3D Monolithic Integration : performance, Power and Area Evaluation for 14nm and beyond

Alexandre Ayres de Sousa

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTE UNIVERSITE GRENOBLE ALPES

Spécialité : **Nano Electronique et Nano Technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Alexandre AYRES DE SOUSA

Thèse dirigée par **Laurent FESQUET** et
co-encadrée par **Olivier ROZEAU et Bertrand BOROT**

préparée au sein du **Laboratoire Techniques de l'Informatique et de la Microélectronique pour l'Architecture des systèmes intégrés (TIMA) et du Laboratoire d'électronique des technologies de l'information (CEA-LETI)**
au sein de **l'École Doctorale électronique, électrotechnique, automatique et traitement du signal (EEATS)**

Intégration monolithique en 3D: étude du potentiel en termes de consommation, performance et surface pour le nœud technologique 14nm et au-delà

Thèse soutenue publiquement le **16 Octobre 2017**,
devant le jury composé de :

Pr. Francis CALMON
Professeur des Universités, INSA de Lyon, Lyon (Président)
Pr. Lionel TORRES
Professeur des Universités, université de Montpellier, Montpellier (Rapporteur)
Dr. Olivier ROSSETTO
Maître de conférences, université Grenoble Alpes, Grenoble (Membre)
Dr. Laurent FESQUET
Maître de conférences, Grenoble INP, Grenoble, (Directeur de thèse)
Dr. Olivier ROZEAU
Ingénieur de recherche au CEA-Leti, Grenoble (Membre Invité)
Mr. Bertrand BOROT
Ingénieur de recherche à STMicroelectronics, Crolles (Membre Invité)

"Every great advance in science has issued from a new audacity of imagination."

--John Dewey

ACKNOWLEDGMENTS

My doctoral thesis took place between 2014 and 2017 in Grenoble, France. It was performed in collaboration among STMicroelectronics in Crolles, the research institute of the French Alternatives Energies and Atomic Energy Commission (CEA-LETI) in Grenoble and the TIMA laboratory from Grenoble-Institute of Technology (Grenoble INP). This environment including enterprise, research institute and academia allowed me to be in the center of the Grenoble semiconductor development "war machine". The level of excellence and competence in the Grenoble ecosystem inspired me.

I would like to thank all people who gave me this fantastic and unique opportunity, and welcomed a Brazilian that had just left the electrical engineering school. In a great surprise I have met nice, intelligent and welcoming people 9000 km away from home. The adventure of coming to France without even knowing the basics in the French language, quickly turned in a pleasant life, in a lovely country. As a CIFRE thesis program, I could experience the importance of academia close to the industry.

The TIMA laboratory and its team amused me by their competence in semiconductor design. I would like to thank the whole team which I had the opportunity to discuss my work several times and receive constructive feedback. A warmhearted thanks to all Brazilian friends working at TIMA to whom I wish the best. I specially thank Rodrigo BASTOS. I had the luck to work with Laurent FESQUET, who gave me an incredible guidance, quickly solved my problems with tight deadlines, always with clear insights and gently.

At STMicroelectronics, my gratitude to Bertrand BOROT, his competence, readiness, technical knowledge and management will be inspiration for the rest of my days. I greatly appreciate the Crolles fast pace, the industrial environment and support for new ideas.

I spent most of my PhD at the LICL laboratory from CEA-LETI. I have gained a lot of advanced process technical knowledge at LETI, even working with design. My gratitude to Louis HUTIN, Perrine BATUDE, François ANDRIEU, Laurent BRUNET and Claire FENOUILLET who were always prompt to discuss and explain ideas. Also, a big thanks to Olivier ROZEAU. I could write a book of the adventures of being advised by Olivier. His excellence, attention to the details, competence and rigor always pushed me to move forward. I am really honored to have worked with him. I also particularly thank Mathilde, Julien, Luca, Lina, José, Julien, Jessy, Remy, Fabien, Vincent and Giulia for the great times spent during my thesis. Really nice people, to whose I wish the best. I can speak French now because the insistence of Julien, Mathilde and Lina who always taught me.

Finally, my deepest gratitude to my family who have supported me unconditionally. Despite of the distance and seeing them only few days per year, I always felt their support which kept me strong in the difficult times. A genuine thanks to my parents Alexandre and Adriene.

Sincerely,

Alexandre Ayres de Sousa

TABLE OF CONTENTS

Table of Contents

Glossary

Acronyms	Meaning
3DCO	3D Monolithic Contact Between Tiers
3DVLSI	Very-Large-Scale 3D Monolithic Integration
ACV	Across-Chip Variations
ASIC	Application-Specific Integrated Circuit
BEOL	Back-End Of Line
BOX	Buried Oxide
CMOS	Complementary Metal-Oxide-Semiconductor
CMP	Chemical Mechanical Planarization
CPP	Contacted Poly Pitch
DES	Data Encryption Standard
DIBL	Drain-Induced Barrier Lowering
DRC	Design Rules Check
DRM	Design Rules Manual
EDA	Electronic Design Automation
EOT	Equivalent Oxide Thickness
EUV	Extreme Ultra-Violet
FD-SOI	Fully Depleted Silicon On Insulator
FEOL	Front-End Of Line
FF	Flip-Flop
FFT	Fast-Fourier Transform
FinFET	Fin Field Effect Transistor
FO	Fan-Out
GAA	Gate-All-Around
GND	Ground
GPU	Graphic Processor Unit
High-k	Relative permittivity higher than SiO_2
HUEBR	Heated Unique and Equal Bonding Reliability
iBEOL	Intermediate Back-End Of Line
IV	Inverter Gate
LDPC	Low-Density Parity-Check
LIDAR	Light Detection and Ranging
LoL	Logic over Logic
LT	Low Temperature Process
LVS	Layout Versus Schematic
MC	Monte Carlo
MDR	Metal Design Routing
MOSFET	Metal-Oxide-Semiconductor Field Effect Transistor
Mx	Metal Layer Level - M1 standards for layer one
NAND	Floating-Gate Transistors Flash Memory
NEMS	Nanoelectromechanical System

NMOS	Metal-Oxide-Semiconductor with an n-type conduction channel
PAI	Pre-amorphization-Assisted Implantation
PDK	Process Design Kit
PDN	Power Delivery Network
PD-SOI	Partially-Depleted Silicon On Insulator
PEX	Parasitic Extraction
PMOS	Metal-Oxide-Semiconductor with an p-type conduction channel
PPA	Performance, Power and Area Metric
RF	Radio Frequency
RO	Ring Oscillator
RSD	Raised Source-Drain
RSNM	Read Static Noise Margin
RTL	Register-Transfer Level
RX	Transistor Active Region
S/D	Source and Drain
SNM	Static Noise Margin
SOI	Silicon On Insulator
SPER	Solid-Phase Epitaxial Regrowth
SPICE	Simulation Program with Integrated Circuit Emphasis
SRAM	Static Random-Access Memory
SS	Subthreshold Swing
STR	Self-Timed Ring
TSV	Through-Silicon Via
UTBOX	Ultra-Thin Buried Oxide
VDD	Positive Supply Voltage
VLSI	Very-Large-Scale Integration
VTC	Voltage Transfer Characteristics
WL	Wirelength
WNM	Write Noise Margin
XOR	Exclusive OR gate

INTRODUCTION TO CHAPTER ONE

"There's a basic principle about consumer electronics: it gets more powerful all the time and it gets cheaper all the time." --Trip Hawkins

The integrated circuit evolution over the years has changed the world. From massive data centers to mobile devices, the electronic is in our life.

Chapter one discusses how the above quote is possible. The historical perspective of making smaller devices, namely the transistor scaling is reviewed by discussing the Moore's Law. Over the last 50 years the scaling trend has been achieved, decreasing the costs per transistor.

The MOSFET transistor operation and its figure of merits are briefly discussed. The MOS is the fundamental brick in the digital nanoelectronics, and increasing its performance translate in a better circuit efficiency. Quantifying transistor characteristics for digital operation is necessary to determine the best circuit behavior.

Dennard's Law is an insight into the transistor performance boost by the miniaturization. By scaling parameters such as widths, lengths, thickness the transistor performance is enhanced. Coupled to Moore's Law, the outcome is the cited quote.

The 3D integration concept and state of the art is shown in this chapter. As the scaling is approaching to the atomic level size, the process complexity is escalating. The 3D integration, or stacking transistor in several tiers is presented as an alternative to the traditional scaling. The parallel 3D integration is briefly discussed, and then the state of the art of 3D sequential is reviewed.

Chapter One – Introduction to 3DVLSI

1.1 Introduction to 3DVLSI

1.1.1 CMOS Scaling

In the 1960's electronic circuits suffered a major transformation, the circuitries were in a transition from discrete elements to fully integration on a single die. Gordon Moore noted that the number of transistor was increasing year by year, and possibly the trend could continue through the following years, as no fundamental physical barrier was in sight. The trend was first proposed for the next ten years, or until 1975 as illustrated in Figure 1.1.1.1a.

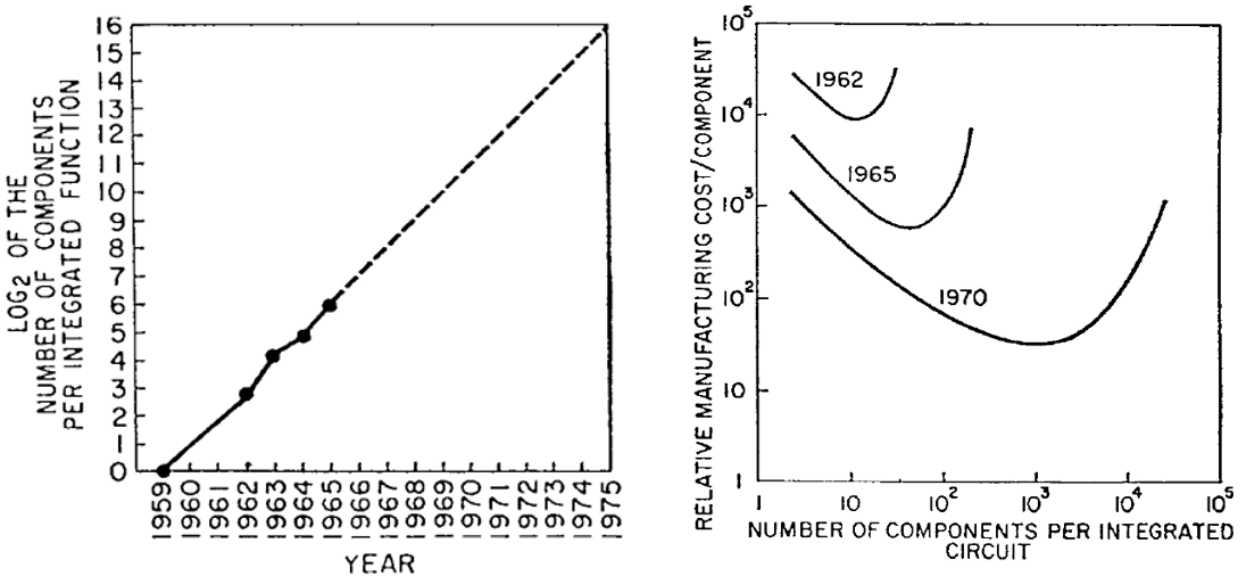


Figure 1.1.1.1 "Cramming more components onto integrated circuits". On the left, (a) Moore's transistor count projection through the years. On the right, (b) the optimum number of components to decrease the cost. During the years the optimum cost is reduced and the optimum transistor count is higher. [Moore 1998].

In the same publication, Moore argues that the cost of the circuit is inversely proportional to the number of components. This holds true until a certain point, where the huge number of components increment the circuit complexity and reduces the process yield, thus increases the cost. The outcome is an **optimum number of components for the minimum cost**. Moore's then predicts **the cost failing over the years**, especially for the optimum number of components in an integrated circuit as illustrated in Figure 1.1.1.1b. Miniaturization and technology evolution are cited as the main reason to reducing costs. The main messages of the paper are the beginning of a new era using integrated circuit, instead of the previous discrete circuits; the benefits of using integrated circuits, such as reliability, increased circuit complexity and utility, and most importantly the cost reduction proportioned by the miniaturization. Those arguments form the base of the well-known Moore's law: over the years the cost per transistor is reduced, because the advancements in the technology and miniaturization. The original paper cites the miniaturization pace of doubling the transistor count every two years, thus the transistor count in Figure 1.1.1.1a is in the logarithm scale. Another interesting discussion in the publication is the process yield, which is said to improve as high as economically possible, only needing engineering efforts. This confirms the central key message of the paper: the cost of integrated circuits.

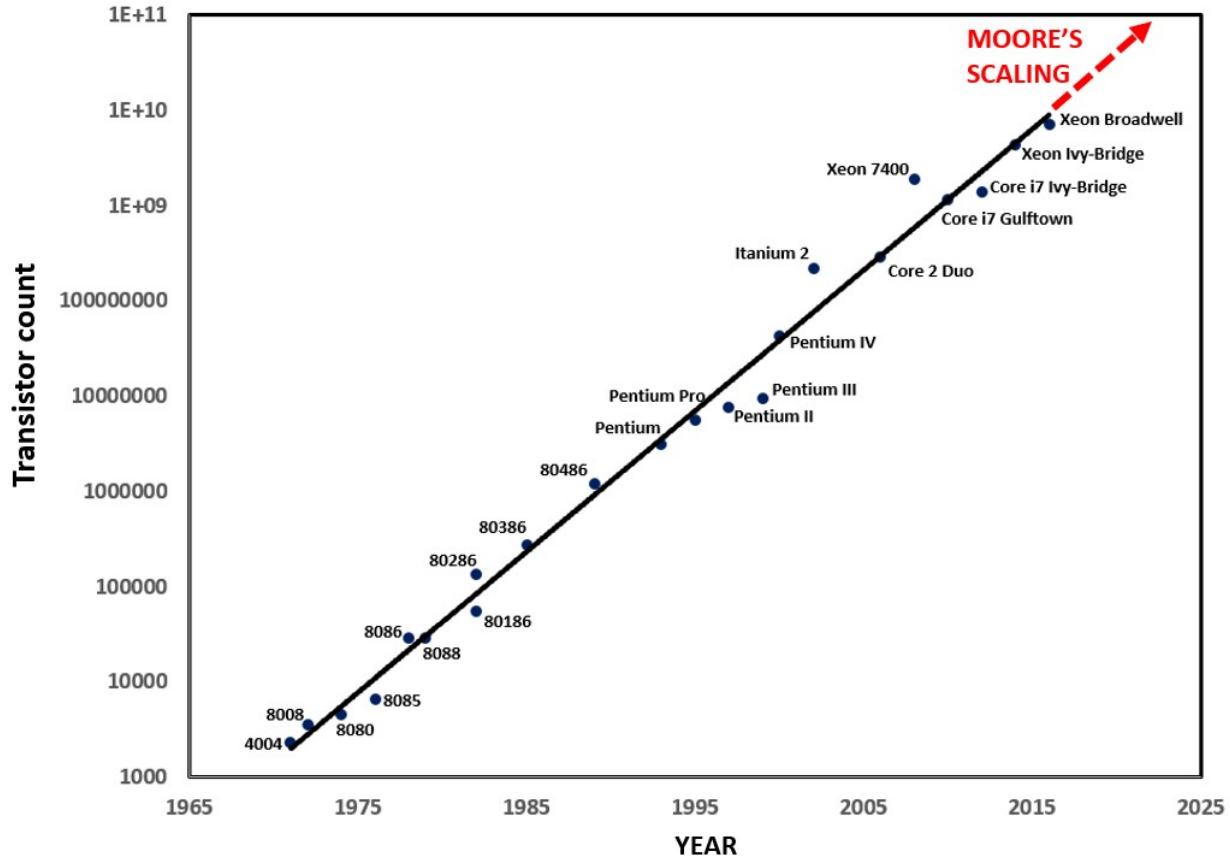


Figure 1.1.1.2 Transistor count in logarithm over the years for several Intel processors.

Moore had only data from the previous six years, and made the prediction of transistor count for the next ten years, or until 1975. It turns out, that the proposed miniaturization pace proposed by Moore held true for at least 52 years. The scaling of transistors is occurring for five decades, and the transistor count should increase in the next years, continuing the Moore's scaling trend. After many years of success, the Moore`s prediction became more than a classic paper, it turned into a law. Indeed, **the transistor count doubling each eighteen or twenty-four months is the big roadmap followed by the industry**, especially for the advanced logic integration. The transistor dimension scaling has been scaled as shown in Figure 1.1.1.2, often depicted as the Moore's Law. Despite of transistor count being the imminent result of the scaling, the reason that Moore's Law is alive after so many years is the transistor cost, as proposed in the original paper. The cost per transistor is decreasing for each node, as pictured in Intel data in Figure 1.1.1.3. As suggested by Moore, this figure shows that more transistors can be crammed for a given area, and despite the cost increasing in each node due to process complexity, **the final price per transistor is decreasing**. This trend is expected to continue at the 7nm node by Intel. **A major caveat in the data is about the yield and process maturity**. Those factors are not transparent, so the analysis may consider a very good process yield and the cost per transistor after a long time of mass production, where the process maturity is good enough to deliver high yields, and optimized performance.

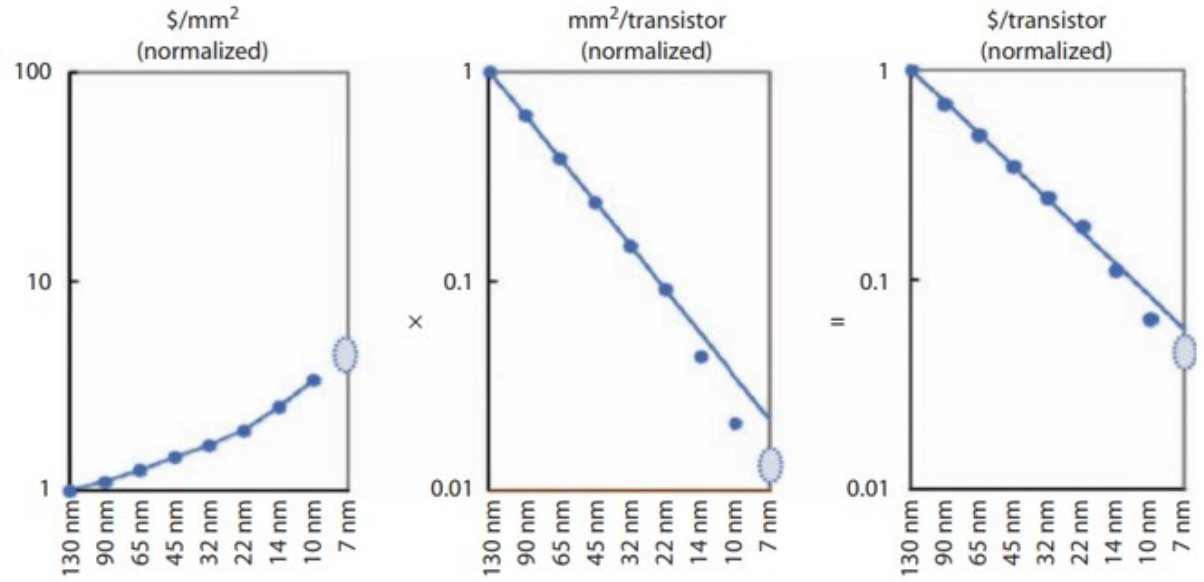


Figure 1.1.1.3 Intel data for the relative transistor cost from the 130nm to 7nm node. The cost per transistor is decreasing in a logarithm scale.

1.1.2 MOSFET Device Overview and Typical Figures of Merit

The Metal-Oxide-Semiconductor Field Effect Transistor abbreviated by MOSFET is the structure created using the three elements (MOS) as the name suggests, in order to create a layer of free carriers by applying of electrical field in the semiconductor. The transistor has four terminals, the gate which controls the semiconductor channel, the source and drain, and finally a body terminal connecting the semiconductor mass. As consequence, **the field is able to control the current density of these free carriers, either electrons or holes, effectively changing the semiconductor conductivity between the drain and the source**. The semiconductor can be classified as N-type or P-type depending on the majority of electric carriers. The source and drain are doped regions accordingly to the MOSFET type (NMOS or PMOS) and opposite to the channel type, forming a NPN or PNP structure respectively. The typical bulk structure is illustrated in Figure 1.1.2.1a as well the contacts in Figure 1.1.2.1b FD-SOI MOSFET transistor. In schematic abstraction level, the NMOS is represented as Figure 1.1.2.1c.

