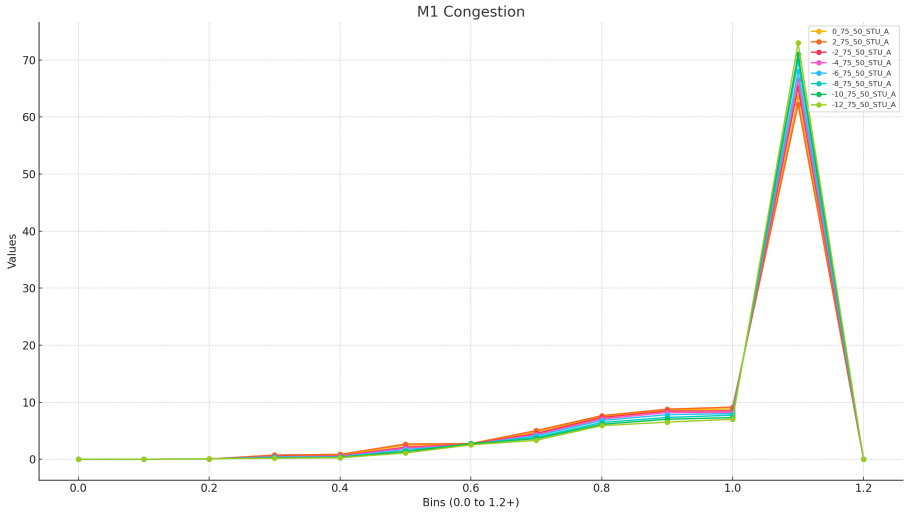

Figure 4.7: Congestion at M1 as function of Area

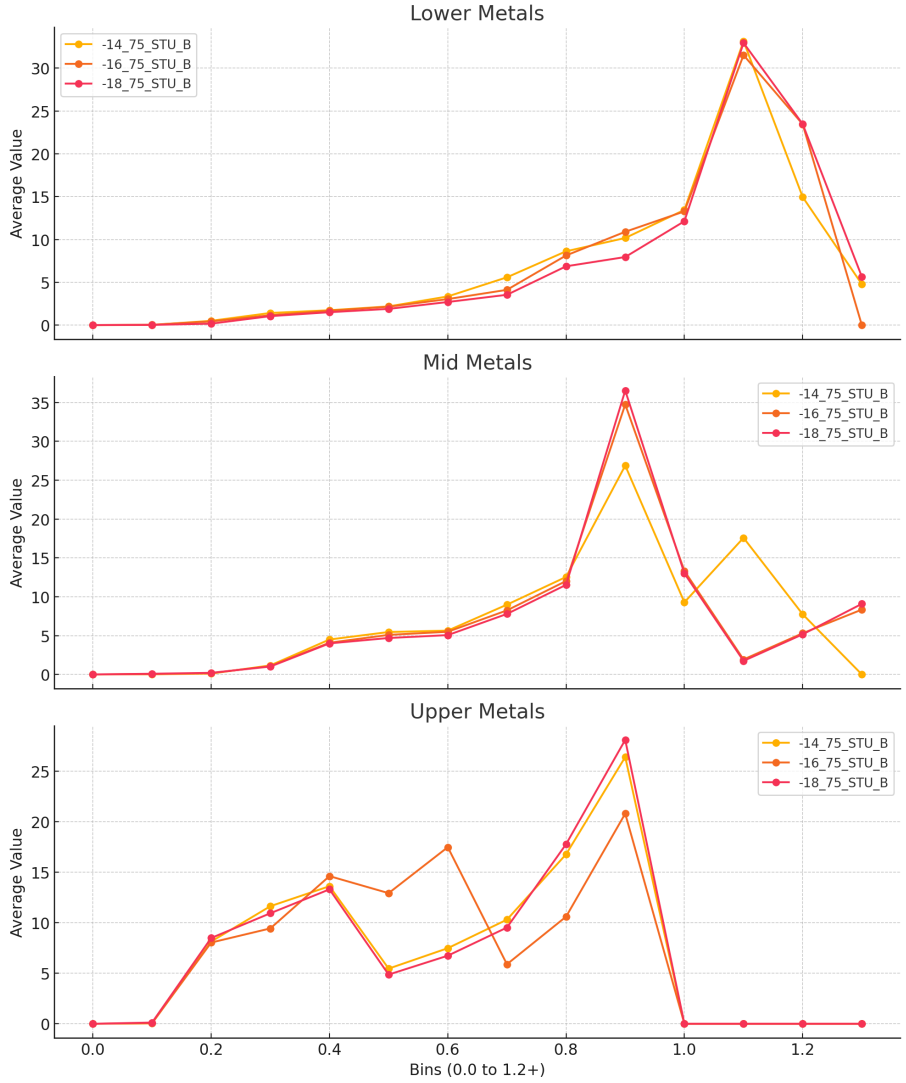

Figure 4.8: Metal Congestion for the 75_STU case

4.1.3 wirelength

An important metric to keep into consideration across different designs is wirelength. wirelength directly impacts propagation delays when signal routing is considered, parasitic Resistance and Capacitance dependence on wirelength can be modeled as:

$$\begin{cases} R = \rho \frac{L}{A} \\ C = \alpha L \end{cases}$$

$$\implies \text{Delay} \approx R \cdot C \propto L^2$$

wirelength also impacts power consumption, because since capacitance increases so does dynamic power:

$$P_{dyn} = \alpha C V^2 f$$

Hence, wirelength minimization is usually aimed by both the designer and the EDA tool.

In figure 4.9 the normalized wirelength values for all the test cases are grouped. The difference between the two extremes is $\approx 11\%$. Two trends can be identified:

- wirelength decreases with area: as the cells are more packed the paths are shorter and so the connections. This trend actually has not to be taken for granted, if the design is shrank too much and congestion rises, some connections will be forced to detour the congested area resulting in a longer point to point connections. In this test cases the latter has not been observed even for the extreme cases in terms of area shrink. A contribution is also given from the power grid, whose length and width scales linearly with the boundary edges.

- wirelength decreases with PDN scheme relaxation: Is clear from the bar-chart that more relaxed PDN schemes have lower wirelengths. This was expected since it is straightforward that a denser power grid will "consume" more wire with respect to a more relaxed one. At this point is not easy to quantitatively estimate the relative weight in terms of wirelength savings for signal and power routing.

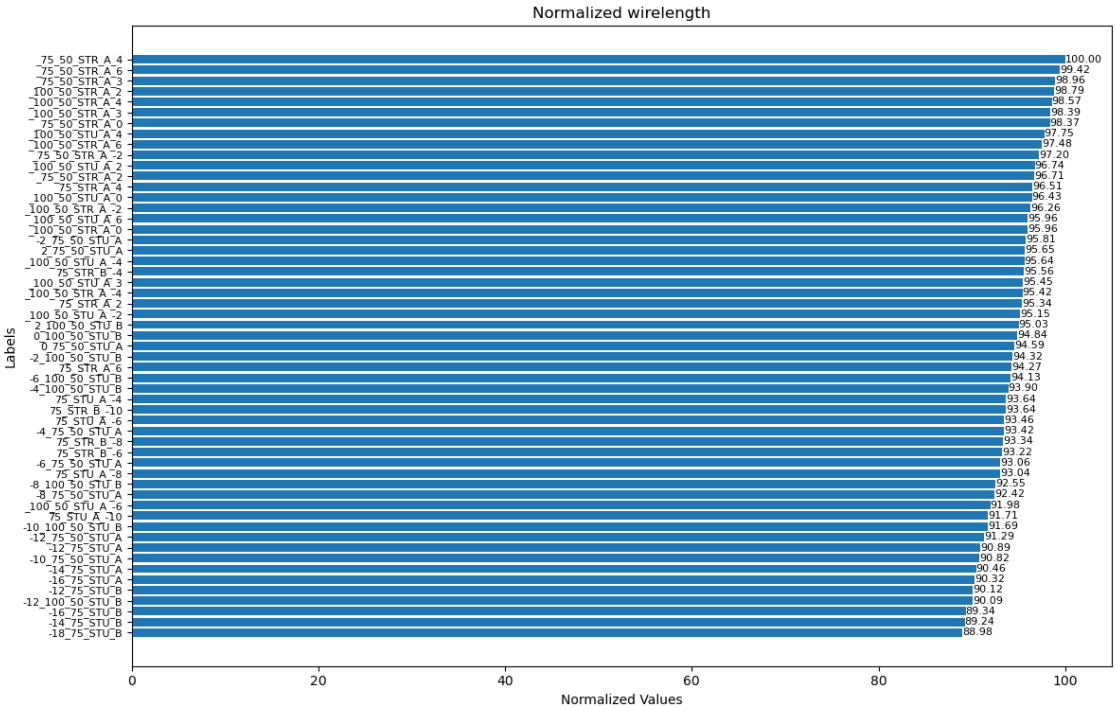

Figure 4.9: Normalized wirelength values

In figure 4.10 the barchart presents the trend of normalized wirelength values vs area, for the 100_500_STU_B case. In this example the trend is monotonic, with a difference in wirelength of $\approx 5.5\%$ over a 14% area difference.

