16FFP (FINFET Plus) Design Challenges, Mitigations and Solutions

Dissertation submitted in
partial fulfillment of the requirements for
the award of the degree of

M. S.
IN
VLSI ENGINEERING
BY
CH.SHARATH
14181J5016

JNTUH, Kukatpally, Hyderabad

Learn. Explore. Excel.

VEDA IIT
Plot No 90, Road No 2
Banjara Hills, Hyderabad, AP
Pin 500034

JNTUH, Kukatpally, Hyderabad

VEDA IIT
Plot No 90, Road No 2
Banjara Hills, Hyderabad, AP
Pin 500034

DECLARATION

I, CH.SHARATH, hereby declare that the project work entitled, 16FFP (FINFET Plus) DESIGN CHALLENGES, MITIGATIONS AND SOLUTIONS, submitted by me in partial fulfillment of the requirements for the award of M.S. degree in VLSI Engineering by Jawaharlal Nehru Technological University, Hyderabad. This is the bonafide work done by me during the period July 2015 to July 2016, under the guidance and supervision of Industry guide Mr. J.NAGAMAHESH & Internal guide Prof. K. SUBBARANGAIAH. The results embodied in the report have not been submitted to any other University or Institution for the award of any degree or diploma.

CH.SHARATH

JNTUH, Kukatpally, Hyderabad

Learn. Explore. Excel.

VEDA IIT
Plot No 90, Road No 2
Banjara Hills, Hyderabad, AP
Pin 500034

CERTIFICATE

This is to certify that the project work entitled **“16FFP(FINFET Plus) DESIGN CHALLENGES, MITIGATIONS AND SOLUTIONS”** is submitted by **CH.SHARATH (Roll No: 14181J5016)**, in partial fulfillment of the requirements for the award of **M.S. degree in VLSI Engineering** to **Jawaharlal Nehru Technological University, Hyderabad**, during July 2014 to July 2016. This is the bonafide work done by him under our guidance and supervision. The results embodied in this report have not been submitted to any other University or Institution for the award of any degree or diploma.

Project Guide in the Institute:
Dr. K. Subbarangaiah
Professor and Director
VEDA IIT
Banjara Hills
Hyderabad 500034

Head of the Institute
Dr. K. Subbarangaiah
Professor and Director
VEDA IIT
Banjara Hills
Hyderabad 500034

Project Guide in the consortium company:
Mr. J.NAGAMAHESH,
Senior Engineer (Physical Design)
Soctronics Technologies Pvt. Ltd
Banjara Hills, Hyderabad 500034

The project report has been approved as it satisfies the academic requirements in respect of the project prescribed for the said degree.

Internal Examiner

External Examiner

Acknowledgments

I would like to express my deep sense of gratitude to **Prof. K.Subbarangaiah**, Director of VEDA IIT, for providing me with an opportunity for carrying out this project and for his support, encouragement and inspiration.

I would like to thank **Soctronics Technologies Pvt. Ltd.,** for giving me an opportunity to carry out the project. I also express my sincere thanks to **Mr. J.NAGA MAHESH, Senior Engineer (Physical Design), Soctronics Technologies,** for his tremendous support and valuable guidance.

I also thank VEDAIIT faculty **Mr. D.Chakrabarthy and Mr. Sathish Khanna, and Mr. Kiran Rapaka, Project Manager, Soctronics Technologies Pvt Ltd,** for their valuable guidance in exploring and learning the concepts and implementing the same in the project.

In addition to this, I also sincerely thank all the staff members of VEDA IIT and my project colleagues for supporting me to successfully complete this project.

(CH.SHARATH)

Table of Contents

List of Figures

List of Tables

ABSTRACT

As Moore's law extends, Technology keeps shrinking towards lower nodes. When we moved to below 28nm process the planar technology (MOSFETS) hits the device scaling limit, leakage power accounts for an increasingly larger portion of the total power consumption. The FINFET technology introduces a paradigm shift addressing this specific limitation, where it has not only a better scalability, but also better short-channel characteristics and a more effective way to suppress the leakage .The FINFET technology brings a significant performance increase compared to planar technology.

Race to outperform & fueled by performance benefits, has drawn attention to 16nm FINFET PLUS (16FFP) technology node. This has brought in new complexity and alien challenges to timing, area and power parameters of the design. It is comprehensive study of preroute to postroute gaps, Crosstalk analysis – fixing, optimum power recovery implementation, Double Pattern Loops (DPTs) and congestion mitigations.

In this paper I talk about distinguished methodologies used to address the design challenges at PnR level. These methodologies have shown improvement which means a serious business for the size/complexity of contemporary designs.

Full dissertation is available with VEDA IIT as a soft copy.

CHAPTER 1

INTRODUCTION

Throughout the momentous history of semiconductors, everything from computer hardware to multifunction mobile devices, Moore's law has remained the same. The number of transistors on a given area of silicon doubles every two years. With the foundries developing advanced process nodes and their consumers unquenchable demand for more functionality, the industry has satisfied Moore's Law. The transistor count on today's advanced multicore processors is reaching the 3 billion range – a long way from the 6800 processor of the mid 1970s that had just 5000 transistors.

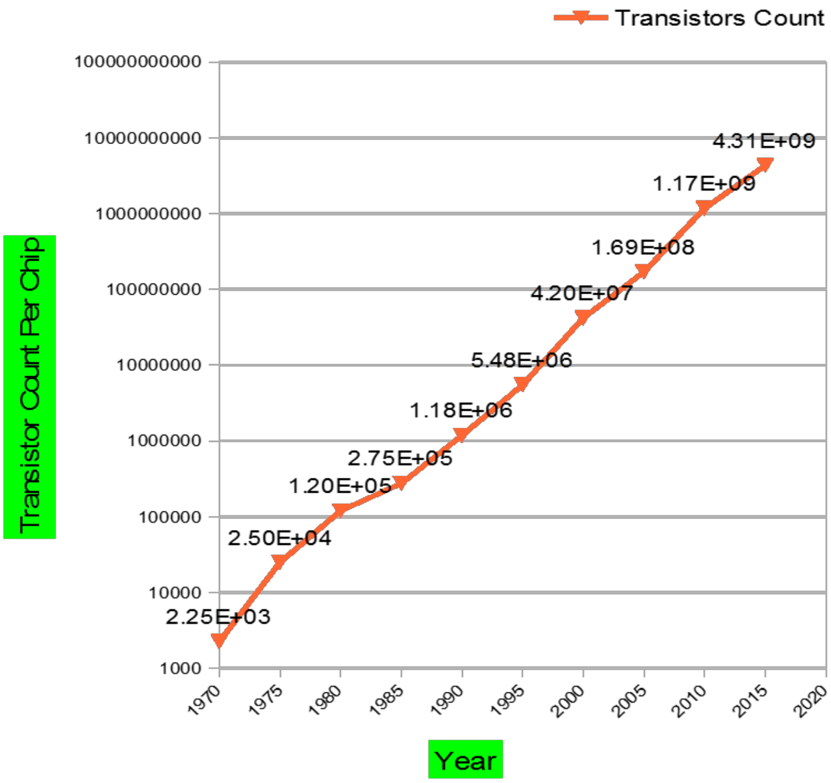


Figure 1.1 Moore's law

Moore's Law started as an observation in 1965 by Dr. Gordon Moore that number of transistors per square inch will double every two years. To continue the progression of Moore's law the transistor gate geometry (process technology node) needs to shrink every 18-24 months. This requires innovative process technology and massive R&D spending on semiconductor foundries. There have been several speculations in the past for demise of Moore's law due to the limits of physical transistors dimensions and manufacturing complexity.

1.1 Need to meet the Design Parameters

Design parameters like,

- Performance
- Power
- Area
- Cost
- Time to market

have not changed since the beginning of the integrated-circuit industry.

In fact, Moore's law is all about optimizing those parameters by driving to the smallest possible transistor size with each new technology generation. As the Moore's law extends, the computing products continue to shrink dimensions and improve performance. The infrastructural cost of developing next-generation products and chip manufacturing processes is increasing with each generation.

1.2 challenges faced to meet the Design parameters

As we approached the sub nanometer range with the 90nm node and beyond, static leakage became an important factor such that while every new process generation may have doubled the gate density, it also doubled the amount of leakage current.

As process technologies further continued to shrink towards 20nm, it became impossible to achieve a similar scaling of certain device parameters, particularly the power supply voltage, which is the dominant factor in determining dynamic power.

Additionally, optimizing for one variable such as performance evidently lead to unwanted compromises in other areas like power.

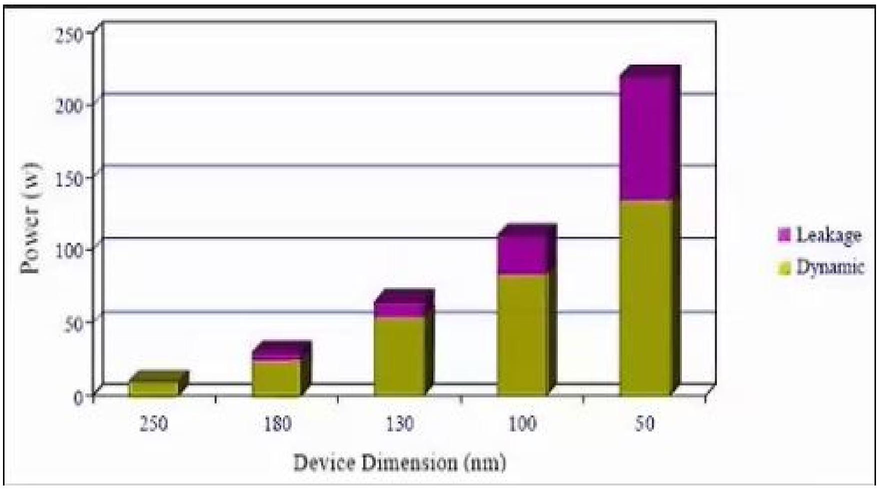


Figure 1.2 Total power vs Device dimension(nm)

1.2.1 Power Dissipation in Digital Designs

Total power for any design consists of dynamic power and static power.

Dynamic power consumption is the power dissipated when device is active or signals are changing values. The dynamic power of a circuit consists of:

- Switching Power
- Internal Power

Switching power: The power dissipated by charging and discharging the load capacitance.

Internal power: Internal power is any power dissipated within the cell. The major part of internal power is short circuit power. Short circuit power is caused by the short circuit current that occurs during the time when a CMOS gate is switching and inside the gate both transistors are on.

Psw (switching power) = C * Vdd2 * Fclock

Psc (Internal power) = Tsc * Vdd * Ipeak * Fclock

As Tsc is very short we can take Pdyn = C * Vdd2 * Fclock

Static power consumption is the power dissipated by a gate not changing state. Static power is mostly determined by leakage current and grows dramatically with shrinking nodes. The main source of leakage power is the sub-threshold leakage current. It is dependent on the gate threshold voltage (Vth). The Vth is controlled by adjusting the doping concentration under the gate itself. Other possibilities to influence the Vth are gate oxide thickness variation and back biasing. But on the other hand these parameters are often fixed because of the performance goals for the design/technology.

Pleakage = Vdd * Ilk

The issue of leakage current could be mitigated by the use of high voltage threshold dopants at the expense of device performance, or through the use of advanced design techniques, such as power gating or multivoltage islands. Controlling current leakage when the transistors are switched off is important to preserve battery life or minimize power consumption in computer and mobile applications that spend most of their time in an idle state.

Economics also plays a vital role in determining whether to move to and when. If chips can take advantage of the increased density to provide more functionality, then it generally made sense to move to the next node, even if mask and process costs were higher.

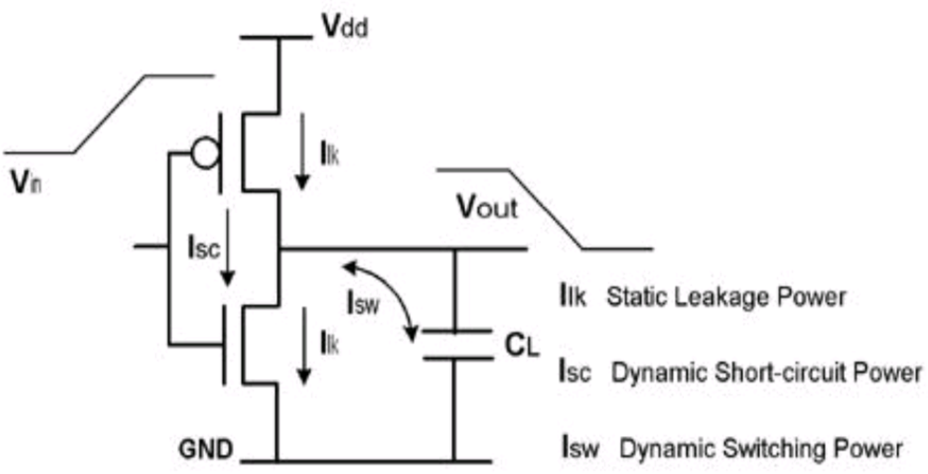


Figure 1.3 Power Dissipation in CMOS

This was the case when designs moved from 65nm to 45/40nm and then again to 28nm. However, as it moved to the 20nm process node, there has been a new set of challenges, including double patterning and very leaky transistors due to short channel effects. Both are nullifying the benefits of transistor scaling if planar transistor technology is used.

The move from 28nm to 20nm was unappealing economically to many, since it didn't provide the same level of performance and area gains as observed with previous generations.

1.3 FINFETs, extending the Moore’s law

The advanced geometry planar FET technologies, such as 20nm, the source and the drain intrude into the channel, making it easier for leakage current to flow between them and making it very difficult to turn the transistor off completely. While the planar FET may have reached the end of its scalable lifespan, the industry has found a fitting alternative in FINFETs for next generation of advanced processes.

FINFET technology had its origins in the 1990s, when DARPA were keen to fund research possibilities of replacements to the planar transistor. A UC Berkeley team led by Dr. Chenming Hu proposed a transistor with a new structure which would reduce leakage current.

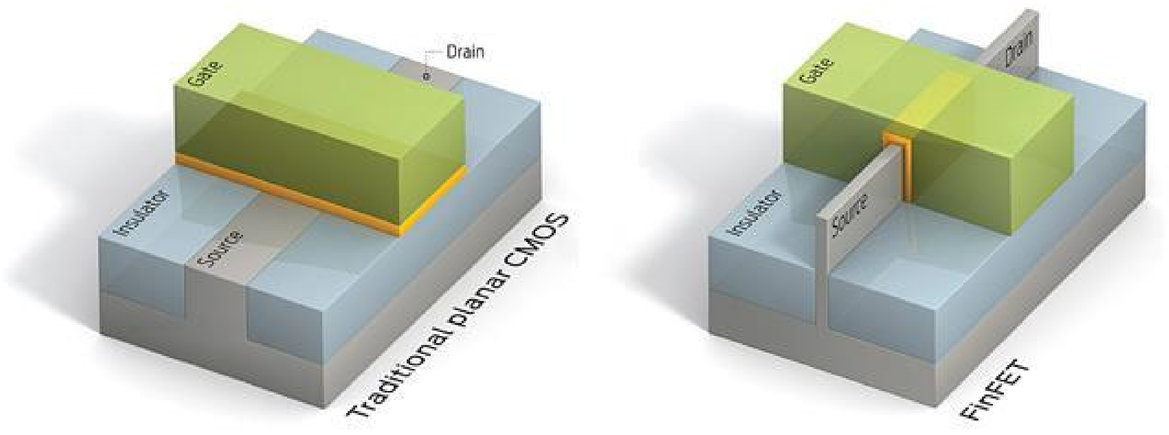


Figure 1.4 Traditional planar CMOS vs FINFET

They suggested that a thin-body MOSFET structure would control short-channel effects and subdue the leakage by keeping the gate capacitance in closer proximity to the whole of the channel.

FINFETs are named so because they are 3d structures that projects out of the substrate and resemble a fin. The 'fins' form the source and drain, effectively providing more volume for the same area than a planar transistor. The gate wraps around the fin, providing better control of the channel and allowing very little current to leak through the body when the device is in the 'off' state. This, in turn, enables the use of lower threshold voltages and results in better performance and power.

FINFET technology introduces a paradigm shift addressing this specific limitation, where it has not only a better scalability, but also better short-channel characteristics and a more effective way to suppress the leakage The FINFET technology brings a significant performance increase compared to planar technology at the same time it has introduced a new challenge pertaining to high input cell capacitances on FINs.

With this new reality, the dynamic power has become the dominant factor in the total power calculation and there is also a new behavior regarding the cells selection for faster timing closure.

In the planar technology, dynamic and leakage power were almost at par. The main power reduction technique then, was to minimize the usage of low VT cells in order to optimize the total power.In the FINFET technology, the standard cell leakage is not anymore a significant contributor of the total power and at the same time the gate capacitance of FIN structures contribute to the dominance of the dynamic power.

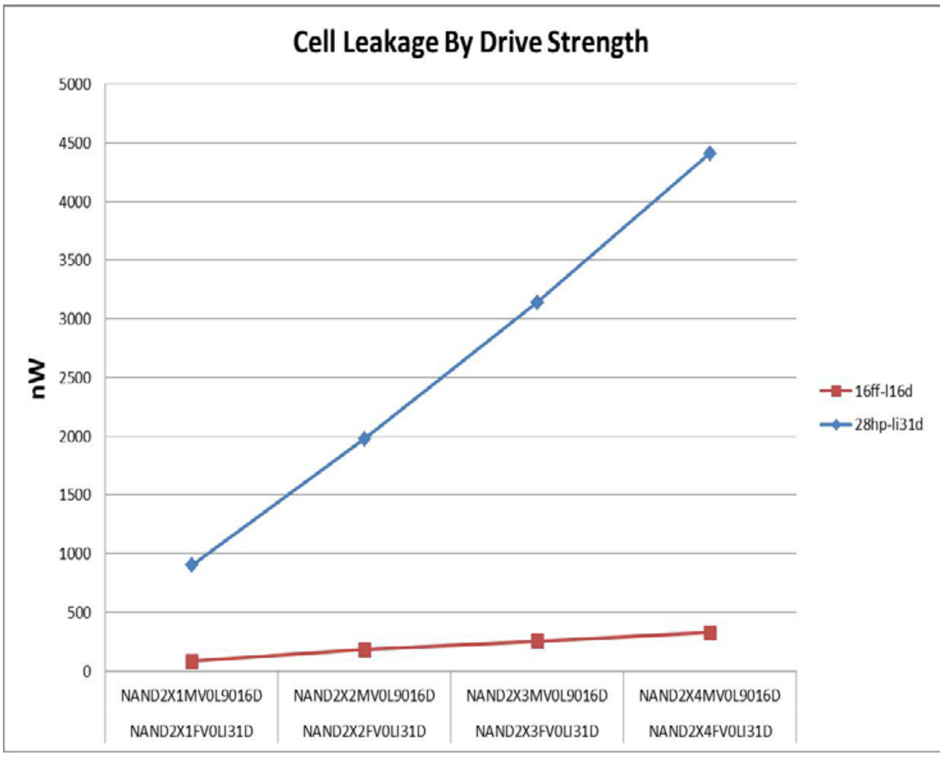


Figure 1.5 Leakage difference between 16ff & 28nm

The 20nm process node was necessary to help build the foundation for the advanced FINFET processes. With the smaller device geometries, traditional lithography/optical manufacturing techniques no longer have the required resolution where double patterning – using litho-etch-litho processing – becomes necessary. The number of manufacturing design rules has increased significantly and these have to be handled by various EDA tools, such as place and route, physical verification, and extraction.

The industry's experience with 20nm has paved the way for an easier transition to FINFET processes. Many of the tool improvements can still be applied, but the handling of FINFETs does require a few more changes; for example, SPICE BSIM-CMG models had to be created to add the 3D effects. It is also true that, with 3D transistors, capacitance becomes a primary concern. EDA tools must build in high resistance interconnect optimization in order to mitigate these capacitive effects. Layer awareness is also essential to provide optimal metal layer assignment during routing of the design.

Another interesting trend to note is the cost variation as the size decreases. As the size decrease lower than 28nm, there is a price increase unlike the gradual price decrease with the size reduction upto 28nm.

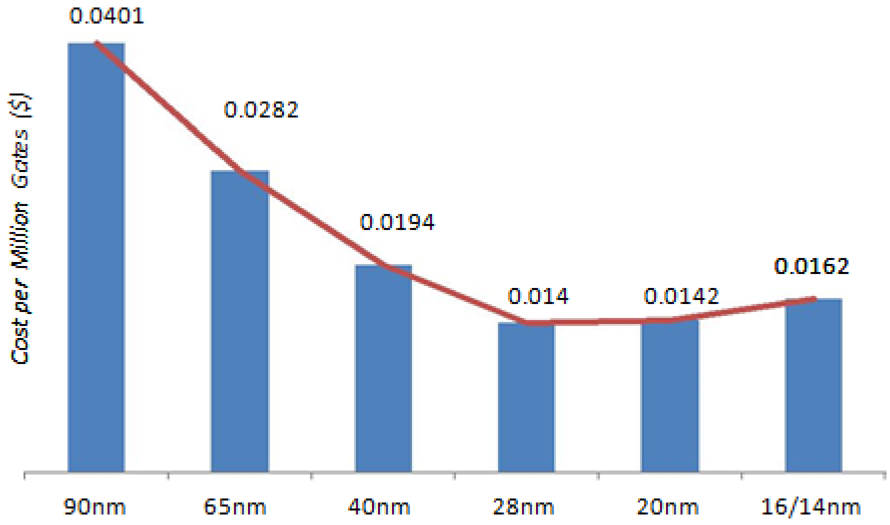


Figure 1.6 Cost per million gates vs Feature size

Although FINFET processes may seem new, development of the technology itself has been in progress for almost a decade. The industry has worked together to make the shift to an advanced new process node as seamless as possible, with minimal impact to current design methodologies. Consumer appetite for new functionality remains high and the move to designing with FINFET process technologies will help fill that need and keep Moore's Law very much alive.

1.4 Introduction to 16nm node

In last two years there has been significant developments on 16 FF . The technology has revolutionized the structure of transistor. This revolution has brought in new advantages of performance, power, cost. Meanwhile there were many flow and methodology updates for physical design engineers. Still the technology suffers from challenges of congestion, timing, DRC, power .

Similar challenges were there in earlier technology nodes also. But as we have scaled down further , and also change the structure of the transistor , there are various factors which are new.

Primarily ,

1. Layer_Resistance ,
2. Double Pattern Layers(DPT)
3. Crosstalk

takes control of these challenges.
Layer resistance, and Crosstalk has made the timing closure a bit harder.

The traditional approaches could not adapt to the new challenges. The methodology for the discussed challenges takes each problem conceptually, and is based on rectifying the root cause. The approach is based on ICC and PrimeTime platform. Gravity of issues , extent of mitigation and detailed analysis of the results is discussed. Performance and quality of result is not impacted. There is a tradeoff of Run time with better correlation gap . The approach comes with a limitation where the subchip owner has to be careful of caveats and pitfalls of the approach.