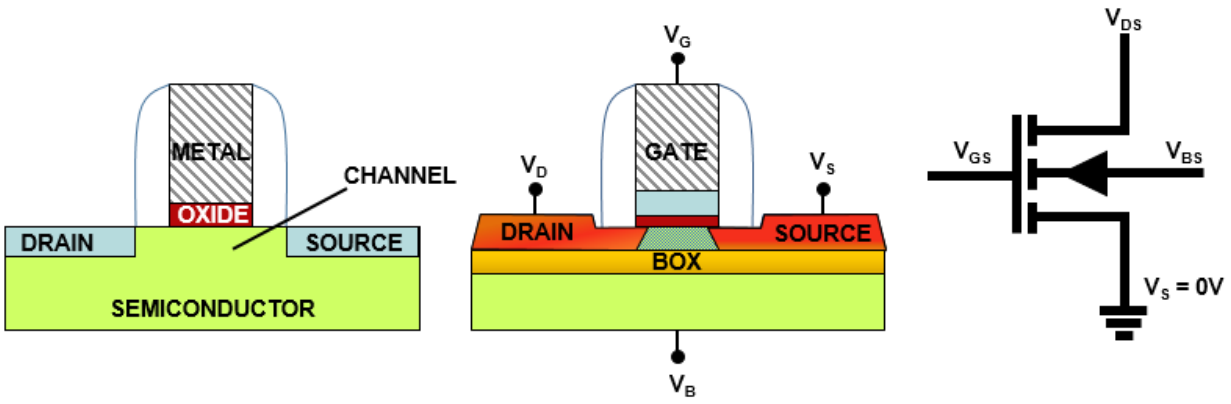


Figure 1.1.2.1 (a) Typical Bulk MOSFET; (b) FD-SOI MOSFET with 4 terminals; (c) Schematic representation of NMOS.

The gate field effect can modulate the charge concentration in the channel region, this effect is possible due to the MOS capacitance structure. For a NMOS, when gate voltage bias is applied, after a certain level (called threshold voltage) it can create an inversion layer in the P-substrate, meaning that substrate-gate oxide interface is populated with electrons, and then an electrical current can flow from drain to source. In an ideal condition, no current flows when the gate bias is under the threshold voltage. **In real devices, a small current can flow even if the gate is turned off, and this effect is called leakage current**. The leakage arises from the junction's carrier recombination as well as the tunneling effects, such as gate-oxide leakage [J. Chen 1987]. The bulk terminal can modulate the voltage across the semiconductor substrate, modifying the channel charges concentration, thus providing a V_T shift in order to decrease the leakage or increase the performance. The typical MOSFET top-view for design layout is shown in Figure 1.1.2.2. The active region, or the complete transistor bulk is seen in green. The gate stack is simplified as strip shown in red. The source and drain contacts are in white and blue. This is a typical representation for planar transistor, either bulk or FDSOI. With FinFETs the design layout may differ, as the Fins are shown in the layout, and the bulk region may not be represented. The gate length L_G was the smallest feature size in the transistor, and historically was used to name the transistor node, despite of this not holding true anymore.

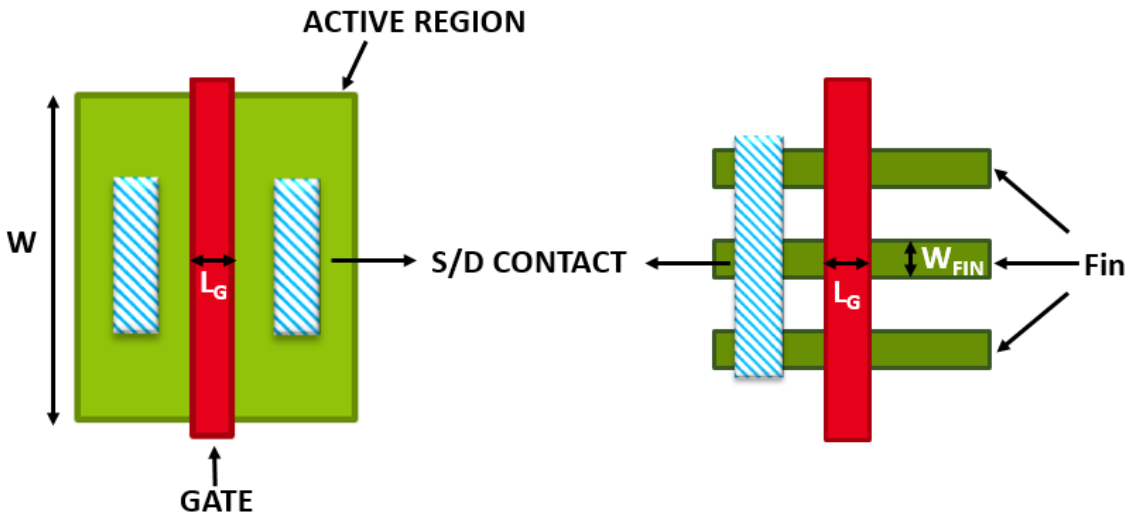


Figure 1.1.2.2 Simplified top view of transistor layout for (a) Bulk and FDSOI; (b) FinFET

The operation of MOSFET transistor depends on the combination of voltages applied in its four terminals and the analysis is done regarding the current between the drain to source, referenced as $I_{D.}$ In a general manner, the MOSFET can operate in analog, radio-frequency (RF) and digital domains. The following figures of merit will focus on the digital operation, which the main goal is to switch between the logic states.

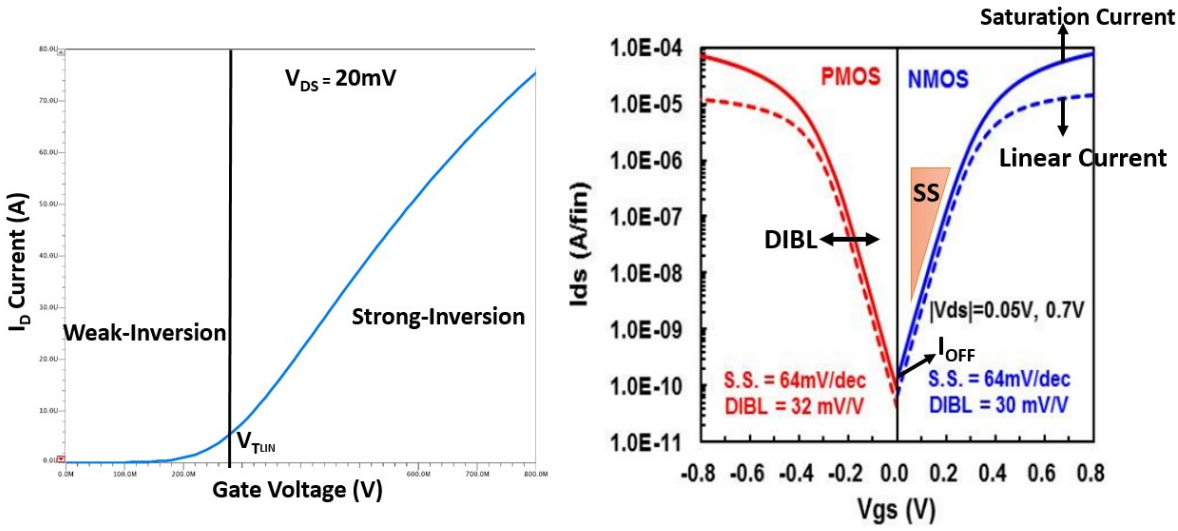


Figure 1.1.2.3 (a) Transistor I_D current versus gate voltage for 14nm FDSOI; (b) 10nm FinFET for SRAM figures of merit for I_D in logarithm scale [S. Y. Wu 2016].

The classic benchmarking parameters are illustrated in Figure 1.1.2.3b by plotting the logarithm I_D versus the gate voltage, in this case for PMOS and NMOS. The parameters are the I_{OFF}, $I_{D\ SAT}$, $I_{D\ LIN}$, Subthreshold Swing (SS), and the Drain-Induced Barrier Lowering (DIBL). The DIBL is given by (1.1) considering the

difference of the threshold voltage for a given delta in VDD. The inverse of the slope of the I_D curve is called SS, and it is given by (1.2).

$$DIBL = \frac{\Delta V_T}{\Delta V_{DD}} \quad (1.1)$$

$$SS = \frac{\partial V_{GS}}{\partial log I_D} \quad (1.2)$$

The SS has a minimum value of 60mV/dec for a conventional silicon device, using T=300K and $C_{OX} >> C_{DEP}$, where C_{OX} is the gate-oxide capacitance and C_{DEP} is the depletion layer capacitance. The DIBL is due to the short channel effect, where the S/D (source and drain) junctions create a superposed depletion laterally under the channel, thus reducing the gate electrostatic control and lowering the transistor V_T. By applying a bias in the drain, this effect is enhanced, thus there is an I_D curve difference between the linear and saturation modes. The DIBL is calculated by the difference of V_{TSAT} and $V_{TLIN.}$ Besides the MOSFET figures of merit, the final circuit is often evaluated using the **P**erformance, **P**ower and **A**rea **(PPA) metric**. The first two are directly linked to costumer usability, in the sense of speed, mobility for battery powered devices, emitted heat and power consumption cost. The area is inherently tied to the device cost.

1.1.3 Dennard's Scaling

Moore's scaling became a *de facto* standard in the industry over the years, and it was marketed as the node number metric, representing the transistor minimum feature size, or the gate length. This trend is shown in Figure 1.1.3.1. Starting at 0.25 µm node, the transistor gate length does not represent anymore the real gate size, but due to marketing reasons, the expected node scaling x0.7 trend was kept in the node name.

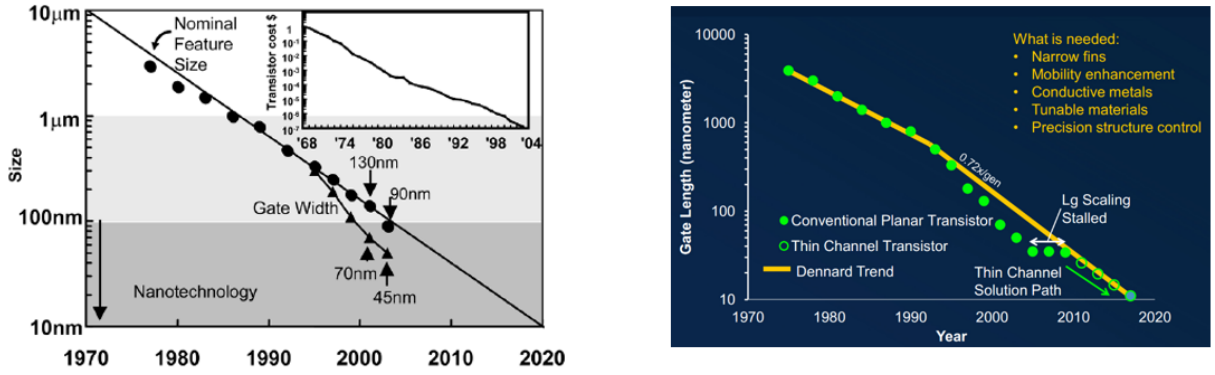


Figure 1.1.3.1 Transistor gate length feature size over the years. On the left, (a) Intel data and transistor cost [Thompson 2004]; (b) Applied Materials data extrapolation to 2020.

Besides the gate length scaling, the transistor parameters were also being scaled. This overall scaling trend was noted by Robert H. Dennard in 1974 as follows in Figure 1.1.3.2 [Dennard 1974], proposing a performance increase due to scaling. The parameters are scaled by a factor κ.

Device or Circuit Parameter	Scaling Factor
Device dimension t_{ox}, L, W	$1/\kappa$
Doping concentration N_a	κ
Voltage V	$1/\kappa$
Current I	$1/\kappa$
Capacitance $\epsilon A/t$	$1/\kappa$
Delay time/circuit VC/I	$1/\kappa$
Power dissipation/circuit VI	$1/\kappa^2$
Power density VI/A	1

Figure 1.1.3.2 Dennard's Scaling parameters to increase circuit performance. [Dennard 1999]

Dennard's law supposes a constant power density for a given area, thus the voltage and current are scaled by a factor κ^{-1} while area of a given device is reduced by a factor κ^2. The area reduction for a given element, such as metal interconnection, is also κ^2 while the dielectric insulating distance decreases by a factor κ, hence resulting in parasitic capacitance reduction by a factor κ^{-1}. Circuit performance wise, the Moore's Scaling is highly beneficial. As noted by Dennard, the several scaling parameters increase the device performance as illustrated in Figure 1.1.3.3 for scaling the spacing between adjacent gates (also defined as Contact Poly Pitch: CPP). **By scaling the devices, it reduced the overall costs, but it also increased the**

circuit performance. In other words, doing more powerful devices made them cheaper. This is sometimes referred as the "**scaling free lunch**" or "**golden age of scaling**". The outcome is strongly opposed to conventional product engineering: usually to make something better, it becomes more expensive.

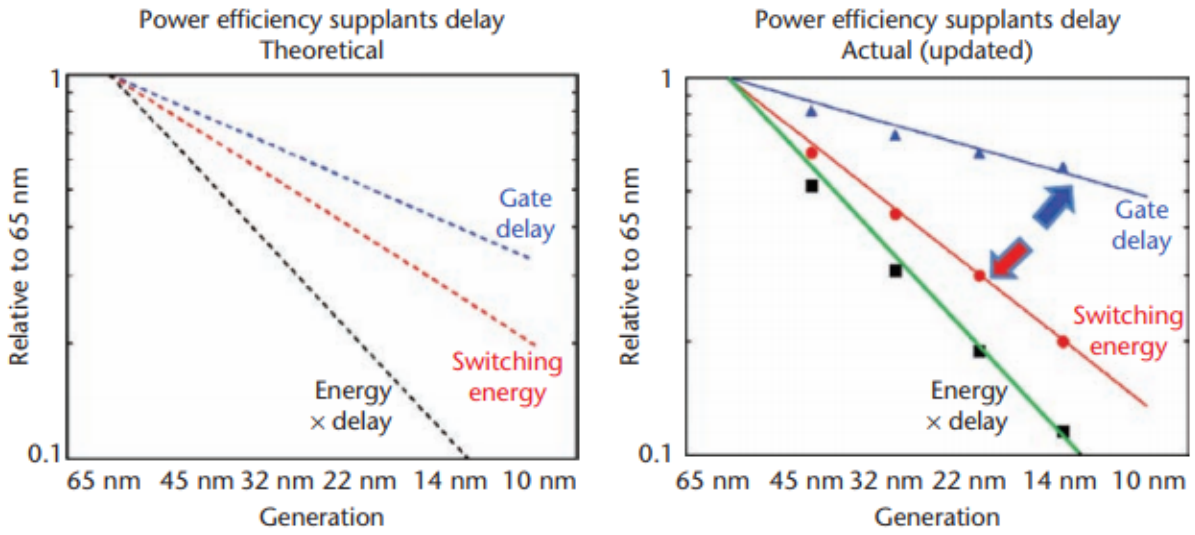


Figure 1.1.3.3 Performance increase over the years (reducing gate delay). At the same time, the switching power was reduced. [Gargini 2017]

1.1.4 2000's Technical Advances on Scaling

The scaling met some difficulties to deliver the expected circuit performance in the years 2000's. New device solutions were proposed and implemented, increasing the process complexity rather than a simple parameter scaling as in the previous years. A brief of major technical solutions is presented in this section.

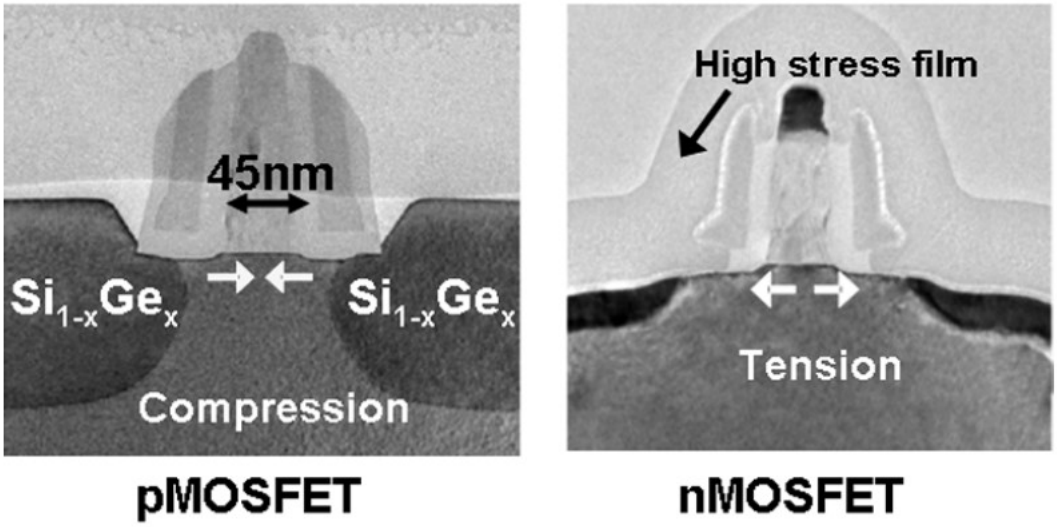


Figure 1.1.4.1 In the 90nm node, the channel mechanical strain was introduced to increase the transistor performance. [Thompson 2004, 90-].

In the 90nm node, circa 2004, for the first time the devices featured the **mechanical strain engineering**. The strained silicon is achieved in PMOS by using SiGe in S/D, hence the mismatch between the SiGe and the Si in the channel creates a compressive stress in the P channel improving the hole mobility. In the NMOS channel the tension stress is necessary, thus in Intel process it was achieved by using a high stress nitride-capping layer as in Figure 1.1.4.1. The tensile stress in N channel increases the electron mobility.

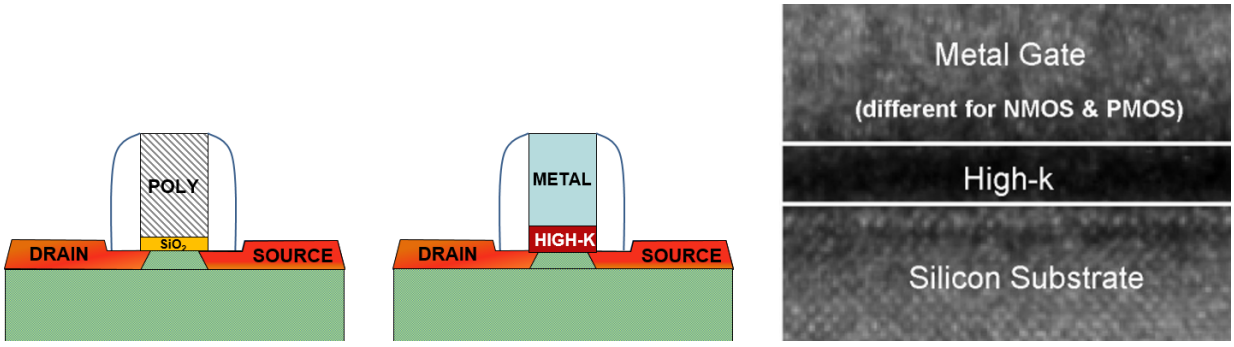


Figure 1.1.4.2 In the 45nm node, the gate-oxide was changed from the typical SiO_2 to HfO_2 High-K isolation. [Mistry 2007]

Continuing the scaling, at the 45nm node another important feature has been introduced: the **high-k materials for the gate-oxide** as pictured in Figure 1.1.4.2. Following the Dennard's scaling, the gate oxide

thickness was reduced over the years in order to increase the Cox, and consequently the transistor performance. However, the thickness required in the 45nm would be lower than 2nm following the scaling trend, and consequently the device would become too leaky, because of the carrier tunneling through the very thin gate oxide. The solution was to use a **higher permittivity material** than SiO_2 (ε=3.9). The HfO_2 was chosen, with ε>10, allowing a higher Cox, while being thicker than the equivalent SiO_2 film, hence **reducing the gate leakage** by 25x [Mistry 2007]. After this technology, the Equivalent Oxide Thickness (EOT) definition was widely adopted. It straightforwardly compares a given high-k to SiO_2 theoretical thickness to achieve the same capacitance. The EOT definition is described in 1.3.

$$EOT = thickness_{HighK} \frac{\varepsilon_{SiO_2}}{\varepsilon_{HighK}} \qquad (1.3)$$

Along with gate-oxide change in the 45nm node, some processes started using a **gate-last integration** in the 22nm node. In this process flow the gate high-k is deposited after the removal of dummy gates and before the metal gate electrodes. This integration avoids the thermal stress in the gate oxide during the S/D annealing. Another solution used in the 2000's is the raised source and drain (RSD) starting from the 90nm node. This solution reduces the source and drain contact resistance by increasing the contact height, decreasing the S/D sheet resistance, which is in series with the channel resistance, thus increasing the transistor performance. Several process solutions can achieve the raised S/D, for example the selective epitaxial growth [H.-J. Huang 2001].

1.1.5 Rise of new MOSFET architectures

After the 28nm node, the standard bulk transistor was difficult to scale even using the improved process developed in the 2000's. Emergent transistor architectures were employed, and are currently in research. Those **architectures are focused on improving the gate electrostatic control over the channel**. The industry has been biased over three major architectures illustrated in Figure 1.1.5.1, namely the FD-SOI, FinFETs and Nanowires.