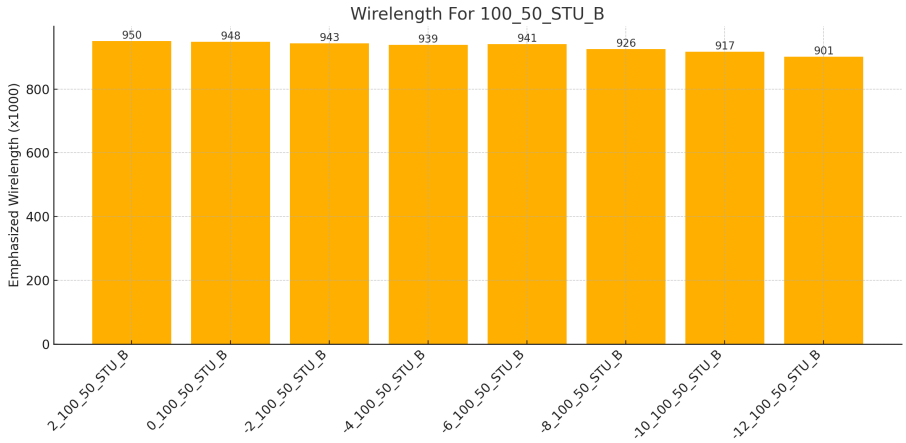

Figure 4.10: Wirelength vs area shrink for 100_500_STU_B

In figure 4.11 the barchart compares wirelength values for the same area but different PDN scheme. In this case 100_50_STR_A vs 100_50_STU_A, the difference is just in the presence of Stubs or Stripes at M1.

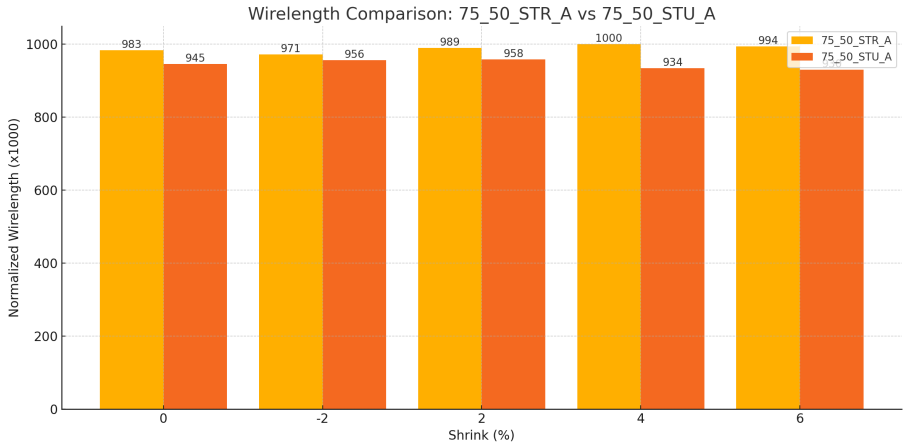

Figure 4.11: wirelength vs area shrink for 100_50_STR_A vs 100_50_STU_A

4.2 Timing

Despite not being the focus of this research, the main STA metrics will be compared in order to have a more complete picture.

In figure 4.12, the heatmap shows the different values of WNS (Worst negative slack) across the different PDN and area configurations.

$$\text{Slack} = \text{Required arrival time} - \text{Actual arrival time}$$

$$WNS = \min(\text{slack})$$

A negative slack implies a timing violation, the WNS evaluates the largest among all violations. This metric is often used in sign-off phase to identify critical paths that needs intervention.

From the WNS heatmap there is not a clear pattern that can be quickly recognized. This is expected since WNS evaluate a single path delay that can be affected by several factors other than area, making the dependence strongly non linear.

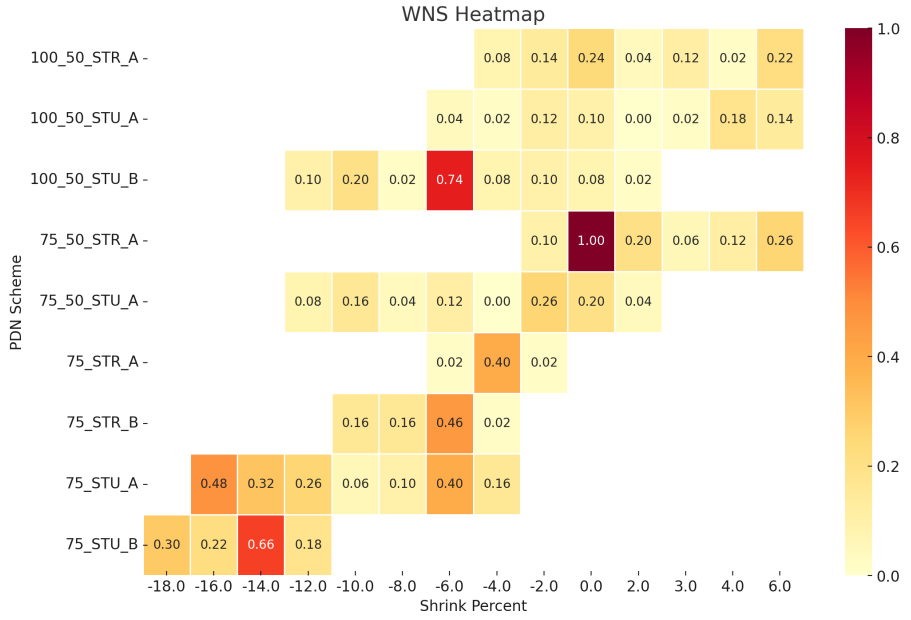

Figure 4.12: WNS vs PDN scheme vs area heatmap

Another important metric is the TNS (Total negative slack), defined as the sum over all the negative slacks:

$$\text{TNS} = \sum_i \text{slack}_i$$

This metric gives a more complete picture of the timing performance of the design, since it actually cumulates over all the paths. In figure 4.13 an heatmap of TNS vs PDN scheme vs area is presented.

In this case a relation with area is more evident, still mild. This is still reasonable, TNS captures a more complete picture timing-wise, indeed the highest values are located in the low area region of the graph, meaning that area reduction is actually worsening timing. At the same time the placing and routing tools tune their actions

also considering timing constraints and effort. For this kind of designs, timing performance are critical and not expandable, hence the timing effort is usually set to high. Concluding, since the tools have an adapting behavior with respect to timing, trying the best to meet the constraints, no immediate considerations can be retrieved from the analysis of this data.

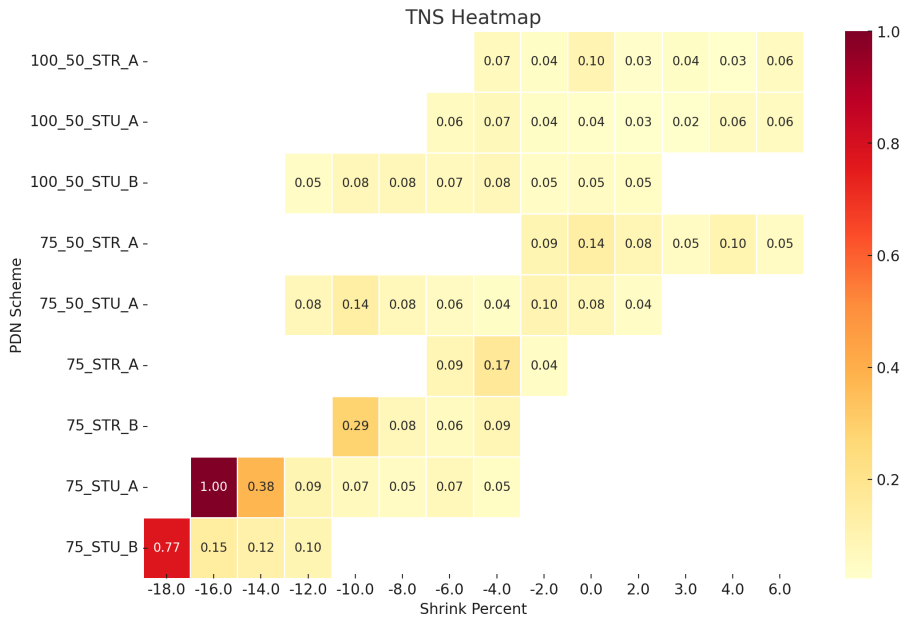

Figure 4.13: TNS vs PDN scheme vs area heatmap

The last metric to be discussed is the NFE (Number of failing instances) 4.14. It is the sum of all instances (paths) having negative slack.

$$\text{NFE} = \sum_{i} \text{Failing instance}_i$$

As expected is strongly tied with TNS, indeed shows an almost identical pattern. This indirectly means that the slack values across the paths are uniformly distributed and widespread.

What can also be noted from the heatmap is that near the breaking point of the design (lowest area value for each one) the derivative with respect to area strongly increases, sign that the routing tool is not capable of handling the constraint anymore.

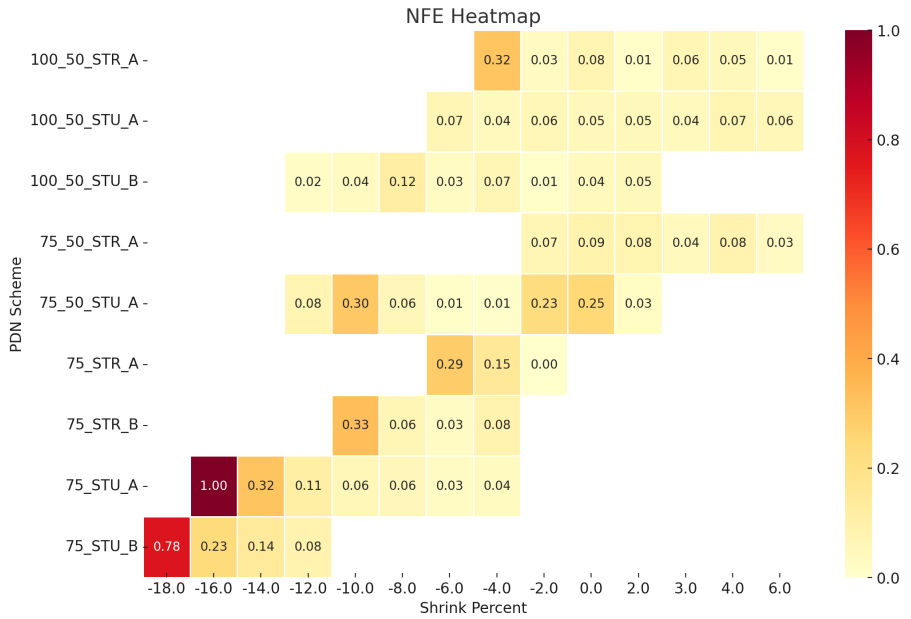

Figure 4.14: NFE vs PDN scheme vs area heatmap

To summarize the analysis, in figure 4.15 a correlation matrix is presented. The values confirms the qualitative considerations:

- area vs WNS : 0.23, very low correlation.
- area vs TNS : -0.53, moderate (negative) correlation.
- area vs NFE : -0.52, moderate (negative) correlation.
- TNS vs NFE : 0.93, high correlation.