1.5 Thesis Organization

Chapter 2 presents Introduction about FINFET Technology, What is need of this technology, what are the advantages of FINFET technology compare to planar technology. What are challenges faced interms of Design and verification in Physical Design. Chapter 3 introduces the stages involved in traditional ASIC flow.

Chapter 4 presents the new methodology and implementation of the congestion mitigation flow. Chapter 5 explains about PreRoute-to-PostRoute timing correlation and metal layer stack resistance variation.

Chapter 6 presents what are the challenges we faced while fixing DRCs and how to deal with DPTs. Chapter 7 presents the power recovery flow. Chapter 8 concludes this work with a summary and list of contributes.

CHAPTER 2

FINFET TECHNOLOGY

2.1 What is a FINFET?

A FINFET is a MOSFET with the channel elevated so the gate can surround it on three sides. This FINFET is a transistor design, first developed by Chenming Hu and colleagues at the University of California at Berkeley, which attempts to overcome the worst types of short-channel effect encountered by deep submicron transistors, such as drain-induced barrier lowering (DIBL).

These effects make it harder for the voltage on a gate electrode to deplete the channel underneath and stop the flow of carriers through the channel – in other words, to turn the transistor Off. By raising the channel above the surface of the wafer instead of creating the channel just below the surface, it is possible to wrap the gate around up to three of its sides, providing much greater electrostatic control over the carriers within it.

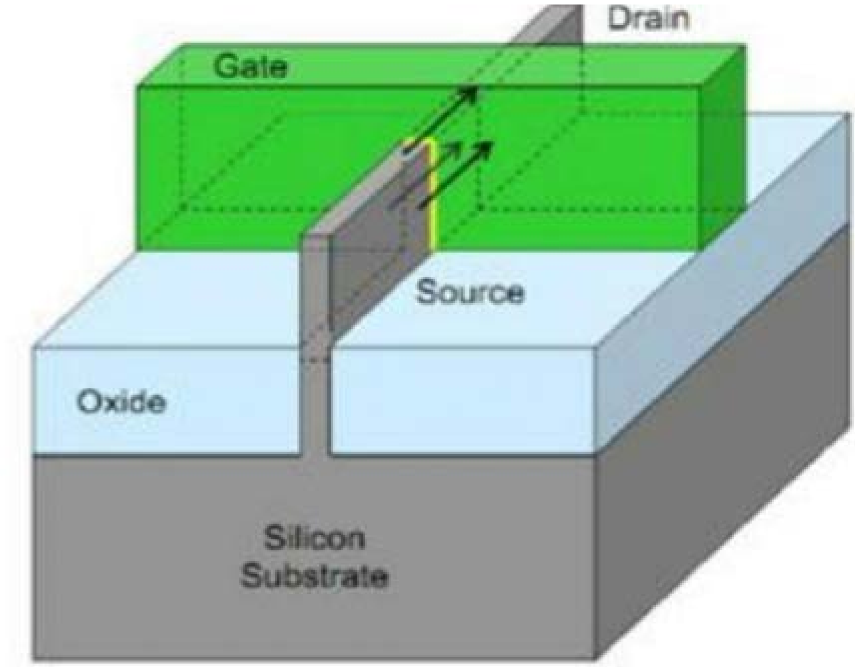


Figure 2.1 3-D Tri-gate transistor form conducting channels on three sides of a vertical fin structure providing fully depleted operation

FINFETs provides a new pathway for Moore's Law beyond 20nm as they have much better performance and reduced power consumption compared to planar transistors. A 16nm/14nm

FINFET process can potentially offer a 40-50% performance increase or a 50% power reduction compared to a 28nm process.

The next few years should be very interesting as the benefits of this technology are seen in products from smart phones to servers. Although it has numerous benefits, the move to FINFETs comes with quite a few new challenges such as design-rule complexity and skyrocketing resistance, new Layout Proximity Effects. Routers face difficulty to connect efficiently to pins on standard cells. Furthermore, extracting parasitic from FINFETs is significantly different from regular planar CMOS devices.

Thus FINFET processes should be made as transparent and smooth as possible for the designers. To achieve this, Semiconductor industries need to work behind the scenes to ensure that the tools understand and model the complexities involved.

FINFETs are named so because they are 3d structures that projects out of the substrate and resemble a fin. The 'fins' form the source and drain, effectively providing more volume for the same area than a planar transistor

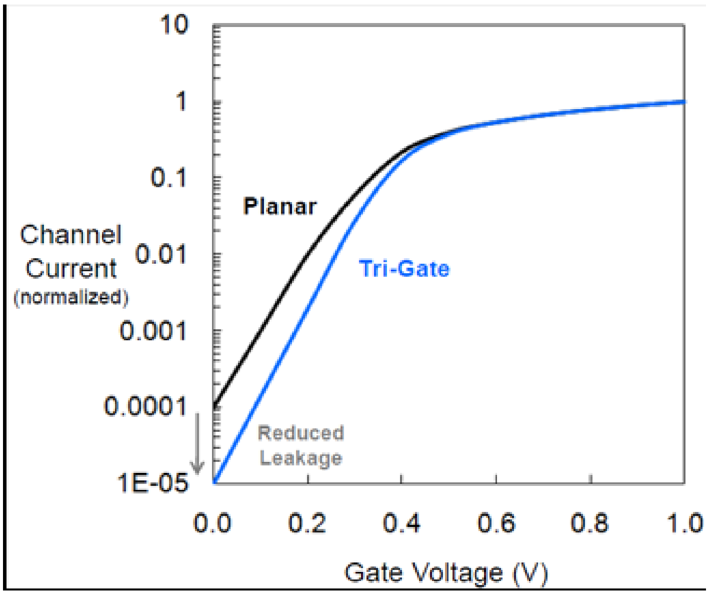


Figure 2.2 Channel current vs Gate voltage for planar and trigate

The gate wraps around the fin, providing better control of the channel and allowing very little channel current to leak through the body when the device is in the 'off' state. This, in turn, enables the use of lower threshold voltages and results in better performance and power.

2.2 MOSFET vs FINFET

The below figures are simplified depictions of a planar FET and a FINFET respectively. In the planar FET a single gate controls the source-drain channel. Such a gate does not have good electrostatic field control away from the surface of the channel next to the gate, resulting in leakage currents between source and drain even when the gate is off. By contrast, in the FINFET the transistor channel is a thin vertical fin with the gate fully "wrapped" around the channel formed between the source and the drain.

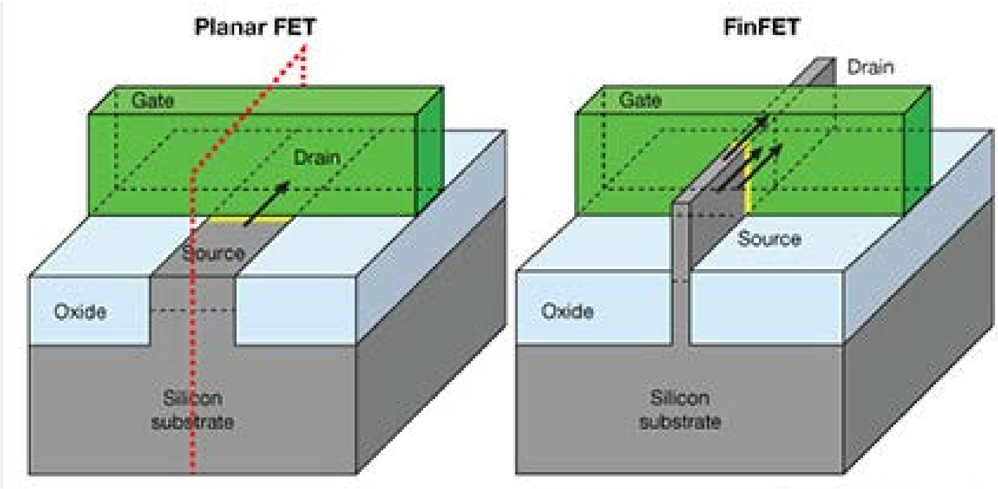


Figure 2.3 MOSFET vs FINFET

The gate of the FINFET can be thought of as a "multiple" gate surrounding the thin channel. Such a multiple gate can fully deplete the channel of carriers. This results in much better electrostatic control of the channel and thus better electrical characteristics.

2.3 Need of FINFET

Since the fabrication of MOSFET, the channel length of the device has been shrinking constantly so as to fabricate compact and fast devices. The following parameters related to MOSFET highlight the need for smaller, compact devices and explain why the MOSFET is not the suitable choice for the same. The shorter section of the gate electrode is known as the length and the longer section is called the width.

As the channel length of a MOSFET reduces, the short-channel effects increase. The short-channel effects are attributed to two physical phenomena:

a. The limitation imposed on electron drift characteristics in the channel

b. The modification of the threshold voltage due to the shortening channel length.

There are five different distinguishable short-channel effects :

1. Drain-induced barrier lowering

The two depletion layers merge as a result of depletion region surrounding the drain which extends to the source. This leads to occurrence of punch through. Punch through can be reduced with the help of thinner oxides, larger substrate doping, shallower junctions, and also with longer channels.

2. Surface scattering

As the channel length becomes smaller, the longitudinal electric field component increases, and the surface mobility becomes field-dependent. The carrier transport in a MOSFET is confined within the narrow inversion layer. The surface scattering causes reduction of the mobility. The electrons find it difficult to move parallel to the interface. This is necessary so that the average surface mobility is about half as much as that of the bulk mobility. Surface scattering are the collisions suffered by the electrons which are accelerated toward the interface.

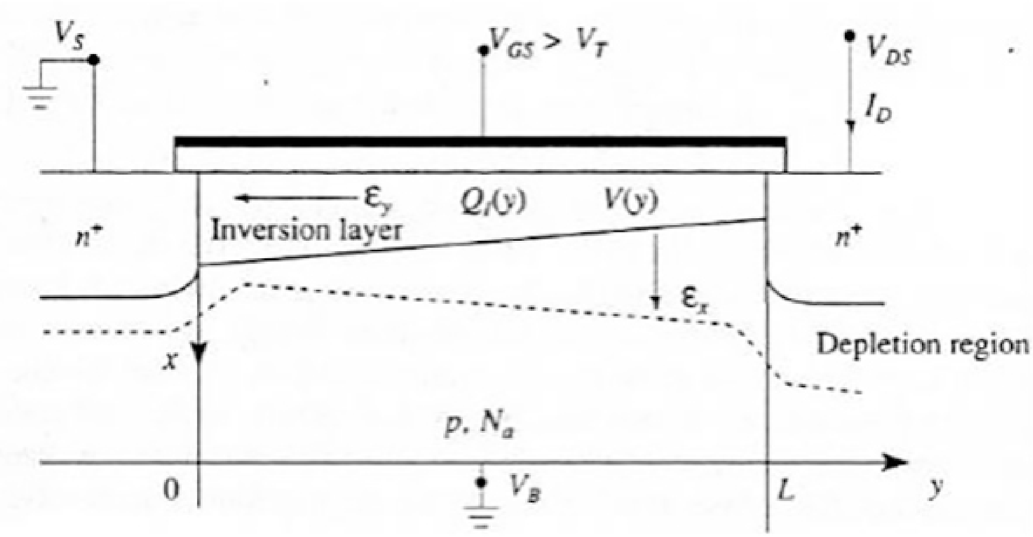


Figure 2.4 Surface scattering

3. Velocity saturation

Velocity saturation reduces transconductance in saturation mode. When a strong electric field is applied, carrier velocity reaches maximum value known as saturation velocity. When this occurs, the state of the transistor is known as velocity saturation. Velocity saturation is caused by the increased scattering rate of highly energetic electrons, primarily due to optical phonon emission. This effect increases the transit time of carriers through the channel.

4. Impact ionization

This is usually occurs due to the high velocity of electrons in presence of high longitudinal fields that can generate electron-hole (e-h) pairs by impact ionization. Impact ionization occurs due to the impact on silicon and ionizing of the electron-hole pairs. It happens as follows: usually, most of the electrons are attracted by the drain, while the holes enter the substrate to form part of the parasitic substrate current. Also, the region between the source and the drain can act like the base of an N-P-N transistor.

The source plays the role of the emitter and the drain that of the collector. If the holes are collected by the source, and the corresponding hole current creates a voltage drop in the substrate material of the order of .6V, the normally reversed-biased substrate-source P-N junction will conduct. Then electrons can be injected from the source to the substrate, similar to the injection of electrons from the emitter to the base. They can gain enough energy as they travel toward the drain to create new eh pairs. The situation can worsen if some electrons generated due to high fields escape the drain field to travel into the substrate, thereby affecting other devices on a chip.

5. Hot electron effect

Hot electron effect occurs when electrons or holes gain high kinetic energy due to the presence of high electric field within a semiconductor device. Hot electrons are more probable than hot holes since they have higher mobility to begin with. Hot carriers get injected or trapped in certain areas and cause undesirable device behavior and degradation hence give rise to hot carrier Effects.

As the size of the devices is scaled down, the electric field of the channel increases. This leads to the high field region near the drain terminal occupying a large fraction of the channel length. This leads to the hot electron effect which in turn degrades the device parameters with time. This effect creates obstacles while scaling down the device.

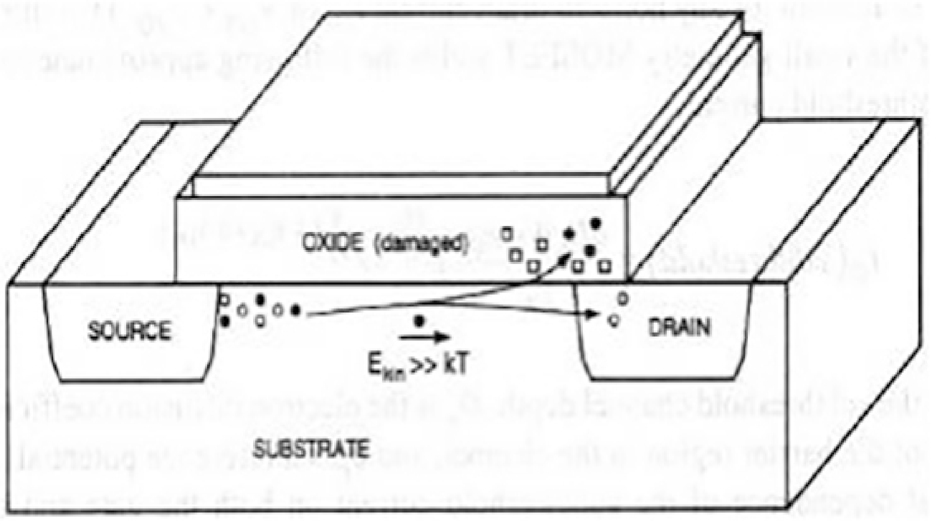


Figure 2.5 Hot electron effect

2.4 The FINFET Promise

Due to its many superior attributes, especially in the areas of performance, leakage power, intra-die variability, low voltage operation (translates to lower dynamic power), and significantly lower retention voltage for SRAMs, FINFETs are replacing planar CMOS as the device of choice.

Leading foundries estimate the additional processing cost of 3D devices to be 2% to 5% higher than that of the corresponding Planar wafer fabrication. FINFETs are estimated to be up to 37% faster while using less than half the dynamic power or cut static leakage current by as much as 90%.

FINFETs also promise to alleviate problematic performance versus power tradeoffs. Designers can run the transistors faster and use the same amount of power, compared to the planar equivalent, or run them at the same performance using less power. This enables design teams to balance throughput, performance and power to match the needs of each application.

2.5 FINFET advantages

At any one technology node the FINFET has several advantages over its planar counterpart including, but not limited to:

- Very good electrostatic control of the channel, meaning the channel can be “choked off” more easily. FINFETs boast a near-ideal sub-threshold behavior (associated with leakage), something that’s not easy to achieve in planar technology without considerable effort.

- Greatly reduced short channel effects. The short channel effects in planar technology are complex and have a significant impact on gate length variations and, therefore, on electrical performance.
- High integration density, 3D, thanks to vertical channel orientation delivers more performance per linear “w” than planar even after the isolation dead-area between the fins is taken into account.
- Smaller variability, especially variability resulting from random dopant fluctuation primarily due to doping-free or low doping channels.
- The immediate and obvious advantage of FINFETs that the effective width Of the channel becomes:

$$Weff = 2Hsi + Wsi$$

So that current carrying capability of the transistor will increase.

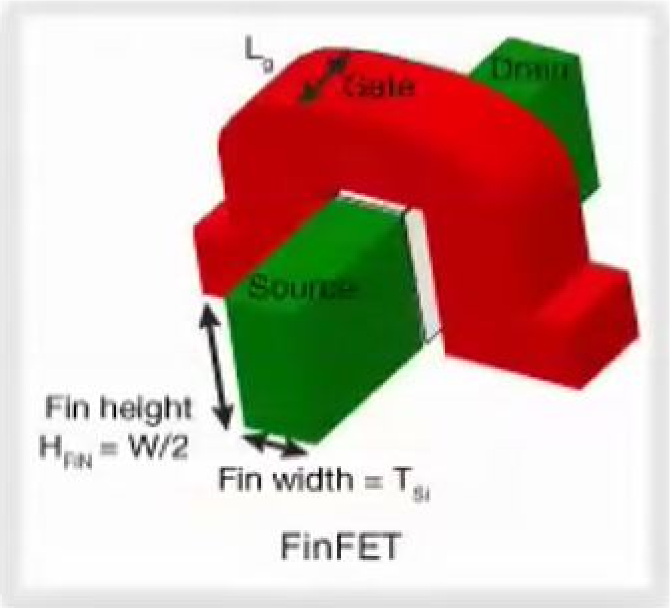


Figure 2.6 FINFET electrical dimensions

2.6 FINFET as an Opportunity for IP Design

Design metrics of performance, power, area, cost, and time-to-market (opportunity cost) have not changed since the inception of the IC industry.

Designing in FINFET broadens the design window. Operating voltage continues to scale down, significantly saving on dynamic and static power. Additionally, short channel effects are significantly reduced, decreasing the guard-banding needed to deal with variability, and performance continues to improve (compared to planar at an identical node). In fact, at very low power supply voltages, the performance advantage of the FINFET compared to its planar equivalent widens due to the superior gate control of the channel in the FINFET even at low voltages. For memory designers, an added advantage of FINFET is the significantly lower retention voltage of FINFET-based SRAM compared to that of planar.

One additional feature that eases the transition from designing in planar to designing in FINFET is the fact that the back-end of the process is essentially the same for both, and therefore a significant amount of the design flow associated with the back-end remains intact.

2.7 FINFET design: The challenges

FINFET is a significantly more complex device to model. Accurate FINFET parasitic extraction is more complicated. Generating good, yet compact SPICE models is also more challenging than for planar devices. For most design activities the aforementioned complexities are transparent to the designer. However, there still remain many design optimization challenges for the circuit designer utilizing FINFET.

FINFET has a lower DIBL / SS (sub-threshold swing) that is a desirable characteristic as far as leakage is concerned. On the other hand the undoped (or very lightly doped) and practically fully-depleted channel renders the use of body biasing techniques commonly used in planers less effective, making alternatives necessary.

The finite granularity of the fin width “W” and the limited range of freedom in channel length for a given architecture make optimizing analog as well as digital design more complex. Granted that many fins can be “ganged” together to generate a desired “W”, still “L” and “W” are not exactly free continuous parameters. This is because FINFETs are 3D structures, and reining in etching variability for the high-aspect ratio processes with non-uniform pitches or locally varying pitches may be a problem. Thus FINFETs have a significant numbers of restricted design rules (RDR).

2.7.1 Resistance/Capacitance (RC) Extraction Tools

The 3D nature of FINFETs and the multiple fins making up the transistors introduce a large number of new parasitic resistance and capacitances to be considered, modeled and extracted from the FINFET-based designs.

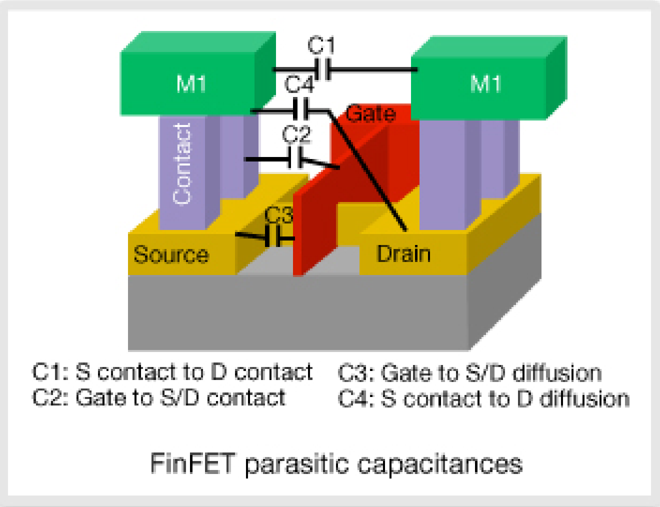

C1
C4
M1
M1
Gate
C2
Contact
C3
Source
Drain
C1: S contact to D contact
C2: Gate to S/D contact
C3: Gate to S/D diffusion
C4: S contact to D diffusion
FinFET parasitic capacitances

Figure 2.7 FINFET Parasitics

The interconnect modeling of semiconductors has been standardized in the open source Interconnect Technology Format (ITF). This format has recently been extended to add the FINFET requirements.

Synopsys StarRC extraction tool has been enhanced to support the new ITF models and is extensively used in the extraction of FINFET-based designs. StarRC is certified by leading FINFET foundries and is the industry standard for signoff extraction.

2.7.2. Physical Verification Tools

Physical verification is another area affected by FINFET technology. The new runsets used by the physical verification tools are used to verify Logic versus Schematic (LVS) correctness, and Design Rule Checks (DRCs). FINFETs require LVS enhancements to support recognition of these new devices in the layout and enable parameter extraction and identification of proximity effects. Other LVS enhancements include new source-drain resistance calculations. A number of new design rules have been introduced including fin-to-fin spacing and fin widths.

Synopsys’ IC Validator physical verification product has been enhanced to support LVS and DRC for FINFETs. It is currently being used for the development of FINFET- based designs and IP.