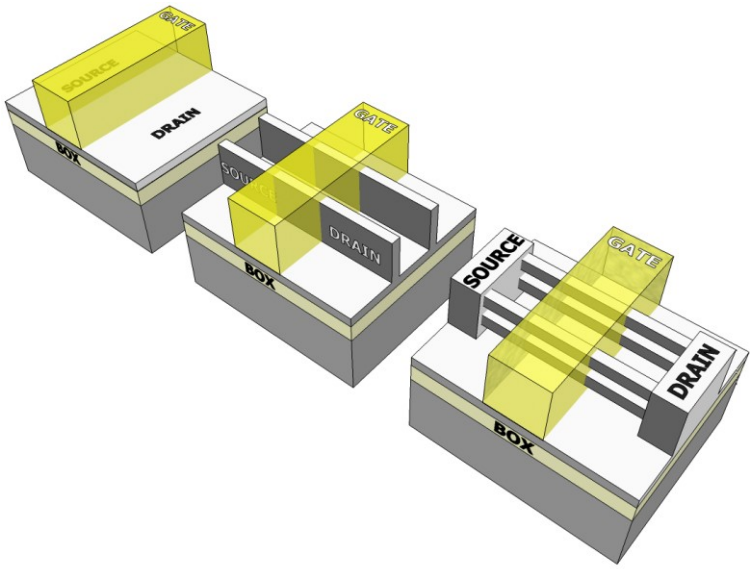


Figure 1.1.5.1 Conceptual transistor archichetures in burried oxide for: (a) FD-SOI; (b) FinFET; (c) Nanowires.

FD-SOI stands for Fully Depleted Silicon On Insulator, and was a technological substrate process evolution from PD-SOI (Partially Depleted version). The Buried Oxide (BOX) limits the bulk leakage, and the S/D extends to BOX, minimizing junction area, leakages as well as the capacitances [Cristoloveanu 1999]. The PD-SOI silicon film thickness (T_{SI}) over the BOX forming the channel is not thin enough to create a depletion charge fully controlled by the gate as illustrated in Figure 1.1.5.2a. In fact, a quasi-neutral zone is present under the formed channel, and if the body connection is floating some undesirable effects may occur, such as the kink effect (causing hysteresis in I_D-V_G curves, and potentially latching-up the transistor). Employing a thin T_{SI} can suppress those undesirable effects because the channel becomes fully depleted 1.3as shown in Figure 1.1.5.2b. The channel can be controlled by the usual gate, but also from the potential under the BOX, namely the back-gate. The FD-SOI advantages over BULK and PD-SOI are the lower leakage, no latch-up risk, and the lower variability due to lower doping in the channel, as well the increased electrostatic control.

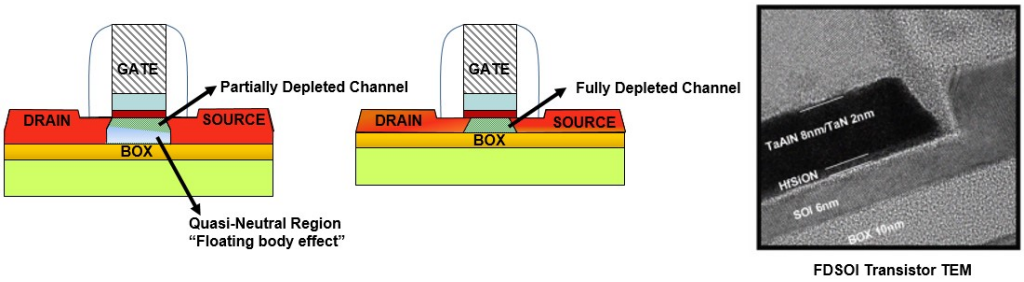


Figure 1.1.5.2 (a) PD-SOI concept; (b) FD-SOI concept; (c) FD-SOI TEM cross section image [Weber 2010].

A T_{SI} of 6nm is shown in Figure 1.1.5.2c for a 28nm FD-SOI transistor for the 28nm node. An ultra-thin SOI wafer process was developed to achieve low variability The SmartCut™ process is described in Figure 1.1.5.3. The core of this process is to use a handling wafer bonded to another wafer on which an oxidation layer is done (and later will become the BOX). Then, hydrogen atoms are implanted to a certain depth, and a splitting is done at this level, followed by chemical-mechanical planarization (CMP) to flatten uniformly the silicon thickness.

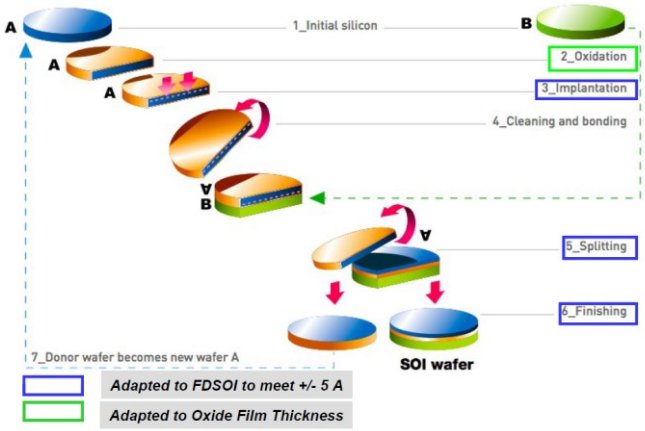


Figure 1.1.5.3 Ultra-thin SOI wafer process for mass production, SmartCut™. [Schwarzenbach 2011]

Besides the FD-SOI, another transistor architecture was developed and committed to mass production: FinFETs. In this architecture, the transistor channel is no longer planar. During the process, the silicon is etched to create the "fins", or the 3D channel. Then the gate stack is fabricated around the channel. This **effectively increases the gate electrostatic control for a normalized planar area**. **Indeed, the effective gate width is calculated as two times the fin height plus the fin width**. This technology is also called tri-gate, as the gate controls the channel by three different faces. The excellent electrostatic control as shown in Figure 1.1.5.4 increases the transistor performance per area, compared to the planar devices.

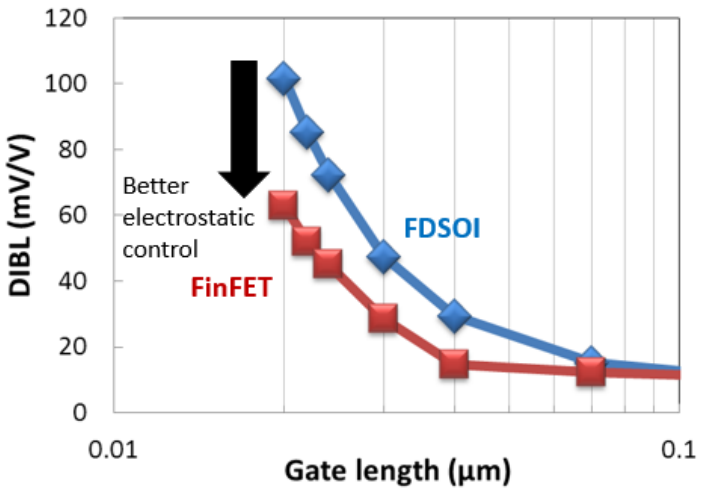


Figure 1.1.5.4 Channel electrostatic control for FinFET and FDSOI. The higher W_{EFF} for FinFET results in a better electrostatic control, especially for short gates [Rozeau 2015].

After the introduction of the first FinFET family in the 22nm, the following nodes are scaling by reducing dimensions, and more importantly by increasing the gate W_{EFF}, which translates in an increase of the fin height as illustrated in Figure 1.1.5.5.

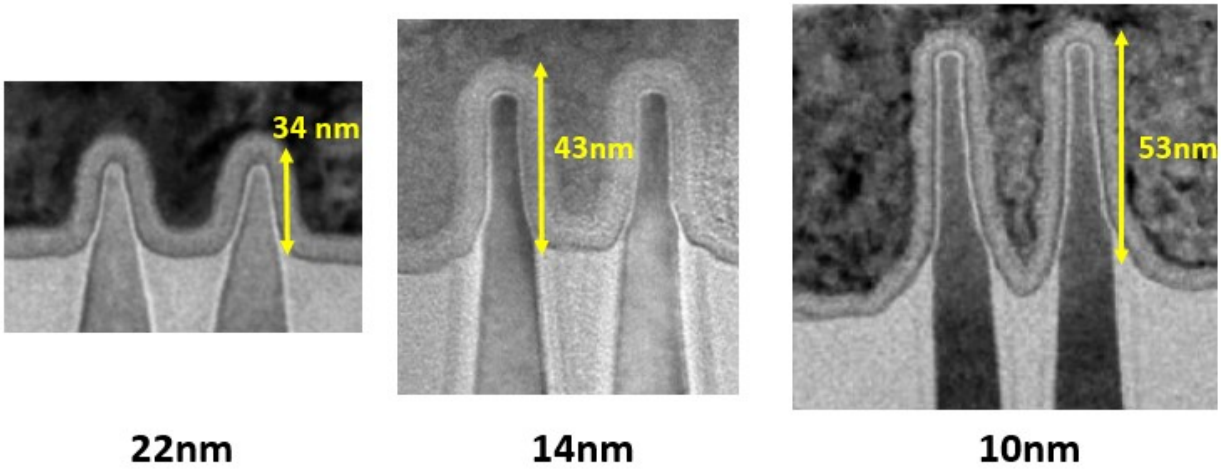


Figure 1.1.5.5 FinFET scaling over the nodes. Each node the fin height is increasing to increase the channel W_{EFF} (data from Intel).

As the time of this publication (2017), the 10nm is the current node. The 7nm probably will be the last node employing the FinFET architecture due to the difficulty to scale any longer the fin height, limited by the aspect ratio. The Gate-All-Around (GAA) architecture is proposed and fully researched as a natural advancement from fins. In those transistors, the fin is transformed in a square, and then the gate encloses its four faces. **The higher W_{EFF} per area allows an even better channel electrostatic control** as illustrated in Figure 1.1.5.6, reducing the DIBL (higher gate control for a given V_{DS} polarization) and the SS (faster transition among off state to saturation).

	10nm		7nm	5nm
	FDSOI	FinFET	FinFET	GAA
DIBL (mV/V)	100	62	78	70
Slope (mV/dec)	84	72	79	71

Figure 1.1.5.6 Transistor DIBL and SS compared for different architectures [Rozeau 2015].

Due to the channel format, those transistors are also called nanowires or nanosheets. The first nanowires were proposed as omega gates for less complex process integration, as pictured in Figure 1.1.5.7b. **The GAA can be stacked to increase even further the performance per area** as illustrated in Figure 1.1.5.7b. The nanowires using a cross section of a rectangle, are named nanosheets. Those architectures **should be delivered in the 4nm** as illustrated in Figure 1.1.5.8. Hence, the scaling trend should continue, at least for six years more. The downside is the process complexity, as for the example the introduction of EUV (extreme ultraviolet) lithography, which increases the lithography resolution by employing a shorter wavelength (13.5nm).

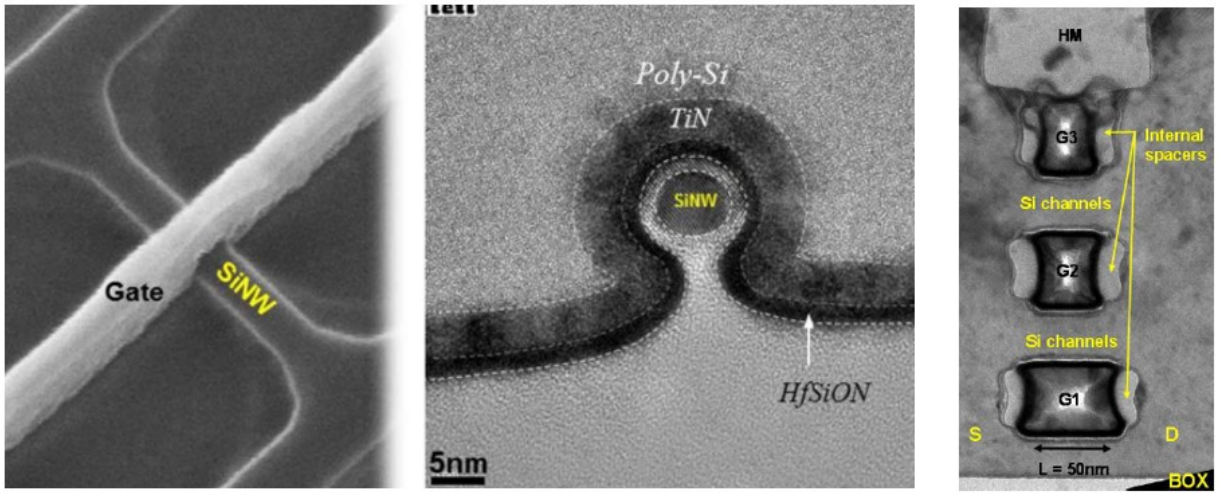


Figure 1.1.5.7 (a)Nanowire TEM image.(b) The omega gate encloses the channel [Barraud 2012]. On the right, (c) The GAA; granting exceptional channel electrostatic control [Ernst 2008].

Summary of TSMC's Major Future R&D Projects

Project Name	Description	Risk Production (Estimated Target Schedule)
7nm logic platform technology and applications	4th generation FinFET CMOS platform technology for SoC	2017
5nm logic platform technology and applications	5th generation FinFET CMOS platform technology for SoC	2019
3D IC	Cost-effective solution with better form factor and performance for SiP	2016 ~ 2017
Next-generation lithography	EUV lithography and related patterning technology to extend Moore's Law	2016 ~ 2019
Long-term research	Specialty SoC technology (including new NVM, MEMS, RF, analog) and transistors for 5nm node and beyond	2015 ~ 2019

Samsung Foundry Roadmap – June 2017

Project Name	Description	Estimated Schedule
8nm	10nm FinFET shrink	2018
7nm	First introduction of EUV lithography	2017~2018
6nm	7nm shrink, second generation using EUV	2019~2020
5nm	Last FinFET node scaling	2019~2020
4nm	First introduction of Nanosheets	2020~2022

Figure 1.1.5.8 Foundries roadmap. In the top, (a) TSMC expected technology development. In the bottom, (b) Samsung Roadmap, displaying nanosheets introduction at 4nm node. (Data from TSMC and Samsung).

1.2 3D Integration as More than Moore's Alternative

1.2.1 Motivation and Concept

The standard integration presented in the previous sections, is also known as monolithic. The transistors are done and interconnected in a sequential process. The transistor steps are done, in what is called the front-end of the line process (FEOL). The interconnections are done above the FEOL level, by stacking metal levels separated by dielectric isolation material, forming the back-end of the line (BEOL). Both FEOL and BEOL are done using the same mask alignment marks. Despite of 3D transistor architecture being employed such as FinFETs, the overall process rests in a planar fashion, since all transistors are done side by side. The scaling reduces the BEOL metal dimensions and the contacted poly pitch (distance between adjacent gates) as shown in Figure 1.2.1.1.

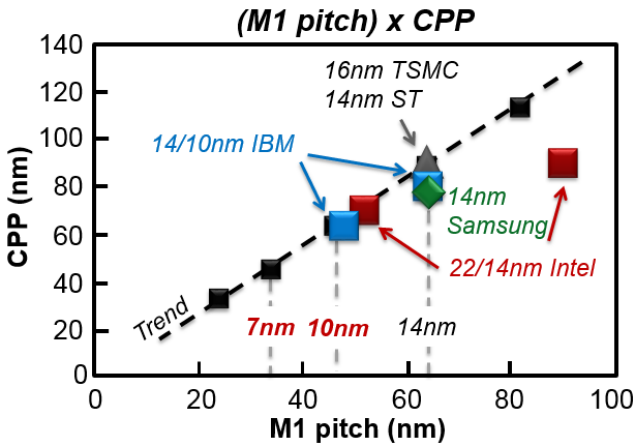


Figure 1.2.1.1 CPP and Metal pitch scaling for the first metal level over the transistors in planar devices. Each node decreases even further the dimensions [Rozeau 2015].

The scaling increases the BEOL parasitic capacitance and resistance as illustrated in Figure 1.2.1.2. Those effects will be discussed in detail on Chapter Two, however it demonstrates a limit for scaling. **The circuits will achieve one point, where the BEOL parasitic elements will become a physical barrier**, and 3D integration is an excellent alternative.

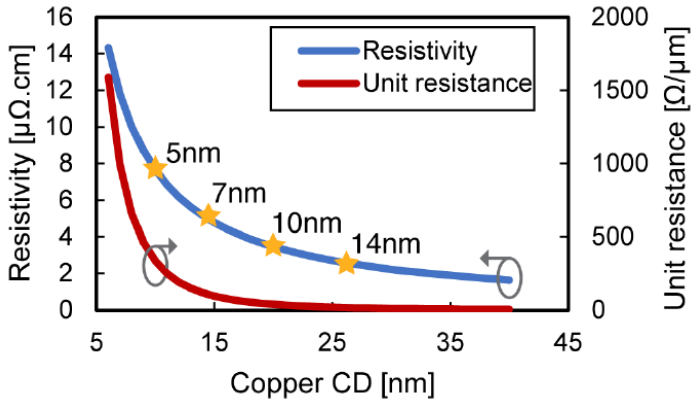


Figure 1.2.1.2 Metal resistance increase due to reduced dimensions. The resistivity also increases due to reduced mean free path for electrons. [Huynh-Bao 2017]

Given the increasing difficulty to continue the Moore's Law, as transistor dimensions are achieving atomic scales, stacking devices in several tiers can be a solution to do more than scaling. Indeed, several solutions are already used for mass production, especially for memory and imaging devices [Fujii 2012; Knickerbocker 2012]. The Moore's Law as discussed previously is about reducing the cost per transistor. In this area, the 3D integration is also effective. The scaling can bring up the total cost, because of the difficulty to reduce dimensions even more, the cost of high precision machines, and the non-recurring costs of process and design development. By **employing the 3D integration**, the machines and the process from planar nodes can be reused, thus the investment is lower. The total **count of transistor per area increases with the number of stacked tiers**. Therefore, in each tier the area is lower compared to the similar planar circuit, translating in **lower interconnections length and increased circuit performance**. A planar process performance can effectively boost up by using the 3D scheme. The foundry doing this transition potentially has a lower cost, as the development and non-recurring **costs are lower than researching scaling of a new node from the scratch**. The **3D integration also allows the function integration in the same chip**, for example, using a process optimized for logic in one tier, while another tier is optimized to analog or sensing functions.

The 3D circuits are divided into three categories:

- **2.5D Integration:** Circuit dies are fabricated separately and bonded or soldered to an interposer. The interposer is generally a silicon part with no active devices (without transistors), and its main function is to provide the interconnection for its attached dies. It can be though as a backplane connecting the bus of several dies.
- **Parallel TSV Integration:** The circuit is done in different wafers, and then the wafers are bonded together. The circuit in different tiers are connected using Through-Silicon Vias (TSVs). This approach limits the via size, because the bonding phase has a certain alignment precision, limiting the via pitch, and via surface area has to be big enough to deliver an acceptable aspect ratio.
- **Sequential Integration:** The whole circuit is processed in a continuous manner. After the first circuit tier process, the following top process will use the previous alignment marks. The top process has to be special in order to not affect the already built tier, meaning that the top thermal budget has to be limited. The main advantage of this integration over the parallel integration is the via density connecting the two tiers, as the alignment is secured by the process. 3D sequential stacked transistors are shown in Figure 1.2.1.3.

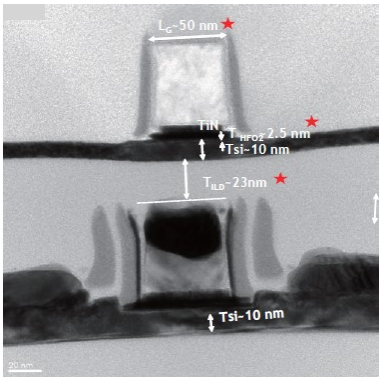


Figure 1.2.1.3 3D Sequential Integration in two tiers. [Batude 2011]

1.2.2 TSV – Parallel Integration

Using distinct process for different circuit functions can be cost effectively, as the process can be focused for a determined design, and the circuit can be processed in parallel; giving the chance to test the circuits before bonding in the final package and reducing the production time. The 2.5D integration uses dummy backplanes to provide connections between its soldered dies. An interesting example of 2.5D interposer and TSVs is illustrated in Figure 1.2.2.1, effectively forming a 3-tier package.