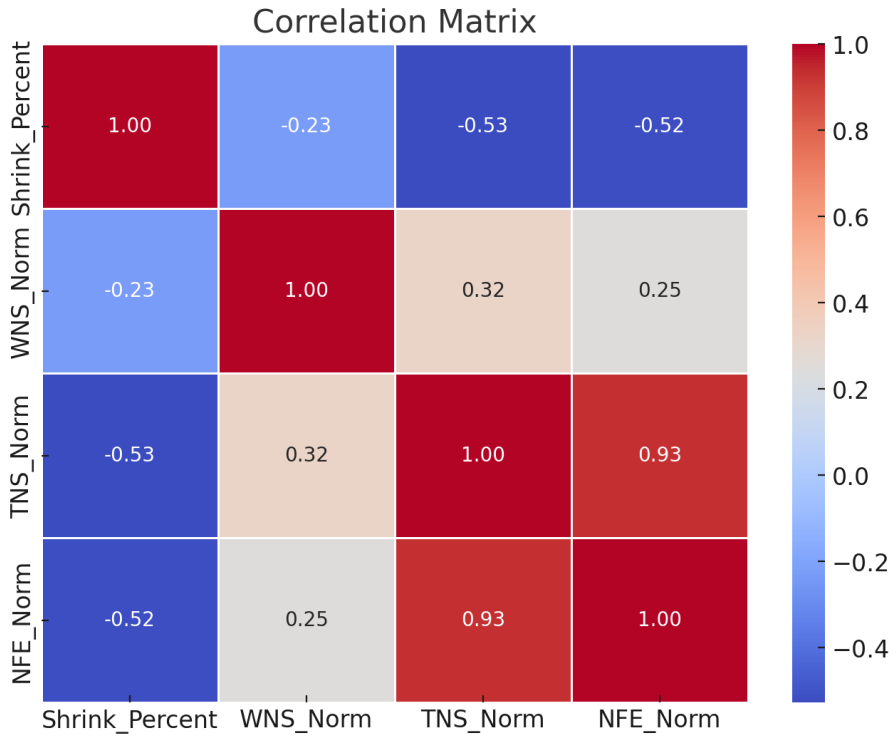

Figure 4.15: Correlation matrix of timing metrics vs area

4.3 Design Rules check violations

One of the main criteria used during the analysis to evaluate the designs is the DRCs count. An internal threshold was set empirically to discriminate which designs were successful and which were not. In figure 4.16 the graphs shows the normalized DRC count for each different trial. The trend shows a proportional increase in DRCs with shrink amount, and an explosion at the design breaking point.

Some outliers are present: 75_STU_A_-6, 100_50_STU_B_-6, 75_50_STU_A_-6. These cases presents a very high DRCs counts outside the trend. The interesting aspect is that all of them share the same area value. This could indicate some kind of geometrical dependence that triggers the routing tool, leading it to generate an high number of DRCs. This problem was not investigated further because considered out of the scope of the research.

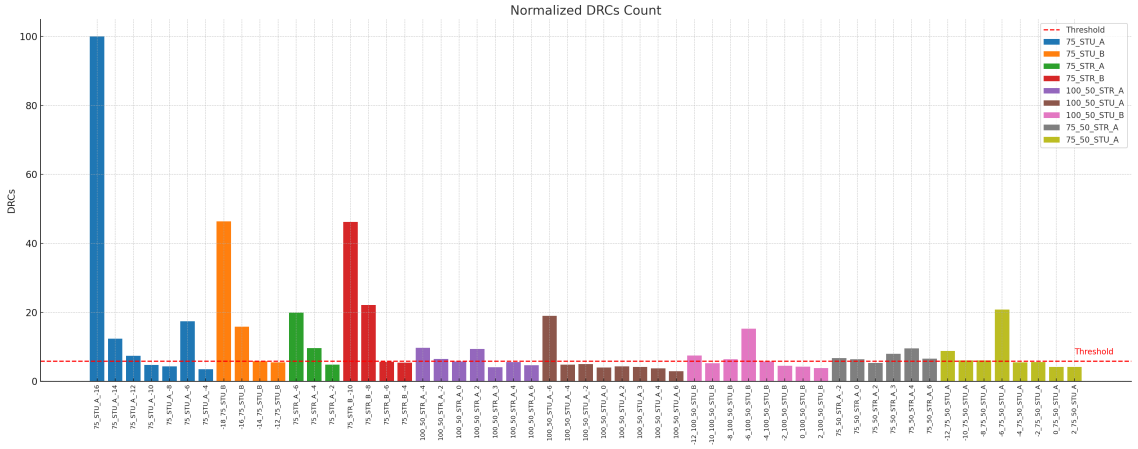

Figure 4.16: Normalized DRCs count vs PDN scheme vs area shrink

In figure 4.17 all the trials that have a DRCs count below the imposed threshold are grouped. As said, such designs are considered successful. In table 4.1, the results are filtered considering the best achieved shrink value (smaller area) for each type of PDN scheme. From this point on, the analysis will focus on this subgroup, evaluating the benefits and downsides fro every specific case.

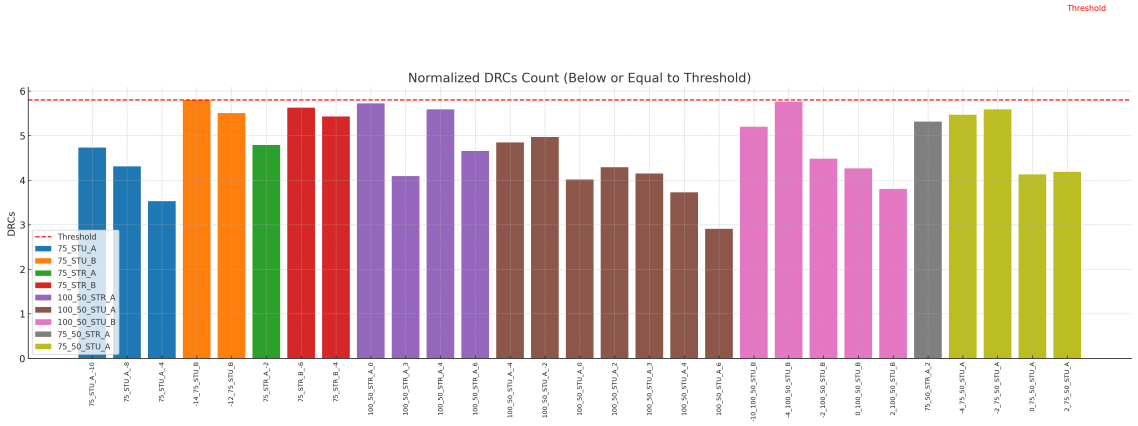

Figure 4.17: Normalized DRCs count - below threshold results

First of all, it is worth noticing that between the best (**75_STU_B**) and the worst (**75_50_STR_A**) result in terms of area, there is a 16% difference (visual in figure 4.18. This underlines the big impact that the PDN scheme has on the area figures of the design. The order in terms of area, follows qualitatively the expectations based on the grade of strictness of the PDN scheme.

Stubs designs dominates the chart, confirming to be a decisive factor, also the choose of the via stack density (A vs B) has non negligible impact.

Comparing the relative differences:

- **STU vs STR** : Up to 8% area difference for the case **75_STU_B** vs **75_STR_B**
- **A vs B** : Up to 6% area difference for the case **100_50_STU_B** vs **100_50_STU_A**
- **STR_A vs STU_B** : Up to 12% area difference for the case **75_STU_B** vs **75_STR_A**
- Pitch : Considering pitch, equal configurations (STR_A) ranks in order of shrink potential **75 single spacing**, **100_50** and **75_50**.

PDN scheme	Best shrink value
75_STU_B	-14%
75_STU_A	-10%
100_50_STU_B	-10%
75_STR_B	-6%
100_50_STU_A	-4%
75_50_STU_A	-4%
75_STR_A	-2%
100_50_STR_A	0%
75_50_STR_A	+2%

Table 4.1: Best shrink vs PDN scheme

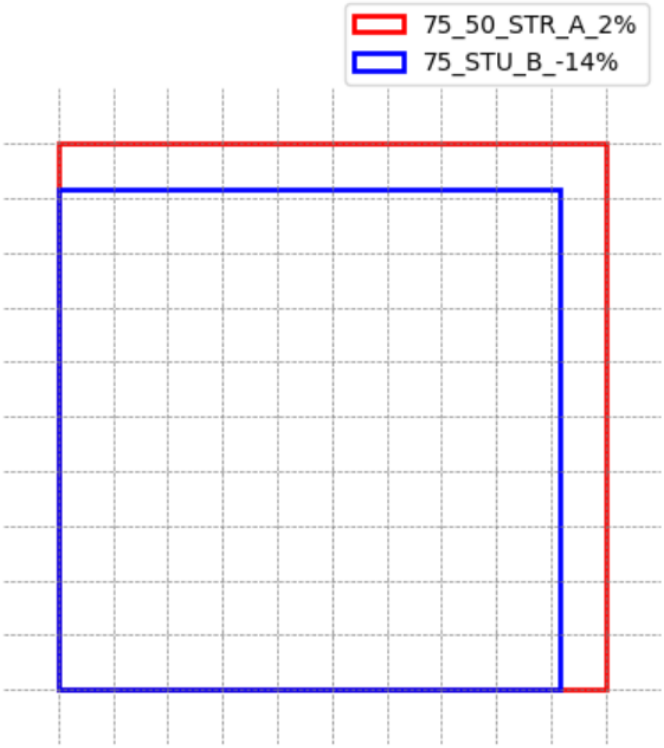

Figure 4.18: Visual comparison between the best and worst result in terms of area

4.3.1 utilization factor

The Utilization factor is defined as the ratio of the area occupied by the standard cells to the total available core area, it is usually expressed as a percentage. It is a critical metric to evaluate the efficiency in terms of Area of the design.

$$\text{Utilization Factor} = \frac{\text{Area occupied by standard cells}}{\text{Total core Area}}$$

Usually:

- **Low utilization** ($< 50\%$): Indicates underutilized space, may lead to inefficient area usage but provides additional flexibility for routing and buffer insertion.
- **High utilization** ($> 80\%$): Indicates high packing density, can cause routing congestion, timing violations and power issues. Often lead to unroutable designs.
- **Optimal range**is typically considered between 60% and 80%, to balance area efficiency and routing feasibility.