CHAPTER 3

Traditional ASIC Flow

3.1 Traditional Flow

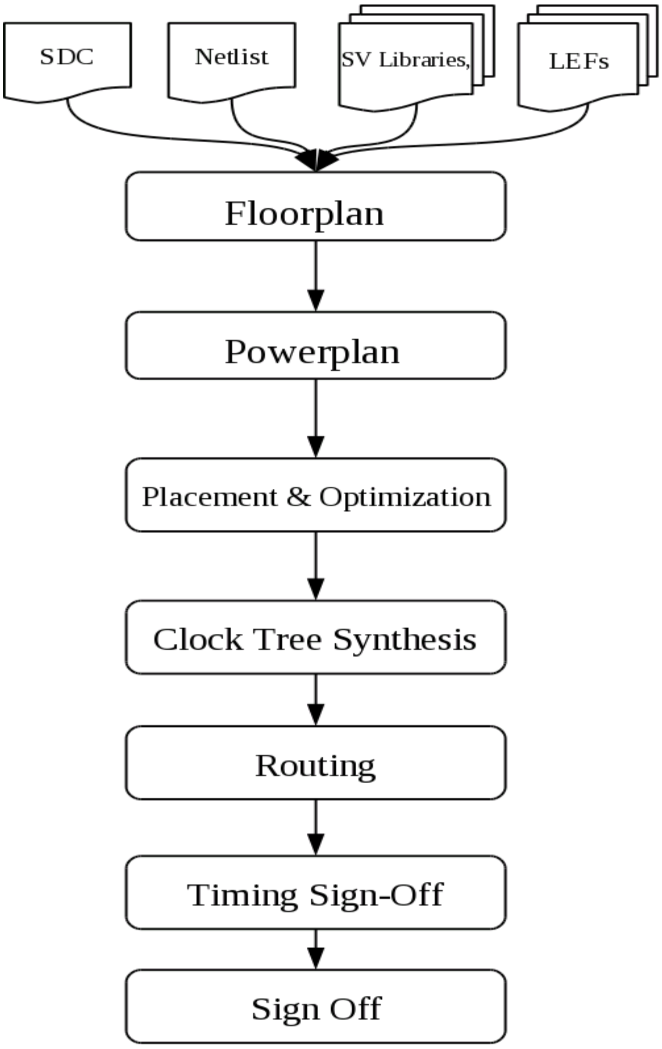


Figure 3.1 Traditional ASIC Flow

3.2 Floorplanning

Floorplanning is the art of any physical design. A well thought-out floorplan leads to an ASIC design with higher performance and optimum area. Floorplanning can be challenging in that it deals with the placement of I/O pads and macros as well as power and ground structures. Before one proceeds with physical floorplanning one needs to make sure that the data used during the course of physical design activity is prepared properly. Proper data preparation is essential to all ASIC physical designs in order to implement a correct-by-construction design.

The floorplanning problem is to plan the positions and shapes of the modules at the beginning of the design cycle to optimize the circuit performance with respect to the following:

- Chip area
- Total wire length
- Delay of critical path
- Routability
- others
  - e.g., noise, power dissipation, etc

Floorplanning a chip or block is an important task of physical design in which the location, size and shape of soft modules, and the placement of hard macros are decided. Depending on the design style or purpose, floorplanning can also include row creation, I/O pad or pin placement, bump assignment (flip chip), bus planning, power planning and more.

Floorplanning is very important when preparing the design for timing closure and detailed routing. Floorplanning in conjunction with placement and trail route is an iterative process. Floorplanning usually starts by preplacing blocks, modules and sub-modules according to the prepared floorplan. All other modules or blocks not in the prepared floorplan are left outside the chip area.

3.3 Power Planning

The next step is to plan and create power and ground structures for both I/O pads and core logic. The I/O pads power and ground buses are built into the pad itself and will be connected by abutment.

For core logic, there is a core ring enclosing the core with one or more sets of power and ground rings. A horizontal metal layer is used to define the top and bottom sides, or any other horizontal segment, while the vertical metal layer is utilized for left, right, and any other vertical segment. These vertical and horizontal segments are connected through an appropriate via cut.

Next consideration is to construct the standard cell power and ground that is internal to the core logic. These internal core power and ground buses consist of one or two sets of wires or strips that repeat at regular intervals across the core logic, or specified region, within the design. Each of these power and ground strips run vertically, horizontally, or in both directions.

Figure illustrates these types of power and ground connections.

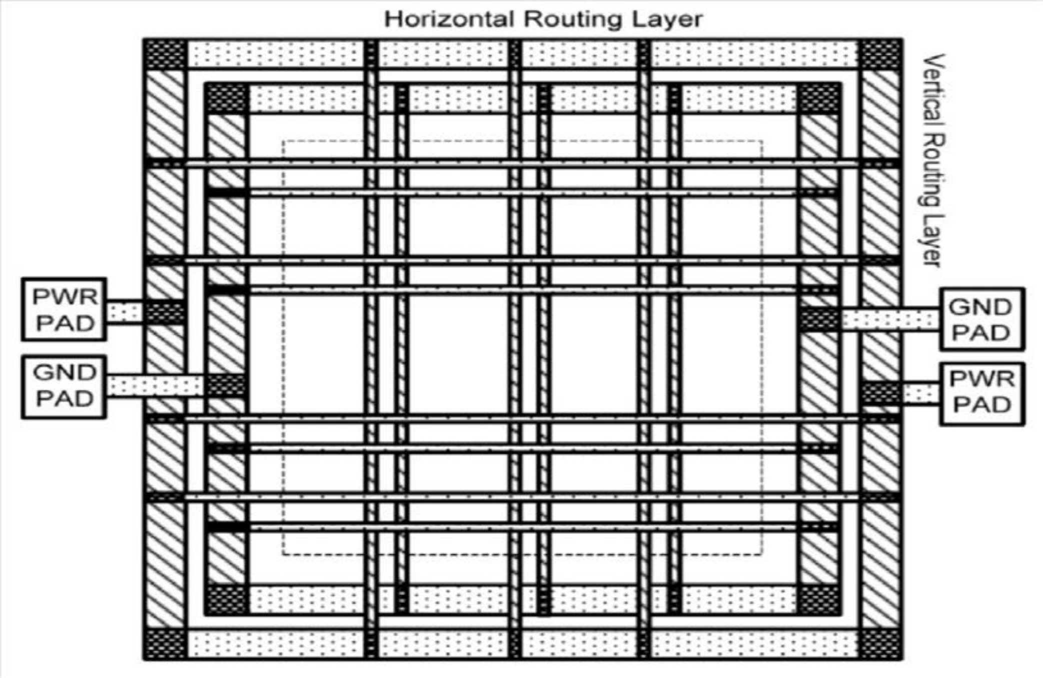


Figure 3.2 Power Plan

If these strips run both vertically and horizontally at regular intervals, then the style is known as power mesh. The total number of strips and interval distance is solely dependent on the ASIC core power consumption.

As the ASIC core power consumption (dynamic and static) increases, the distance of power and ground strip intervals increases. This increase in the power and ground strip intervals is used mainly to reduce overall ASIC voltage drop, thereby improving ASIC design performance. In addition to the core power and ground ring, macro power and ground rings need to be created using proper vertical and horizontal metal layers. A macro ring encloses one or more macros, completely or partially, with one or more sets of power and ground rings.

Another important consideration is that when both analog and digital blocks are present in an ASIC design, there is a need for special care to insure that there is no noise injection from digital blocks or core into the sensitive circuits of analog blocks through power and ground supply connections.

Much of this interference can be minimized by carefully planning the power and ground connections for both digital core and analog blocks. There are several methods to improve the noise immunity and reduce interference.

The most effective method is to decouple the digital and analog power and ground by routing the digital power/ground (DP and DG) and analog power/ground (AP and AG) supply connections separately as shown in figure.

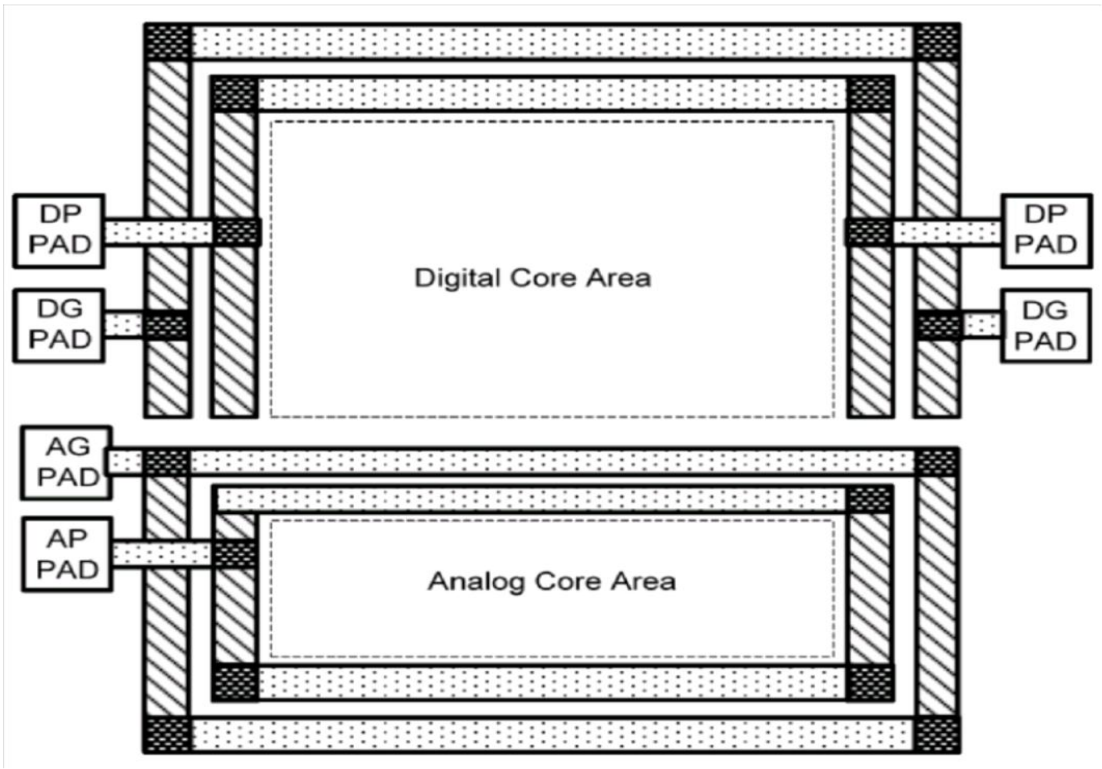


Figure 3.3 Power pad Connections

3.4 Placement

The goal of standard cell placement is to map ASIC components, or cells, onto positions of the ASIC core area, or standard cell placement region, which is defined by rows. The standard cells must be placed in the assigned region (i.e. rows) such that the ASIC can be routed efficiently and the overall timing requirements can be satisfied. Standard cell placement of an ASIC physical design has always been a key factor for achieving physical designs with optimized area usage, routing congestion, and timing behavior.

In the early days of physical design, the total area for placing standard cells consisted of the area required for the standard cell rows and the area required for channel routing. With advancement in place-and-route tools, standard cell channel routing has almost vanished because all place-and-route tools today are capable of routing over the standard cells. Over-standard-cell routing utilizes all empty space above the standard cells. This allows physical designers to create an ASIC that is as compact as possible without creating extra channels for routing purposes. With the disappearance of routing channels, the routing congestion problem has become more important. During standard cell placement, excessive congestion resulting in a local shortage of routing resources must be avoided. In over standard-cell routing , the objective of most place-and-route tools has been to utilize all the available core area to prevent routing overflow. This routing overflow accounts for an increase in ASIC device size and results in performance degradation.

Standard cell placement may be thought of as an automatic process that requires less physical designer intervention. However, a number of design constraints that can be applied during standard cell placement to achieve optimal ASIC design with respect to area, performance, and power. These constraints can be congestion, timing, power, or any combination thereof. Most place-and-route tools use a two-step approach to place standard cell instances. These steps are global and detail placement. The objective of the global placement algorithm is to minimize the interconnect wire lengths, whereas the objective of the detail placement algorithm is to meet design constraints such as timing and/or congestion, and to finalize the standard cell placement.

3.4.1 Global Placement

When the floorplan is first created, standard cells are in a floating state. This means that they are placed arbitrarily in the ASIC core and have not been assigned to a fixed location within the standard cell rows. At this time one can partition the standard cell area and assign a group of cells to these partitions, or simply group a set of standard cells. Almost every place-and-route tool supports cluster and region options. These two options are used to guide placement algorithms during standard cell placement.

Cluster refers to a group of standard cells that, during placement, are placed near each other. The location of the cluster is undefined until all standard cells have been placed. This option is mainly used to control the closeness of timing-critical components during placement and resembles a module definition in the structural netlist. Since the development of placement algorithms (e.g. interconnect driven), this option has been rarely used except in very special cases. An example of a standard cell cluster is shown in Figure.

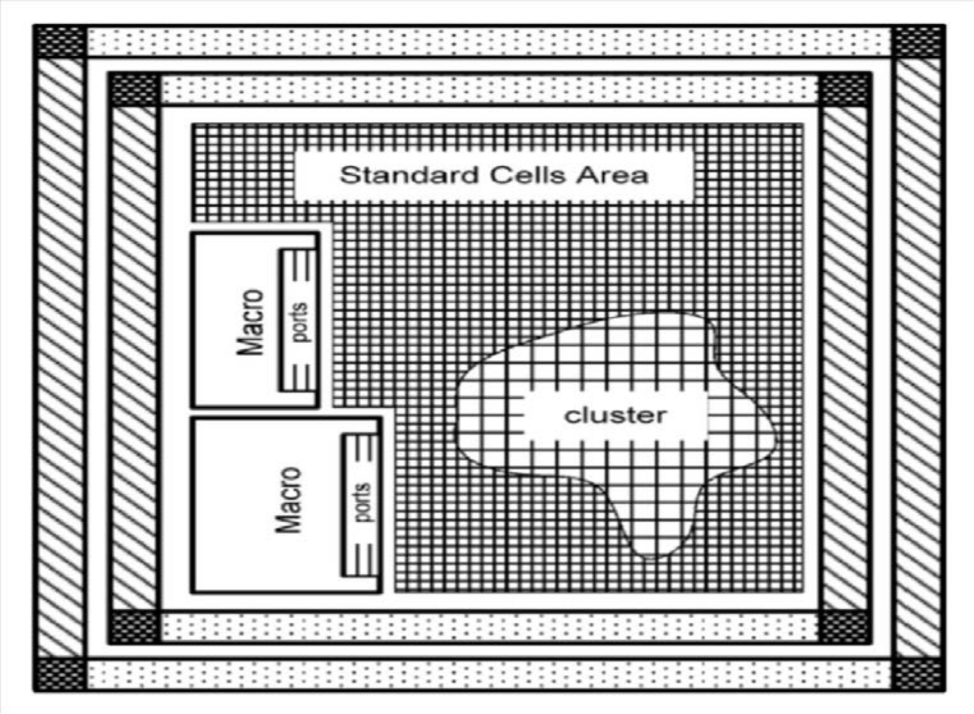


Figure 3.4 Cluster Placement

Region is very similar to the cluster method with the exception that the location of the region is defined prior to standard cell placement. The way this option is implemented is that a cluster or group of standard cells is created and then assigned to a particular area on the core ASIC. Regions can be soft or hard. Soft region, is physical constraint where logical module is assigned to a location in the core and boundary of the region, it is subject to change during standard cell placement. Hard region, however, is more rigorous than the soft region and defines a physical partition for modular design. It has boundaries that prevent standard cell crossing during placement. Using the hard region option, one must define the location as well as the shape of the region. This option is used primarily for timing related issues such as grouping clock, voltage, or threshold voltage domains.

In addition, a region can be exclusive or non-exclusive. An exclusive region only allows standard cells assigned to the region to be placed within the region. On the other hand, a non-exclusive region will allow standard cells that do not belong to the region to be placed within it. A hard region (or an exclusive region with a predefined physical boundary) might be used to enforce a floorplan consisting of separate blocks. This approach is useful for dividing the ASIC core area into regions that have different functions or physical aspects.

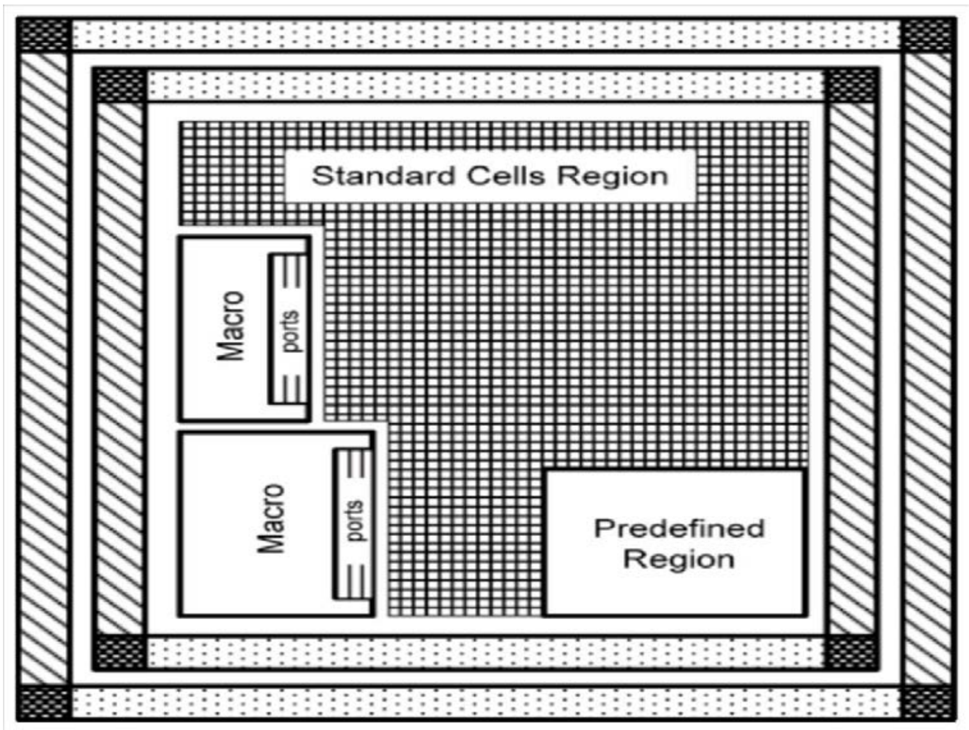


Figure 3.5 Regioned Placement

For example, a hard region can be used to partition the ASIC core area so one region has a different voltage from the rest of the design or other regions as shown in Figure.

After clusters and regions are defined, global placement algorithms begin distributing standard cell instances uniformly across the available ASIC core and use a method of estimation to minimize wire lengths. During this time, ASIC design is recursively partitioned along alternatively horizontal and vertical cut lines, and standard cell instances are assigned to rectangular bins, or slots, taking partitioning into account. Then, instances in each bin will be moved across each cut line in order to minimize the number of connections between each partition.

The procedure of partitioning and moving standard cell instances across cut lines terminates when certain stop criteria are satisfied. After design partitioning is completed, a legalization step is executed to remove any standard cell overlap and fit current placement into row structures. These types of global placement algorithms are classified as partition-based. There are two main cost functions associated with partition-based algorithms: to reduce the total wiring or routing length and to distribute the standard cell instances homogeneously in the ASIC core area such that optimal equilibrium among vertical and horizontal routing is achieved.

3.4.2 Detail Placement

Once all standard cell instances are placed globally, a detail placement algorithm is executed to refine their placement based on congestion, timing, and/or power requirements.

Congestion refinement or congestion-driven placement is more beneficial to ASIC designs with very high density, and the objective of the detail placer is to distance standard cell instances from each other such that more routing tracks are created among them. The quality of congestion placement directly relates to how well the global placer partitions the design and could have a negative impact on the device size and performance.

For minimal device size, one may use more routing layers. In determining the total number of routing layers to use, it is imperative to consider the trade-off between increasing the device size and using extra routing layers. In some instances, it may be more economical to increase the device size rather than adding extra routing layers (i.e. extra mask).

Timing-driven placement algorithms have been classified as either net or path based. Net-based schemes try to control the delay on a signal path by imposing an upper-bound delay or by assigning a weight to each net. Path-based approaches apply constraints to delay paths of small sub-circuits (the disadvantage of path-based algorithms is the fact that it is impossible to enumerate all paths within a design). The major challenge in timing-driven placement is to optimize large sets of path delays without enumerating them in the ASIC design. This optimization is accomplished by interleaving weighted connectivity-driven placements with timing analysis that annotates individual instances, nets, and path delays with design constraint information.

To meet these types of design constraint, various placement techniques have been proposed or used. The most well-known detail placement method is simulated annealing. Not only is simulated annealing efficient, it can also handle complex design constraints. Simulated annealing is a simulation-based placement technique and is used as an iterative improvement algorithm during the detail placement process. The objective of this procedure is to find an optimal or near-optimal placement for each pre-placed standard cell instance.

3.5 Clock Tree Synthesis (CTS)

The clock signal is generated external to the chip and provided to the chip through the clock entry point or the clock pin. Each functional unit which needs the clock is interconnected to the clock entry point by the clock net. Each functional unit computes and waits for the clock signal to pass its results to another unit before the next processing cycle. The clock controls the flow of information within the system. Ideally, the clock must arrive at all functional units at precisely the same time. In this way, all tasks may start at the same time and data can be transferred from one unit to another in an optimum manner. In reality, the clock signals do not arrive at all functional units simultaneously. The maximum difference in the arrival time of a clock at two different components is called clock skew. Clock skew forces the designer to be conservative and use a large time period between clock pulses, that is, lower clock frequency. The designer uses the clock period which allows for logical completion of the task as well as some extra time to allow for deviations in clock arrival times. If the designer can be provided a guarantee that the maximum deviation of the clock arrival time is small, then faster clocks can be used. The smaller the deviation, the faster the clock. Thus, controlling the deviation of signal arrival time is the key to improving circuit performance.

In order to reduce the deviation of the arrival times of the clock at two different components we need to build a buffer tree for the clock. This process of building up a buffer tree for the clock net is called as clock tree synthesis (CTS). CTS also helps in reducing the load capacitance seen on the clock root pin/port due to the high fanout for the clock which subsequently helps in reducing the transition time on the clock net. The buffers used for clock tree have equal rise and fall times which would help in achieving the required timing on the clock net.

Not just the above reasons are sufficient for the clock tree requirement. Given a source and n sinks clock tree synthesis must connect all n sinks to the source by an inter-connect tree so as to minimize:

- Clock Skew
- Clock insertion Delay
- Total wire length
- Noise and coupling effect

- ➔ Power dissipation
- ➔ Transition time

Some of the terminologies related to CTS are as follows:

- o **Insertion delay** is a measure of time it takes the clock to propagate from the root of the tree to the leaf cells
- o **Skew** is the measure of the difference of delay between the minimum and maximum time it takes the clock to reach the leaf cells
  - ▪ A pair of registers are sequentially-adjacent if only combinatorial logic (no sequential elements) exist between the two cells.
- o **Effective skew** is the measure of the difference of delay between the minimum and maximum time it takes the clock to reach the leaf cells which are sequentially-adjacent.