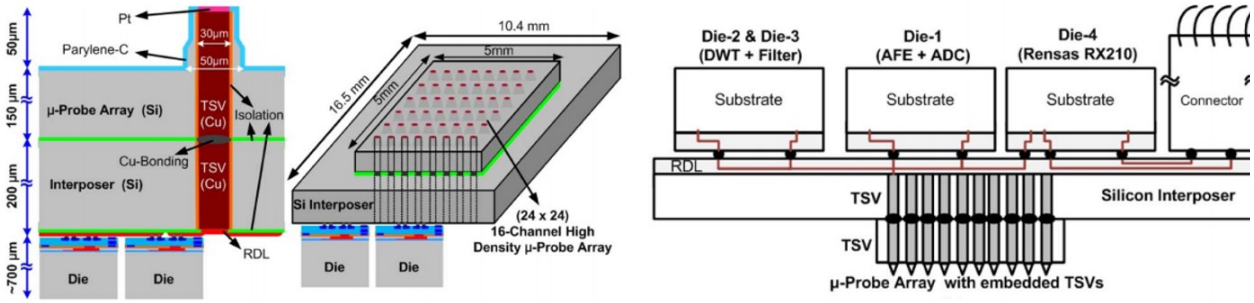


Figure 1.2.2.1 Integration using 2.5D interposer and TSVs. [P. T. Huang 2014]

A parallel integration using TSVs can be done in two main flavors: face-to-face, where the wafers are bonded with one in reverse position, meaning the both back-end are together; another way is to do the face-to-back, where one wafer is bonded on the back of the active region of another wafer as shown in Figure 1.2.2.2a. The face-to-back has at least three main process approaches: via-first (TSV processed before the FEOL), via-middle (TSV are done after the FEOL and before the BEOL) and the via-last (TSVs manufactured after the FEOL and BEOL) [Hsieh 2012]. The TSV bonding step can be done at wafer metal level, dielectric level or both, forming a hybrid bonding [K. N. Chen 2011] as illustrated in Figure 1.2.2.2b. The bonding at metal levels is done by parallel contact pressure and heat, usually using cooper (hence the name Cu-Cu bonding). If the wafers are bonded at dielectric level, the oxide-to-oxide gluing is done by fusion bonding of hydrophilic surface, thus the oxide surface should be ultra-clean.

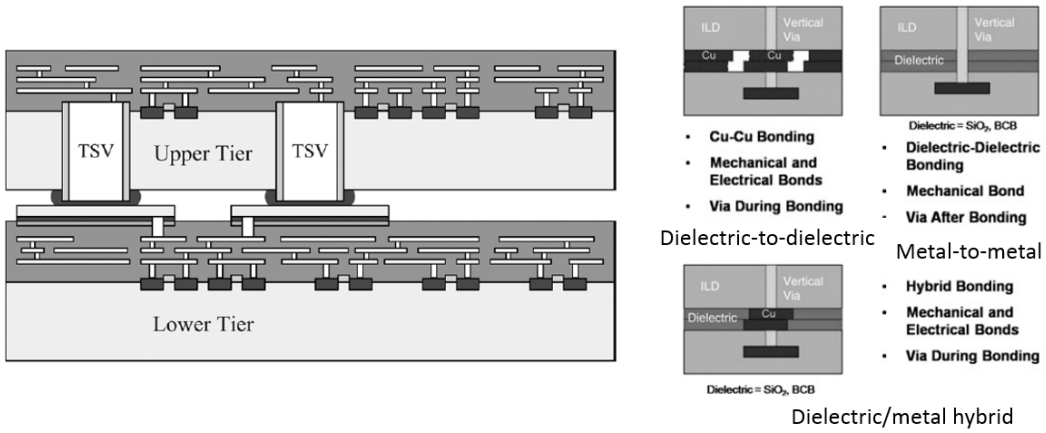
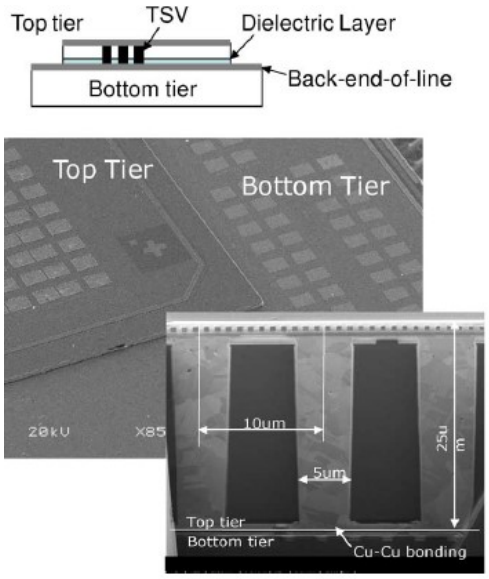


Figure 1.2.2.2 On the left, (a) TSV packaging schematic for face-to-back [Hsieh 2012]. On the right, (b) Wafer bonding process schemes for TSV manufacture [K. N. Chen 2011].

Finally, the wafers can be bonded by using polymers which are compliant to silicon oxide, reducing the requirements of surface cleanness. As the alignment is the key factor during the process, the TSV pad size has to be set accordingly to the alignment limitations; in other words, the TSV pad size has to be large enough to compensate the misalignment. The TSV aspect ratio, the proportion between the contact area and its height has to follow a determined constant in order to be compliant to the process, avoiding defects due to deep etching and metal filling.



TSV diameter	5um
Min. TSV pitch	10um
Top tier die thickness	25um
TSV Isolation layer thickness	120n
Dielectric layer thickness	0.7um

Figure 1.2.2.3 3D parallel integration using TSV Cu-Cu bonding with dimensions. [Plas 2011]

The TSV process featuring Cu-Cu bonding is illustrated in Figure 1.2.2.3 as well the TSV dimensions and pitch. The TSV size benchmarking is shown in Figure 1.2.2.4. The x-axis represents the TSV size while the y-axis shows the aspect ratio, both in log-scale. The ideal TSV has a small diameter size, and a small aspect ratio, because it decreases the via overhead, and the via length reducing the parasitic capacitances. As seen in Figure 1.2.2.4, the TSV diameter size still higher than 1µm, limiting the maximum via density between tiers. Moreover, the TSV technology can suffer from defects, such as internal voids, structural damage or misalignments as shown in Figure 1.2.2.5a. The high via size makes it area costly to employ double vias in order to increase the circuit reliability, in case of one via fails during the process. In Figure 1.2.2.5b a study shows the percentage of chips without failures, depending on the TSV via count and the number of tiers. As expected, by increasing the number of tiers and the number of TSV contacts, the yield gets worse.

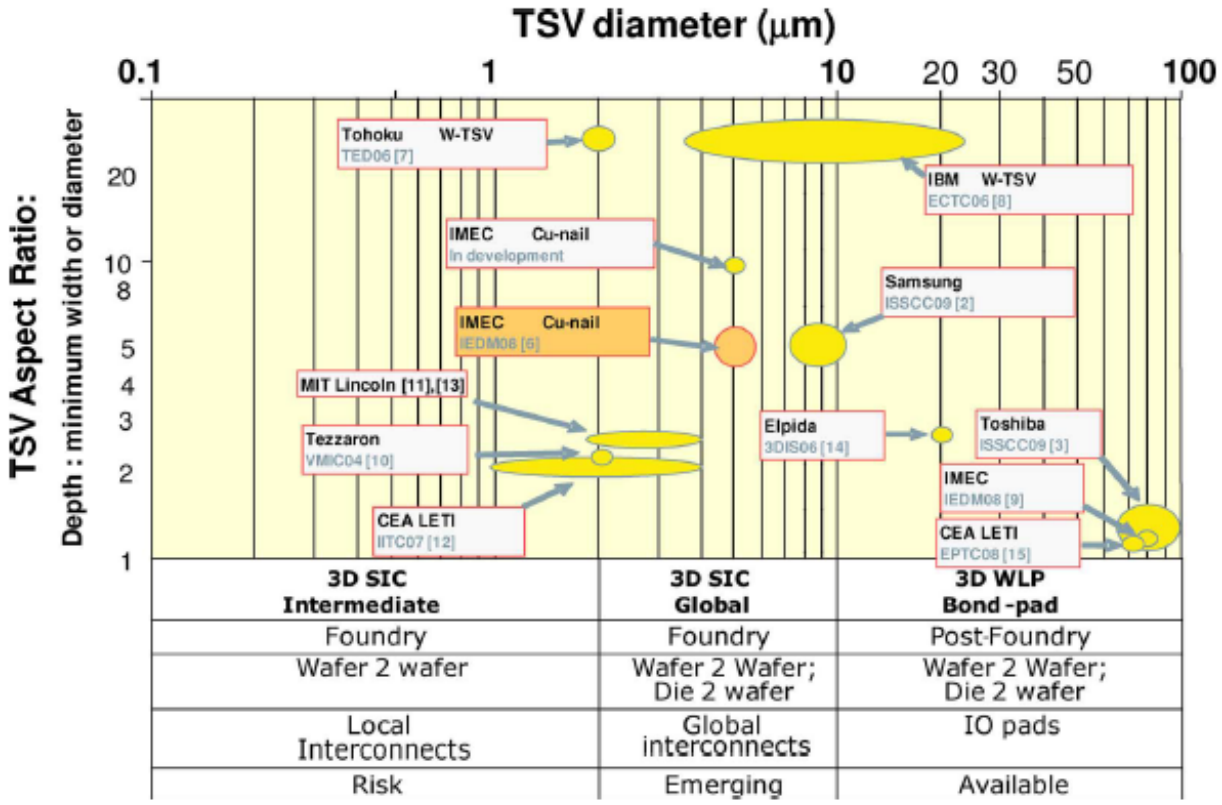


Figure 1.2.2.4 TSV contact size for given aspect ratio. [Plas 2011]

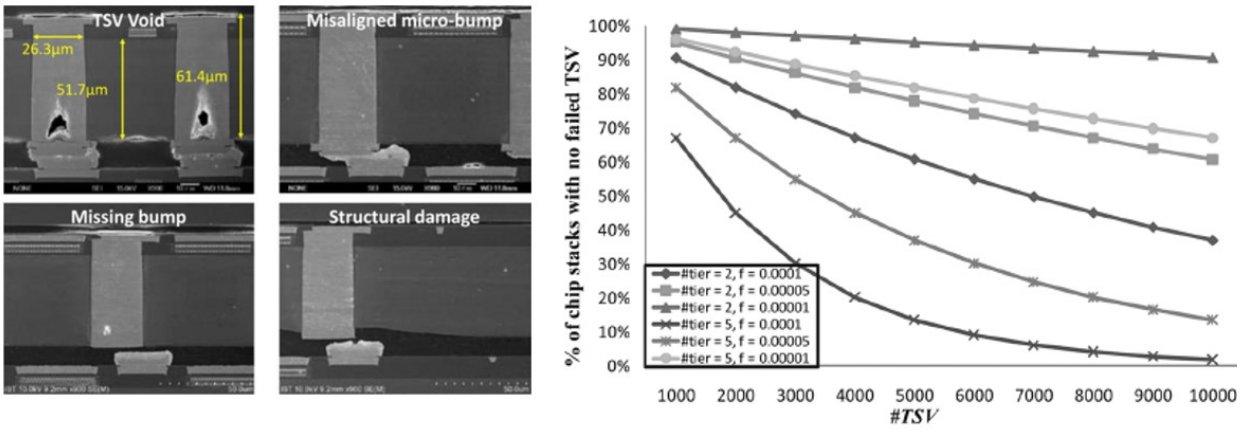


Figure 1.2.2.5 On the left, (a)TSV sources of defects [Lin 2013] (b)On the right, the circuit realibity considering the TSV count [Hsieh 2012].

As result of limited TSV density, the design workaround usually takes benefits from memory blocks close to the processing unit, or even stacking memory tiers, as memory I/O can be reduced to a bus and address lines, being capable to overcome the TSV limitation [Lee 2014].

1.2.3 Monolithic 3D Sequential Integration – State of the Art

To overcome the limitations of the TSV process, the 3D sequential processing was proposed as illustrated in Figure 1.2.3.1. As its tiers are aligned over the process, the vias connections through the tiers (3DCO) can have a higher density, as well sizes and pitch comparable to Metal 1 vias, allowing a very flexible placement during the design phase.

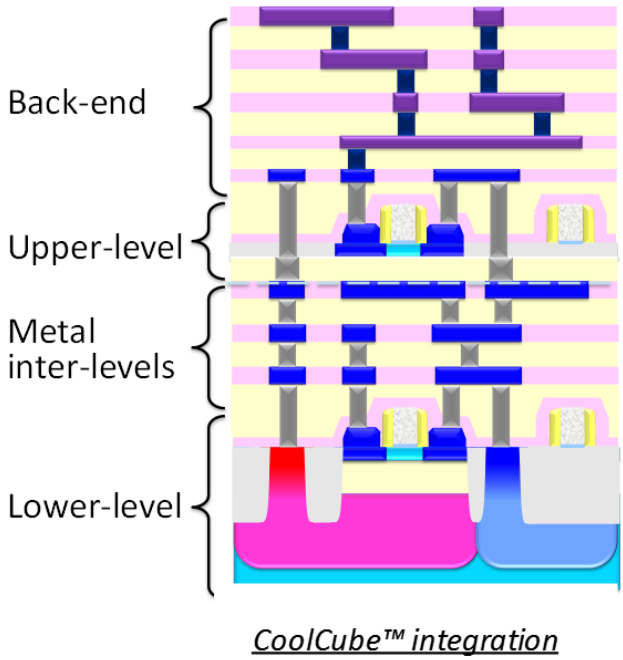


Figure 1.2.3.1 CEA-LETI 3D sequential integration process CoolCube™. The bottom tier has routing metals called intermediate back-end (iBEOL).

The ultra-high density of 3DCO gives more flexibility to 3D sequential designs, as the circuit can have more interconnections among tiers, and the area penalty is minimal allowing the placement of double vias to increase the reliability. A comparison between layouts using 3D parallel and sequential is shown in Figure 1.2.3.2.

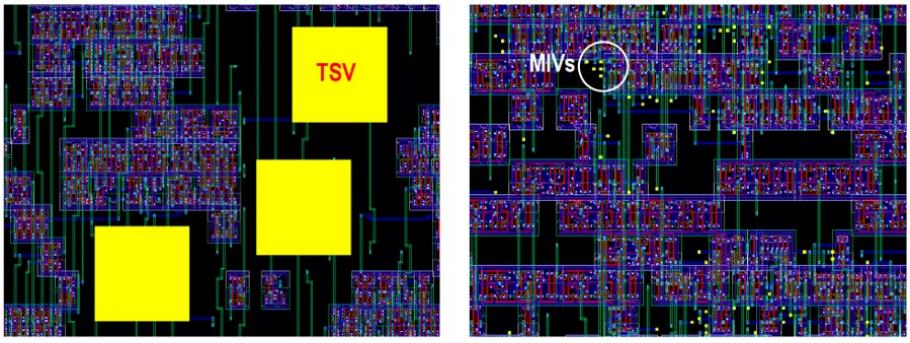


Figure 1.2.3.2 Layout view comparing TSV to 3D sequential via placement [Liu 2012]. The MIV stands for Monolithic Integrated Vias, which in this thesis is often referenced as 3DCO.

The density of 3CDCO is tremendously superior compared to the TSV density as illustrated in Figure 1.2.3.3, mainly due to alignment precision. The alignment also allows a small 3DCO size, reducing the via height to keep a certain aspect ratio. Hence, the 3DCO parasitic elements are decreased such as capacitance, because the small size reduces the lateral area decreasing the parasitic capacitances.

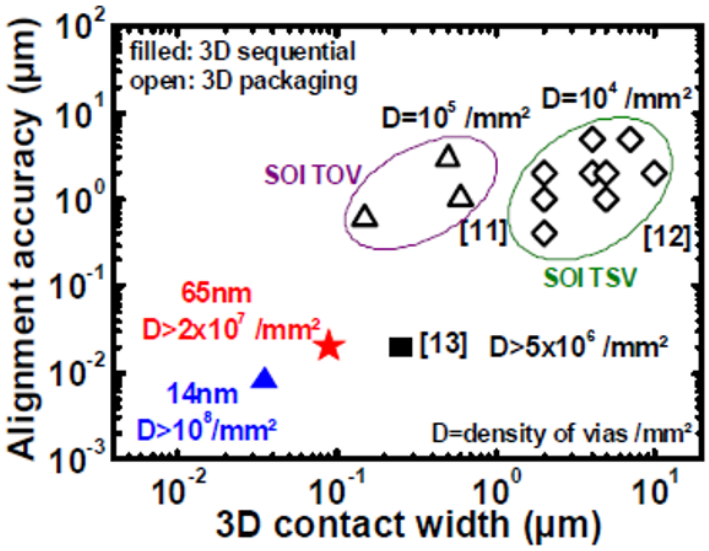


Figure 1.2.3.3 TSV to 3DCO comparison for different processes. [Brunet 2016]

In a design perspective, **the contact density among the tiers defines the maximum granularity**, or in other words, in which design level the circuit can be partitioned between tiers. In Figure 1.2.3.4 the design granularity level is shown as four main flavors. The parallel integration is suitable for low-density contacts, enabling entire core or logic blocks 3D integration. On the other hand, the sequential process enables all the granularity scale, including logic gates (CMOS over CMOS) and transistor over transistor integrations.

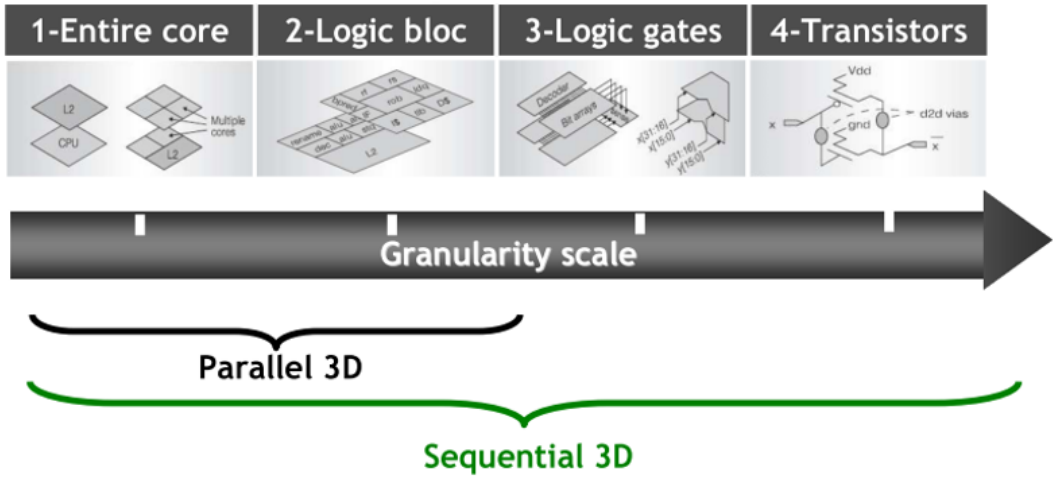


Figure 1.2.3.4 Nanowire TEM image. The gate encloses all the channel; granting exceptional channel electrostatic control. [Batude 2014]

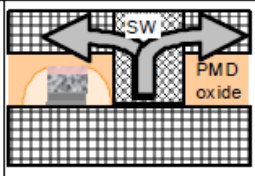
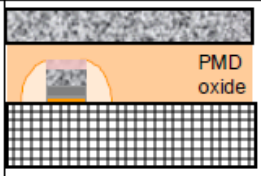
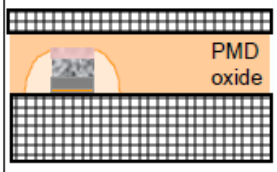
	Seed window (SW)	Poly-Si	Wafer bonding
Description			
Density	limited due to SW	Same than bottom level	Same than bottom level
Crystalline quality	Defect in SW region with controlled location	Random defects location	Perfect quality ~SOI supply quality
Thickness control	10s nm range	nm range	Å range
layer orientation	same orientation	random orientation for top substrate	different orientation possible

Figure 1.2.3.5 Comparison of 3D sequential process manufacturing strategies. [Batude 2011].

The **monolithic process has several approaches** in order to manufacture the top tier as categorized by [Batude 2011] seen in Figure 1.2.3.5. The main goal of the process is to achieve a good silicon lattice quality, and protect the already built tier by controlling the thermal budget. The seed window method creates a path for silicon seed from bottom wafer in order to recrystallize the top wafer. A second approach creates the transistor without the monocrystalline silicon in the top. The epitaxy-like silicon is done by a laser annealing; however the silicon layer is polycrystalline. Although the polycrystalline limits the transistor performance, the process is potentially low-cost. The process integration using laser annealing is shown in Figure 1.2.3.6 from [Shen 2013; T. T. Wu 2015].

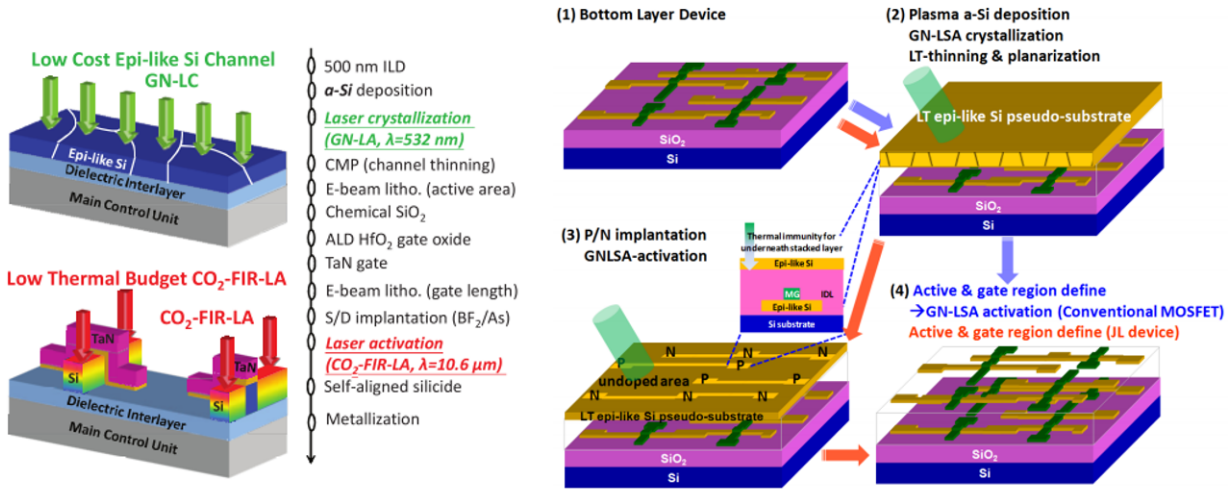


Figure 1.2.3.6 Process flow for 3D sequential integration using Poly-Si laser recrystallization [Shen 2013; T. T. Wu 2015].