Utilization factor has a double utility, it can be used at the floorplanning stage to preliminary evaluate the area required by a block, or like in this case to discriminate different designs in terms of area efficiency at post-route stage.

In figure 4.27 the plot presents the **Normalized** utilization values for the different designs. Utilization grows inversely proportional to area, almost linearly. This is expected since the area boundary diminishes but the netlist area remains the

same. The small fluctuations for which close area value have slightly different utilization ,are due to the actual physical implementation. The total actual cell area could vary due to buffer insertion and different std cell sizing, since each design has slightly different timing necessities.

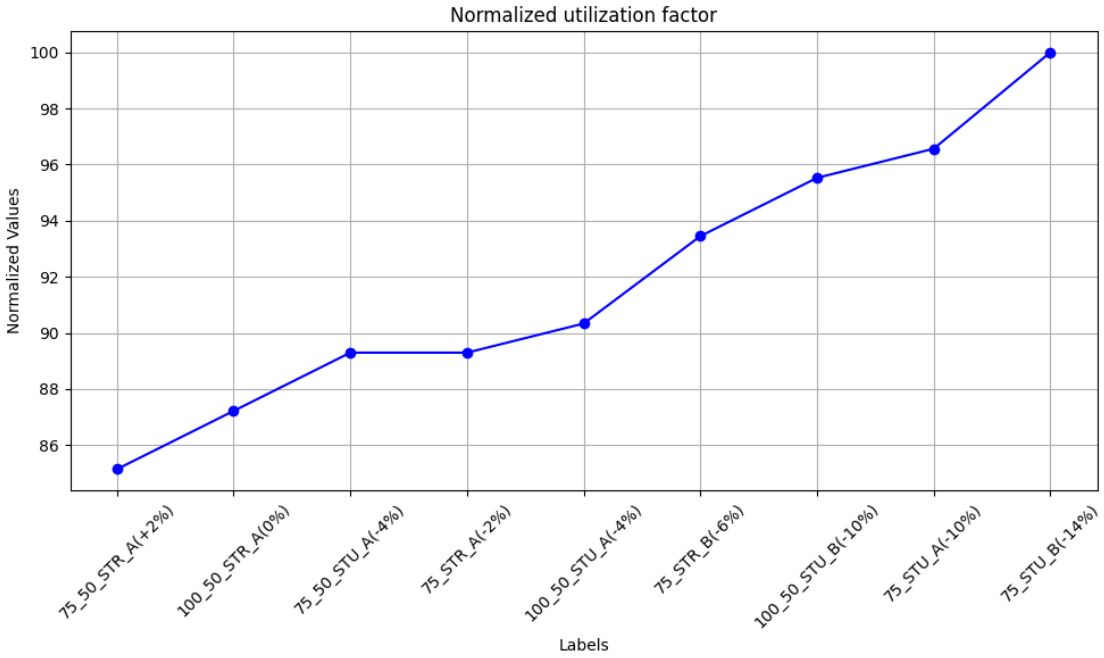

Figure 4.19: Utilization factor vs area shrink

4.4 Leakage

Leakage power has become a critical design constraint in advanced technology nodes, particularly as devices scale down to nanometer dimensions. Variations in physical implementations—such as placement strategies, cell library choices, and routing optimizations—can lead to significant differences in leakage characteristics, even for the same logical design. These variations impact not only power efficiency but also thermal performance and overall reliability.

In figure 4.20 a bar chart illustrates the normalized leakage results for all the selected designs. Most of the designs shows similar values, two outliers are present: **75_STU_A(-10%)** has a leakage figure $\approx$ 30% higher than the average, while **75_STR_B(-6%)** has $\approx$ 14% less leakage.

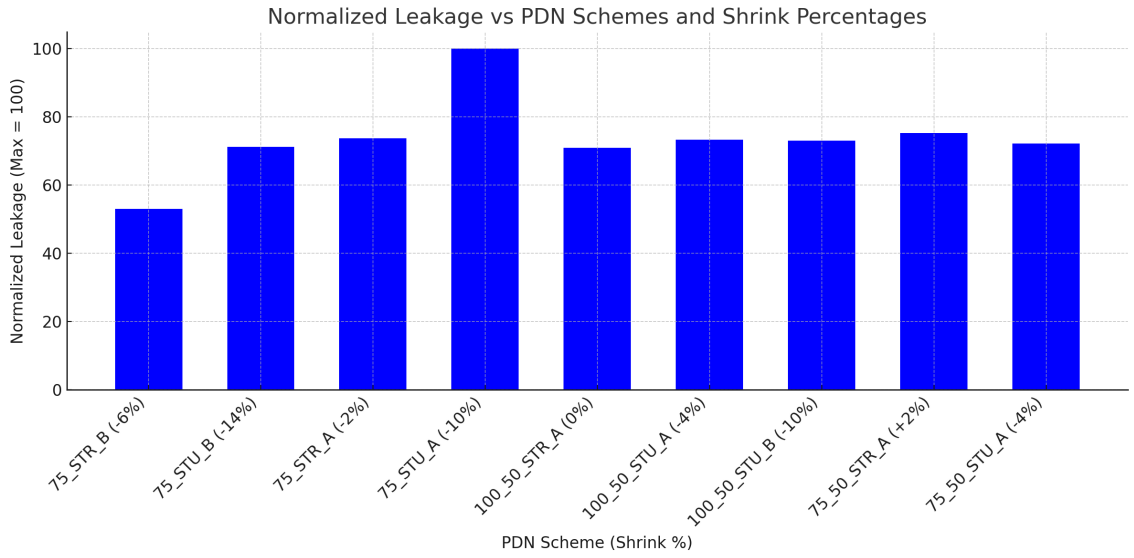

Figure 4.20: Leakage figures for beast shrink value results

Modern standard cell libraries include multi-threshold voltage (multi-Vt) cells to balance performance, power, and leakage in physical design. The threshold voltage impacts both switching speed and leakage current, making it a key parameter in power optimization strategies. The leakage current depends exponentially on the threshold voltage as:

$$I_{leak} = I_0 \cdot e^{\frac{V_{gs} - V_{th}}{nV_T}}$$

$$V_T = \frac{kT}{q}$$

Where:

- I_0: Pre-exponential constant dependent on device parameters.
- V_{gs}: Gate-to-source voltage.
- V_{th}: Threshold voltage.
- n: Subthreshold swing coefficient (process-dependent, typically 1–2).

- V_T: Thermal voltage, given by $\frac{kT}{q}$.
- k: Boltzmann constant ($1.38 \times 10^{-23}\, J/K$).
- T: Temperature in Kelvin.
- q: Charge of an electron ($1.6 \times 10^{-19}\, C$).

and so the static power consumption $\implies$

$$P_{static} = V_{DD} \cdot I_{leak}$$

Where:

- P_{static}: Static power dissipation (Watts).
- V_{DD}: Supply voltage (Volts).
- I_{leak}: Leakage current (Amperes).

Due to the exponential dependency of I_{leak} on V_{th}, small changes in the latter have a tangible impact on the leakage figure, for example:
Changes of V_{th} by $\pm 5\%$ and $\pm 10\%$ would lead to the following results :

Change in V_{th}	Relative Leakage Change
+5%	-32.06%
-5%	+47.18%
+10%	-53.83%
-10%	+116.61%

Table 4.2: Relative Leakage Change for V_{th} Variations

1. Increasing V_{th} (+5% and +10%) significantly reduces leakage current, as leakage depends **exponentially** on V_{th}.
2. Decreasing V_{th} (-5% and -10%) causes a sharp increase in leakage, with a 10% reduction resulting in a **116.61% rise** in leakage current.
3. The example highlight the sensitivity of leakage power to small variations in V_{th}, underscoring the importance of careful threshold voltage selection and the use of **multi-Vt libraries** for leakage optimization.

At the same time changes on the threshold voltage significantly impact timing: The delay (t_{delay}) of a MOS transistor is inversely proportional to the drain current (I_{on}), which depends on V_{th} as follows:

$$t_{delay} \propto \frac{C_{load} \cdot V_{DD}}{I_{on}}$$

The ON current (I_{on}) is given by:

$$I_{on} \propto (V_{DD} - V_{th})^{\alpha}$$

where:

- C_{load}: Load capacitance (F)..
- α: Empirical factor (1–2 based on process node).