The following figure shows the general clock structure:

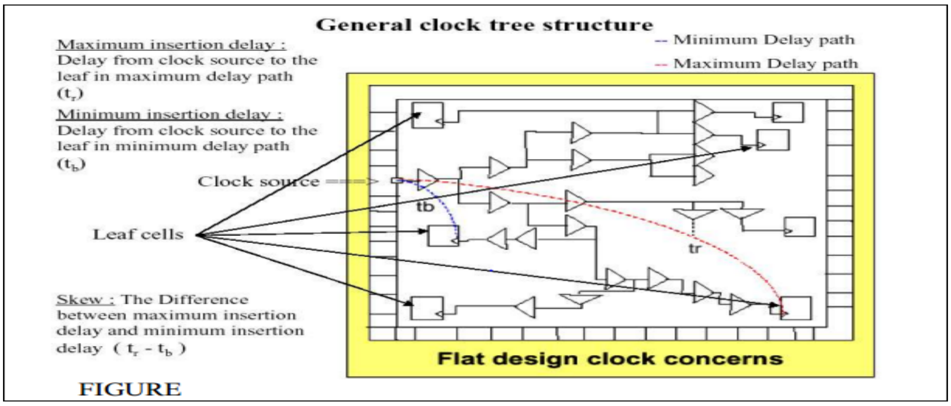

General clock tree structure
Maximum insertion delay :
Delay from clock source to the leaf in maximum delay path
(t_r)
Minimum insertion delay :
Delay from clock source to the leaf in minimum delay path
(t_b)
Clock source
Leaf cells
Skew : The Difference between maximum insertion delay and minimum insertion delay (t_r - t_b)
FIGURE
-- Minimum Delay path
-- Maximum Delay path
tb
tr
Flat design clock concerns

Figure 3.6 Clock Tree Structure

3.6 Routing

After completion of standard cell placement and power analysis, the next phase is to route the ASIC design and perform extraction of routing and parasitic parameters for the purpose of static timing analysis and simulation. As ASIC designs are getting more complex and larger, routing is becoming more difficult and challenging. It is possible for routing to fail to complete, or to take an unacceptable amount of execution run time. Besides the routing algorithms, the factors which influence the routability of a given ASIC are the layout of standard cells style, a well-prepared floorplan, and the quality of standard cell placement.

Routing algorithms are mainly classified as channel or over-the-cell based routers. Channel-based routing was used in the early days of ASIC physical design. This was because semiconductor factories were not able to process large numbers of metal layers (e.g. two or three). Thus with a limited number of routing layers, all connections were restricted to the area between cells or around macro blocks such as memories.

Fundamentally, channel, or river-based, routers use reserved space between standard cell rows (routing channels) and feed-through (dedicated routing areas inside the standard cell layout) to perform routing between instances as shown in Figure.

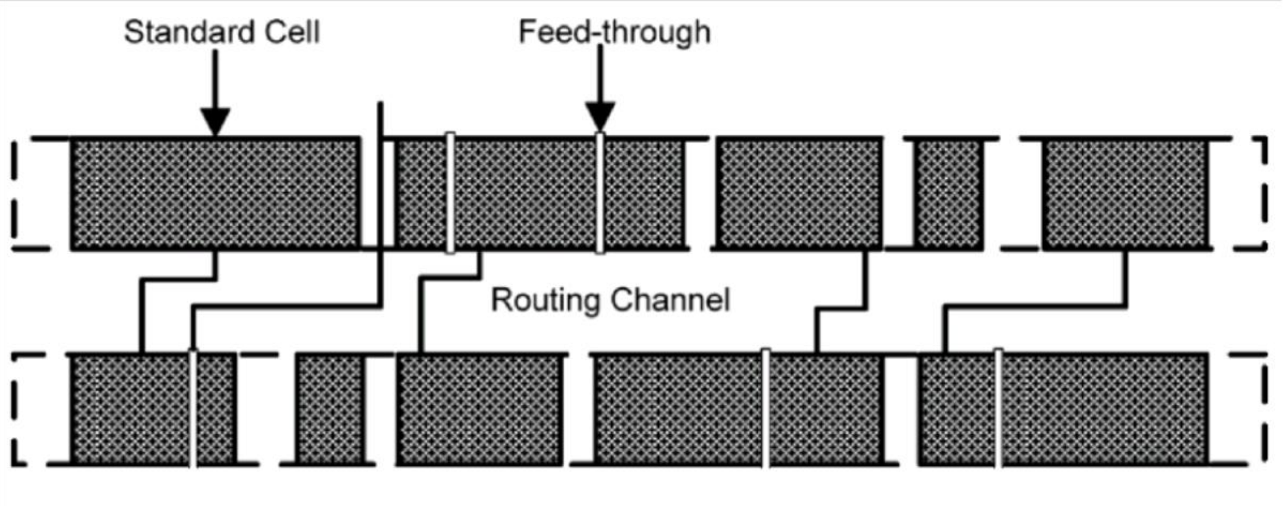


Figure 3.7 Routing

With recent improvements in semiconductor processes and increasing numbers of routing layers, the routing channel and standard cells with feed- through have been eliminated, and over-the-cell based routing is widely utilized by many physical synthesis and place-and-route engines during the physical design of ASIC devices.

Due to the inherent complexity of ASIC designs and the very large numbers of interconnections associated with them, the overall routing is performed in three stages:

- ➔ special routing
- ➔ global routing
- ➔ detail routing

3.6.1 Special Routing

Special routing is used for standard cells, macro power, and ground connections. Most special routers use line-probe algorithms. The line-probe method uses line segments to connect standard cells, macro power, and ground ports to ASIC power and ground supplies.

Line-probe routers use a generated connectivity list of sources and targets (the generated connectivity list can be port-to-port, port-to-line, or line-to-line) or connection. The line segments are used to perform routing according to the connectivity list starting from the target and tracing the line segment until the source is reached. These generated line segments do not pass through any obstructions. If they were to pass through an obstruction, they might create an unfixable design rule violation. Therefore, one needs to make sure that there are no obstructions where the power and ground ports are located.

In connecting macro power and ground ports to the main power and ground nets, line-probe routers use the size of the ports to set the width of the power and ground segments. This automatic width setting may not be adequate for current density considerations, and one may need to create more power and ground ports to satisfy the current density requirements.

3.6.2 Global Routing

Global routing is the decomposition of ASIC design interconnections into net segments and the assignment of these net segments to regions without specifying their actual layouts. Thus, the first step of the global routing algorithm is to define routing regions or cells (i.e. a rectangular area with terminals on all sides) and calculate their corresponding routing density. These routing regions are commonly known as Global Routing Cells, or GRC, as shown in Figure.

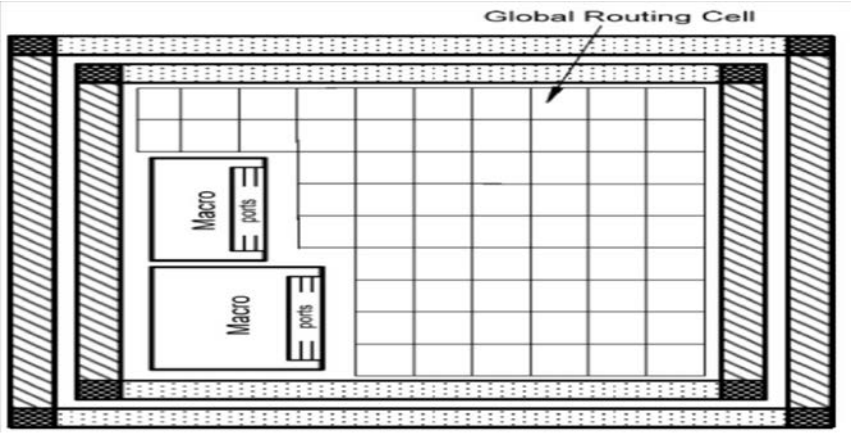


Figure 3.8 Global Routing Cells in Core area

The density, or capacity, of these cells is defined as the maximum number of nets crossing the routing regions and is a function of the number of routing layers, cell height in vertical or horizontal direction, minimum width, and spacing of wire.

Global routing uses a graph to model the interconnection networks. The vertices of the graph denote the standard cell ports. The edges of the graph correspond to connections between two ports within routing cells and among routing cells themselves. The graph is constructed by means of region assignment and assignment of each net to a pin on the boundary of the routing region. After global routing is performed, the pin locations will be determined such that the connectivity among all standard cells in the ASIC core area is minimal. Almost all global routers report the design routability statistic using overflow or underflow for Global Routing Cells (GRC), which is the ratio of routing cells capacity and the number of nets that are required to route a given routing cell for all vertical and horizontal routing layers. The GRC statistic is a very good indication of wiring congestion and shows the number of nets needed to route a region versus the available number of routing layers. For an ASIC design to be routed completely without any design rule violations, this number needs to be less than one.

Global routing algorithms generate a non-restricted route (i.e. not a detail route) for each net in the design and use some method of estimation to compute wire lengths and extract their corresponding parasitics.

3.6.3 Detail Routing

The objective of detail routing is to follow the global routing and perform the actual physical interconnections of ASIC design. Therefore, the detail router places the actual wire segments within the region defined by the global router to complete the required connections between the ports. Detail routers use both horizontal and vertical routing grids for actual routing. The horizontal and vertical routing grids are defined in the technology file for all layers that are being used. The detail router can be grid- based, gridless-based, or subgrid-based.

Grid-based routing imposes a routing grid (evenly spaced routing tracks running both vertically and horizontally across the design area) that all routing segments must follow. In addition, the router is allowed to change direction at the intersection of vertical and horizontal tracks as indicated in Figure. The advantage of grid-based routing is efficiency. When using a grid-based router, one needs to make sure that the ports of all instances are on the grid.

Otherwise, they can create physical design rule errors and will be difficult to resolve with the router.

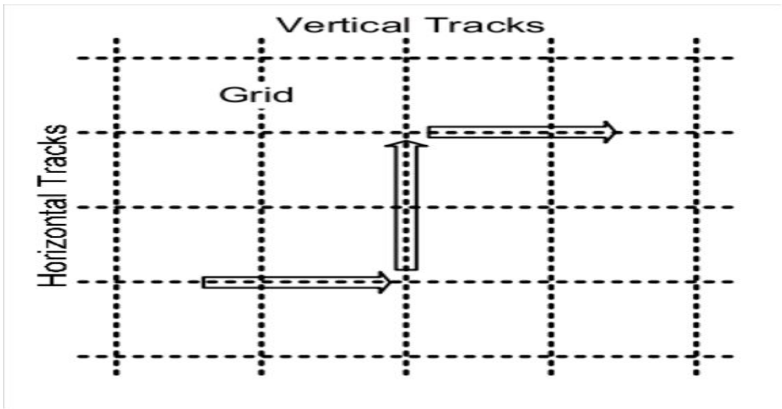


Figure 3.9 Tracks For Routing

Grid less-based (or shape-based) routers do not follow the routing grid explicitly, but are dependent on the entire routing area and are not limited by grid痴 restrictions. They can use different wire widths and spacing without routing grid requirements. The most fundamental problem with this type of router is that they are very slow and can be very complicated. The subgrid-based router brings together the efficiency of grid-based routers with the flexibility of the gridless-based routers. The subgrid-based router follows the normal grid similar to the grid-based router. However, a subgrid-based router considers these grids only as guidelines for routing and is not required to use them, as illustrated in Figure.

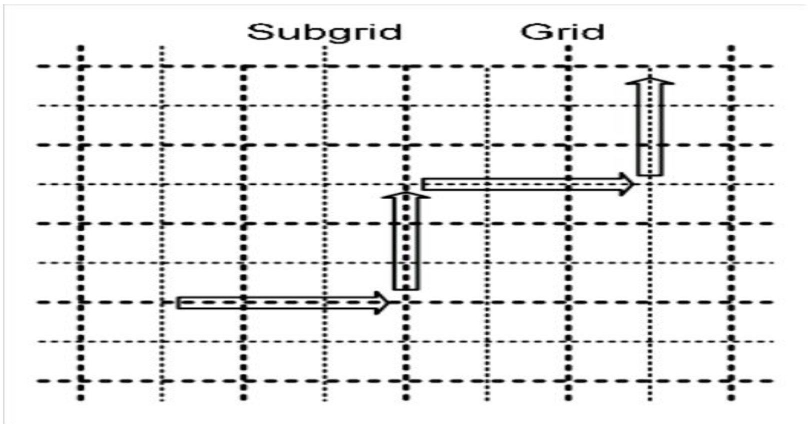


Figure 3.10 Sub-Grid Based Routing

3.6.4 Post Route Optimization

Routed design is sometimes required to be further optimized from timing point of view. This step is possible only after all the routing violations are taken care of. Optimization involves the following operations:

- ➔ Buffer Insertion
- ➔ Buffer Deletion
- ➔ Upsizing a cell
- ➔ Downsizing a cell

3.6.5 Shielded Routing

Routed design is sometimes required to be further optimized from cross-talk effect point of view. If two nets are running in parallel, and one of the net driver cells has high drive strength, then there is a possibility for the cross-talk delay or noise or both to be observed on the weak driver net. The high drive strength driven net is called as an aggressor. The low drive strength driven net is called as a victim. In order to reduce the cross-talk effect, the methods followed can be:

- ➔ Increased wire spacing
- ➔ Downsizing the aggressor driver cell
- ➔ Upsizing the victim driver cell
- ➔ Inserting buffer on the victim
- ➔ Placing a shield wire, which is grounded or connected to VDD, between the aggressor and the victim

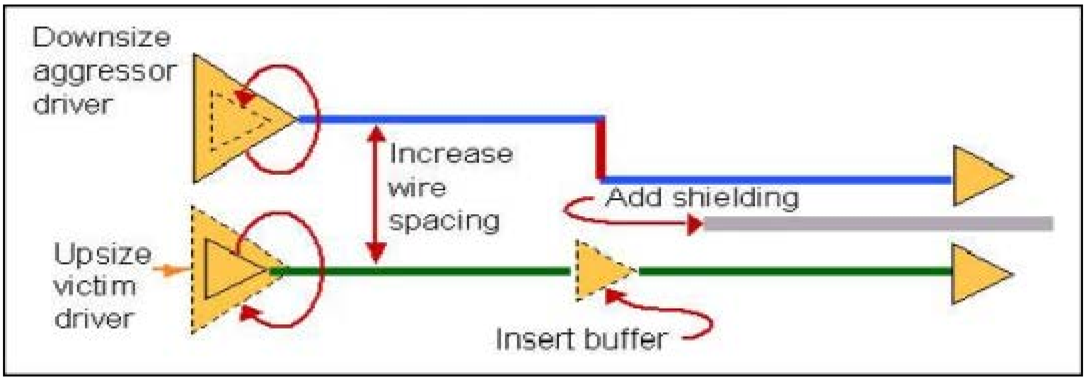


Figure 3.11 Shielding Techniques

3.6.6 ECO Routing

Minor routing corrections can be carried out on a routed design through ECO routing. The time taken should be a fraction of the total routing time. This type of routing is done when there is less than 10% change in the total routing. For example, if the timing violations are fixed by buffer insertion or buffer deletion or upsizing or downsizing the cells after routing and these changes are very few then an ECO (Engineering Change Order) routing can be performed for the design to be completely routed.

3.6.7 Route Only Specific Area

If a design has very few DRC violations after routing, then it may be sufficient to reroute only in the area where there are DRC violations. In such cases area routing is very useful because this will save a lot of run time.

3.7 Timing Analysis and Fixing

The goal of timing analysis is to verify that a design meets timing requirements under a specified set of timing constraints, such as arrival and departure times, operating conditions, slew rates, false paths, and path delays. Performing timing analysis lets you determine how fast a design can run without incurring timing violations. You can use the results of timing analysis to fine tune and debug the speed-limiting, critical paths in a design.

You can perform timing analysis using Cadence and Synopsis constraint formats and timing libraries, such as TLF, lib, and Stamp Models. The Timing Analyzer can calculate instance-based delays and include interconnect RC loads. It calculates the signal propagation delay through each instance and net combination then generates delay, slew rate, and net load information in standard delay format (SDF).

The delay information in the technology library applies to the timing arcs from input ports to output ports of each cell and the corresponding wire delays. The cell delays and the wire delays are expressed as a function of the physical characteristics of the nets in the design, such as wire capacitance and wire resistance.

The basic timing checks carried out for synchronous signals are setup, hold and transition time checks. For a synchronous signals, these are recovery, removal and transition time checks. In addition to these the maximum capacitance violations are also checked for.

There are four different timing paths in a design. They are

(i) Design input to register

(ii) Register to Register

(iii) Register to Design output

(iv) Design input to Design output

Setup time for a flip-flop is the minimum amount of time before the clock edge for which the data on the data path of the flip-flop should be stable. This is one of the two factors which ensures data to be correctly registered into the flip-flop. The setup time constraint restricts the maximum delay in a timing path. The setup time violating paths are fixed by reducing the path delays. Delay can be reduced by restructuring the logic from a complex logic cell to simple logic cells or by deleting the unnecessary buffer cells or by using a high drive strength cells for driving high load capacitances or by moving the cells close together. Setup time of the timing paths are checked and fixed using the worst case design corner.

Hold time for a flip-flop is the minimum amount of time after the clock edge for which the data on the data path of the flip-flop should be stable. This is the other factor which ensures data to be correctly registered into the flip-flop. The hold time constraint restricts the minimum delay in a timing path. The hold time violating paths are fixed by increasing the path delays. Path delays can be increased by adding buffer cells into the path. But addition of buffers in one path should not violate setup time in another path. Hold time of the timing paths are checked and fixed using the best case design corner.

Transition time of a signal can be either rise time or fall time. Rise time of a signal is the time taken by the signal to rise from 10% of VDD to 90% of VDD. Fall time of a signal is the time taken by the signal to fall from 90% of VDD to 10% of VDD. The designer requires that worst case transition time in the design should not exceed the maximum limit. Inorder to fix transition time violations, the driver of the violating net can be upsized or a buffer can be added on that net. Transition time of the signal nets are checked and fixed using the worst case design corner.

3.8 Parasitic Extraction

The process of extracting interconnect information in any one of the parasitics format is called “Parasitic Extraction". In this process, the interconnects are modeled in terms of RC or RLC pi networks and is written out in Spice like syntax. The STA (Static Timing Analysis) tools use this file to carry out min-max type STA on the routed design, where min STA is done for hold time check, max STA is done for setup and transition time checks and typical STA is done for al l the three timing checks.

Some of the STA tools are PT - Prime Time from Synopsis, PEARL- from Cadence and Envisia- from Cadence.

To carry out post-route STA, the interconnect information has to be extracted from the layout in any one of the following form:

- ➔ SPF- standard parasitic format
    - ➢ RSPF-reduced standard parasitic format
    - ➢ DSPF-detailed standard parasitic format
- ➔ SPEF- standard parasitic extraction format

The following figure shows an example of how 2 different interconnects can be modeled in RLC and C networks:

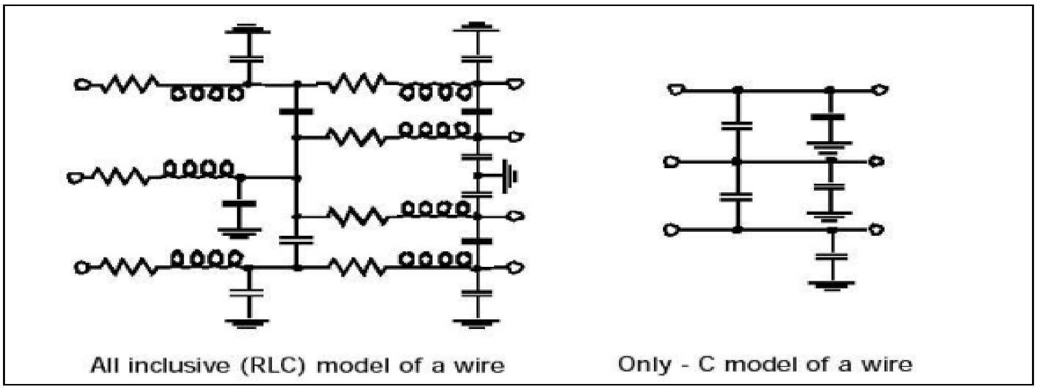


Figure 3.12 Extraction models

3.8.1 Extraction Basics

As process technologies shrink, five different major effects start to happen:

- ➔ Device(gate) delays decrease, due to the thinning gate oxide and reduced gate length.
- ➔ Inter connect resistance increases, because of shrinking wire widths.
- ➔ Vertical heights of interconnect layers increase, in an attempt to offset increasing interconnect resistance.
- ➔ The area component of interconnect capacitance no longer dominates. Lateral (sidewall) and fringing components of capacitance start to dominate the total capacitance of the interconnect.
- ➔ Interconnect capacitance dominates total gate loading.

With reduced feature size, design density is increasing. With advanced technologies and increasing complexities the die size is increasing and the percentage of global interconnects in a design is also increasing. But the interconnects are not sizing in the same way as transistors. To keep resistance almost the same, the higher layers wires are becoming taller. This is however, increasing the capacitance. Because of long lengths and higher frequency of operation, inductive effects are becoming prominent. In order to gauge the amount of delay added by the interconnects, there is increasing need to model them as RC networks or RLC networks.

Total delay of a path includes (as shown in the following figure):

- ➔ Device delay
- ➔ Interconnect delay-can be estimated through modeling
- ➔ Slew rate

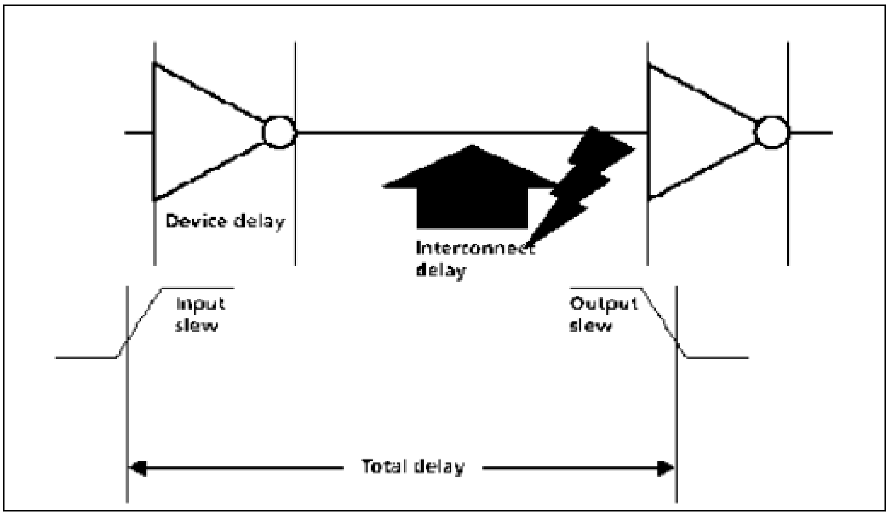


Figure 3.13 Total Delay Calculation

Older technology used wide wires. These wires had more cross-sectional area and so had less resistance& more capacitance. So early parasitic extraction models consisted of only capacitance wire models and the resistance is considered to be negligible as seen in the following figure:

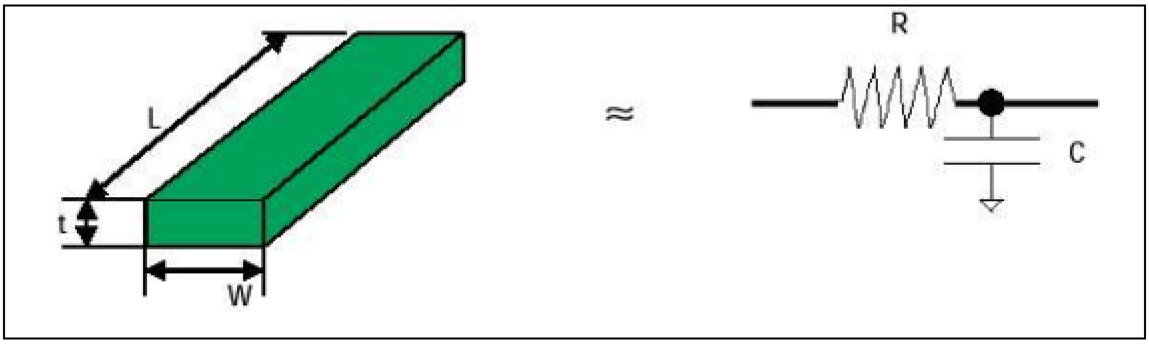


Figure 3.14 Capacitance Model

However, with technology scaling, the width of the wire is reduced and the resistance of the wire is no longer negligible. Also the number of global wires have increased. Only-C parasitic model is not enough for modeling the interconnects. Parasitic models consisting of lumped RC is good enough approximation, but distributed RC models are better than lumped RC models.