Finally, the process can be done by **wafer bonding over an already processed tier**. In this approach, the top wafer has already high-quality silicon. Thanks to its high alignment precision the process can reach ultra-fine pitch connection between tiers. As a result, **ultra-high 3D contact densities** (10^8 via/mm2 for a 14nm node stacking) is achieved [Batude 2011]. 2-tier integration with intermediate back-end (iBEOL) and 3D contacts (3DCO) is shown in Figure 1.2.3.7.

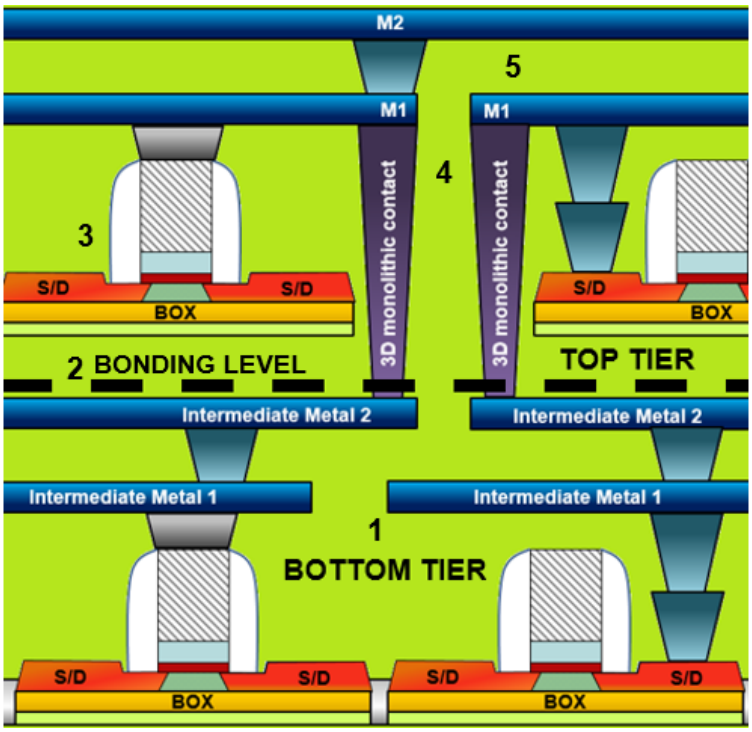


Figure 1.2.3.7 3D stack with sequential integration featuring homogenous process technologies, and intermediate back-end.

The challenge of such integration is to obtain a top level with similar bottom level performances; but with a limited thermal budget process. Depending on the considered technology and node, the maximal sustainable thermal budget by the bottom layer will differ. The bottom wafer is a standard process with routed metal layers (iBEOL). **The bottom devices have to resist to the top highest process temperatures**. For example, [Deprat 2016], [Fenouillet-Beranger 2014] show how a conventional 950°C annealing degrades the bottom tier performance. This restricts the **thermal budget** for the top tier, where all the front-end processes are limited to 500ºC during 2 hours for the 14nm FD-SOI technology, or more specifically **a time-temperature window** [Batude 2015]; hence in literature, this is referred as **low temperature (LT)**. The bottom silicide is enhanced to become stable for this range of temperatures [Deprat 2016]. Tungsten is used as **iBEOL** metal due to the contamination risks in case of wafer breaking during the top processes in front-end machines. The LT monolithic process currently achieves the high performance in 14nm thanks to Solid Phase Epitaxy Regrowth (SPER)[Batude 2015] while keeping low variability [Pasini 2016]. The 3DCO are done sequentially after the top LT process using standard tungsten plug in oxide, connecting iBEOL to the final BEOL, which is a standard Cu/Low-k back-end. This process integration has been demonstrated on 300 mm wafers with one metal line of Tungsten in iBEOL [Brunet 2016] and its process flow is illustrated in Figure 1.2.3.8.

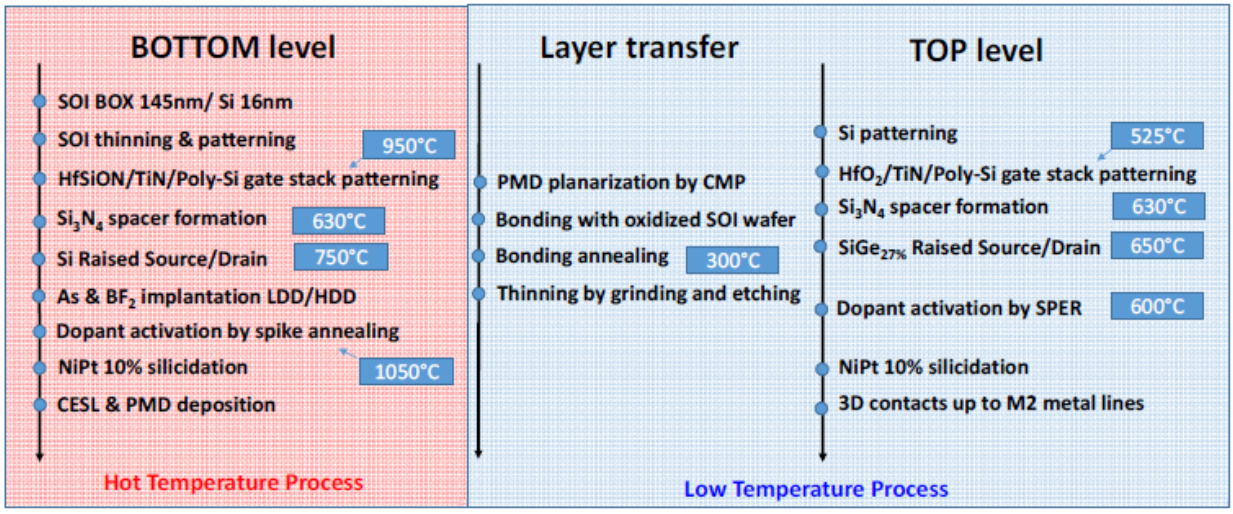


Figure 1.2.3.8 3D sequential design flow for CoolCube™ process using wafer bonding and maximum temperature for critical steps [Brunet 2016].

The laser annealing can help the low temperature process as well, even for the wafer bonding integration. As the laser can concentrate high power in a small area, the heat is enough to do the dopant activation instead of SPER process, but the heat diffusion in a very short time is not sufficient to degrade the tier already built. This is an active topic in research for low temperature process as seen in Figure 1.2.3.9. For certain laser energy; the light can activate the junction. However, as the top tier is done over the already processed circuit, the bottom circuit may also contain back-end metals for interconnections routing, which is called intermediate back-end (iBEOL). The iBEOL metals can reflect some of the laser energy, influencing the heating process, as the power density per area is critical. Hence, the iBEOL metal pattern can change the outcome of the laser annealing [Fenouillet-Beranger 2016].

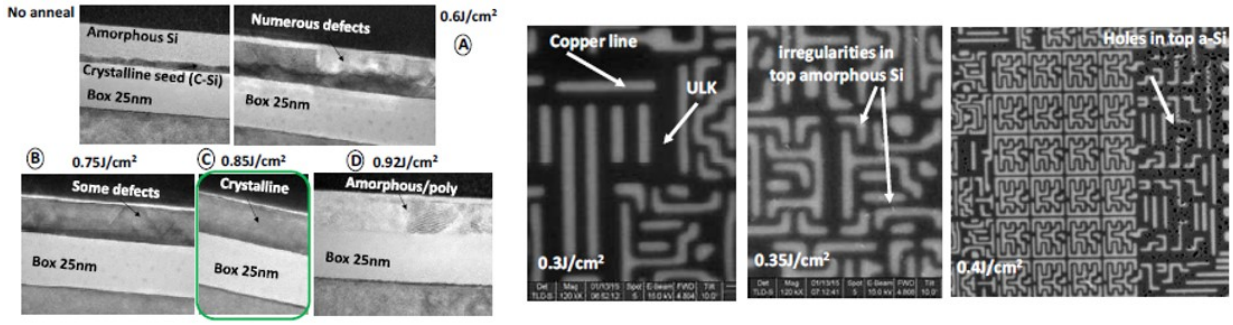


Figure 1.2.3.9 Laser annealing for junction activation for low temperature 3D process. [Fenouillet-Beranger 2016].

The 3D CoolCube™ sequential integration is illustrated in Figure 1.2.3.10, in this case without iBEOL. The 3DCO vias between tiers shows excellent alignment precision.

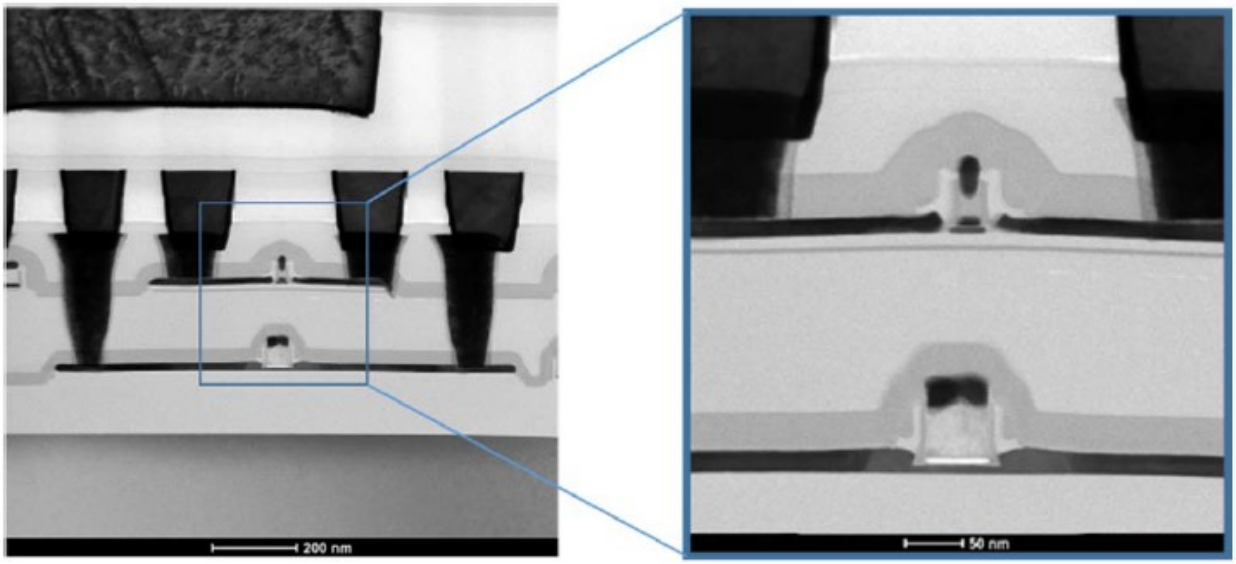


Figure 1.2.3.10 3D sequential integration using bonded SOI wafer. The 3DCO processed after the bonding yields a high alignment precision [Brunet 2016].

The ultimate research goal is to produce low temperature stacked transistors with the same performance of standard planar process without degrading the bottom transistors, as illustrated in I_{ON} vs I_{OFF} figures of merit in Figure 1.2.3.11; showing the bottom MOSFET thermal stability and the possibility of high performance transistor using low temperature process.

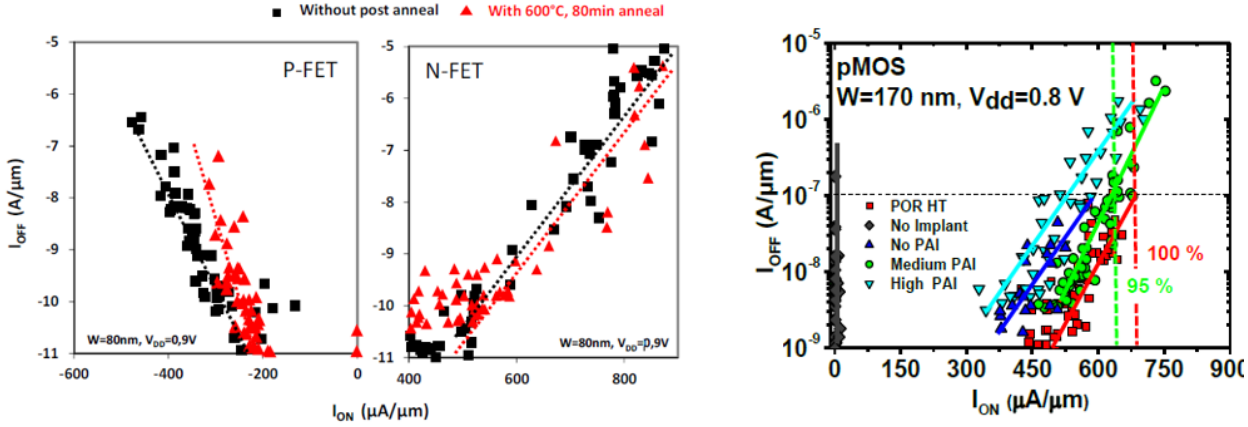


Figure 1.2.3.11 Typical I_{ON} vs I_{OFF}. On the left, (a) The bottom performance before the top process in black and after a thermal budget in red (Batude et al., 2014). On the right, low temperature process with different implant doses for the S/D compared to reference planar process in red [Pasini 2016].

The 3D process described uses the FDSOI as transistor architecture. Other processes have already been demonstrated to achieve 3D sequential integration featuring heterogeneous technologies [Shulaker 2014],[Irisawa 2014]. Indeed, the **3D monolithic process can feature different technologies in each tier in order to use the best potential of a given technology** for a determined application, as illustrated in Figure 1.2.3.12. Such integration, can use FinFETs for high performance logic, and FDSOI on the top tier for

RF communications. This solution is infeasible and cost prohibitive in a planar monolithic integration, otherwise different processes have to be managed in the same masks.

Figure 1.2.3.12 Homogenous 3D integration featuring different transistor architectures in each tier [Batude 2014].

1.3 Thesis Objectives

As the Moore's Scaling is approaching to materials physical limits, new solutions have been proposed: either continuing miniaturization and exploring new materials or doing more than scaling and adding more features to the circuit, such as integrating logic and sensors in the same monolithic circuit.

Evaluating a technology proposal among the crowd is the main goal of this thesis. Specifically, the 3D Sequential Integration (3DVLSI) is assessed in several design aspects, such as Performance/Power/Area (PPA) and Variability for digital circuits.

Most of the state of the art presented in Chapter One for 3D sequential circuits is based on advanced process. The automated design tools for 3DVLSI are under development. In this context, this thesis also provides guidelines for EDA development and process performance. The analyses are based on SPICE simulations.

In a succinct way, the work done consists in:

- 3D environment evaluation using SPICE and Full-Custom layouts
- 3D Contacts operation assessment in final circuit performance
- Area overhead and solutions for 3DVLSI
- Guidelines for EDA tools development and process performance guidance
- Assessment of variability – Global and Local variation impact in circuit figures of merit
- Across-Chip Variations (ACV) modeling and its effects in 3D sequential circuits

The thesis has been divided into two distinct parts, the first one is focused on circuit design while the second part scrutinize the process variability impact on design.

1.4 Chapter Conclusion

The miniaturization of circuits has been presented as the main engine of the semiconductor industry research and development. The scaling trend that effectively took place over the last 50 years is discussed as the Moore's Law and illustrated also as the roadmap of the industry.

The scaling reduced the transistor cost per node as suggest by Moore's paper. Coupled to miniaturization, the better transistor performance due to reduce dimensions was noted by Dennard in 1975. Over the years, the circuits became cheaper and more powerful.

The traditional scaling, reducing dimensions and thicknesses finished in the 2000's. Due to the current leakage and the need to increase device performance, new technical solutions were employed. Finally, in the 2010's the transistor architecture was complete revamped to continue the trend. As the scaling is approaching to atomic scales and becoming more difficult, new solutions like 3D integration were proposed. Stacking tiers can bring benefits to circuits considering the PPA and augment the function capabilities (such logic integration with sensors).

Semiconductor 3D integration can be classified as parallel and sequential integration. Both state of the art for the process has been illustrated in Chapter One. The parallel integration has the advantage to use a standard planar process like, having the wafers bonded after, with minor modifications to create the through silicon vias (TSV) to contact tiers. Due to the bonding step, the misalignment between wafers is a limiting factor and this issue is mitigated by a large TSV size in order to grant the contact. This creates a huge area overhead, limiting the connection density between the tiers. The problem can be completely avoided by stacking the tiers aligned, in a 3D sequential process. This enables an ultra-high 3D contact density among the tiers. The main disadvantage of such an approach is the top processes with a limited thermal budget, in order not to damage the already processed tier. The state of the art of low temperature process for 3D integration has been discussed in Chapter One. With ultra-high density 3DCO and high transistor performance for both tiers, the sequential 3DVSLI is shown as a perfect candidate for digital logic circuits to continue the scaling in a more than Moore's flavor.

REFERENCES

Barraud, S., R. Coquand, M. Casse, M. Koyama, J. M. Hartmann, V. Maffini-Alvaro, C. Comboroure, et al. 2012. "Performance of Omega-Shaped-Gate Silicon Nanowire MOSFET With Diameter Down to 8 Nm." *IEEE Electron Device Letters* 33 (11): 1526–28. doi:10.1109/LED.2012.2212691.

Batude, P., C. Fenouillet-Beranger, L. Pasini, V. Lu, F. Deprat, L. Brunet, B. Sklenard, et al. 2015. "3DVLSI with CoolCube Process: An Alternative Path to Scaling." In *2015 Symposium on VLSI Technology (VLSI Technology)*, T48–49. doi:10.1109/VLSIT.2015.7223698.

Batude, P., B. Sklenard, C. Fenouillet-Beranger, B. Previtali, C. Tabone, O. Rozeau, O. Billoint, et al. 2014. "3D Sequential Integration Opportunities and Technology Optimization." In *IEEE International Interconnect Technology Conference*, 373–76. doi:10.1109/IITC.2014.6831837.

Batude, P., M. Vinet, B. Previtali, C. Tabone, C. Xu, J. Mazurier, O. Weber, et al. 2011. "Advances, Challenges and Opportunities in 3D CMOS Sequential Integration." In *2011 International Electron Devices Meeting*, 7.3.1-7.3.4. doi:10.1109/IEDM.2011.6131506.

Batude, P., M. Vinet, C. Xu, B. Previtali, C. Tabone, C. Le Royer, L. Sanchez, et al. 2011. "Demonstration of Low Temperature 3D Sequential FDSOI Integration down to 50 Nm Gate Length." In *2011 Symposium on VLSI Technology - Digest of Technical Papers*, 158–59.

Brunet, L., P. Batude, C. Fenouillet-Beranger, P. Besombes, L. Hortemel, F. Ponthenier, B. Previtali, et al. 2016. "First Demonstration of a CMOS over CMOS 3D VLSI CoolCube #x2122; Integration on 300mm Wafers." In *2016 IEEE Symposium on VLSI Technology*, 1–2. doi:10.1109/VLSIT.2016.7573428.

Chen, J., T. Y. Chan, I. C. Chen, P. K. Ko, and C. Hu. 1987. "Subbreakdown Drain Leakage Current in MOSFET." *IEEE Electron Device Letters* 8 (11): 515–17. doi:10.1109/EDL.1987.26713.

Chen, K. N., and C. S. Tan. 2011. "Integration Schemes and Enabling Technologies for Three-Dimensional Integrated Circuits." *IET Computers Digital Techniques* 5 (3): 160–68. doi:10.1049/iet-cdt.2009.0127.

Cristoloveanu, S. 1999. "SOI: A Metamorphosis of Silicon." *IEEE Circuits and Devices Magazine* 15 (1): 26–32. doi:10.1109/101.747564.

Dennard, R. H., F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc. 1974. "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions." *IEEE Journal of Solid-State Circuits* 9 (5): 256–68. doi:10.1109/JSSC.1974.1050511.

Dennard, R. H., F. H. Gaensslen, Hwa-Nien Yu, V. L. Rideout, E. Bassous, and A. R. Leblanc. 1999. "Design Of Ion-Implanted MOSFET's with Very Small Physical Dimensions." *Proceedings of the IEEE* 87 (4): 668–78. doi:10.1109/JPROC.1999.752522.

Deprat, F., F. Nemouchi, C. Fenouillet-Beranger, M. Cassé, P. Rodriguez, B. Previtali, N. Rambal, et al. 2016. "First Integration of Ni0.9Co0.1 on pMOS Transistors." In *2016 IEEE International Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC)*, 133–35. doi:10.1109/IITC-AMC.2016.7507708.