Summarizing:

1. Leakage-Timing Trade-off: Lowering V_{th} reduces the delay (t_{delay}) by increasing I_{on}, improving performance. However, this comes at the cost of higher leakage current (I_{leak}), as shown below:

$$t_{delay} \downarrow \quad \Rightarrow \quad I_{on} \uparrow \quad \Rightarrow \quad I_{leak} \uparrow$$

2. High-Vt vs Low-Vt Cells: - High-Vt cells: Lower leakage but slower switching (higher t_{delay}). Suitable for non-critical paths. - Low-Vt cells: Higher leakage but faster switching (lower t_{delay}). Suitable for timing-critical paths.

3. Design Optimization: Multi-threshold libraries provide flexibility by allowing designers to balance timing and leakage across different paths, optimizing performance and power.

Figure 4.21 shows the voltage threshold distribution for the standard cells of each design. In this case only **Low vth** and **Ultra low vth** cells have been used.

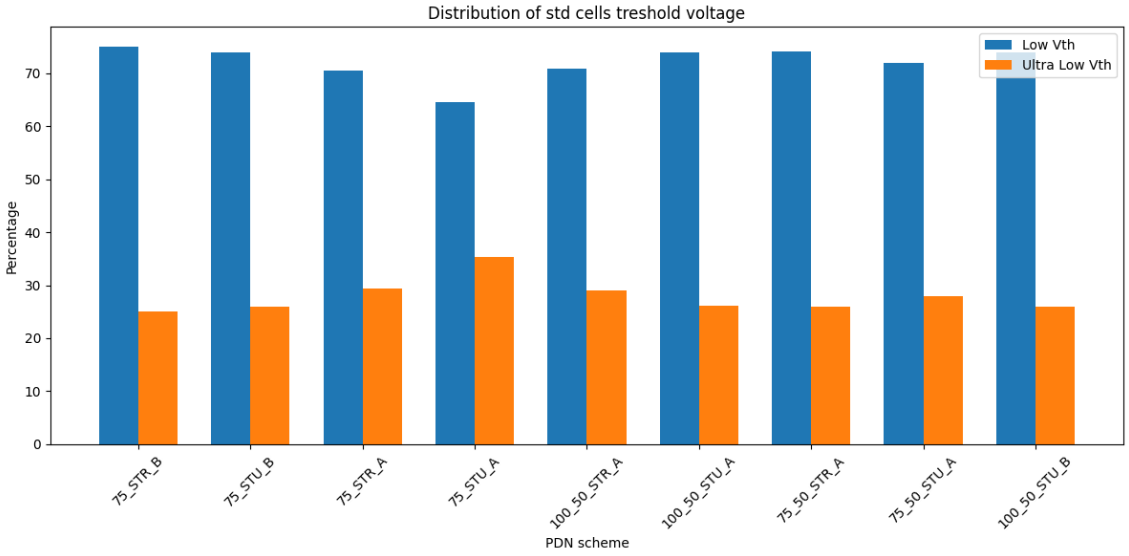

Figure 4.21: Threshold voltage standard cell distribution across all the designs

Modern advanced designs, targeting performance are more prone to the use of such cells for several factors:

- **Design aspects:
- Modern designs (e.g., CPUs, GPUs, AI accelerators) operate at multi-GHz frequencies, demanding low-delay logic paths
- Aggressive scaling policies leads to more stringent timing necessities.
- **Technological aspects:

- Increase in parasitics of interconnections: Narrower interconnects and thinner metal layers lead to higher resistance (R) and capacitance (C), leading to higher propagation delays.
- Shrinking transistor dimensions weaken gate control, increasing leakage and lowering the effective drive current for a given threshold voltage.

Results of figure 4.21 shows a strong correlation with the voltage threshold distribution and the actual leakage figures:

- **75_STR_B** Has an higher count of Low-vth cells (less leaky then ultra low) and presents the lowest leakage figure.
- **75_STU_A** Has an higher count of Ultra-Low-vth cells (more leaky) and presents the highest leakage figure.

The reasons behind the difference in the distributions could be attributed to several factors. Either high congestion is causing more timing needs in this specific designs, or the tool triggered a different strategy due to multi factorial dependencies.

4.5 IR drop

In modern VLSI designs, particularly at advanced technology nodes (e.g. $< 7nm$), power integrity is a critical concern. Among the key challenges affecting power delivery networks (PDNs) is IR drop—the voltage drop caused by resistive losses in the power delivery network as current flows through it.

As supply voltages continue to scale down in pursuit of lower dynamic power consumption, the available noise margins shrink, making designs increasingly sensitive to IR drop. Even minor variations in delivered voltage can lead to timing violations, functional failures, or increased leakage currents, particularly in timing-critical paths.

IR drop refers to the voltage drop experienced along the power (V_{DD}) and ground (V_{SS}) paths due to the resistance (R) and current (I) in the power grid:

$$V_{drop} = I \cdot R \tag{4.1}$$

This drop reduces the effective supply voltage seen by the cells, impacting switching speeds and signal integrity.

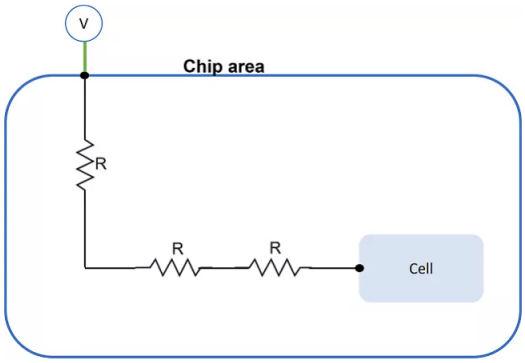

Figure 4.22: IR drop

Aggressive technology scaling has exacerbated IR drop issues due to:

- Increased transistor densities and lower operating voltages require **higher currents** to maintain performance.
- Narrower interconnects lead to **higher resistances** in the power grid, increasing the likelihood of IR drop.
- Lower supply voltages (e.g., 0.8V–0.6V) result in **tighter noise margins**, making circuits more sensitive to voltage variations.
- The use of **power gating** introduces regions with **variable current consumption**, leading to localized IR drop hotspots.

IR drop is typically divided into two categories:

4.5.1 Static IR Drop

- Caused by **leakage currents** (e.g. gate leakage, subthreshold leakage, biasing) across the chip during steady-state operation.

- Results from **resistive losses** in the power grid.

Result for mean static IR drop are shown in figure 4.23.

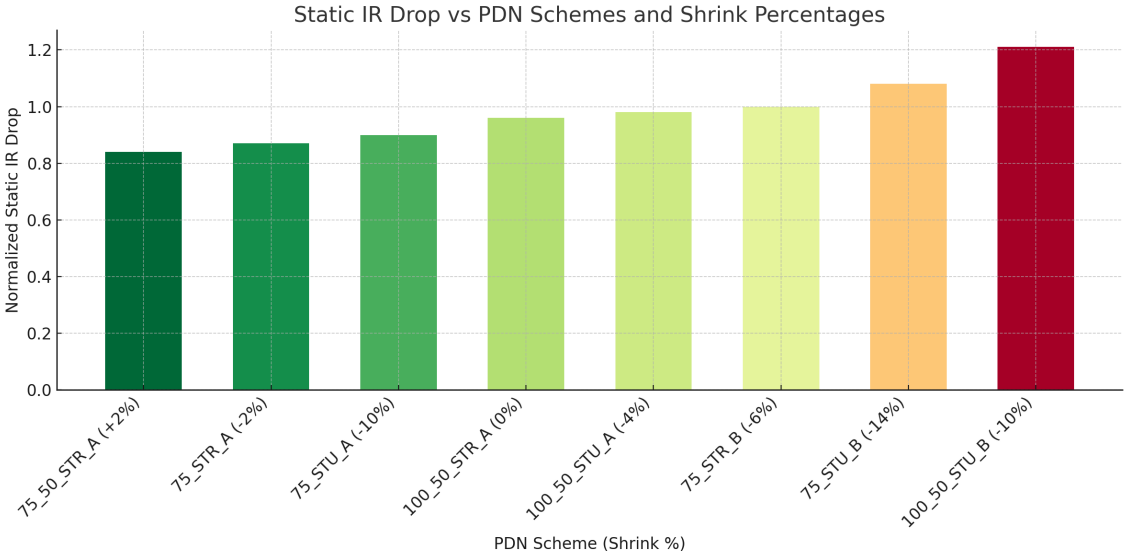

Figure 4.23: Static IR drop results vs PDN scheme

The difference between the best (**75_50_STR_A**) and worst (**100_50_STU_B**) case is $\approx 45\%$. The trend is expected, more stringent PDNs suffer less from static IR drop, having a denser grid leads to lower effective resistance and more uniform current distribution.

In figure 4.24 a scattering diagram shows how the appears in the Static IR drop vs Shrink plane. The aim of this chart is to evaluate the optimization level of each design.

Designs that fall closer to the lower-left corner of the plot exhibit low IR drop and high shrinkability, indicating an optimal trade-off between power integrity and scaling potential. Conversely, points further to the upper-right corner indicate designs where mild scaling has resulted in poor power integrity.