3.8.2 Resistance Calculation

Resistance of a metal wire is directly proportional to the length of the wire and inversely proportional to the cross-sectional area of the wire. The following formula is used for resistance calculation of any metal wire:

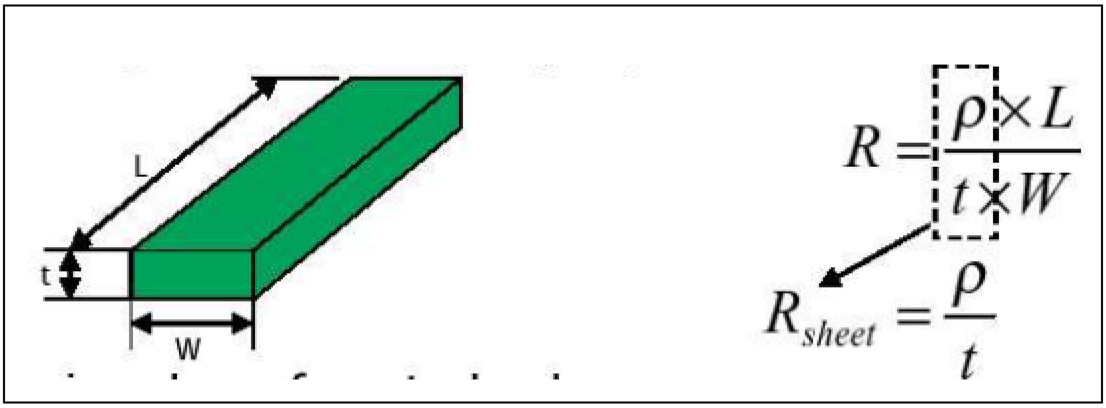
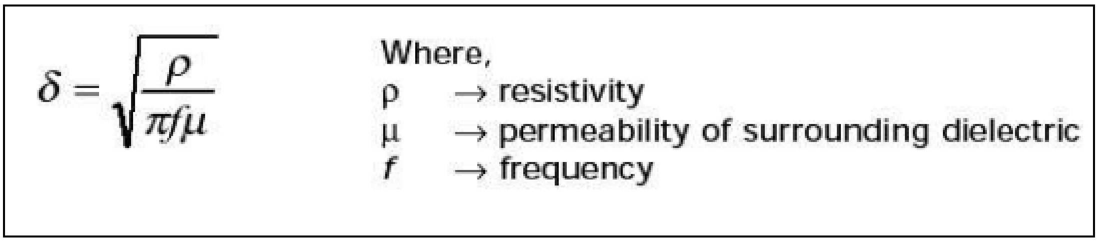

L
t
W
$R = \frac{\rho \times L}{t \times W}$
$R_{sheet} = \frac{\rho}{t}$

Figure 3.15 Resistance Calculation Model

In a given metal layer for a technology the thickness of the wire is fixed and the resistivity of the metal is fixed. Therefore - R sheet is fixed. One can calculate the resistance by multiplying the sheet resistance withL/W.

R = Rsheet* L/W

Resistance is not necessarily purely linear. Nonlinear component exists due to " skin effect" where resistance becomes frequency dependent. At high frequency, due to skin effect current tends to flow primarily on the surface of the conductor. This causes an exponential fall in current density with depth of the conductor. Skin depth “δ" is defined as the depth where the current falls off to a value e-1 times of its nominal value. The following formula gives the skin depth:

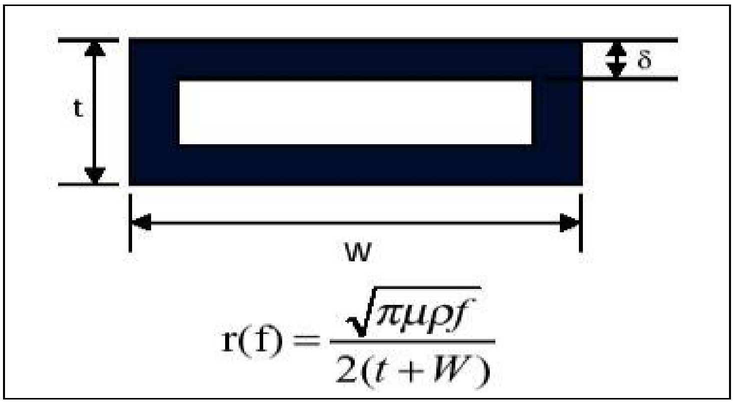


$$\delta = \sqrt{\frac{\rho}{\pi f \mu}}$$

Where,
ρ → resistivity
μ → permeability of surrounding dielectric
f → frequency

Figure 3.16 Resistivity equation

Skin effect can be approximated by assuming that the current flows uniformly in an outer shell of the conductor with thickness “δ". Increased resistance causes more attenuation of the signal flowing through it. The following figure illustrates the skin effect.

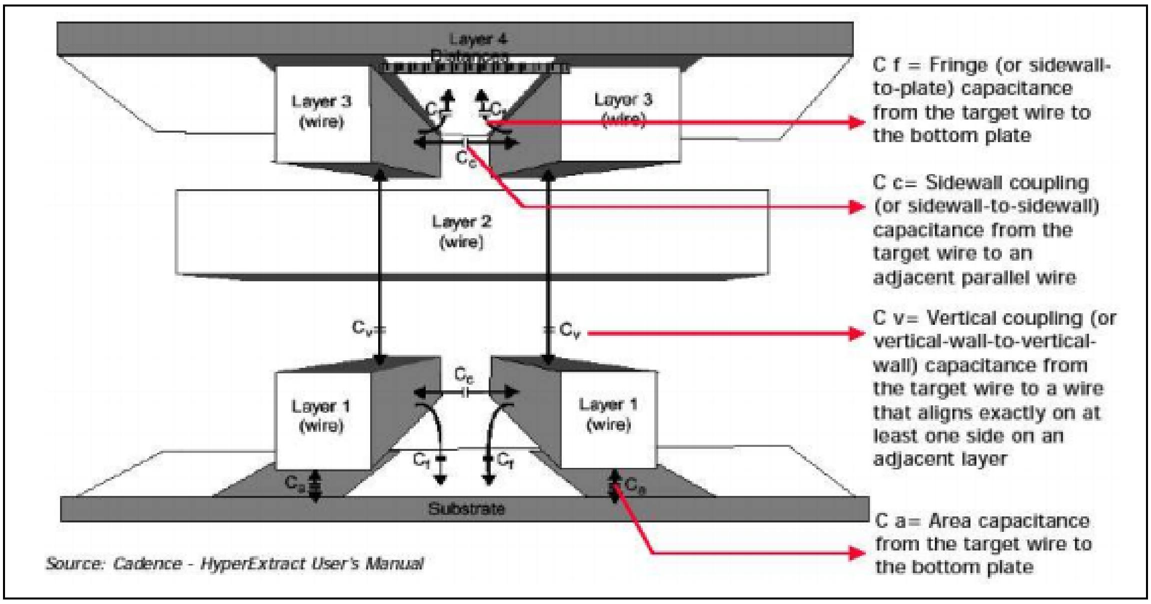


$$r(f) = \frac{\sqrt{\pi \mu \rho f}}{2(t + W)}$$

Figure 3.17 Resistance model

3.8.3 Capacitance Calculation

Capacitance of a wire is a function of shape of the wire, distance to surrounding wires and distance to the substrate. Estimating capacitance is a non-trivial task and is a subject of active research. To get an accurate estimate, electric field solvers (2D or 3D) are used. Solving fields is slow and will take ages for estimating capacitance of the whole chip. So various assumptions and approximations are used to get quick estimates. There can be 4 types of capacitance formations in a design which are illustrated in the below figure:



C f = Fringe (or sidewall-to-plate) capacitance from the target wire to the bottom plate

C c= Sidewall coupling (or sidewall-to-sidewall) capacitance from the target wire to an adjacent parallel wire

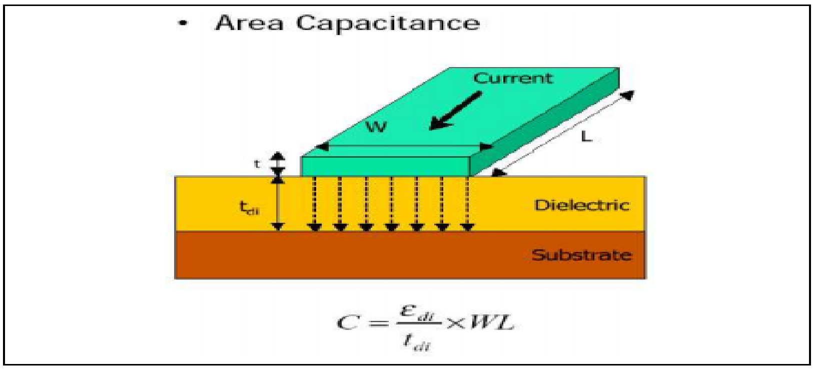
C v= Vertical coupling (or vertical-wall-to-vertical-wall) capacitance from the target wire to a wire that aligns exactly on at least one side on an adjacent layer

C a= Area capacitance from the target wire to the bottom plate

Source: Cadence - HyperExtract User's Manual

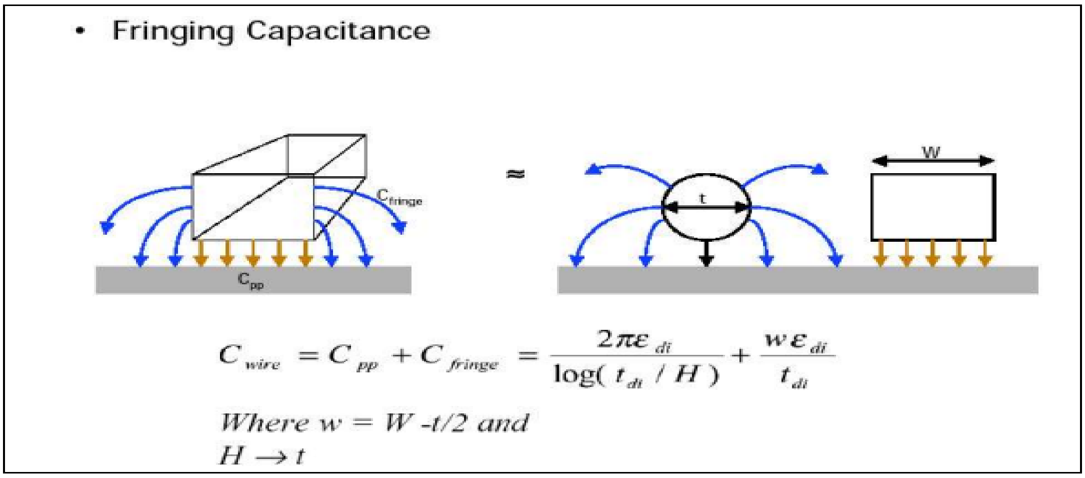
Figure 3.19 Capacitance Calculation

In the above figure Metal layer 1 is orthogonal to Metal layer 2 and Metal layer 2 is orthogonal to Metal layer 3. The following figures illustrate the calculation of area and fringe capacitances:



$$C = \frac{\varepsilon_{di}}{t_{di}} \times WL$$

Figure 3.20 Area calculation



- Fringing Capacitance

$$C_{wire} = C_{pp} + C_{fringe} = \frac{2\pi\varepsilon_{di}}{\log(t_{di}/H)} + \frac{w\varepsilon_{di}}{t_{di}}$$

Where $w = W - t/2$ *and*
$H \rightarrow t$

Figure 3.21 Fringing capacitance

3.9 Cross Talk Analysis

Cadence's Celtic is the tool used for Sign-off Cross-Talk analysis. Crosstalk is the undesired electromagnetic coupling between signal lines that causes functional failures and delay variation. Because it is not practical to check a full chip (with millions of cell instances) for crosstalk effects through dynamic simulation, CeltIC employs a static analysis technique, which enables practical checking of crosstalk effects on a chip-widebasis.

The analysis of cell instances is based on calculating the worst-case noise waveforms on each primary input and cell instance output node. The noise immunity of a cell instance is verified given the noise waveforms appearing at each cell instance input. For noise on delay analysis, CeltIC can employ SPICE-like simulations with transistors, if necessary, to accurately calculate the switching waveforms. CeltIC assumes all worst-case crosstalk scenarios unless prohibited by logic or timing constraints.

The CeltIC features introduce nets called victims and aggressors. Victims are nets that suffer from noise effects. A victim net is the net that is typically used for noise calculations. The adjacent nets that contribute to the noise are the aggressors, also known as attackers. Parallel adjacent lines on the same layer are the major contributors to crosstalk. Crosstalk can be attributed to the following:

3.9.1 Crosstalk-induced functional failures

The following figure shows how a noise pulse (or glitch) induced by an aggressor (attacker) net causes a logic-level glitch on a victim net and its receivers. This produces unintended logic transitions on receivers, resulting in repeated functional failures

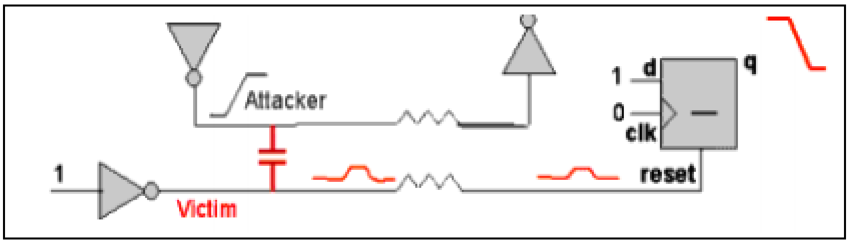


Figure 3.22 Cross Talk Analysis

The following figure shows how to test glitches to determine if they will be rejected by their receiving latches or flip-flops. The input is tested against the filtering threshold.

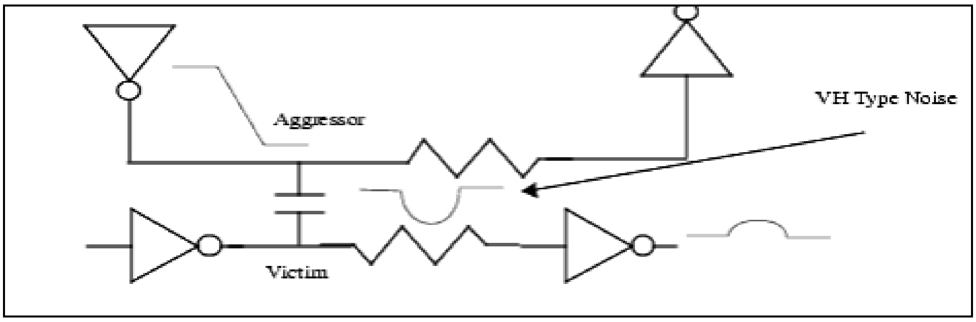


Figure 3.23 Cross Talk Types

The following figure shows how to compute for all possible aggressors to calculate maximum delta voltage for a net and then test against the threshold.

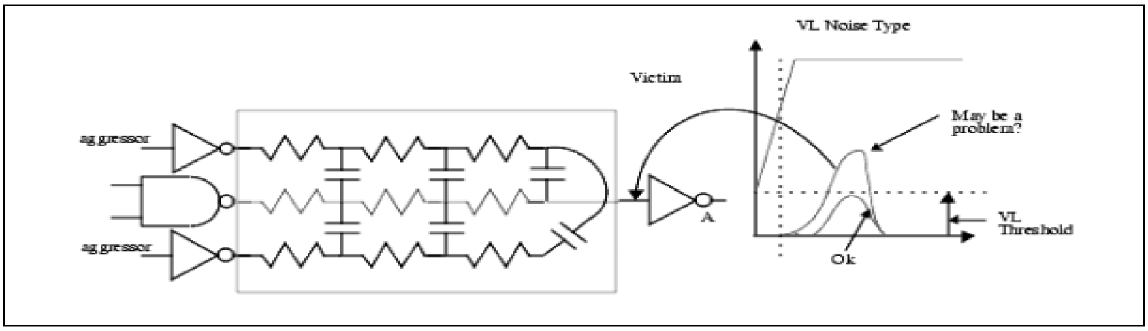


Figure 3.24 Cross Talk Types

3.9.2 Crosstalk-induced delay changes (noise-on-delay effects)

Coupling capacitance results in delay changes. The total capacitance seen by a signal on the victim net depends on the switching direction of neighboring nets. The delay is either increased or decreased. The following figure shows how to analyze and generate an incremental SDF file for speed-up and slow-down interconnect delays.

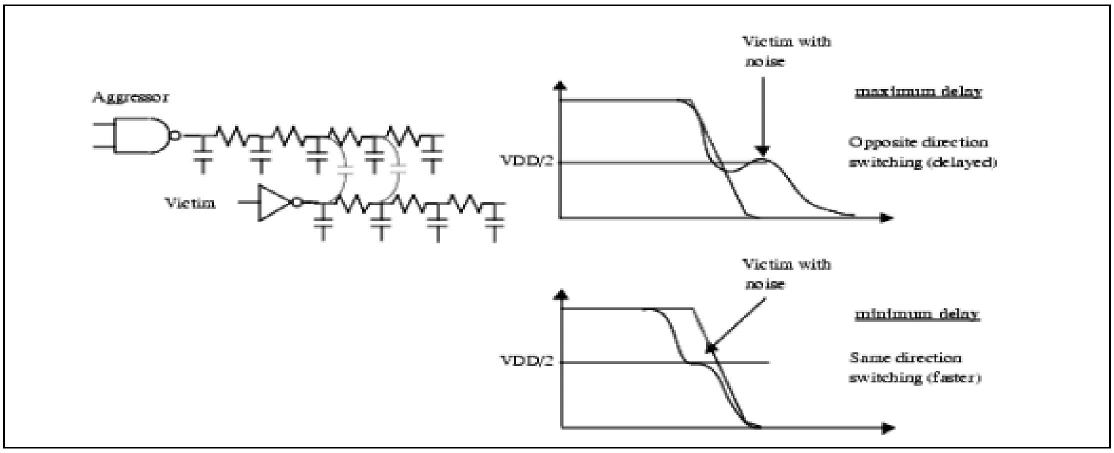


Figure 3.25 Crosstalk-induced Delay Changes

3.10 Physical Verification

The primary objective of physical verification is to check the ASIC layout against process rules provided by semiconductor foundries to insure that it can be manufactured correctly. Dramatically decreased device sizes, with respect to the number of transistors and routing layers involved in the net result for physical verification of larger layout databases, need to be verified against rules that are more complex. Verifying this amount of data puts strain on the software and computer infrastructure. Therefore, to improve physical verification efficiency and debugging procedures, one recommendation is to make sure that the ASIC is physically designed and implemented by employing a correct-by-construction methodology. Several checks are performed during physical verification. Mainly these checks are

➔ Layout Versus Schematic (LVS)

➔ Design Rule Check (DRC)

➔ Electrical Rule Check (ERC)

LVS verification examines two electrical circuits to see if they are equivalent with respect to their connectivity and total transistor count. One electrical circuit corresponds to transistor level schematics, or netlist (reference), and another electrical circuit is the result of the extracted netlist from the physical database. Through the verification process, if the extracted netlist (layout) is equivalent to the transistor level netlist, they should function identically.

Finally, but most importantly, it is recommended to begin the verification process during an early stage of the design to insure that the physical database is correct. Under the assumption that most of today physical design tools are capable of producing error free place and routed designs, the most common sources of physical errors arise during the floorplanning stage and most often are related to power and ground connections. Power and ground shorts, and/or opens, impact device (transistor) recognition during LVS verification, which then leads to an extremely long execution time.

DRC is considered a prescription for preparing photo-masks that are used in the fabrication of the ASIC design. The major objective of the layout or design rule checking is to obtain optimal circuit yield without design reliability losses. The more conservative the design rules are, the more likely it leads to correctly manufactured ASIC designs. However, the more aggressive the design rules are, the greater the probability of yield losses.

Another aspect of DRC allows one to check for DFM (Design For Manufacturability) such as contact/via overlaps and end-of-line enclosures. The DFM rules are considered optional and are provided by the silicon manufacturer. From a yield perspective, it is beneficial to check the ASIC physical design for DFM rule violations and then correct the errors as much as possible without influencing the overall die area.

ERC verification is intended to verify an ASIC design electrically. In comparison with LVS that verifies the equivalence between the reference and extracted netlists, ERC checks for electrical errors such as open input pins or conflicting outputs. A design can pass LVS verification, but may fail to pass ERC checks. For example, if there is an unused input in the reference netlist, then the extracted (routed) netlist will also contain the same topology. In this case, the result of the LVS process will be correct by matching both circuits, whereas the same circuits will cause an error during ERC verification based on the fact that a floating gate can lead to excess current leakage.

The ERC verification process is considered to be a custom verification rather than a generic verification. The user can define many electrical rules for the purpose of verification. These rules can be as simple as checking for floating wires, or more complex, such as identifying the number of PWELL- or NWELL-to-substrate contacts, latch-up, and ESD.

CHAPTER 4

CONGESTION ISSUES, MITIGATION

4.1 What is congestion ???

If the no. of routing tracks available for routing in a one particular area is less than the required routing tracks then the area said to be congested. There will be a limit for no. of nets that can be routed through particular area.