Ernst, T., L. Duraffourg, C. Dupre, E. Bernard, P. Andreucci, S. Becu, E. Ollier, et al. 2008. "Novel Si-Based Nanowire Devices: Will They Serve Ultimate MOSFETs Scaling or Ultimate Hybrid Integration?" In *2008 IEEE International Electron Devices Meeting*, 1–4. doi:10.1109/IEDM.2008.4796804.

Fenouillet-Beranger, C., P. Acosta-Alba, B. Mathieu, S. Kerdilès, M. P. Samson, B. Previtali, N. Rambal, et al. 2016. "Ns Laser Annealing for Junction Activation Preserving Inter-Tier Interconnections Stability within a 3D Sequential Integration." In *2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 1–2. doi:10.1109/S3S.2016.7804375.

Fenouillet-Beranger, C., B. Previtali, P. Batude, F. Nemouchi, M. Cassé, X. Garros, L. Tosti, et al. 2014. "FDSOI Bottom MOSFETs Stability versus Top Transistor Thermal Budget Featuring 3D Monolithic Integration." In *2014 44th European Solid State Device Research Conference (ESSDERC)*, 110–13. doi:10.1109/ESSDERC.2014.6948770.

Fujii, H., K. Miyaji, K. Johguchi, K. Higuchi, C. Sun, and K. Takeuchi. 2012. "x11 Performance Increase, x6.9 Endurance Enhancement, 93% Energy Reduction of 3D TSV-Integrated Hybrid ReRAM/MLC NAND SSDs by Data Fragmentation Suppression." In *2012 Symposium on VLSI Circuits (VLSIC)*, 134–35. doi:10.1109/VLSIC.2012.6243826.

Gargini, P. A. 2017. "How to Successfully Overcome Inflection Points, or Long Live Moore's Law." *Computing in Science Engineering* 19 (2): 51–62. doi:10.1109/MCSE.2017.32.

Hsieh, A. C., and T. Hwang. 2012. "TSV Redundancy: Architecture and Design Issues in 3-D IC." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 20 (4): 711–22. doi:10.1109/TVLSI.2011.2107924.

Huang, Hsiang-Jen, Kun-Ming Chen, Tiao-Yuan Huang, Tien-Sheng Chao, Guo-Wei Huang, Chao-Hsin Chien, and Chun-Yen Chang. 2001. "Improved Low Temperature Characteristics of P-Channel MOSFETs with Si1-xGex Raised Source and Drain." *IEEE Transactions on Electron Devices* 48 (8): 1627–32. doi:10.1109/16.936576.

Huang, P. T., S. L. Wu, Y. C. Huang, L. C. Chou, T. C. Huang, T. H. Wang, Y. R. Lin, et al. 2014. "2.5D Heterogeneously Integrated Microsystem for High-Density Neural Sensing Applications." *IEEE Transactions on Biomedical Circuits and Systems* 8 (6): 810–23. doi:10.1109/TBCAS.2014.2385061.

Huynh-Bao, T., J. Ryckaert, Z. Tökei, A. Mercha, D. Verkest, A. V. Y. Thean, and P. Wambacq. 2017. "Statistical Timing Analysis Considering Device and Interconnect Variability for BEOL Requirements in the 5-Nm Node and Beyond." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* PP (99): 1–12. doi:10.1109/TVLSI.2017.2647853.

Irisawa, T., K. Ikeda, Y. Moriyama, M. Oda, E. Mieda, T. Maeda, and T. Tezuka. 2014. "Demonstration of Ultimate CMOS Based on 3D Stacked InGaAs-OI/SGOI Wire Channel MOSFETs with Independent Back Gate." In *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, 1–2. doi:10.1109/VLSIT.2014.6894395.

Knickerbocker, J. U., P. S. Andry, E. Colgan, B. Dang, T. Dickson, X. Gu, C. Haymes, et al. 2012. "2.5D and 3D Technology Challenges and Test Vehicle Demonstrations." In *2012 IEEE 62nd Electronic Components and Technology Conference*, 1068–76. doi:10.1109/ECTC.2012.6248968.

Lee, D. U., K. W. Kim, K. W. Kim, H. Kim, J. Y. Kim, Y. J. Park, J. H. Kim, et al. 2014. "25.2 A 1.2V 8Gb 8-Channel 128GB/S High-Bandwidth Memory (HBM) Stacked DRAM with Effective Microbump I/O Test Methods Using 29nm Process and TSV." In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 432–33. doi:10.1109/ISSCC.2014.6757501.

Lin, Y. H., S. Y. Huang, K. H. Tsai, W. T. Cheng, S. Sunter, Y. F. Chou, and D. M. Kwai. 2013. "Parametric Delay Test of Post-Bond Through-Silicon Vias in 3-D ICs via Variable Output Thresholding Analysis." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 32 (5): 737–47. doi:10.1109/TCAD.2012.2236837.

Liu, C., and S. K. Lim. 2012. "A Design Tradeoff Study with Monolithic 3D Integration." In *Thirteenth International Symposium on Quality Electronic Design (ISQED)*, 529–36. doi:10.1109/ISQED.2012.6187545.

Mistry, K., C. Allen, C. Auth, B. Beattie, D. Bergstrom, M. Bost, M. Brazier, et al. 2007. "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-Free Packaging." In *2007 IEEE International Electron Devices Meeting*, 247–50. doi:10.1109/IEDM.2007.4418914.

Moore, G. E. 1998. "Cramming More Components Onto Integrated Circuits." *Proceedings of the IEEE* 86 (1): 82–85. doi:10.1109/JPROC.1998.658762.

Pasini, L., P. Batude, J. Lacord, M. Casse, B. Mathieu, B. Sklenard, F. P. Luce, et al. 2016. "High Performance CMOS FDSOI Devices Activated at Low Temperature." In *2016 IEEE Symposium on VLSI Technology*, 1–2. doi:10.1109/VLSIT.2016.7573407.

Plas, G. Van der, P. Limaye, I. Loi, A. Mercha, H. Oprins, C. Torregiani, S. Thijs, et al. 2011. "Design Issues and Considerations for Low-Cost 3-D TSV IC Technology." *IEEE Journal of Solid-State Circuits* 46 (1): 293–307. doi:10.1109/JSSC.2010.2074070.

Rozeau, O., J. Lacord, S. Martinie, Anouar Idrissi-El Oudrhiri, M. A. Jaud, S. Barraud, T. Poiroux, M. Vinet, and J. C. Barbé. 2015. "Performance Benchmarking of Device Architectures for the Sub-10nm CMOS Technologies." In *The 5th International Workshop on Nanotechnology and Application (IWNA))*.

Schwarzenbach, W., X. Cauchy, F. Boedt, O. Bonnin, E. Butaud, C. Girard, B. Y. Nguyen, C. Mazure, and C. Maleville. 2011. "Excellent Silicon Thickness Uniformity on Ultra-Thin SOI for Controlling Vt Variation of FDSOI." In *2011 IEEE International Conference on IC Design Technology*, 1–3. doi:10.1109/ICICDT.2011.5783188.

Shen, C. H., J. M. Shieh, T. T. Wu, W. H. Huang, C. C. Yang, C. J. Wan, C. D. Lin, et al. 2013. "Monolithic 3D Chip Integrated with 500ns NVM, 3ps Logic Circuits and SRAM." In *2013 IEEE International Electron Devices Meeting*, 9.3.1-9.3.4. doi:10.1109/IEDM.2013.6724593.

Shulaker, M. M., T. F. Wu, A. Pal, L. Zhao, Y. Nishi, K. Saraswat, H. S. P. Wong, and S. Mitra. 2014. "Monolithic 3D Integration of Logic and Memory: Carbon Nanotube FETs, Resistive RAM, and Silicon FETs." In *2014 IEEE International Electron Devices Meeting*, 27.4.1-27.4.4. doi:10.1109/IEDM.2014.7047120.

Thompson, S. E., M. Armstrong, C. Auth, M. Alavi, M. Buehler, R. Chau, S. Cea, et al. 2004. "A 90-Nm Logic Technology Featuring Strained-Silicon." *IEEE Transactions on Electron Devices* 51 (11): 1790–97. doi:10.1109/TED.2004.836648.

Weber, O., F. Andrieu, J. Mazurier, M. Cassé, X. Garros, C. Leroux, F. Martin, et al. 2010. "Work-Function Engineering in Gate First Technology for Multi-VT Dual-Gate FDSOI CMOS on UTBOX." In *2010 International Electron Devices Meeting*, 3.4.1-3.4.4. doi:10.1109/IEDM.2010.5703289.

Wu, S. Y., C. Y. Lin, M. C. Chiang, J. J. Liaw, J. Y. Cheng, C. H. Chang, V. S. Chang, et al. 2016. "Demonstration of a Sub-0.03 um2 High Density 6-T SRAM with Scaled Bulk FinFETs for Mobile SOC Applications beyond 10nm Node." In *2016 IEEE Symposium on VLSI Technology*, 1–2. doi:10.1109/VLSIT.2016.7573390.

Wu, T. T., C. H. Shen, J. M. Shieh, W. H. Huang, H. H. Wang, F. K. Hsueh, H. C. Chen, et al. 2015. "Low-Cost and TSV-Free Monolithic 3D-IC with Heterogeneous Integration of Logic, Memory and Sensor Analogy Circuitry for Internet of Things." In *2015 IEEE International Electron Devices Meeting (IEDM)*, 25.4.1-25.4.4. doi:10.1109/IEDM.2015.7409765.

PART ONE: DESIGN

INTRODUCTION TO CHAPTER TWO

3D sequential integration, or 3DVLSI is the proposed alternative to the Moore's Law scaling. The stacking of transistors in several tiers can increase the circuit density for a given area, and optimizations in back-end interconnections can improve the circuit performance and reduce the power usage. This chapter explores the 3D design environment focusing ultra-dense logic circuits, and benchmark several aspects of the 3D monolithic circuits.

At the present time (2017); logic circuits such as Graphic Processor Units (GPUs) have more than 15 billion transistors. In order to achieve an architecture with so many elements, a well-established design methodology is used in planar circuits. In this chapter, those techniques are synthetized in a brief introduction. The 3DVLSI design tools are under development; but are mainly based in planar as an inheritance. The work has been done using a 3D Process Design Kit (PDK), constantly having upgrades as the technology is under evolution. The technology hypotheses and their implications in the design are discussed completing the chapter introduction.

Design-wisely, the high density of 3D Contacts (3DCO) through the tiers allows several integration implementations. Solutions such as CMOS over CMOS and Transistor over Transistors are discussed and compared. This work is largely based on the design of full custom circuits, as placement and routing design tools are not commercially available.

Chapter Two – Transistor Level 3D Design

2.1 VLSI Digital Design Flow

2.1.1 Overview in Planar Design Flow

The automated design flow in Very-Large Scale Integration (VLSI) is composed by various stages. As a natural evolution during the years, the abstraction level has increased in order to manage complex circuits having billions of transistors.

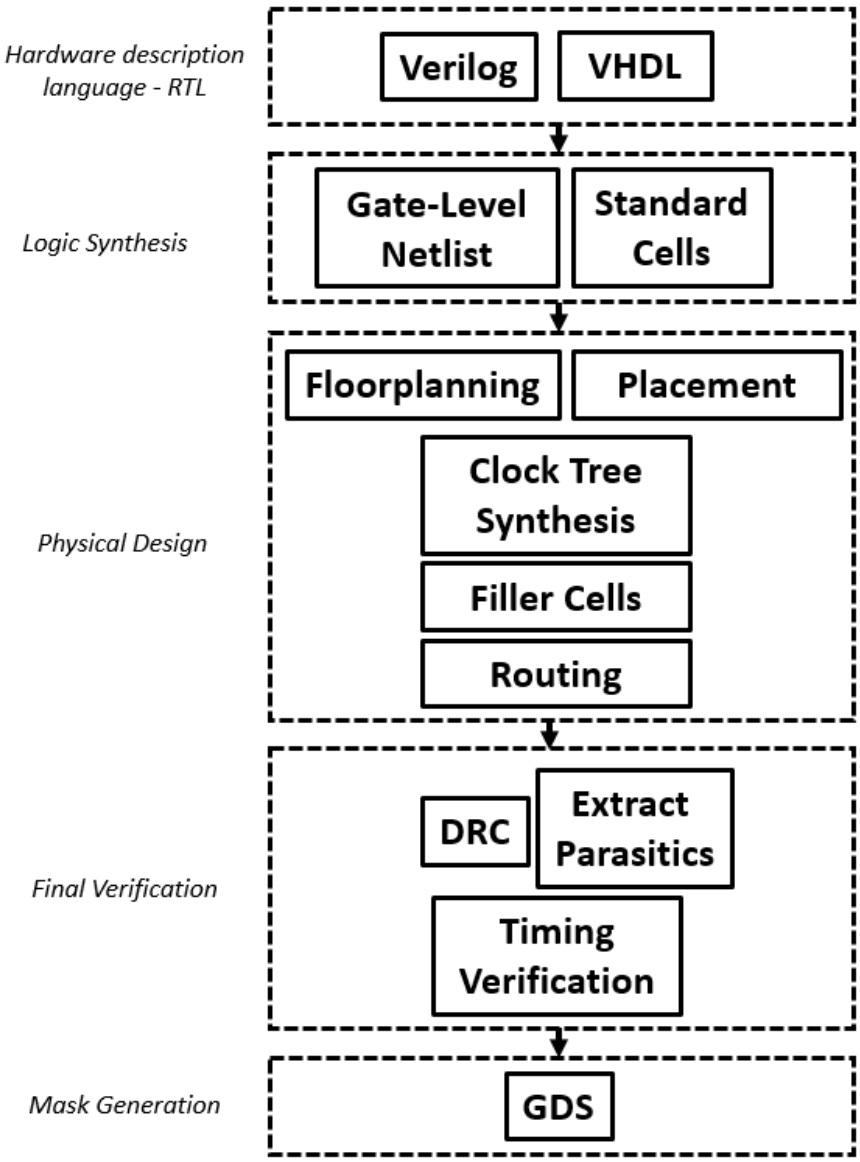


Figure 2.1.1.1 Usual digital design flow methodology in advanced planar nodes.

The big picture of the digital design flow, divided into five macroscopic steps, is shown in Figure 2.1.1.1. The circuit design is done by coding at high abstraction level, such as Register-Transfer Level (RTL) using VHDL or Verilog languages. The combinational and sequential logic are described by the code and later are synthetized, entering in the second design step: the logic synthesis. The synthesis tool transforms the RTL circuit description into a gate-level netlist. This tool optimizes the logic implementation and the timing

analysis tool checks the fulfillment of the signal timing assumptions (at least for synchronous circuits). The standard cells are made of several views including symbols, HDL code, electrical schematics and physical drawings of the logic gates (this is not an exhaustive list), and are extracted from a library to be associated to the netlist cells. The third step is defined by the physical design; the tools firstly allow the definition of a floorplan, then place and optimize the gate position of the netlist. Once the gates are placed, the clock tree synthesis can start. The clock tree synthesis tool is able to modify the gate placement in order to guaranty the clock design constraints in the circuit, ensuring an appropriate drive for the flip-flops and a minimal skew. The unused space between standard cells is completed by filler cells in order to guaranty performance and reliability. The physical design step is finalized by the routing tool, which connects together the standard cells with back-end metal rails and vias forming the interconnection network. Usually, the routing algorithm targets the shortest possible wirelength. Final verifications are done to check the possible design errors (Design rules for manufacturing) and if the circuit design works as expected. Then the process masks can be fabricated for the silicon production.

2.1.2 3D Design Flow

Today, 3D Design Flows based on the classical planar flow, with some modifications able to accommodate the multi-tier physical design.

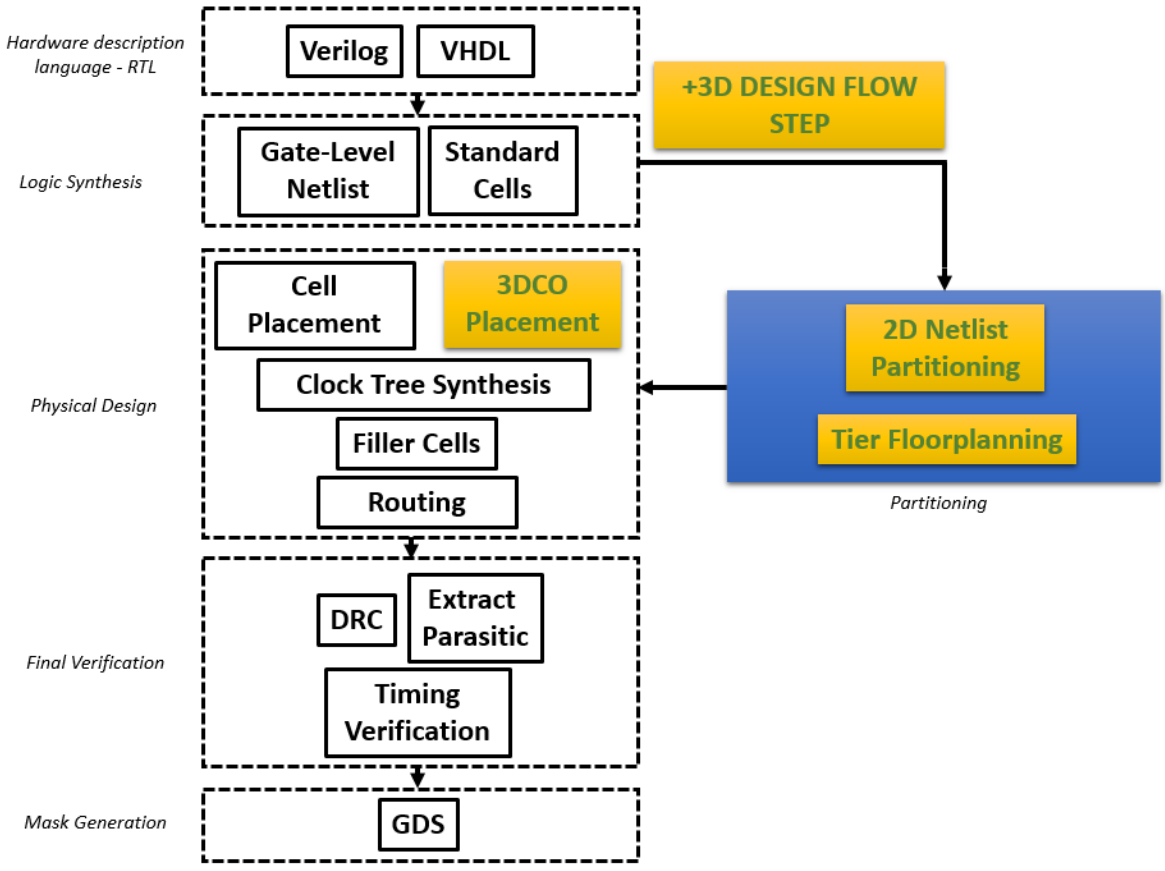


Figure 2.1.2.1 Modified digital design flow methodology for 3DVLSI integration.

An additional design step of partitioning is introduced as shown in Figure 2.1.2.1. The 2D netlist is divided into tier using a given separation strategy. Each tier has its own floorplan, and then the 3DCO placement is an additional step compared to the standard planar physical design. A variation in the 3D flow, could be the introduction of a parameter to explicit the circuit tier in the RTL description or in the gate-level netlist.

2.1.3 Design Flow with EDA

The state of the art of 3DVLSI electronic design automation tools (EDA) is presented in this section. As 3D sequential integration focuses very large digital circuits, with billions of transistors, the automated design tools are a necessity and the interest is quickly increasing in this topic as the number of publications. The prominent works focus in the netlist partitioning in order to create a full 3DVLSI EDA environment.