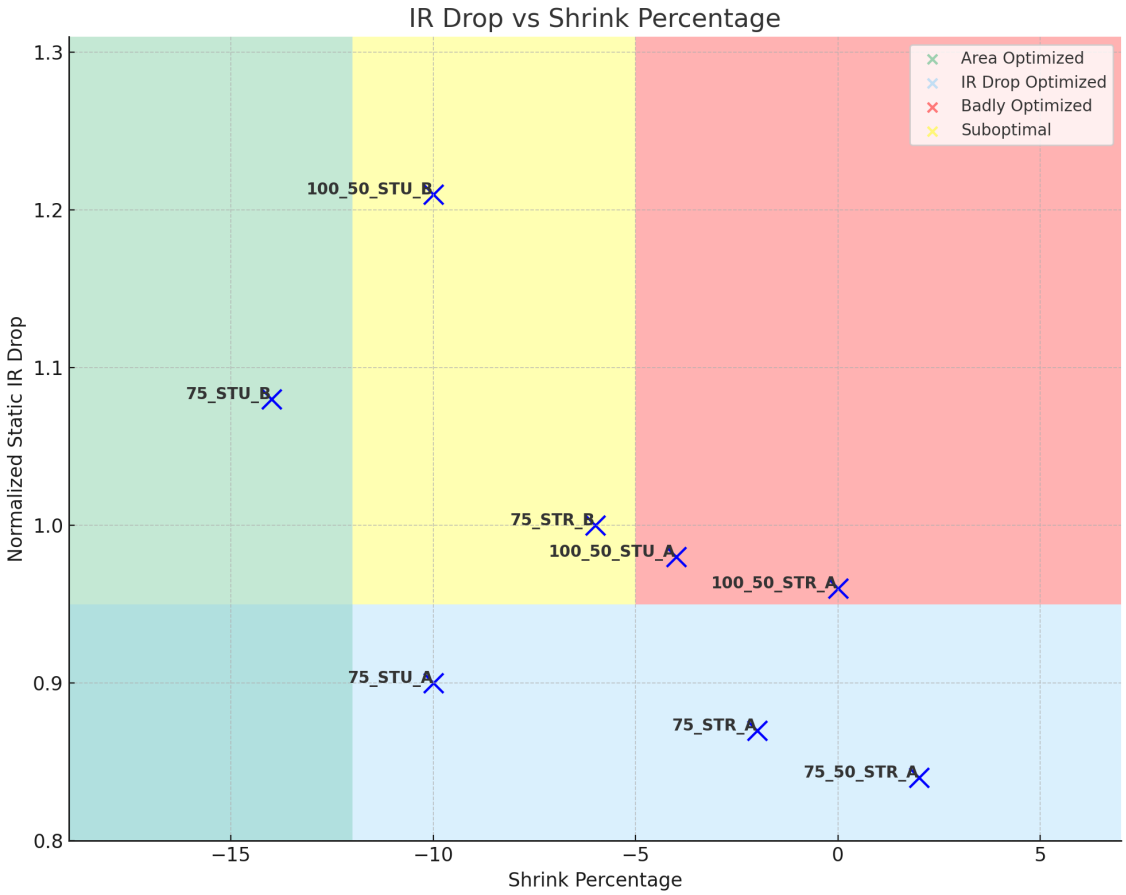

Figure 4.24: IR drop vs Area optimization chart

4.5.2 Dynamic IR Drop

- Caused by **transient current surges** during clock switching or simultaneous switching of large logic groups.
- Amplified by **inductive effects** and **capacitive coupling** in dense designs.

Dynamic IR drop reduces the effective supply voltage (V_{eff}) seen by logic cells during switching events:

$$V_{eff} = V_{DD} - V_{drop} \tag{4.2}$$

Lower V_{eff} reduces the drive current (I_{on}), which impacts the gate delay (t_{delay}):

$$t_{delay} \propto \frac{C_{load} \cdot V_{eff}}{I_{on}} \tag{4.3}$$

A lower V_{eff} leads to slower transitions, which can result in **setup violations** (signals fail to arrive in time for sampling at the clock edge).

It can also cause **hold violations** if slower signals prevent a path from stabilizing before the next clock edge.

Figure 4.25 shows the results of mean Dynamic IR drop for the best four cases in terms of area. Here again, the trend follows the expectations of the qualitative evaluation on how strict the PDN schemes are, being **STU_B** configurations the

most relaxed one. The difference between the worst (**100_50_STU_B** and the best (**75_STR_B** case is ≈ 18%. The other metrics present in the chart are:

- **Outliers count**: Number of instances suffering from a dynamic IR drop > 10% of VDD.
- **Worst case dynamic IR drop**

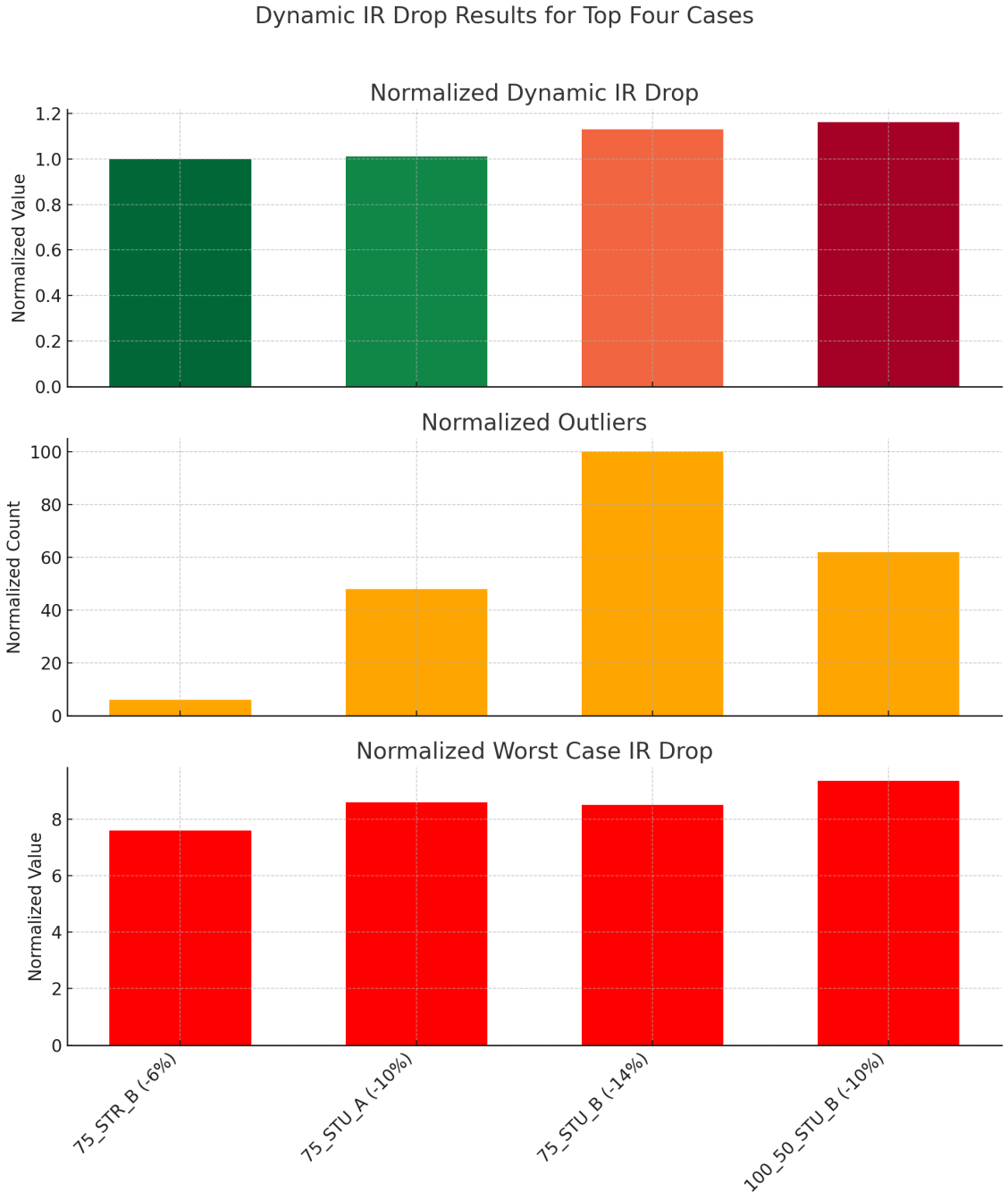

Figure 4.25: Dynamic IR drop results vs top four PDN schemes

To analyze the impact of dynamic IR drop on timing, we consider a scenario where the effective supply voltage (V_{eff}) is 10

- Nominal supply voltage: $V_{DD} = 1.0, V$
- Effective supply voltage: $V_{eff} = 0.9, V$ (10
- Load capacitance: $C_{load} = 1 \times 10^{-15}, F$
- Threshold voltage: $V_{th} = 0.3, V$
- Empirical constant: $\alpha = 1.5$

The gate delay is proportional to the ratio of load capacitance and drive current:

$$t_{delay} \propto \frac{C_{load} \cdot V_{eff}}{I_{on}} \tag{4.4}$$

where the drive current depends on V_{eff} as:

$$I_{on} \propto (V_{eff} - V_{th})^{\alpha} \tag{4.5}$$

- Nominal Delay at V_{DD}:
  Using $I_{on} \propto (V_{DD} - V_{th})^{\alpha}$, the delay is computed as:

$$t_{delay,nominal} \propto \frac{C_{load} \cdot V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \tag{4.6}$$

- Delay at Reduced V_{eff}:
  Similarly, at V_{eff}:

$$t_{delay,eff} \propto \frac{C_{load} \cdot V_{eff}}{(V_{eff} - V_{th})^{\alpha}} \tag{4.7}$$

- Percentage Increase in Delay:

$$\Delta t_{delay} = \frac{t_{delay,eff} - t_{delay,nominal}}{t_{delay,nominal}} \times 100 \tag{4.8}$$

  Substituting values, the delay increases by approximately **13.41%** when V_{eff} is 10% lower than V_{DD}.

This result shows the **nonlinear dependence of gate delay on supply voltage** and highlights how even a **small IR drop** can lead to **timing violations** in high-performance designs.

Figure 4.26 shows the count of paths in the PDN that violates the maximum resistance threshold.

These violations indicate regions in the PDN with insufficient metal density, fewer parallel paths, or inadequate vias, which lead to higher resistive losses.

Areas with multiple path resistance violations are more likely to experience localized **IR drop hotspots**, leading to potential timing violations.