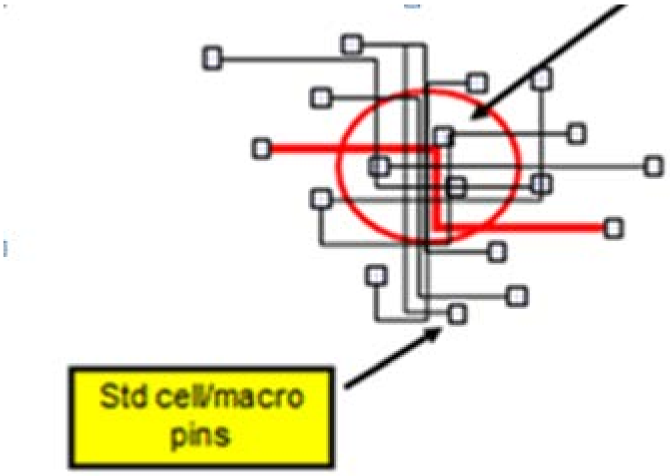


Figure 4.1 Congestion in an area

4.2 Reasons for congestion

- High standard cell density in small area,
- Placement of standard cells near macros,
- High pin density at the edge of macros
- Bad floorplan
- During IO optimization tool does buffering, so lot of cells placed in the core area.

4.3 Congestion is an outcome of two factors here in this node

- More performance and more complexity in the design leads to increase in the utilization.
- Reduced the Routability of lower layers due to double patterning

This had increased deficit of available vs required resources.

4.4 Congestion mitigation flow

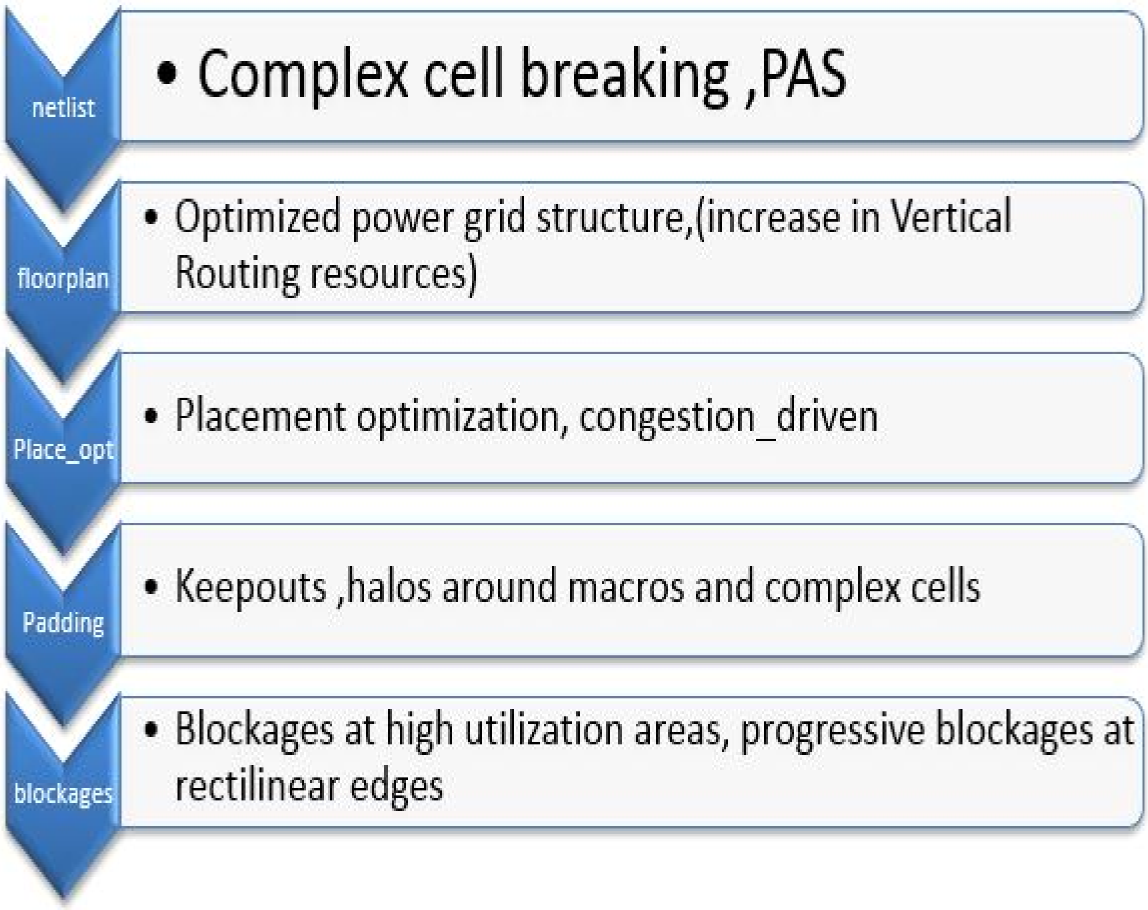

netlist
• Complex cell breaking ,PAS
floorplan
• Optimized power grid structure,(increase in Vertical Routing resources)
Place_opt
• Placement optimization, congestion_driven
Padding
• Keepouts ,halos around macros and complex cells
blockages
• Blockages at high utilization areas, progressive blockages at rectilinear edges

Figure 4.2 congestion mitigation flow

4.4.1 Placement Aware Synthesis (PAS)

At smaller process nodes, chip designers are struggling to meet their aggressive schedules and power, performance, and area (PPA) demands in the ever-so-competitive system-on-chip (SoC) market. One of the most pressing problems designers are facing these days is not knowing how the netlist they produce in synthesis will work out in the place-and-route (P&R) process.

Not only does this lack of predictability impact the design itself, but it also dampens, unnecessarily, quality of life. After all, isn’t it always better when you know that what you created is good – and will allow you to go home at a reasonable time each evening, without worrying that an unknown problem will surface the next day?

At 28nm and below, SoCs are much more complex, making it more challenging than ever to meet PPA targets. Wires dominate the timing at these advanced nodes, so there’s a greater chance of encountering issues such as routing congestion and timing delays. You must cram more transistors into the die, and have to reduce dynamic and leakage power.

So, why are you still doing traditional synthesis?

Physically aware synthesis – the ability to bring in physical considerations much earlier in the logic synthesis process is something that can dramatically improve the design process and significantly shorten the time spent fixing problems. Let’s discuss some key physically aware synthesis techniques that can help you speed up the physical design closure process for your next high-performance, power-sensitive SoC.

Today’s physically aware synthesis technologies bring physical interconnect modelling earlier into the synthesis process to help you create a better netlist structure, one that’s more suitable for today’s P&R tools.

You can start with no floorplan, and allow the synthesis to come up with one. You can give it a very basic floorplan. But the better the floorplan you have, the better you can take advantage of global synthesis optimization with the more detailed physical interconnect. Essentially, you are getting rid of the old logical-physical barrier. You’ll no longer need to, with fingers crossed, wait for your “backend” engineer to say “yay” or “nay.”

Before you can come up with a good floorplan, you need to have a good initial netlist. To create that initial netlist, you can still use physical information and use physical layout estimation (PLE). For this, you just need some basic physical information, such as LEF and cap tables/QRC tech files. The floorplan DEF is optional here.

PAS is a physical modelling technique for capturing timing closure P&R tool behaviour for RTL synthesis optimization. It allows you to create a good initial netlist for floorplanning. And the result? Better timing-power-area balance. PAS

- Uses actual design and physical library info
- Dynamically adapts to changing logic structures in the design
- Has the same runtime as synthesizing with wireload models

Once you have a good initial netlist, you can create a good initial floorplan. Previously, this floorplan was used for P&R stages, and not in synthesis. But now, you can use this floorplan to allow the synthesis engine to “see” long wires before actually building the logic gates for the improved, physically aware netlist.

- Choose placement aware synthesized netlist to improve global connections. Break complex cells to improve pin accessibility and improve connections.
- Generally pins are routed in lower layers(M1,M2), when we want to connect those pins, we will get congestion at that area.

4.4.2 Optimized power grid structure

- Using Optimized Power Grid Structure we can increase routing resources, so that Congestion will reduce.
- With this Optimized Power grid structure , mainly we can increase vertical resources. (in M3,M5 layers only)

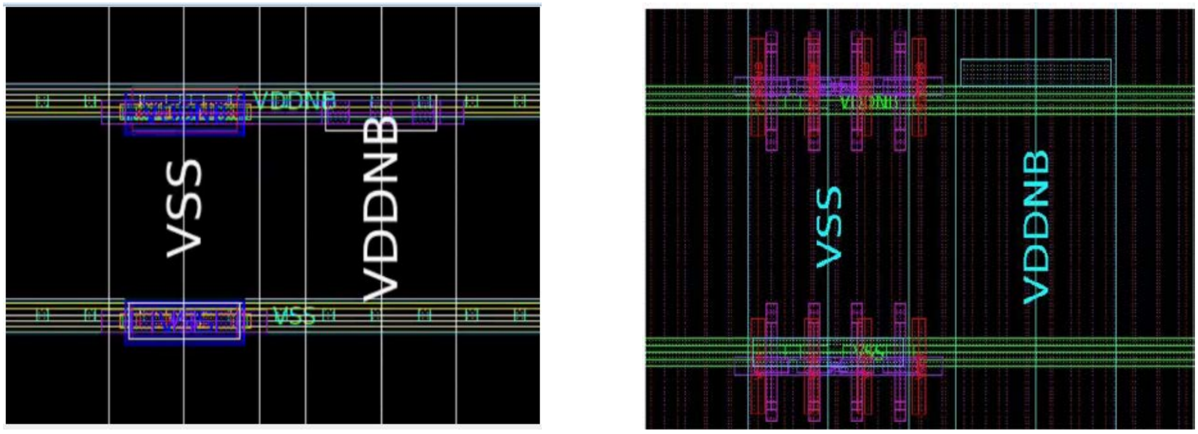


Figure 4.3 normal power grid structure vs optimized power grid structure

4.4.3 Placement optimization, congestion_driven

Rerun the fast placement with congestion driven option (congestion drive placement). Modify physical constrains such as adjust cell density in congested areas. Because higher cell density cause for congestion. modify floorplanning such as moving macros, change core shape/size, move pins to give enough room for routing.

4.4.3.1 Congestion Driven Placement

Congestion driven placement is perform to reduce the congestion.

During congestion driven placement, the cells (higher cell density) which caused for congestion are spread apart from each other by setting congestion effort high.

- Set_congestion_options –max_util <value>
- The -max_util option specifies how densely cells can be packed in uncongested regions to relieve congestion in other regions.
- Congestion options can be design-wide or regional.
- Use the -coordinate option to specify two corners of the bounding box of the area to be affected by the command. Otherwise, the command affects the whole chip.

Congestion needs to be analyzed after placement and the routing results depend on how congested your design is. Routing congestion may be localized. Some of the things that you can do to make sure routing is hassle free are:

Use/modify proper blockage i.e. soft and hard blockages, macro padding(halos) are used proper locations to minimize the congestion near macros.

4.4.4 Padding

Halo is the region around the boundary of fixed macro in the design in which no other macro or standard cells can be placed. Halo allows placement of buffers and inverters in its area.

Macro-padding

Macro padding or placement halos around the macros are placement blockages around the edge of the macros. This makes sure that no standard cells are placed near the pin outs of the macros, thereby giving extra breathing space for the macro pin connections to standard cells.

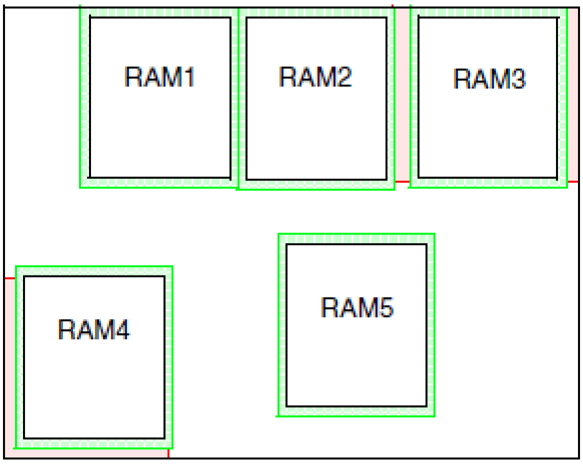


Figure 4.4 Macro padding

Cell padding

Cell Padding refers to placement clearance applied to std cells in PnR tools. This is typically done to ease placement congestion or reserve some space for future use down the flow.

For example typically people apply cell padding to the buffers/inverters used to build clock tree, so that space is reserved to insert DECAP cells near them after CTS.

- Cell Padding is just like halo (keepout margin) specified for cells or instances in design based on congestion.
- keepout margin can varied on selected cells/Instances and it can be soft/hard.
- Keeping the placement of cells out of such regions avoids congestion and net detouring and produces better QoR.

  set_keepout_margin -type hard all_macros -outer {10 0 10 0}

Specifying halos at all sides of the macro , or at sides consisting of Input/Output pins, prevents cell placement at that Input/output pins of macros.

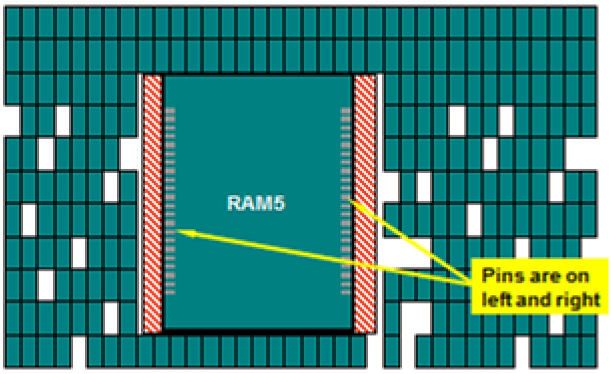


Figure 4.5 padding on both sides of macro

4.4.5 Placement blockages

The utilization constraint is not a hard rule, and if you want to specifically avoid placement in certain areas, use placement blockages. Blockages are specified locations where placing cells are prevented or blocked. These act as guidelines for placing standard cells in the design. Blockages will not be guiding the placement tool to place standard cell at some particular area, but it won't allow placement tool to place standard cell at specified locations.

Maximum Utilization constraint (density screens): Some tools let you specify maximum core utilization numbers for specific regions. If any region has routing congestion, utilization there can be reduced, thus freeing up more area for routing.

- Blockages are specified locations where placing cells are prevented or blocked.
- These act as guidelines for placing standard cells in the design.

4.4.5 Soft (Non-buffer) blockage

- Soft blockage specifies a region where only buffers can be placed. That means standard cells cannot be placed in this region. It blocks(prevents) the placement tool from placing non-buffer cells such as standard cells in this region.
- Create_placement_blockage –type soft –bbox { 23 45 67 70}

4.4.5 Hard blockage

- Hard blockage specifies a region where all standard cells and buffers cannot be placed. It prevents the placement tool from placing standard cells and buffers in this region.
- Create_placement_blockage –type hard –bbox { 23 45 67 70}

4.4.5 Partial blockage

- The blockage factor for any blockage is 100% by default. So no cells can be placed in that region, but the flexibility of blockages can be chosen by partial blockages.By specifying 70% partial blockage, only 30% of area is available for placement .
- Create_placement_blockage –type patial –blocked_percentage 70 –bbox { 23 45 67 70}

4.4.5 Progressive blockage

- Adding Different type of partial blockages adjacent to each other ,mainly it is applied at macro corners or rectilinear cuts.

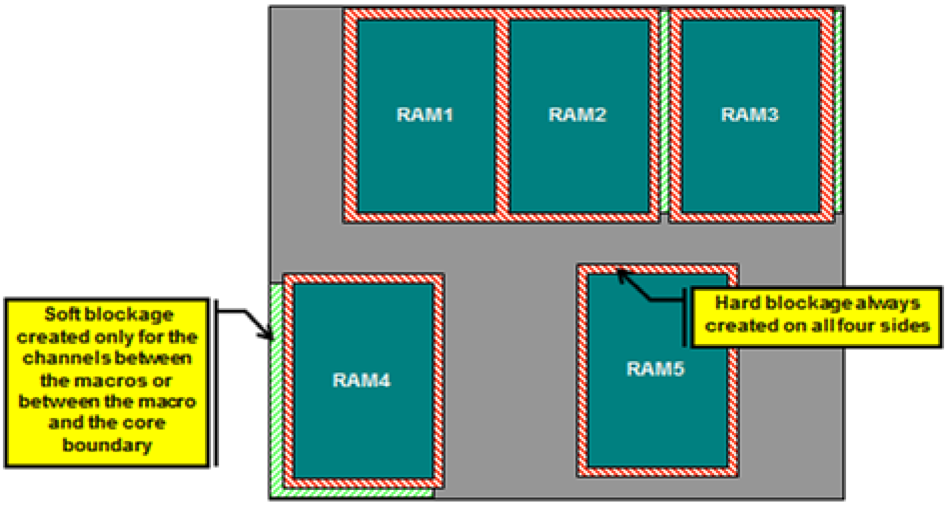

RAM1
RAM2
RAM3
RAM4
RAM5
Soft blockage created only for the channels between the macros or between the macro and the core boundary
Hard blockage always created on all four sides

Figure 4.6 Different types of blockages

4.5 Congestion mitigation flow Results

Design stats	
Macros count	92
Stdcells count	850k
Operating Frequency	930MHZ
Technology	16FF+

Table 4.1 Design stats for congestion mitigation flow

Comparing different congestion reduction methodologies metrics to reduce the congestion

	Default_ flow	Keepouts (halos around macros and complex cells)	Custom blockages (Hard & soft blockages at high density areas , progressive blockages at rectilinear edges)	Mitigation flow
Horizontal OverFlow	14%	10%	6%	2.40%
Vertical OverFlow	18%	14%	10%	3.60%
Total OverFlow	11%	9%	5%	0.15%
Run Time	36Hrs	30Hrs	26Hrs	21Hrs

Table 4.2 Comparing different congestion reduction methodologies metrics to reduce the congestion

CHAPTER 5

PreRoute-to-PostRoute Timing correlation

5.1 Layer-aware optimization

At advanced technology nodes, longer wire lengths and highly resistive metal layers have led to a dramatic increase in interconnect delays. Traditional buffering and upsizing techniques to reduce interconnect delay are no longer as effective due to the area and power impact. To minimize design costs and better predict system performance, upfront and accurate pre-route parasitic estimation of interconnects is necessary during the implementation flow.

As we move to 28nm and below, metal resistance varies significantly (~5X-100X) across routing layers, providing both challenges and opportunities for accurate interconnect delay estimation. In this article, we review techniques that take advantage of resistance variation to reduce buffering and provide tighter post-route correlation to enable better performance prediction.

The different techniques are compared for their benefits and limitations, and an optimal solution is proposed. We end by highlighting some results that illustrate measurable benefits from using layer-aware optimization.

5.2 Growing interconnect dominance

The demand for ‘all-in-one’ devices has led to tremendous on-chip integration. With increasing functionality per mm2, designs today are heavily interconnect dominated. Figure below shows that total interconnect length in ASICs has been doubling with every other technology node (65nm to 32nm to 16nm).

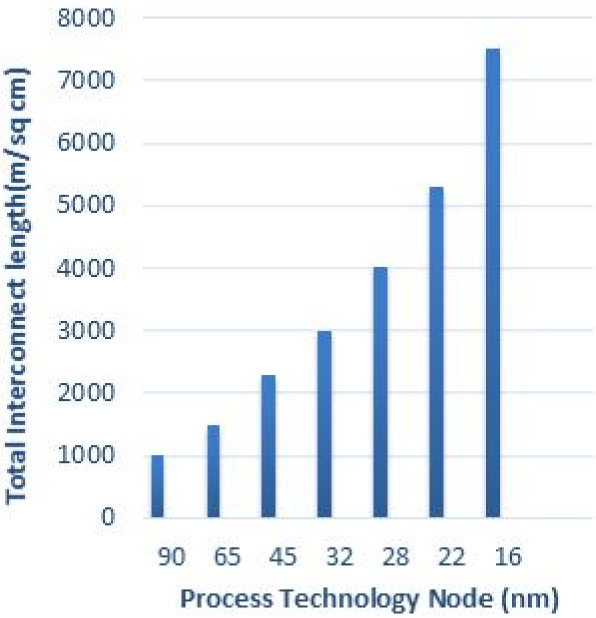
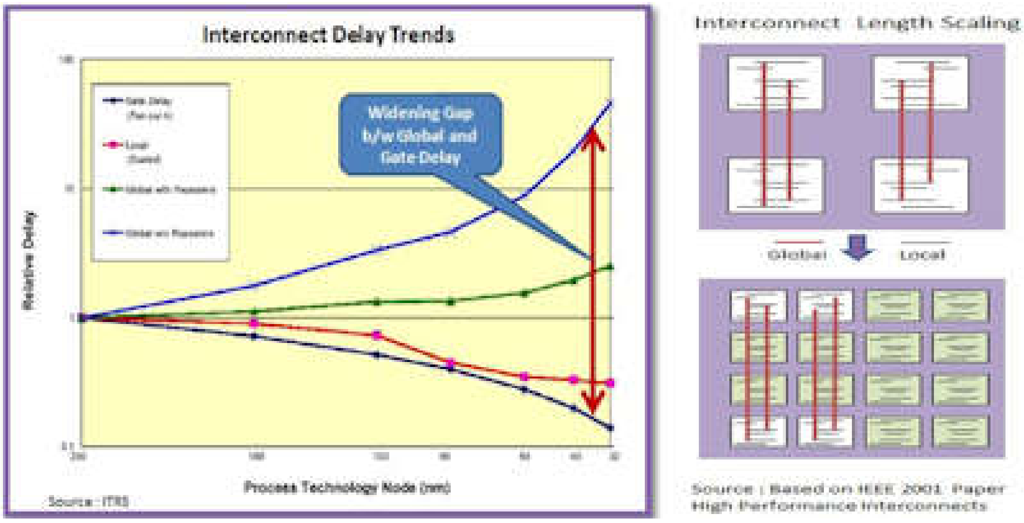


Figure 5.1 Interconnect length trends

With more on-chip interconnects, wire delays begin to dominate gate delays. At lower nodes, shrinking feature sizes reduce gate delays significantly but have the opposite impact on interconnect delays. When wires get thinner and the spacing between them decreases, parasitic effects are more pronounced, resulting in increased interconnect delays. Shorter or local interconnects scale in length; hence the delay increase is minimal. However, global interconnects that span the chip does not scale as well due to increased integration and nearly constant die sizes. They remain a bottleneck, limiting system performance.



Source : ITRS

Source : Based on IEEE 2001 Paper High Performance Interconnects

Figure 5.2 Increasing dominance of global interconnect delays

Traditionally, interconnect delays are reduced using techniques such as buffering and driver cell upsizing. These approaches are becoming expensive due to area and power overhead. To minimize design costs and better predict system performance, it is necessary to find smarter approaches during implementation to reduce global interconnect delays.