2.1.3.1 Netlist Partitioning

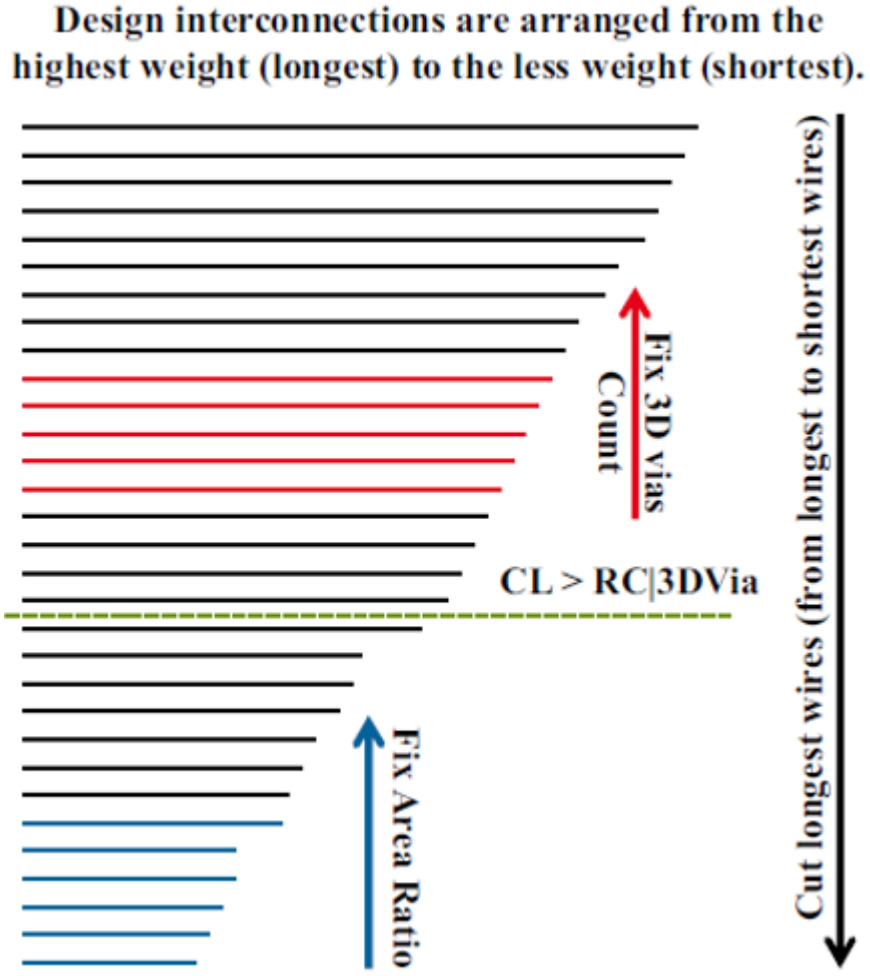


Figure 2.1.3.1 Wirelength evaluation after a 2D placement. Long wires are chosen for partitioning after a set threshold. [Sarhan 2015]

Partitioning is peculiarity of 3D design, which means that the netlist has to be separated into tiers. Several techniques to partition netlist and, later, floor-planning have been published [Panth 2015; Sarhan 2015; Sawicki 2009]. For example, one of those methods proposes to evaluate the wirelength after a 2D gate placement, and then interconnects above a certain threshold are chosen to be cut, meaning that the gates

will be placed into different tiers as illustrated in Figure 2.1.3.2. The length cutoff threshold can be defined by evaluating the maximum number of 3DCO needed. On the other hand, some interconnections can be chosen to stay in only one tier, for pure optimization purposes or in order to balance the tier filling ratio. Notably the **partitioning is able to increase circuit performance by lowering WL** and consequently reducing the interconnection parasitic elements. **In a design focused on power, it can reduce the number of buffers and repeaters because of the reduction of critical paths**.

2.1.3.2 3D IP Blocks

The area ratio between tiers does not necessary needs to be equal, for example, in a two-tier integration the top tier can have 60% of the gates while the bottom tier only have 40% as first proposed by [Sarhan 2015].

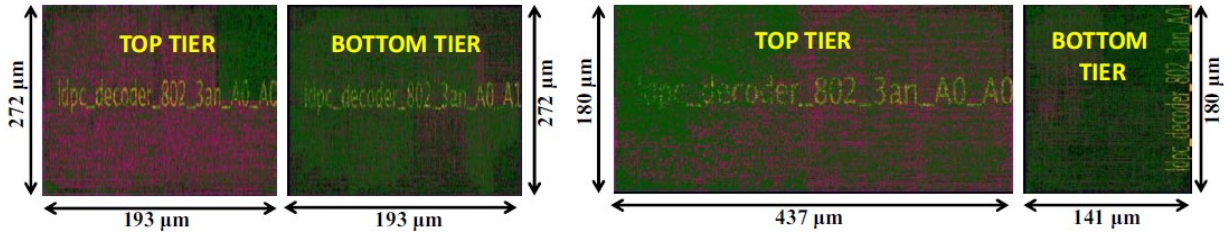


Figure 2.1.3.2 Layout prototype implementation snapshots of 3D LDPC block in two cases. On the left 50-50 partitioning ratio, and on the right, 75-25 partitioning ratio.[Sarhan 2015]

The wirelength optimization can be even further improved by using different partitioning ratios, the, potentially increasing the gains in performance and power. Depending on the design directives, the benefits can outstand the area loss due to unbalanced partitioning. In a complete design, the blank area of the tier with less circuit may be used by secondary elements, such as decoupling capacitors; another possibility is to employ a planar block to fill the gap. The unbalanced partitioning study shows performance and power gains in several circuits as illustrated in Figure 2.1.3.3. The unbalanced placement and routing is compared to the 2D planar case, and another partitioning tool that splits the circuit in a 50-50 balanced area ratio. For some circuit blocks, such as FFT and openMSP, **the unbalanced partitioning can increase the performance even further than a balanced one**. In such circuits, the long interconnections can be optimized, as well the short interconnections which stay on the same tier. In these initial proposed methods, the routing is based on the 2D routing, and the gains shown for 3DVLSI have to be considered as a first step. Indeed, with further works in the 3D EDA, a true 3D commercial tool will improve partitioning, floor-planning, placement, PDN and routing. This could unleash the total potential of 3DVLSI for very dense and complex logic circuits. Also, the cost of a 3DCO interconnection was thought to be penalizing, thus several proposed methods have a feature to limit the 3D vias. In this thesis, we do not need to limit the 3D vias number because the proposed 3DCO is optimized enough to preserve the signal integrity for analog effects such STR Charlie effect [Fesquet 2014], and high speed digital logic, and this is discussed in the section 2.4.

		No. Standard Cells	Area (µm²)		TWL (µm)	No. M3D-VIA	Max. Perf (GHz)	Perf. Gain (%)	Power @2D Max-Perf. (mW)	Power Gain (%)
			Top	Bottom						
openMSP	2D	6122	11390		91278	**NA**	**1.21**	**NA**	**11.54**	**NA**
	hMetis [14]	5456	6000	6000	89642	**1110**	**1.24**	**2.5%**	**11.53**	**0.1%**
	PAP (31-69)	5466	7875	3530	89630	**3775**	**1.42**	**17.5%**	**11.74**	**-1.7%**
Reconf-FFT	2D	29673	27600		420547	**NA**	**1.33**	**NA**	**52.50**	**NA**
	hMetis [14]	20093	13853	13853	316490	**1158**	**1.45**	**9.0%**	**30.87**	**41%**
	PAP (40-60)	20604	16656	10902	321269	**13894**	**1.64**	**24.0%**	**28.75**	**45%**
LDPC	2D	68179	88800		3277363	**NA**	**1.58**	**NA**	**103.5**	**NA**
	hMetis [14]	56896	52496	52496	1431432	**12511**	**1.74**	**9.9%**	**98.3**	**5.2%**
	PAP (25-75)	56896	78660	25380	965889	**26917**	**1.76**	**11.5%**	**100.7**	**3.0%**
DES-3	2D	106989	176400		1588509	**NA**	**2.15**	**NA**	**185.4**	**NA**
	hMetis [14]	61386	57000	57000	428599	**5439**	**2.35**	**9.2%**	**160.9**	**13.2%**
	PAP (50-50)	61386	57000	57000	357048	**9518**	**2.41**	**12.0%**	**158.2**	**15.0%**

Figure 2.1.3.3 Several blocks PPA comparison using the unbalanced 3D partition for two tiers. Due to reduced wirelength and number of gates, the 3DVLSI can gain in performance of power. [Sarhan 2015]. OpenMSP is a 16-bit microcontroller, FFT is a Fast-Fourier Transform block, LDPC stands for Low-density parity-check code, and DES-3 is a Data Encryption Standard block.

2.1.4 Conclusion and positioning

Advances in the EDA tools are illustrated in this section. The software developments have an inclination to reuse most of the planar design flow in first stance. While this approach may not bring the optimal results, it can deliver circuit PPA gains without major tool development. However, many questions were open during the EDA development, for the example the number of 3DCO in the overall design, as well as the performance impact brought by them. Also, the floorplanning strategy, the 3DCO placement and 3D Power Delivery Network are not enough mature points today. Following those open questions, the next section is organized in a bottom-up approach and discussed the design at transistor level, in order to provide guidelines to EDA tools.

2.2 Bottom-Up Approach for the digital design flow

2.2.1 Full Custom Standard Cell

2.2.1.1 Introduction to CoolCube™

Stacking transistors can have different process definitions. In this thesis, the CEA-LETI 3D sequential process called CoolCube™ is used as reference process. The technical details of this process are depicted in Chapter One. The CoolCube™ has been developed with 3D VLSI logic circuits in mind. A presence of an intermediate back-end level (iBEOL) is necessary to increase the routability for dense circuits. The great advantage of this process is the high-alignment between tiers enabling 3D contact vias between tiers (3DCO) with a small size and pitch, as illustrated in Figure 2.2.1.1 for 14nm BEOL rules. The design environment PDK for 3DVLSI CoolCube™ is described in Appendix A.

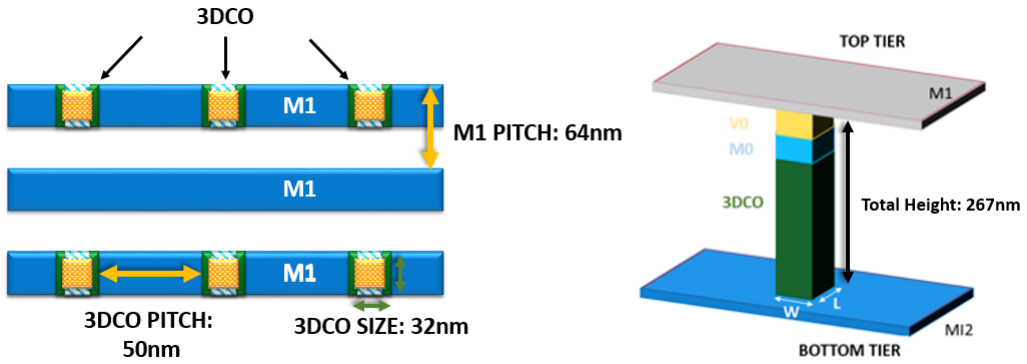


Figure 2.2.1.1 3DVLSI design rules for 14nm integration. On the left, top view from M1 showing the 3DCO dimensions. On the right, concept view of a 3DCO connecting bottom BEOL to the top BEOL.

2.2.1.2 3D Tier and iBEOL

A 3D integration can have many layers of stacked transistor as the designer needs, as the process should be able to build low temperature (LT) transistors over LT transistors. Each layer of transistors is called tier. In the studied 3D integration, each tier has an intermediate back-end. The number of iBEOL layers in each tier is an input from designer, and mainly represents the tradeoff between routability and the cost. An example is shown in Figure 2.2.1.2, displaying an integration with two tiers, one metal layer routing for each tier and a 3DCO connecting the tiers.

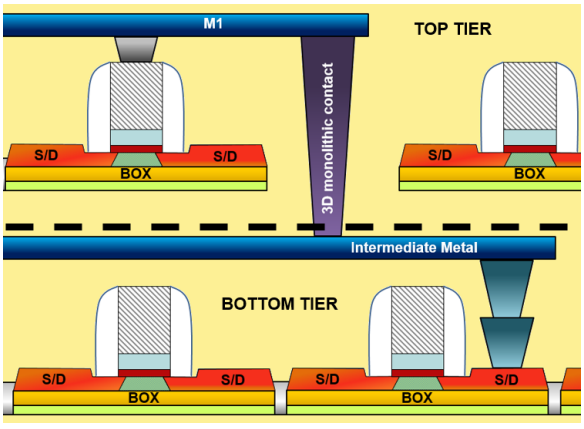


Figure 2.2.1.2 3DVLSI displaying two tiers, iBEOL and a 3DCO. [Ayres 2016]

2.2.1.3 Transistor over Transistor Integration

The transistor over transistor integration targets the **use of unipolar transistors in the same tier**. This means that one tier only has PMOS transistors, while other tier has the NMOS transistors. This opens up a window of possibilities in the process development. In advanced technologies, the PMOS process may differ from the NMOS process, such as in 14nm FDSOI use of SiGe only in the PMOS channel in order to increase the mobility due the increased mechanical stress [Weber 2014]. Thus, one tier can have an optimal wafer substrate in order to explore the strain for unipolar transistors.

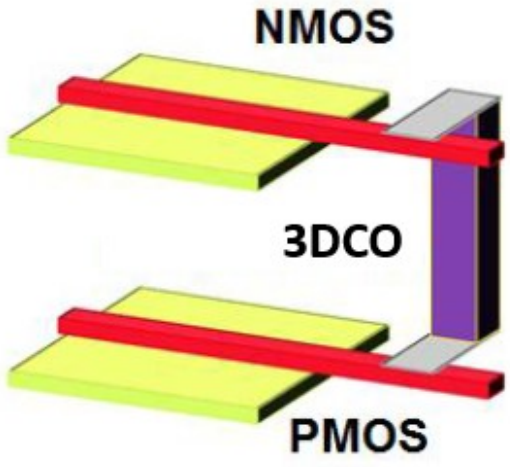


Figure 2.2.1.3 Transistor over Transistor integration. The tier is unipolar (only one type of MOSFET) and the 3D contact is used to form the CMOS logic gate. [Ayres 2015]

In a design perspective, high density local connections grants fine granularity as well as the possibility of stacking transistors over transistors. However due to density penalty caused by the excess of interconnections [Panth 2014], this work is focused on a CMOS over CMOS approach.

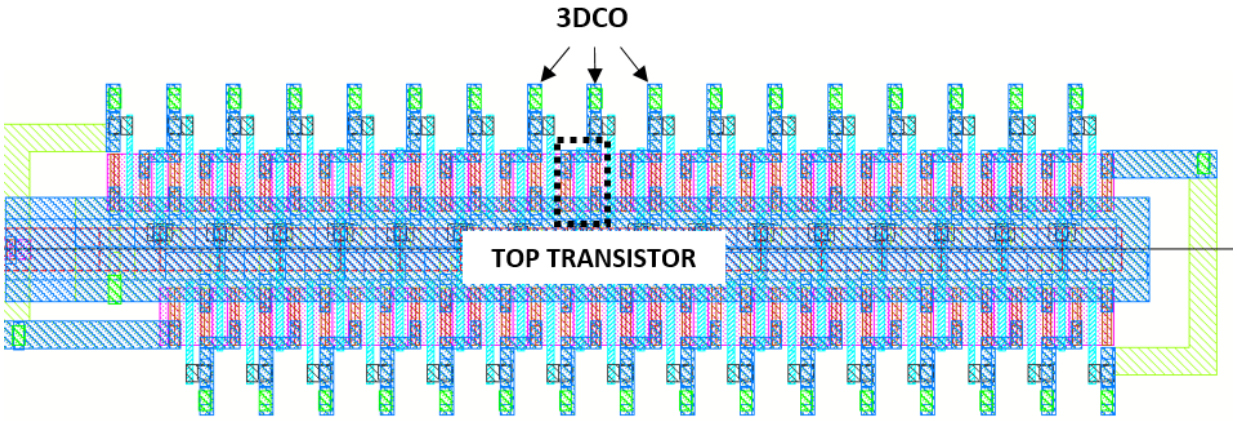


Figure 2.2.1.4 3DVLSI Ring Oscillator partitioned into two tiers using the transistor over transistor integration. Each inverter gate requires one 3D contact, illustrated in green.

Figure 2.2.1.5 Routing obstruction caused by excess of 3DCO vias in yellow. A connection from the bottom transistor up to M2 in top tier needs vias in all layers between M2 and bottom transistor contact. [Billoint 2015]

A ring oscillator layout was developed using the transistor over transistor configuration in the 3D sequential design environment as illustrated in Figure 2.2.1.4. The top tier has only PMOS transistors and bottom tier only NMOS transistors. **Each inverter uses one 3DCO to create the CMOS integration**. Due to the minimum clearance of the 3DCO from the active region, the 3DCOs (green squares) are positioned away from the gate active region, confirming the **density penalty in this type of integration for logic circuits.** Another problem that this integration may face is the congestion caused by the contacts from the bottom transistor to the top transistor. The Figure 2.2.1.5 shows the via wall created by the excess of 3D contacts in one direction, possibly causing **congestion problems during the interconnection routing**.

2.2.1.4 CMOS over CMOS Integration

The CMOS over CMOS integration, also known as 3D gate-level integration, uses both PMOS and NMOS transistors in the tiers. This integration uses less 3DCO compared to transistor over transistor, because the CMOS gates can be formed directly into a given tier, similar to the planar integration. Two inverter gates connect through a 3DCO is illustrated in Figure 2.2.1.6. **A huge advantage of this integration style, is that planar standard cell can be imported straightforwardly to the 3D environment**, with minor modifications. The reuse of 2D cells is a cost-effective approach for industrials.

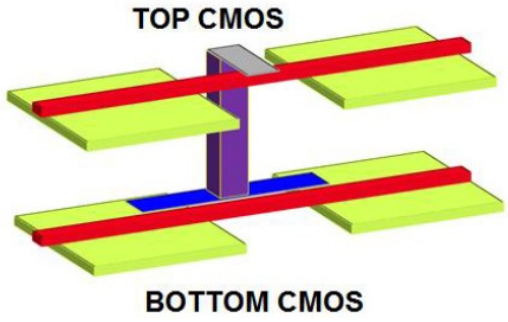


Figure 2.2.1.6 CMOS over CMOS integration. An example of two inverter gates in different tiers connected by a 3DCO.

Another full custom ring oscillator was done, at this time using the CMOS over CMOS integration as illustrated in Figure 2.2.1.7. The inverters are positioned as in planar circuits for the same tier. The bottom inverters are not seen in this figure, because they are placed under the top inverters (bottom tier). The ring oscillator is partitioned into two tiers, and **only two 3DCOs are needed to transpose the ring signal between the tiers**, in order to reduce the 3DCO overhead.

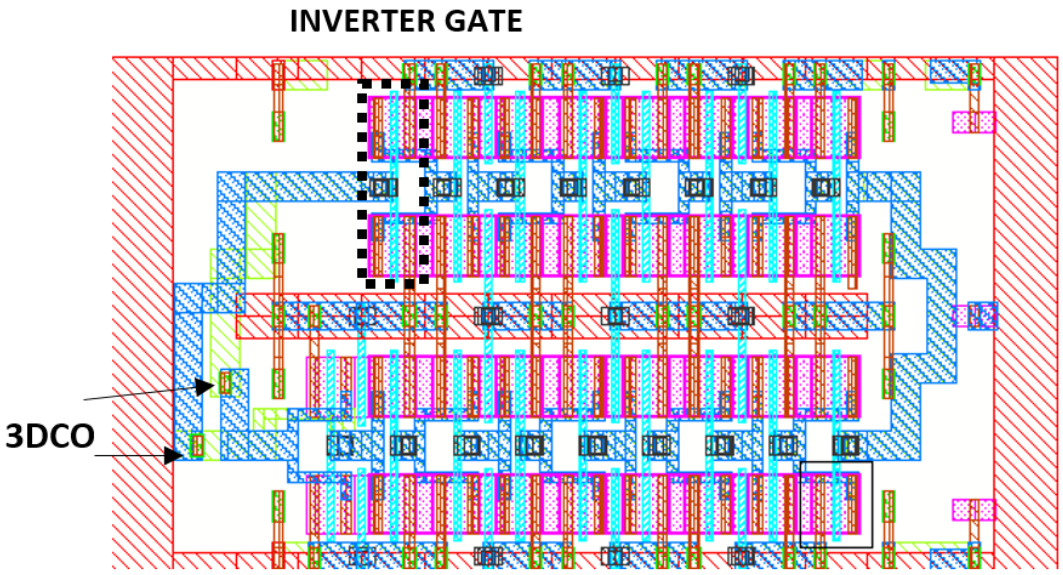


Figure 2.2.1.7 3DVLSI Ring Oscillator layout partitioned into two tiers. The layout uses the CMOS over CMOS style. Only two 3DCO are required to connect the ring structure signal path.

This full custom layout gives a clear view of the positioning of the 3DCOs. The placement of 3DCOs between two inverters, inside the P and N active zone is possible with very aggressive layout design rules. This translates in a tight spacing between 3DCO and poly-gates or active regions. Moreover, only one 3DCO may be placed in this region due to limited area, being not able to meet double vias standard for reliable designs. Another problem due the rule aggressiveness is the occurrence of short circuits. In order to take into account these problems, the **proposed optimal 3DCO placement** is advised in Figure 2.2.1.8. **The 3DCOs are grouped in the size of a standard cell and placed next to the active regions**. In this case, the continuous active region is kept, and the strain optimization remains unchanged**. This guideline also enables a specific and better control of the 3DCO standard cell, allowing high-density DRC rules, as usually done for SRAM cells**. For other technologies such as FinFET or Nanowires transistors, this guideline may be the only possible solution. In the context of back-end EDA tools, it is easy to take advantage of such an approach because the **3DCO placement directives are similar to those of the filler cells. The 3DCO standard cells have just to be placed before the filler cells.** The number of 3D monolithic vias inside a standard cell is defined by the 3DCO pitch. In the presented example, for 14nm design rules at least eight 3DCO can be placed inside the minimal size standard cell, evidencing the high 3D contact density enabled by the sequential integration. The analysis of the suggested integration area gain is done in Figure 2.2.1.9. The worst case 3DCO placement is considered, where only one inter-via per standard filler cell. The area gain is calculated considering FO1 and FO4 inverter standard cell area. Considering the worst-case scenario, the **3DVLSI can double the circuit density (50% area gain)** for approximately 100 FO1 gates per 3DCO via, or 25 FO4 gates per 3DCO vias.