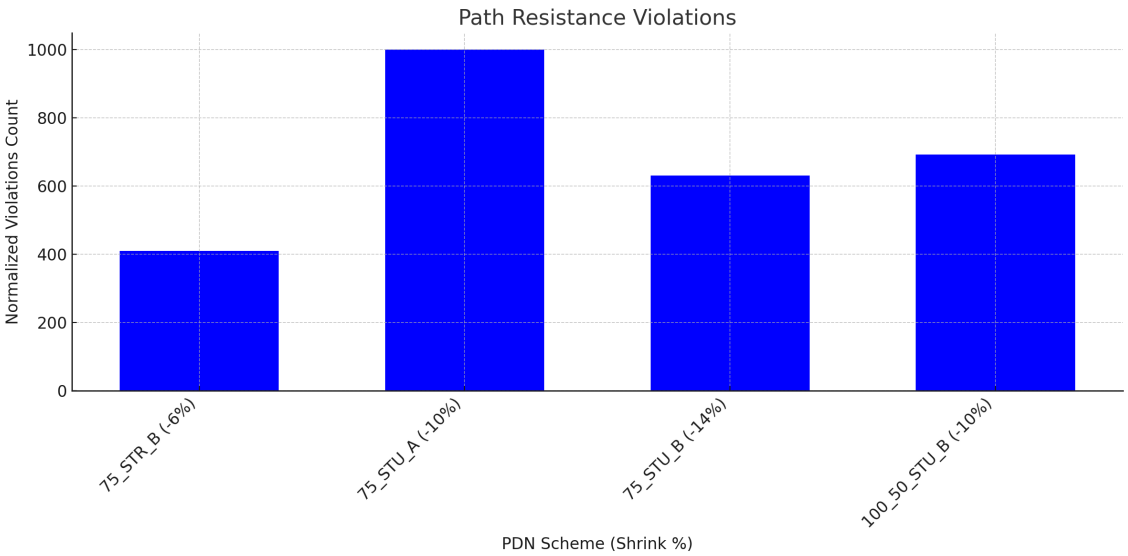

Figure 4.26: Maximum path resistance violations vs top four PDN schemes

4.6 Top results comparison with older tech node

Table 4.3 summarizes the comparison between different PDN schemes implemented in the current technology node (Technode A) and their counterparts in the older technology node (Technode B). Key metrics such as scaling factors, utilization efficiency, and leakage power are analyzed to assess the performance gains and trade-offs achieved with scaling.

4.6.1 Scaling Factor Analysis

The scaling factor indicates the relative size reduction achieved in the newer node compared to the older node.

- 75_STR_B shows a 3% improvement in scaling (0.97x), reflecting modest area optimization with respect to technode B, while preserving the same PDN characteristics.
- The 100_50_STR_A scheme maintains a 1.03 scaling factor, indicating no area benefits over technode B.
- More aggressive scaling schemes, such as 75_STU_B, achieve up to a 12% reduction (0.88x), leveraging a more relaxed PDN density to enable area savings.

  In the ideal scenario, reducing the technology node leads to a proportional decrease in the physical dimensions of the transistors and associated circuit elements. If we assume that the scaling factor is k (where k is the ratio of the new node's minimum feature size to that of the old node), then:

  - **Linear Dimensions:** Each linear dimension is reduced by a factor of k.
  - **Area Reduction:** Since the area is a two-dimensional measure, the ideal area reduction factor is k^2.

  For example, with $k = 0.7$, the expected area reduction is calculated as:

  $$k^2 = (0.7)^2 = 0.49$$

  This suggests that, under ideal scaling conditions, the chip area would shrink to approximately 49% of its original value—corresponding to a 51% reduction in area.

  While the theoretical reduction is significant, practical design challenges often temper these ideal expectations:

  - **Design Rules and Overheads:** Not all circuit components scale uniformly. Elements such as routing channels, bonding pads, and especially PDN structures have fixed dimensions or minimal requirements that do not scale with k. These non-scalable components reduce the overall area savings.

– **PDN Specifics:** The PDN must deliver reliable power across the chip. As nodes shrink, current densities increase, and design constraints for PDN routing (such as minimum widths and spacing) may not scale as aggressively. This can result in a scenario where the effective area reduction is less than the ideal figure.

– **Parasitic Effects and Performance Margins:** Advanced nodes face increased parasitic resistances and capacitances. To mitigate these effects, designers may add extra margins or buffers, further reducing the area efficiency gains expected from scaling.

4.6.2 Normalized Utilization

The utilization metric provides insights into the resource efficiency achieved with each PDN scheme: The trend among the technode A is clear, as stated in previous analysis. Comparing the one to one results with technode B, the latter shows higher utilization figures with respect to A. This is can be attributed to the fact that higher transistor density leads to more complex routing, also metal lines do not scale geometrically as standard cells, this could create bottlenecks preventing to achieve higher utilization rates in Technode A due to higher routing congestion.

4.6.3 Leakage Power Trends

- Designs such as 75_STR_B in Technode A show a 41% reduction in leakage compared to their Technode B counterparts (from 97 to 57),attributed to technological and library improvements.

- Leakage improvements in newer nodes are non-linear, influenced by cell libraries, grid density, and IR drop sensitivity, as observed in previous sections.

PDN Scheme	Scaling Factor	Utilization (%) (Normalized)	Leakage	Dynamic IR Drop
75_STR_B (Technode B)	1.00	100.00	97	N/A
100_50_STR_A (Technode B)	1.03	97.00	100	N/A
75_STU_B	0.88	96.31	77	1.13
75_STU_A	0.93	93.00	108	1.02
100_50_STU_B	0.96	92.00	79	1.17
75_50_STU_A	0.96	86.00	78	1.0
75_STR_B	0.97	90.00	57	N/A
100_50_STU_A	0.99	87.00	79	N/A
75_STR_A	1.01	86.00	80	N/A
100_50_STR_A	1.03	84.00	77	N/A
75_50_STR_A	1.05	82.00	81	N/A

Table 4.3: Results summary Comparison Table

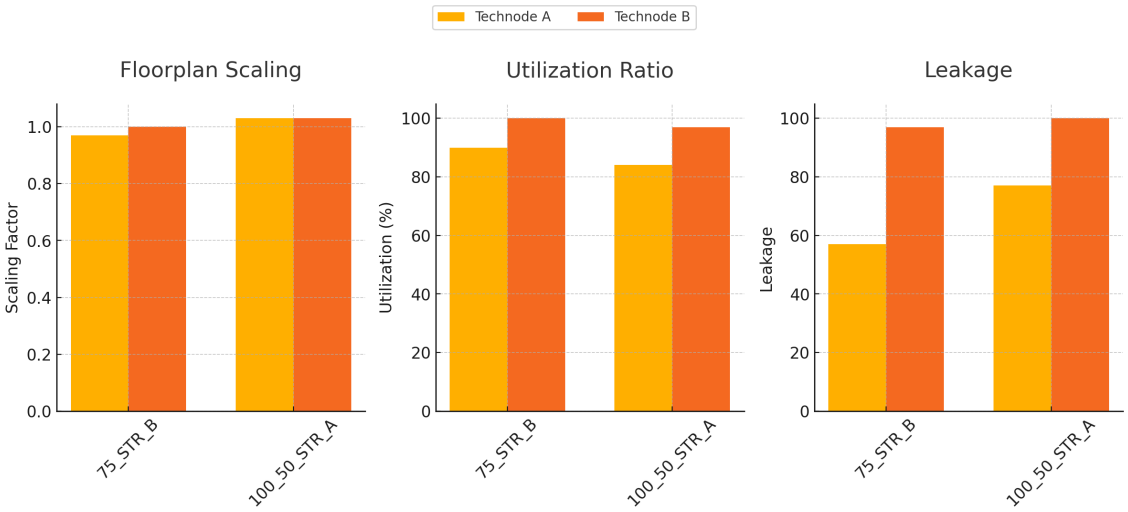

Figure 4.27: One to One comparison - Technode A vs B

Chapter 5

Conclusion

5.1 Moore's Law Slowdown and the Role of PDN in Advanced Nodes

Moore's Law has long served as a benchmark for progress in the semiconductor industry, originally predicting that the number of transistors on a chip would double approximately every two years. This exponential scaling has driven advancements in performance, power efficiency, and overall chip functionality. However, as technology nodes have advanced into the deep submicron regime, the practical realization of Moore's Law has faced significant challenges. Notably, the evolution of the power delivery network (PDN) structures is one such challenge that has reshaped expectations.

In earlier technology nodes, the PDN was relatively straightforward, as lower power densities allowed for simpler designs. Today, however, advanced nodes demand more sophisticated PDN architectures to handle increased current densities and tighter integration. As a result, a design effort must be done in order to retrieve as much scaling potential as possible from technological advancements.

This study concludes that the Power delivery Network design and implementation is of fundamental importance in aggressively scaled technological nodes. Most of the PPA figures are influenced by the PDN structure, as seen from the comparative results, scaling results between different tech-nodes are strongly influenced by the PDN implementation.

5.2 Cost Implications of Area Shrinkage: Impact of PDN Optimization

In the semiconductor industry, the cost of a die is often estimated as a function of its area. The cost depends on the wafer cost, the number of dies per wafer, and the yield. The cost of a die can be estimated using the following formula:

$$\text{Die Cost} \approx \frac{\text{Wafer Cost}}{\text{Number of Dies per Wafer} \times \text{Yield}}$$

The number of dies per wafer is approximated by:

$$\text{Number of Dies per Wafer} \approx \frac{\pi \times (\text{Wafer Diameter}/2)^2}{\text{Die Area}} - \frac{\pi \times \text{Wafer Diameter}}{\sqrt{2 \times \text{Die Area}}}$$

The yield is influenced by defects in the manufacturing process and can be modeled as:

$$\text{Yield} \approx e^{-\text{Defect Density} \times \text{Die Area}}$$

Combining the above, the die cost can be expressed as:

$$\text{Die Cost} \approx \frac{\text{Wafer Cost}}{\left(\frac{\pi \times (\text{Wafer Diameter}/2)^2}{\text{Die Area}} - \frac{\pi \times \text{Wafer Diameter}}{\sqrt{2} \times \text{Die Area}}\right) \times e^{-\text{Defect Density} \times \text{Die Area}}}$$

- Wafer Cost: Cost of the entire wafer.
- Wafer Diameter: Diameter of the wafer (e.g., 300 mm).
- Die Area: Area of a single die.
- Defect Density: Average number of defects per unit area on the wafer.
- Yield: Fraction of functional dies on the wafer.