5.3 Changing profile of metal layer stacks

To meet the growing demands of integration, new process nodes come with an increasing number of routing metal layers. Studies show that the number of metal layers has approximately doubled every decade, reaching a current maximum of 12. Another phenomenon that characterizes the increased number of layers is the fact that each layer does not have the same metal pitch (width).

The metal pitch varies from narrower widths for the lower layers to wider widths for the upper layers. Layer-width variation results in lower metal layers having more routing resources at a higher resistance and upper layers having fewer resources at a much lower resistance. Leveraging this difference in metal layers can help improve interconnect performance.

Below figure illustrates layer variation across process nodes. In earlier technology nodes, only one or two upper metal layers exhibited significantly lower resistance compared to the rest.

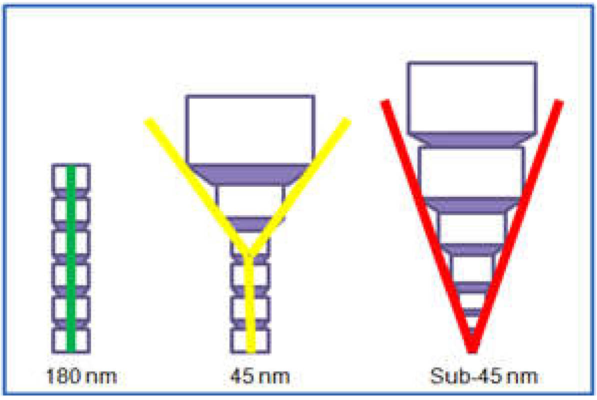
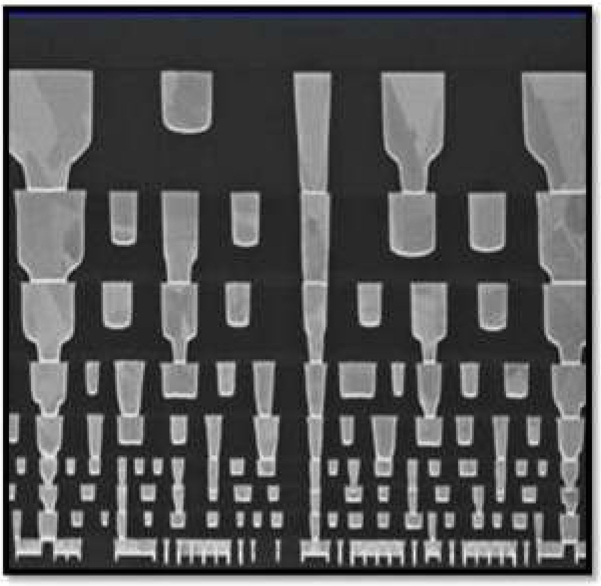


Figure 5.3 Metal layer stack variation across process nodes

By comparison, at 32/28nm the resistance variation across layers can be dramatic, resulting in the telescopic metal stack shown in Figure below.



Source : Intel

Figure 5.4 layer stack cross section(28nm)

The example in below figure plots the unit resistance (R) per layer for 16nm process metal stack, which shows the large variation (~5X-100X) across the metal layers.

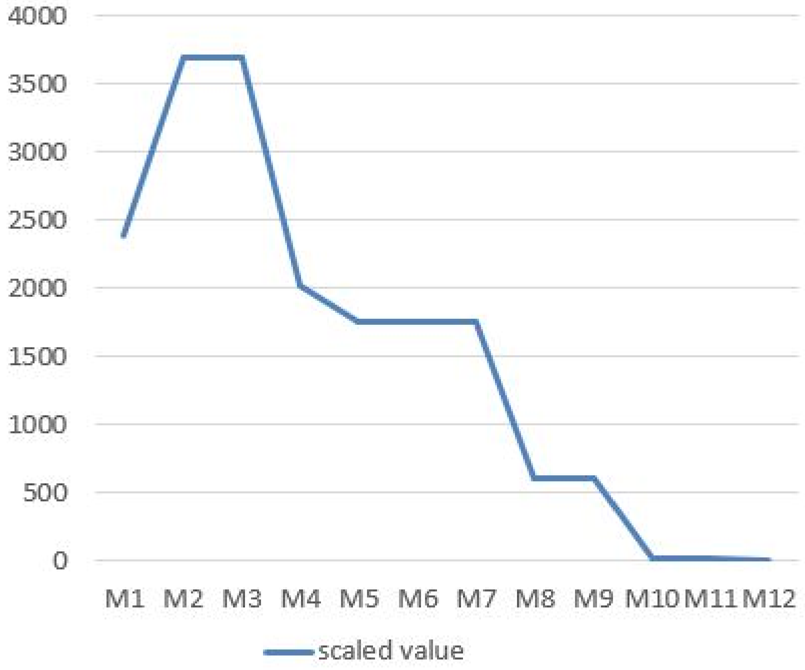


Figure 5.5 Layer resistance variations in 16nm node

5.4 Leveraging Resistance(R) variation during pre-route – existing approaches

Resistance (R) variation provides both challenges and opportunities for upfront and accurate parasitic estimation. Typical pre-route parasitic estimation uses average “R” of all allowable routing layers. Average R works when all layers exhibit uniform resistance. The problem arises when there is a considerable R variation across layers – the effects of which are typically not realized until a net is detail routed. At 16nm, lower layer resistance skews pre-route R calculation to very pessimistic values, requiring increased buffering of global interconnects. Subsequently, during detail route, these nets get routed on upper layers having lower R variation, rendering the increased buffers inserted during pre-route unnecessary.

To summarize, the metal layer(s) used to route a signal at 16nm can have a significant impact on its timing and buffering needs. Judicious use of upper layers can help reduce the impact of interconnect delay and improve system performance.

Routing critical nets on higher metal layers to meet timing is not a new concept. There are several approaches in use today; however, most are iterative and can be error prone and time consuming.

5.4.1 Layer usage control

In this approach, the least resistive upper metal layers are ignored during pre-route and made available only at post-route. The intent is to control upper-layer utilization and prevent routing of non-critical signals on these layers. However, with upper layers completely blocked during pre-route, and the predominant use of highly resistive lower layers, it can lead to over-buffering of critical nets.

5.4.2 Using soft non default rules (NDR)

NDR spacing rules can be used to manipulate the parasitics of critical nets through width and spacing constraints. An iterative, design dependent approach, it requires several trial place and route runs to determine the optimal spacing length thresholds and spacing weights to assign to each rule.

5.4.3 Parasitic scaling

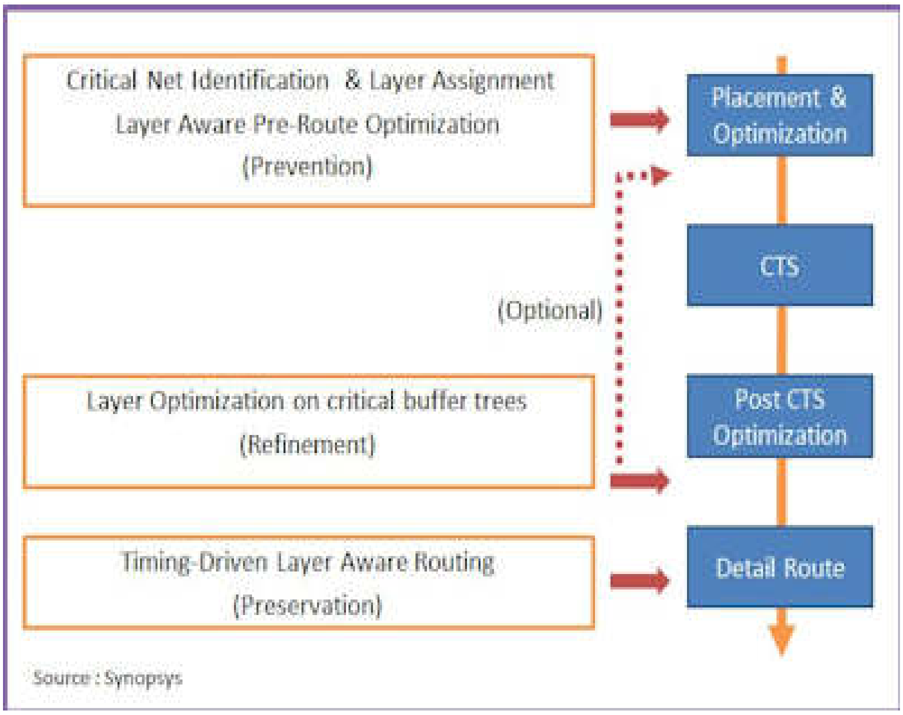
Another approach is the use of scaling factors to control the average R used for pre-route estimation – long nets get scaled to lower R and short nets to higher R, thereby improving the overall interconnect performance. Determining scaling factors is an iterative process relying heavily on the expertise of the user. Furthermore, it requires careful analysis as it can quite easily impact pre-route optimization priorities. Parasitic scaling is a viable option for advanced users, but can be overwhelming for those unfamiliar with this technology.

5.4.4 Global route-based pre-route estimation

This approach involves global routing nets during pre-route. While it eliminates the inaccuracies of averaging R during pre-route, it can be over-engineering for non-critical / short nets. The ideal solution to strike the right balance between QoR and runtime is to have just the critical long nets globally routed.

5.5 Automatic layer-aware pre-route optimization – The smarter approach

At 16nm, there is a need for a more intelligent and automated approach to layer-aware optimization.


Critical Net Identification & Layer Assignment
Layer Aware Pre-Route Optimization
(Prevention)
(Optional)
Layer Optimization on critical buffer trees
(Refinement)
Timing-Driven Layer Aware Routing
(Preservation)
Placement & Optimization
CTS
Post CTS Optimization
Detail Route

Source : Synopsys

Figure 5.6 Layer optimization methodology

An optimal solution should work within the existing place and route flow, as well as provide flexibility for layer promotion at every implementation stage. Above figure details such a methodology, which should consist of three steps:

1. Prevention
2. Refinement
3. Preservation.

- Prevention requires critical net identification and layer assignment during pre-route. Physical and timing properties such as net length, fanout and slack can be used for critical net identification. Flexibility in net pattern selection may be required for designs that are macro-dominated or data-path intensive, or for those with rectilinear floorplans. In such cases, attributes such as congestion, connectivity and aspect ratio can be used to refine net selection
- An optional refinement step can be done post optimization to opportunistically reroute selected buffer trees to the upper layers. It may be needed to further tune quality of results (QoR) for timing-critical designs.
- Finally, the preservation of layer assignment through detail route helps complete the solution.

A phased approach such as the one explained above will result in effective buffer reduction and improved performance. Advanced capabilities such as improved critical net selection and continuous layer assignment evaluation will enable wider adoption of this technology.

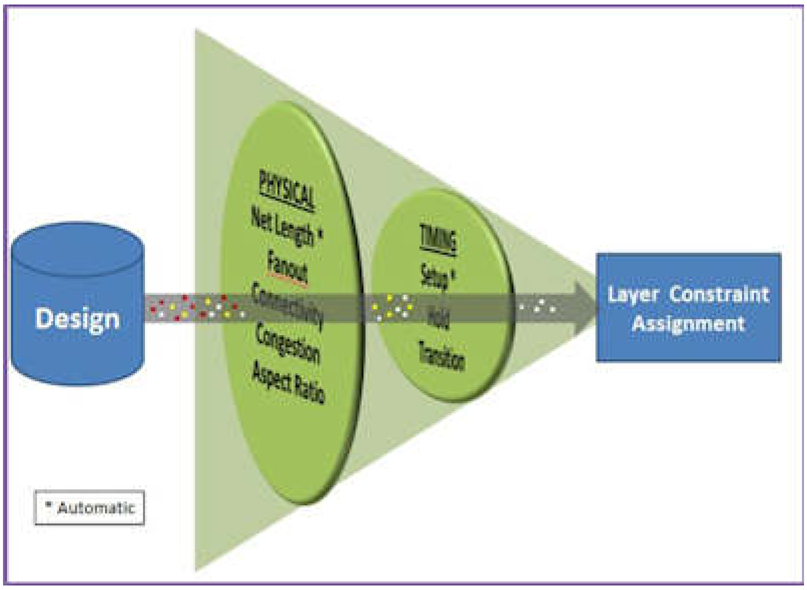
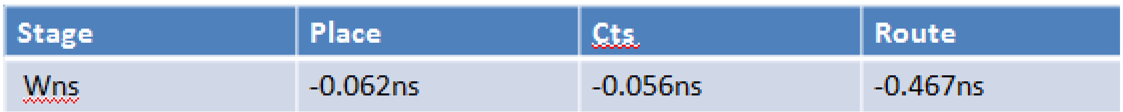


Figure 5.7 Critical net selection process for layer assignment

5.6 PreRoute-to-PostRoute Timing correlation Results

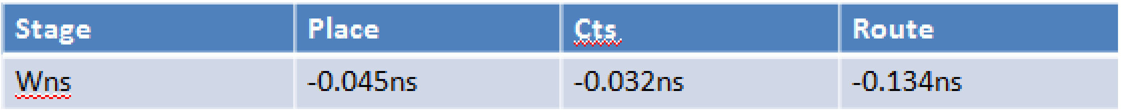
Before Layer-opt :



Stage	Place	Cts	Route
Wns	-0.062ns	-0.056ns	-0.467ns

Table 6.1 Timing results before layer_opt

After Layer-opt :



Stage	Place	Cts	Route
Wns	-0.045ns	-0.032ns	-0.134ns

Table 6.2 Timing results after layer_opt

The above results are for critical in the design (SCLK)

Wns: Worst negative slack

Timing units are ns(nano sesconds)

CHAPTER 6

DRC Challenges

Sub-nano-meter technology gives more advantages to design community. However along with the advantages it also brings in lot of challenges along with it. one of them is DRC complacence. Starting from 20nm lower Metal layers has to be decomposed into two masks and this requirement gave raise to new set of DRC rules called Double Pattering Rules commonly known as DPT rules. Along with these rules, regular spacing and enclosure rules have increased both in numbers and complexity. Routers can only get us to a reasonable closure on DRCs and due to complex DRCs and DPT violations, fixing DRC violations left by router is highly time consuming and manual process. DRC error fixing at 20nm and below nodes are very complicated and error prone. Manual fixing of DRCs will impact tape-out schedules. The DRC challenges seen in 16FF+ nodes are escalated to a new height due to addition of Local Double Pattern loops.

6.1 Why double pattern is needed?

Need for more devices on the same area of silicon is pushing the fabrication mechanism to its limits. At 20nm node, we will not be able to fabricate our designs with existing 193nm wavelength illumination technology. At the same time, Fab houses don’t have the technology to enhance the illumination wavelength to any further to accommodate 20nm drawing challenges. To achieve 20nm node advantages and to keep the current illumination wavelength, fab houses decided to split the design layout into two masks and pass them through serial exposures. This gave raise to a new technology and is called as Double Patterning.

Though double patterning addressed fabrication challenges for 20nm node, it brought new set of complicated Design Rules called DPT rules. DPT rules are tough rules to resolve. Unlike traditional DRCs where area of the DRC violation is limited to couple of polygons, DPT violations can involve hundreds and thousands of polygons.

These complicated rules became big bottleneck for design closure. To accelerate design cycle at 20nm. As the geometry sizes get smaller and are packed more densely in silicon, they pose new challenges for manufacturing. One result is that it is extremely difficult to print 20-nm and below features in one mask exposure with current light sources. This is an effective way to decompose a dense layout mask into two individual masks that contain less complex physical shapes.

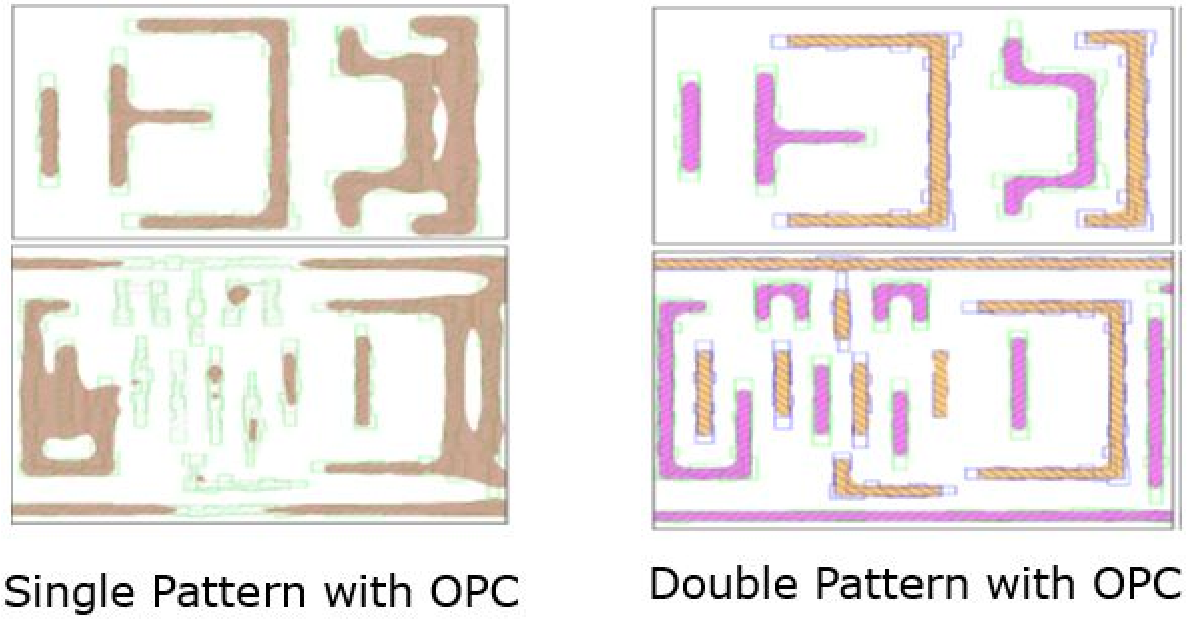


Figure 6.1 Single pattern Vs Double pattern

6.2 What is double patterning ?

- At 20nm and below, critical layers are split into two masks and then combined together after litho/etch process. This technique is called Double Patterning (DP).

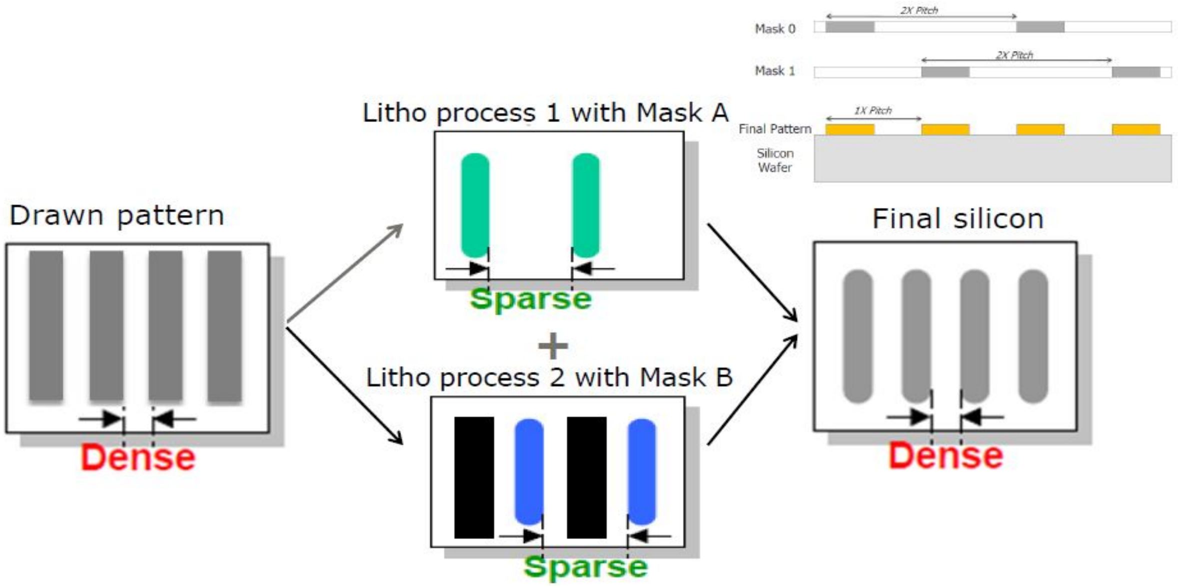


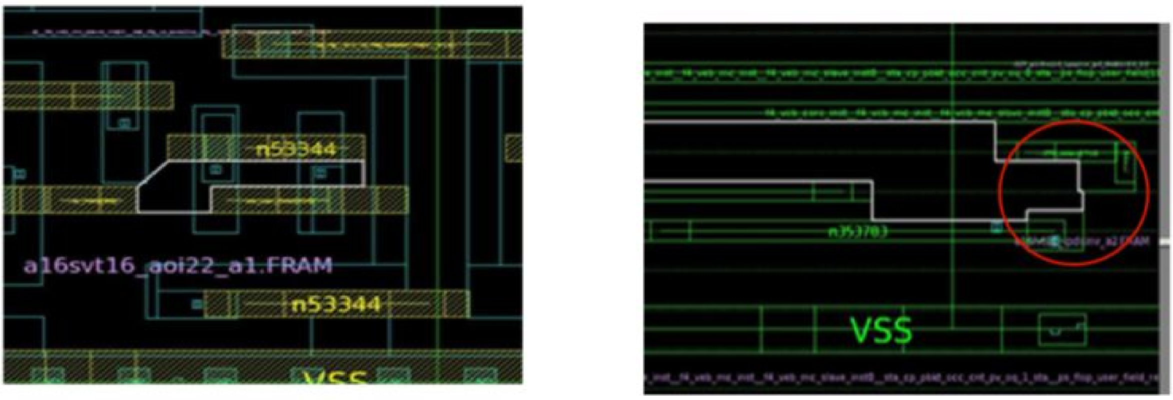
Figure 6.2 Double Patterning technique

6.3 Possibilities for DPT Violation Occurrence

Problem comes when the routing layers fail to get distributed between 2 Masks, whose reasons might be,

- OFF track routing

- Large Via (creating min space)
- Routing in the non-preferred direction



DPT1: Allows sparse routing , eats routing resource

DPT2: Restricts any non-preferred routing

Figure 6.3 Possibilities of DPT violation occurrence

Out of 12 layers used in 16FF+ implementation, 4 layers use double pattern masks. The double pattern masks provide the way denser interconnects but comes with a unique challenge of Local Double Pattern (DPT) Loops DRC. The DPT issue arises when tool is not able to decide on which mask the particular metal shape to be inserted. It also provides significant challenge while doing manual fixes if required.