This analysis provides a rough estimate of the die cost as a function of its area. Actual costs may vary due to additional factors such as process complexity, packaging, and testing.

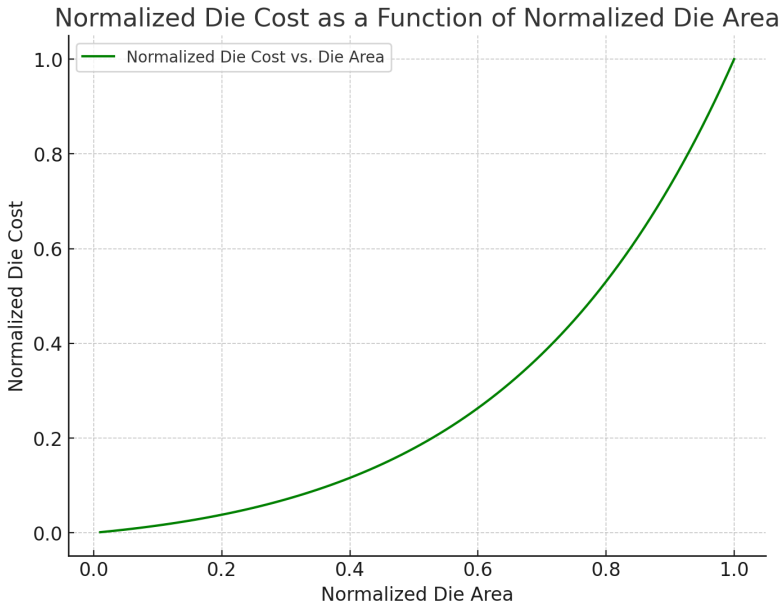

Figure 5.1: Die cost Vs Area

From this simple analysis, it is straightforward that area savings have a huge impact on costs, especially when considering large volumes, as reducing die size directly affects production efficiency and overall manufacturing expenses. The limitations imposed by power delivery networks on further area shrinkage highlight the need for design efforts on the PDN optimization.

5.3 Recommendations for Future Research

While this study has highlighted the significant impact of Power Delivery Network (PDN) design on scaling and area shrinkage in advanced technology nodes, there are several promising avenues for future research that could address the limitations faced by current two-dimensional (2D) integrated circuit (IC) designs:

- Exploration of 3D ICs

  As traditional 2D scaling approaches encounter physical and design constraints, *three-dimensional integrated circuits (3D ICs)* present an exciting opportunity to continue improving performance, power efficiency, and area utilization. By stacking multiple layers of logic and memory vertically, 3D ICs can achieve higher transistor densities without the need for further shrinking in the horizontal plane. However, these designs introduce new PDN challenges, such as delivering power across stacked layers and managing thermal dissipation. Future research should focus on optimizing PDN architectures for 3D ICs to ensure reliable power delivery and heat management. [12] [13]

- Back Powering Techniques

  Another promising area of investigation is *back powering*, a technique that delivers power through the backside of the wafer rather than through the conventional front-side interconnects. This approach could enable more efficient power distribution by freeing up the front side for increased signal routing and reducing the area dedicated to PDN structures. Back powering could also help alleviate some of the limitations associated with high current densities in advanced nodes, making it a key area for research in enabling further area shrinkage and improved power efficiency. It is indeed becoming an industry trend for bleeding edge technologies. [14] [12]

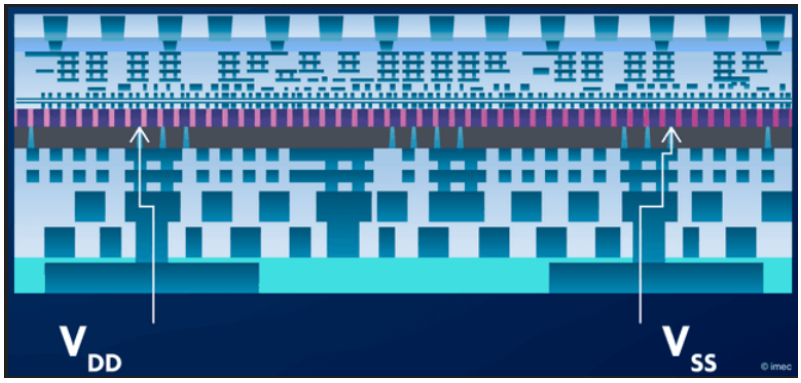

Figure 5.2: Backside power delivery - imec [4]

- Advanced Materials and New Interconnect Technologies

  Research into new materials for interconnects and power delivery components is another critical area. As copper interconnects face increasing resistance and reliability issues at smaller nodes, alternatives like carbon nanotubes or graphene-based materials may offer superior performance. Investigating how these materials can be integrated into PDN structures could pave the way for more scalable and efficient power delivery solutions. [15] [16] [17]

Bibliography

[1] Andrew B. Kahng, Jens Lienig, Igor L. Markov, and Jin Hu. *VLSI Physical Design: From Graph Partitioning to Timing Closure*. Springer, 2022. ISBN 978-3-030-96415-3. doi: 10.1007/978-3-030-96415-3.

[2] WikiChip Fuse. Standard cell rows. `https://fuse.wikichip.org/wp-content/uploads/2019/01/std-cells-rows.png`, 2019. [Online; accessed 2024-12-02].

[3] Wikipedia contributors. Gdsii — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=GDSII&oldid=1240266053`, 2024. [Online; accessed 2024-11-11].

[4] imec. How to power chips from the backside, 2021. URL `https://www.imec-int.com/en/articles/how-power-chips-backside`. Accessed: 2025-03-23.

[5] Yibo Lin, Shounak Dhar Gao, Bei Yu Chen, Igor L. Markov, and David Z. Pan. Dreamplace: Deep learning toolkit-enabled gpu acceleration for modern vlsi placement. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 40(4):748–761, April 2021. doi: 10.1109/TCAD.2020.3003843.

[6] Nicholas A. Lanzillo, Andrew Chu, Rafael Vega, Nancy Perez, Linda Clevenger, and David Dechene. Methodology development to benchmark power delivery designs in advanced technology nodes. *IBM Research Publications*, 2023.

[7] Hosoon Shin, Kyoungin Cho, and Yongchan Ban. Application-driven optimizations of metal stack and pdn (power delivery network) using machine learning framework. In *Proceedings of SPIE*, volume 12954, page 129540B. SPIE, 2023. doi: 10.1117/12.3010937.

[8] Nicholas A. Lanzillo, Andrew Chu, Rafael Vega, Nancy Perez, Linda Clevenger, and David Dechene. Benchmarking power delivery network designs at the 5-nm technology node. *IEEE Transactions on Electron Devices*, 68(12):6311–6316, 2021. doi: 10.1109/TED.2021.3112345.

[9] Shreepad Panth, K. Samadi, Y. Du, and S. K. Lim. Full chip impact study of power delivery network designs in gate-level monolithic 3-d ics. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34(4):624–637, 2015. doi: 10.1109/TCAD.2015.2388751.

[10] Nicholas A. Lanzillo, Andrew Chu, Rafael Vega, Nancy Perez, Linda Clevenger, and David Dechene. Power delivery design, signal routing, and performance of on-chip cobalt interconnects in advanced technology nodes. *IEEE Transactions on Electron Devices*, 68(11):5505–5511, 2021. doi: 10.1109/TED.2021.3106081.

[11] Shreepad Panth et al. A holistic evaluation of buried power rails and back-side power for sub-5-nm technology nodes. *IEEE Transactions on Electron Devices*, 69(7):3812–3818, 2022. doi: 10.1109/TED.2022.3179923.

[12] ASM International. Leading edge technologies: Backside power delivery. 2023. URL `https://dl.asminternational.org/technical-books/monograph/193/chapter/3790358/Leading-Edge-Technologies-Backside-Power-Delivery`.

[13] Applied Materials. Challenges to interconnect scaling at 3nm and beyond. 2021. URL `https://www.appliedmaterials.com/us/en/blog/blog-posts/challenges-to-interconnect-scaling-at-3nm-and-beyond.html`.

[14] Semiconductor Engineering. Backside power delivery gears up for 2nm devices. 2023. URL `https://semiengineering.com/backside-power-delivery-gears-up-for-2nm-devices/`.

[15] Zhican Lin and Zhongliang Pan. Electrical modeling and transmission performance analysis of carbon nanotube–graphene hybrid interconnects. *Microelectronics Journal*, 105:105966, 2023. doi: 10.1016/j.mejo.2023.105966. URL `https://www.sciencedirect.com/science/article/pii/S0026269223002793`.

[16] Uma Sathyakumari and Partha Sharathi Mallick. *Carbon Nanotubes for Interconnects*, volume 300 of *Springer Series in Materials Science*. Springer, 2021. doi: 10.1007/978-3-319-29746-0. URL `https://link.springer.com/book/10.1007/978-3-319-29746-0`.

[17] A. Magnani, M. de Magistris, A. Todri-Sanial, and A. Maffucci. Carbon-based power delivery networks for nanoscale ics: Electrothermal analysis. *IEEE Transactions on Electron Devices*, 69:543–552, 2022. doi: 10.1109/TED.2021.3134563. URL `https://research.tue.nl/en/publications/carbon-based-power-delivery-networks-for-nanoscale-ics-electrothe`.