Other possible DRC violations in this node,

Frozen Routes:

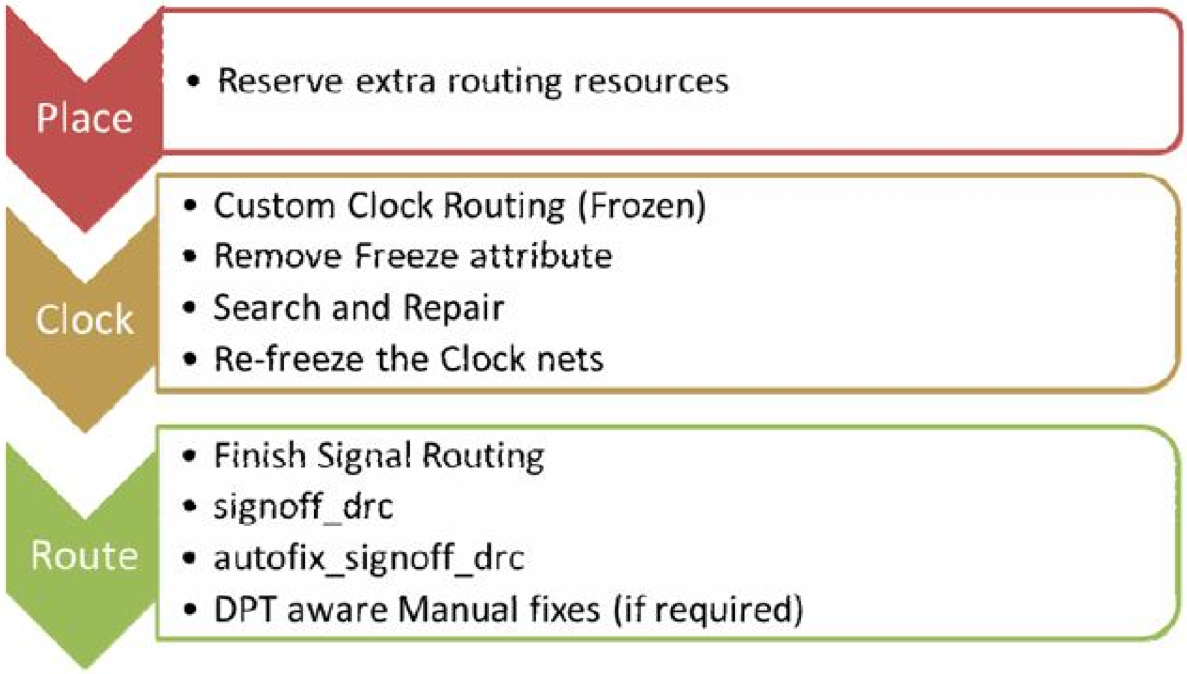
- ❑ Use of custom clock has frozen route and prevents router from addressing the DRCs related to these routes.
- ❑ Issue has been addressed by momentarily removing the freeze attribute from the custom clock nets and doing search and repair loops to remove DRCs before signal routes.

VT-OD Spacing:

- ❑ Though the VT swap related issue is not uncommon in planer technologies, it presents new challenge in 16FF+ node due to size reduction of the cells.

- ❑ The VT swap related issues arise when a small cell gets sandwiched between different VT type cells. The issue has been mitigated by not allowing few small sized cells in the design.

6.4 DRC mitigation flow



Place

- Reserve extra routing resources

Clock

- Custom Clock Routing (Frozen)
- Remove Freeze attribute
- Search and Repair
- Re-freeze the Clock nets

Route

- Finish Signal Routing
- signoff_drc
- autofix_signoff_drc
- DPT aware Manual fixes (if required)

Figure 6.4 DRC mitigation flow

As shown in above flow chart,

- ❑ The DRC challenge has been addressed by reserving extra resources from the beginning, addressing custom clock related DRCs and at the end by employing commands signoff_drc and autofix_signoff_drc to address DPT and other DRC violations.

Also few manual fixed have been done by keeping DPT in the mind.

6.5 DRC mitigation flow results

Default flow

DRC Summary : 2458

Mitigation flow

DRC Summary : 974

CHAPTER 7

Power Recovery Challenge

In the planar technology, dynamic and leakage power were almost at par. The main power reduction technique then, was to minimize the usage of low VT cells in order to optimize the total power.

In the FINFET technology, the standard cell leakage is not anymore a significant contributor of the total power and at the same time the gate capacitance of FIN structures contribute to the dominance of the dynamic power.

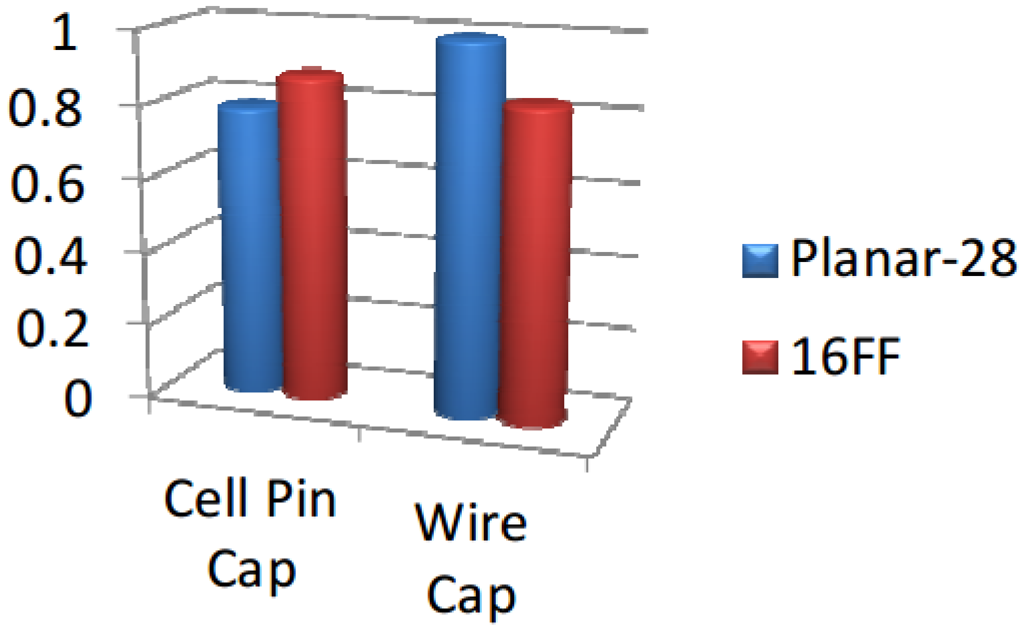


Figure 7.1 Cell Pin Capacitance vs Wire Capacitance

In planar nodes, leakage power was dominant with large focus on low leakage designs. However shift to 3d FINFET has changed that and has brought new challenges in terms of power. FINFET input capacitance is higher for the same gain compared to its planar counterpart, owing to its 3 dimensional structure. This combined with lower leakage due to greater gate control, leads to designs which are dynamic power dominant. It is observed that the gate capacitance is greater than wire capacitance when compared to planar 28 nm in which it was opposite as demonstrated in the above graph.

This chapter looks at:

- ways to improve the standard cells libraries to mask the high cell input capacitance issue while taking full advantage of the intrinsically high performance of the FINFET technology,

- specific design flow to optimize for total power: dynamic & leakage power in the FINFET technology

To control dynamic power in design through physical implementation, experiments suggest that there is a need to control the input pin capacitance. This can be achieved by merging simple cells to complex cells.

❑ This should be done keeping an eye out for congestion which may increase owing to higher pin density, in which case keepouts may be necessary for such cells.

Use of complex cells at synthesis

Here are some of the reasons FINFET benefits complex cells

- FINFET has high cell input capacitance
- Merging two cells for instance requires only one active area, thereby eliminating the minimum distance between the two cells

Merging of two cells

Cell combination in the paths after P&R	Original equations	Simplified equations	New cell name	No. of pins of the cell
NOR3 -> NAND2	~(~(A|B|C)&D)	~(~A&~B&~C&D)	NAND4ABCN	5
NAND2 -> NOR3	~(~(A&B)|C|D)	~(~A|~B|C|D)	NOR4ABN	5

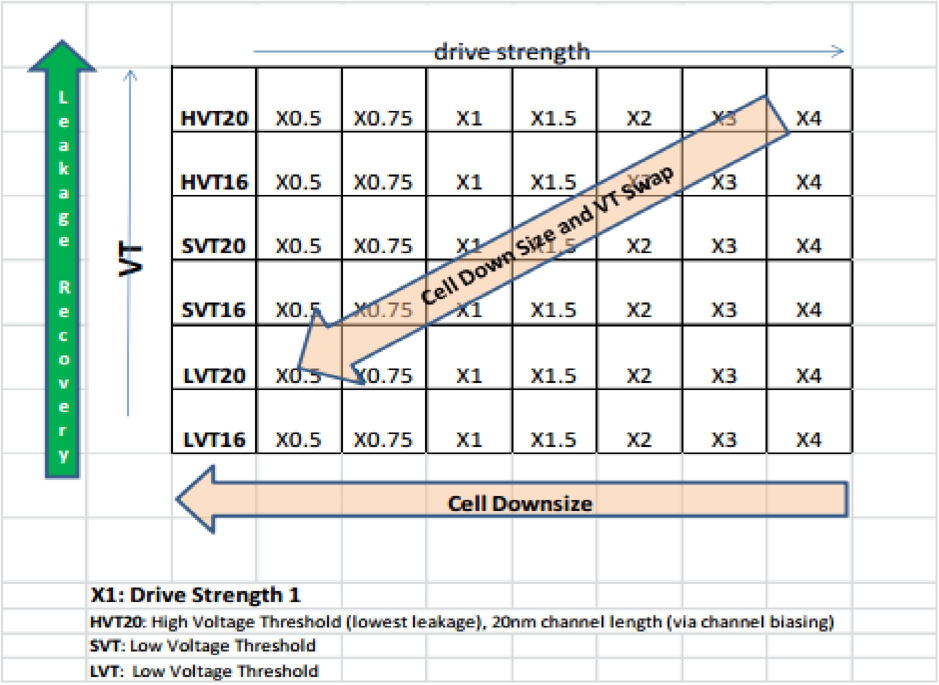
Table 7.1 Example of creation of complex cells

More complex cells were added in the 16ff libraries and were also selected as we will see in the results section. The issue to watch for with complex cells is potential routability issues with high pin count cells. The strategy here was to create a keepout area for cells with more than 8 pins.

7.1 Power Recovery Flow

The power recovery has two steps. The first step is based on total power recovery flow and the 2nd step is the signoff leakage recovery.

In the step, cells are downsized and/or swapped to lower V_{th} cells while preserving timing goals.


drive strength
L e a k a g e R e c o v e r y
VT
Cell Down Size and VT Swap
Cell Downsize

HVT20	X0.5	X0.75	X1	X1.5	X2	X3	X4
HVT16	X0.5	X0.75	X1	X1.5	X2	X3	X4
SVT20	X0.5	X0.75	X1	X1.5	X2	X3	X4
SVT16	X0.5	X0.75	X1	X1.5	X2	X3	X4
LVT20	X0.5	X0.75	X1	X1.5	X2	X3	X4
LVT16	X0.5	X0.75	X1	X1.5	X2	X3	X4

X1: Drive Strength 1

HVT20: High Voltage Threshold (lowest leakage), 20nm channel length (via channel biasing)

SVT: Low Voltage Threshold

LVT: Low Voltage Threshold

Table 7.2 Power Recovery Graph

The timing closed database is loaded into PrimeTime for the total power optimization. Timing analysis is performed at all PVT corners and functional modes and all the PrimeTime sessions are saved.

The search for alternative cells or groups of cells is also performed in PrimeTime. For instance for a given cell in a functional path (NAND4X1SVT), all alternative cells are listed that meet timing, and have reduced total power than the orginal cells in the path. The cell meeting timing with the lowest total power (leakage + dynamic) is selected (for example NAND4X0P5LVT). This process is repeated for all valid timing paths. For runtime reasons, the analysis can also be limited only to paths with a positive timing slack (beyond the timing margin).

An ECO file is exported once the total power optimization is completed.

Since most of swapped cells are not foot-print compatible, routing is not fully preserved there is a route ECO is required

Swapped cells are ECO’ed back in the design (ECO placement and ECO Route). Deleting and rerouting the open nets introduced some unexpected timing changes. A simple ECO route converged. Cells changed in this power recovery step are smaller in size than the previous cells.

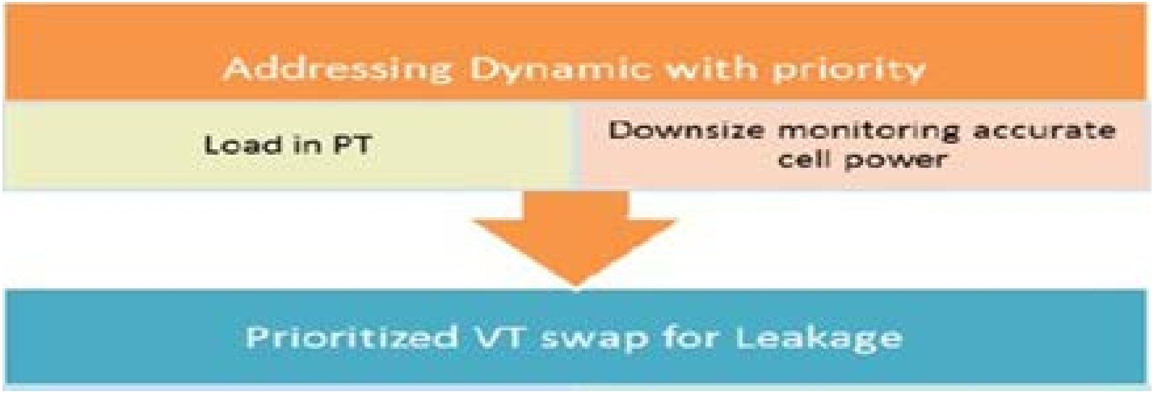

Addressing Dynamic with priority
Load in PT
Downsize monitoring accurate cell power
Prioritized VT swap for Leakage

Figure 7.2 Power Recovery Flow

- Priority is from highest VT to lowest VT ,
- Using a Vt-swap with foot-print compatible cells
- Footprint equivalent V_{th} cell swaps to achieve lower leakage while preserving timing goals
- Signal Routing is preserved
- Modified cells are ECO’ed in the design without any placement or routing changes

7.2 Power Recovery Flow Results

Technologies	28nm 16ff	
#Cells	1.2M	
#Gates	~4.5M	
# SRAMs	96	
# Layers	6+2+2	6 thin layers + 2 intermediate layers + 2 R Layers
Frequencies	800MHz	~40 Logic Levels

Table 7.3 Block Data Used for Analysis

The voltage used for setup timing closure is 0.85V in 28nm and 0.67V in 16nm. The design meets the 800MHz performance in these 2 technologies. One way to reduce power as stated in the introduction is to reduce the supply voltage. The baseline supply voltage in 16nm is 0.8V nominal and 0.72V (-10%) for the setup corner. Dropping the supply voltage to a nominal 0.75V or a setup corner of 0.67V provides ~12% dynamic power reduction assuming that chip capacitance stays constant. This strategy pays off if reducing the voltage does not result into an excessive buffering or a low VT usage.

	VDD for setup time	**#Cells**	**% buffers& Inverters**	**Area Factor**	**Wire Capacitance**	**Cell Pin Capaciatnce**	**Total Capacitance**
28nm	0.85V	1.3M	10.23%	1	3,380nF	2,678nF	6,058nF
16ff	0.67V	1.24 M	9.18%	0.55	2,852nF	2,988nF	5,840nF
					18%+	-11%	-4%

Table 7.4 Design Stats Post Power Recovery

28nm shows~18% more wire capacitance while 16nm shows ~11% more cell input capacitance. 16nm comes with slightly less total capacitance than 28nm. This is mainly due to the fact that the 28nm block has ~5% higher cell count. One of the reasons for the higher cell count is the higher number of buffers and inverters in the 28nm block. Another reason for the reduced cell count in 16nm is the fact that more complex cells are used in 16nm than in 28nm.

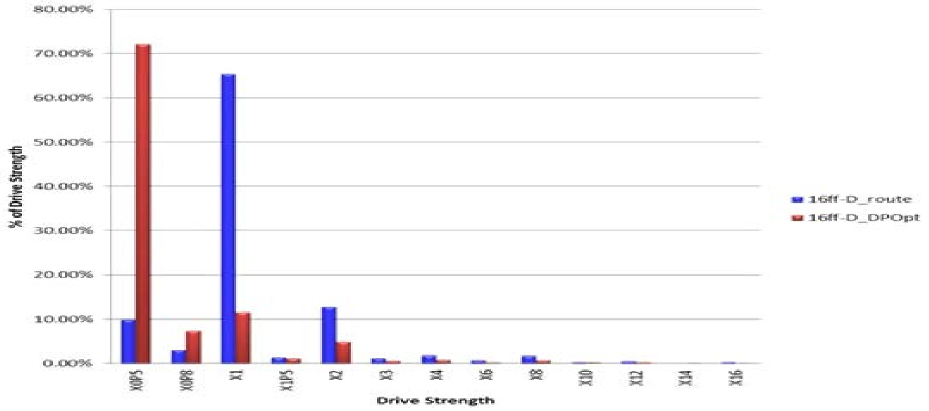


Figure 7.3 16ff Drive Strength Distribution pre-&post power recovery

The majority of cell drive strengths are at X1 (unit drive strength) and after the total power recovery step, the majority of cells moves to the lowest drive strength in the library.

This specific cell downsizing lowers the chip capacitance, and therefore reduces mainly the dynamic power. Over 80% of all cells have drive strength smaller than the unit drive X1 after the total power recovery.

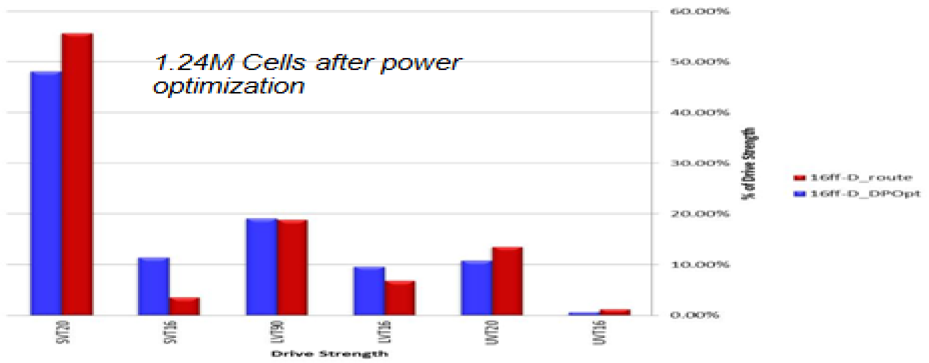


Figure 7.4 16ff Vth Distribution pre-&post power recovery

On the other hand the Vt shift from lower Vt to higher Vt does not appear to be significant as seen in Figure 7.4 In reality there is a significant Vt shift on higher drive strength, for instance X1LVT to X1SVT. But during the total power recovery some of these cells are moved to even lower drive strength but higher Vt as a way to reduce the cell capacitance. This is basically trading some of the leakage power for even lower dynamic power.

The number of cells post power recovery is 1.3M for 28nm and 1.24M for 16ff. The reduction of the number of cells in 16ff is driven by multiple factors. The intrinsic speed of 16ff helps with the reduction of buffers and inverters (10.23% in 28nm, 9.18% in 16ff); ~7% of cells in 16ff are using newly added complex (merged) cells.

It has been observed that primetime is highly efficient in power recovery with the following drive strength distribution before and after power recovery.

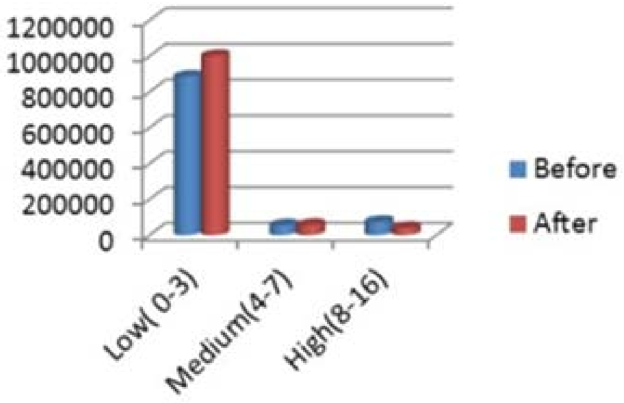


Figure 7.5 Drive strength distribution comparison

CONCLUSION

The FINFET technology has indeed rejuvenated the IC industry in that it has extended the transistor scaling in Moore’s law. We see indeed a tremendous performance increase that can be traded-off for even lower power and lower area .16 FF+ (FINFET PLUS) comes with advantages of power, performance and cost .The challenges due to structure change cannot be ignored. Layer Stack Resistance and crosstalk pose difficulties in timing analysis and closure. The challenges of Layer_opt, optimizations could solve the issue of timing closure and timing correlation gap. While there are several approaches in use today that leverage this resistance variation, an intelligent and automated solution is needed to address existing limitations and provide a more robust flow. With its Automated Layer-Aware Optimization technology, IC Compiler provides a more holistic approach to pre-route parasitic estimation, thereby enabling better performance prediction.

The power recovery flow was updated to adapt with dynamic power dominant technology node, and tool gave pleasant VT distribution and drive distribution. While in the planar technology the recommended approach to reduce power was to minimize the usage of lower VT cells as much as possible even if it meant increasing the drive strength, we see a clear paradigm shift in the FINFET technology With the proposed flow, we can take best out of 16 FF.

REFERENCES

[1]. “Low Power Design in the FINFET Technology”, Benjamin M., SNUG,2014 Santa Clara

[2]. “ICC @2014.12.sp2 user manual”, Synopsys, ICC Manual , 2014

[3]. “FINFET, the promises and the challenges” – Synopsys Insight Newsletter, 2012

[4]. “ PrimeTime and PrimeTime SI User Guide” https://solvnet.synopsys.com/dow_retrieve/latest/ptug/ptug.html

[5]. “Static Timing Analysis for Nanometer Designs”, J. Bhasker, Rakech Chandha, Springer, Published in 2009

[6]. “FINFET Design, Manufacturability, and Reliability” – Synopsys ,2012

[7]. “Power Optimization for FINFET-based Circuits Using Genetic Algorithms”, Jin Ouyang IEEE SOC conference 2008

[8]. “A Low Power Dynamic Circuit Design Technique for Double-Gate FINFET Technology at 32nm”, GLSVLSI 2008

[9]. “High Performance Design Challenges on 16FF+ - Mitigations and Solutions”, - SNUG, 2015, India

[10]. “Designing IP for FINFET technology, The opportunities and challenges” , SNUG ,2013

[11]. “Double patterning lithography: double the trouble or double the fun?” , Paul Zimmerman, SPIE, Jul 2009

[12]. “Multigate device” , https://en.wikipedia.org/wiki/Multigate_device

[13]. “16nm technology” , http://www.tsmc.com/english/dedicatedFoundry/technology/16nm.html

[14]. “PrimeTime – PX for power analysis user guide”