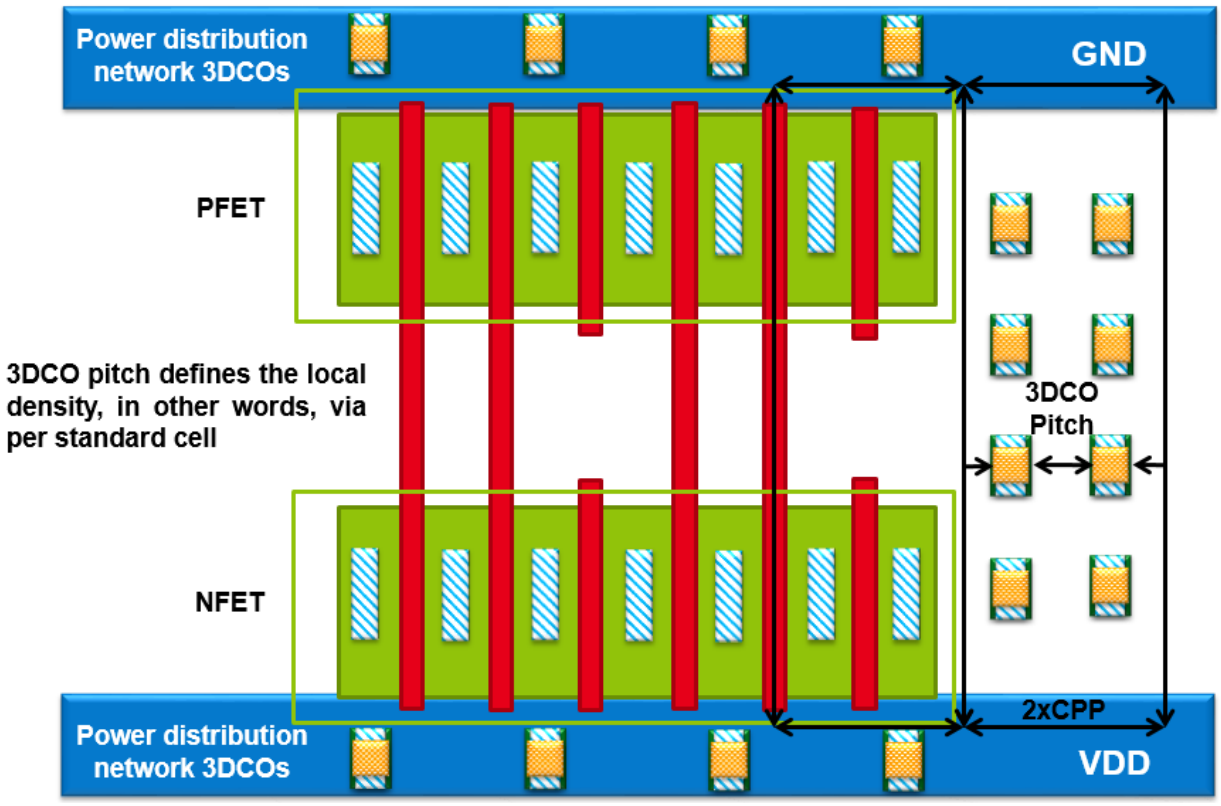


Figure 2.2.1.8 Guidelines on 3DCO placement for 3DVLSI circuits. The 3DCO cannot go through the active region (green). The proposed method treats the 3D interconnections as standard cells.

This illustrates how easily the 3DVLSI CMOS over CMOS can gain in density, even considering the additional overhead due to inter-tie vias.

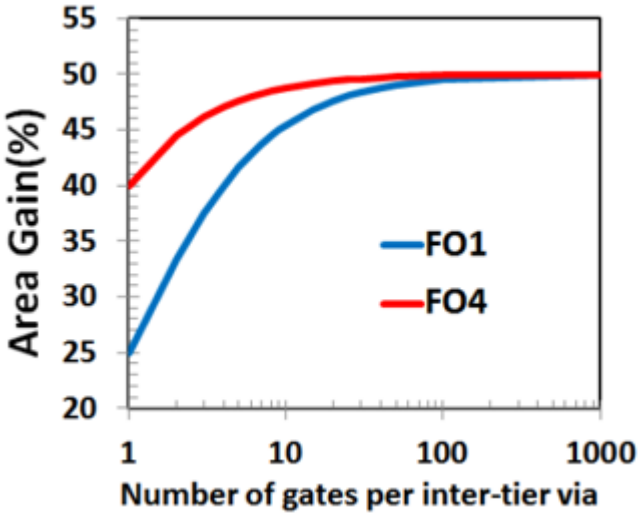


Figure 2.2.1.9 Theoretical area gain using CMOS over CMOS integration regarding the number of gates per 3DCO.

A Self-Timed Ring layout have been done with 64 C-elements gates for a planar integration and 3DVLSI, in which 32 gates are in top and bottom tier. The final layout including the isolation gates has about **1200 transistors**. The area comparison between planar and 3D STR is illustrated in Figure 2.2.1.10. **For bigger circuits the 3D area overhead becomes negligible, and the x0.5 area scaling is possible using 3DVLSI**.

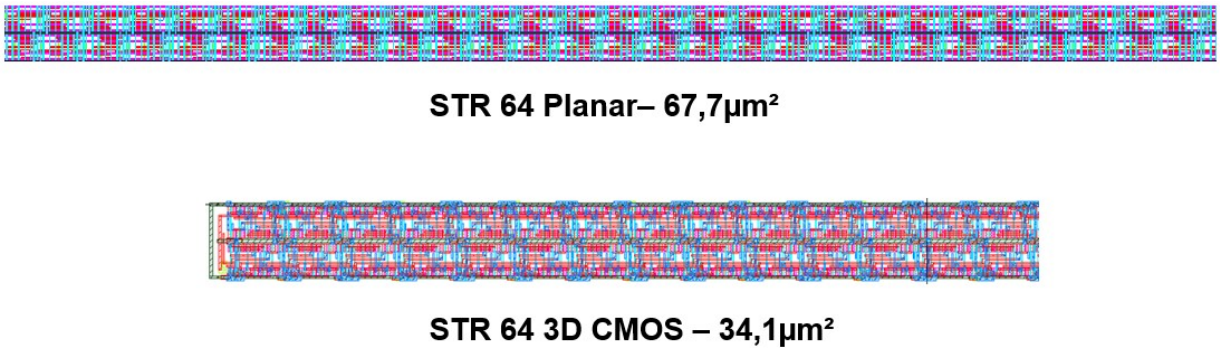


Figure 2.2.1.10 STR64 area comparison between planar and 3D sequential integration.

2.2.1.5 Conclusion

In this section, the CoolCube™ bottom-up approach layout was developed to assess the integration granularity opportunities as well the area gain proportioned by 3DVLSI. The transistor over transistor integration is shown as a possibility, however due to high 3D area overhead the CMOS over CMOS integration is favored, enabling flawlessly 0.5x area increase for circuits using 25 gates per 3DCO in 14nm design rules.

2.3 3D Design Environment

2.3.1 MOSFET Performance and SPICE Models

The SPICE model used in the simulation is LETI-UTSOI2 model [Poiroux 2013, 2015a, 2015b]. All model-cards included in our PDK are based on the performance of the 14nm FDSOI CMOS technology from STMicroelectronics [Weber 2014]. As consequence, the 3DVLSI sequential transistor SPICE models are equivalent on the planar ones. **The parity hypothesis has been taken, meaning that the bottom tier transistor matches the top tier transistor performance**; and both are similar to the planar model. The state of the art process confirms that this hypothesis is valid, and the low temperature process matches the standard planar process [Batude 2015].

2.3.2 Simulation Results

Figure 2.3.2.1a illustrates the model calibration for PMOS and NMOS transistors. A SPICE simulation for NMOS and PMOS I_{DSAT} is illustrated in Figure 2.3.2.1b, showing that the SPICE model is calibrate according the 14nm FDSOI data. The nominal supply voltage for this technology is 0.8 volts.

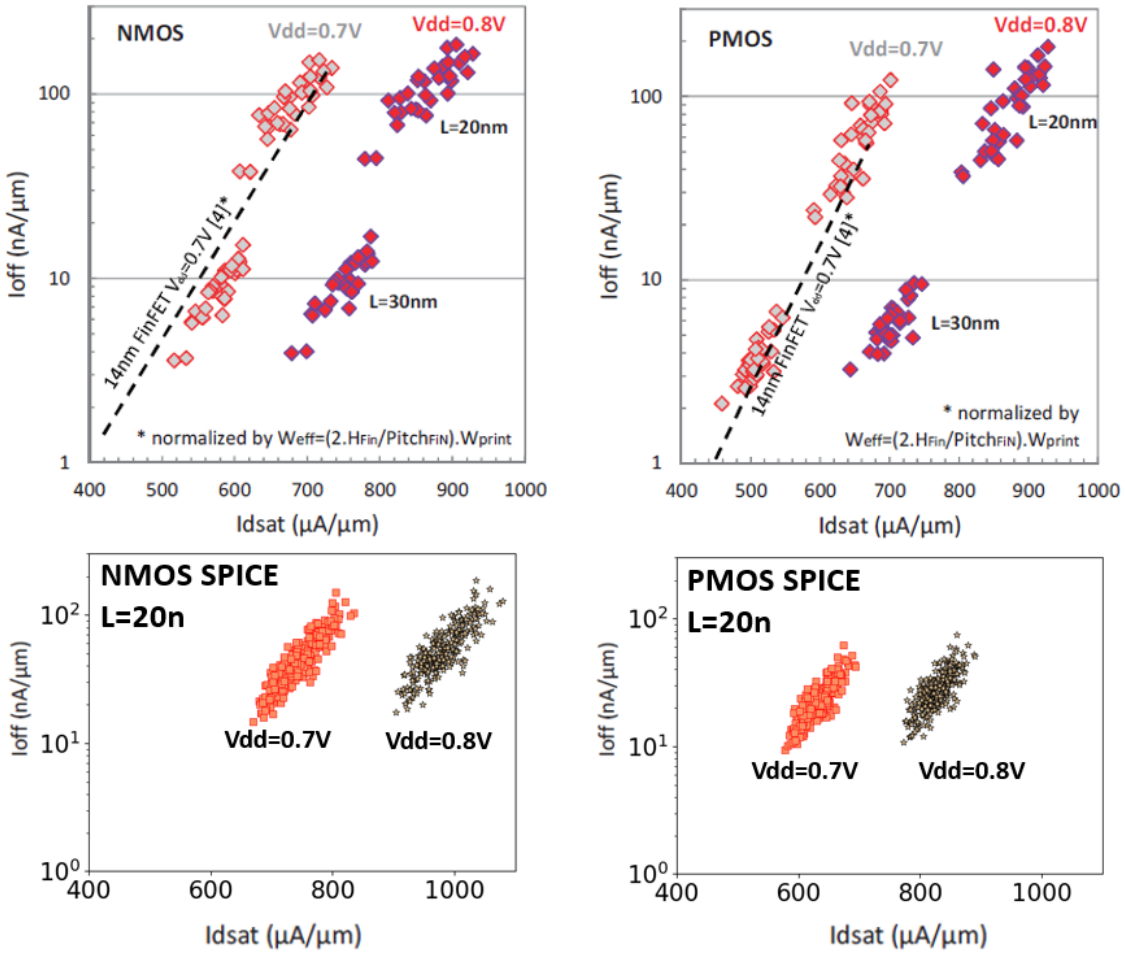


Figure 2.3.2.1 On top, (a) Silicon data for 14nm planar technology [Weber 2015]. On bottom, (b) SPICE model simulations for 14nm using 0.7V and 0.8V as supply voltage.

2.3.3 Parasitic Elements Extractions

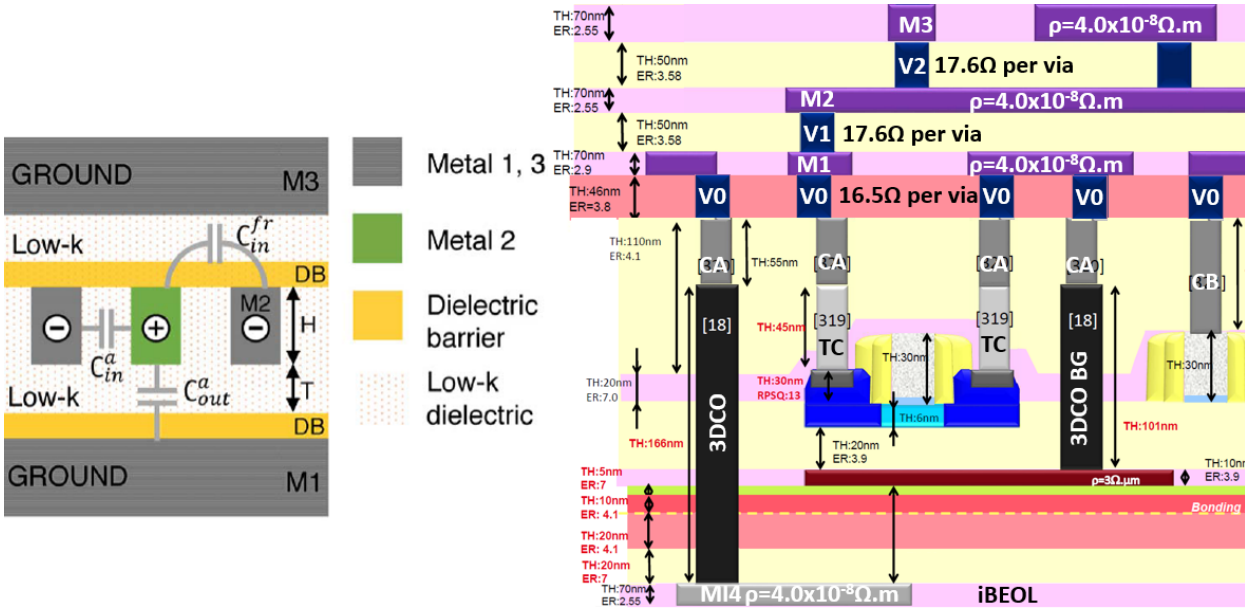


Figure 2.3.3.1 On the left, (a) Back-end cross section, showing capacitances between metal lines. [Huynh-Bao 2017]. On the right, (b) DRM showing upper level tier, with MI4 iBEOL level, top tier FEOL and BEOL.

The layout parasitic extraction **(PEX) is the calculation of parasitic effects from device interconnection, such as the resistance, capacitance** as illustrated in Figure 2.3.3.1a. Those elements affect the circuit timing performance, signal integrity and power consumption. In this chapter the interconnections are assumed to be formed from Cu and low-k dielectrics for BEOL and iBEOL. The 3DCO vias are made of tungsten due to the high aspect ratio. The PEX configuration is based in the technology description, as show in Figure 2.3.3.1b. In Chapter Three, the PEX will be modified assuming different iBEOL characteristics.

2.3.4 Conclusion

The simulation results obtained in this thesis are done using a PDK of a 14nm FDSOI technology including process assumptions for CoolCube™ integration. Indeed, as the 3D sequential process has not achieved a production maturity, the SPICE models are based on planar 14nm FDSOI CMOS technology. Some publications already demonstrate the top LT process matching the standard planar performance. If a different BEOL process solution is used, this will impact the circuit performance and will be discussed in the section 2.5. In Chapter Three, advanced nodes preliminary compact models were used, such as FinFETs and Nanowires transistor architectures.

2.4 Electrical Design Characterization

2.4.1 Full Custom

Full custom circuits were done in the 3D sequential PDK, complying with DRM. The benchmarked circuits are Ring Oscillators, Self-Timed Rings and Full Adders. Those circuits are simple enough to be drawn without EDA synthesis tools, nevertheless they are a representative benchmark for digital design.

2.4.1.1 Ring Oscillators

The inverter gate is the basic brick of the ring oscillator. The inverter is composed by two transistors, a NMOS and a PMOS connected in series as illustrated in Figure 2.4.1.1a.

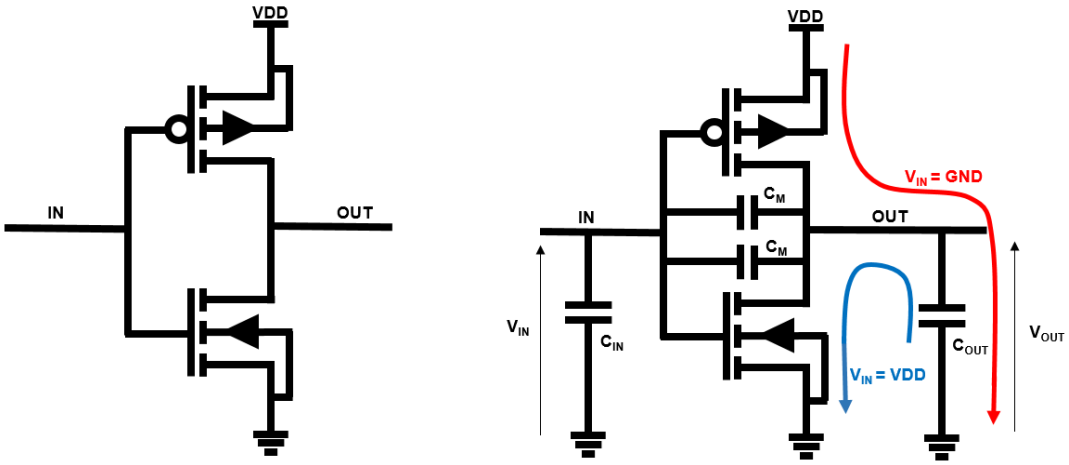


Figure 2.4.1.1 On the left, (a) an inverter gate using a NMOS and PMOS with the bulk terminal connected. On the right, (b) the inverter gate represented with some parasitic capacitances, and the current flow for different input voltages.

The intended operation of this gate is as its name suggests: to invert the input signal, operating in a digital form. In other words, the expected input is always between the supply voltage range, with the lowest voltage, usually the ground or zero volt representing the logical state 0, while the highest voltage, usually the circuit nominal voltage VDD representing the logical state 1. The inverter gate operation is represented in Figure 2.4.1.1b. The input signal drives the **NMOS and PMOS gates which are connected in parallel**. This means that one transistor is opened and while the other one is conducting. For example, if the inverter input voltage V_{IN} is set at the logical state 1, the PMOS will not be conducting. The NMOS will be in saturation regime, effectively connecting the output to the ground as shown in the blue current path. Thus, **the output will be at the logical state 0, the inverted state of the logical input (at 1)**. On the other hand, if the input voltage is at the logical state 0, the PMOS will conduct, as illustrated by the red current path, and the NMOS will be opened. This provides the output connection to the supply voltage, granting the logical state 1 at the output. **The transition speed between the logical states depends on the MOSFET subthreshold swing, the effective current drive I_D, and the parasitic elements**. For example, the capacitances will limit the maximal voltage variation over the time for a given current, and the resistances will limit the effective current, both slowing down the gate.

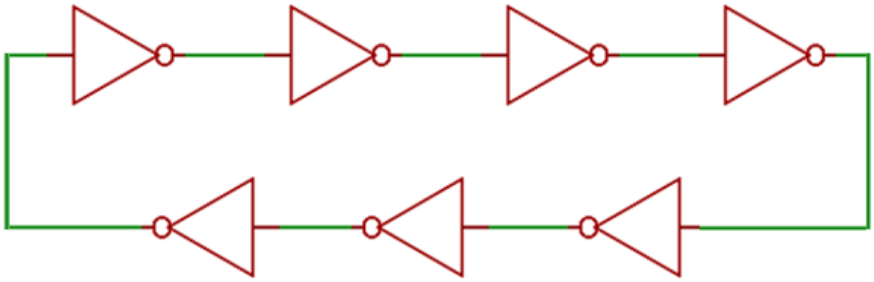


Figure 2.4.1.2 Typical Ring Oscillator schematic, in which triangles represents inverters.

The **ring oscillator is a circuit composed by an odd number of inverters connected in series** forming a ring chain as illustrated in the schematic of Figure 2.4.1.2. This circuit is in constant oscillation because the odd number of inverters. The inverter at a given position will start with a logic state 0. After this pulse propagation through the chain, it will arrive at the same node inverted, because of the odd number of inverters. The oscillation frequency depends on the number of inverters in the chain, as each inverter increases the ring delay. This circuit is very useful to evaluate technology processes, because it is simple to design and check the logic functionality and performance. The RO can be further adjusted to a given fan-out configuration as illustrated in Figure 2.4.1.3. The fan-out is a measure to indicate the gate number able to be connected to the output of a given gate. For example, a fan-out three inverter, has to drive three inverters connected to its output. The RO are usually done in FO3 or FO4 configuration to become closer to real circuit implementations, where the fan-out optimization is a well-known problem [Singh 1990].

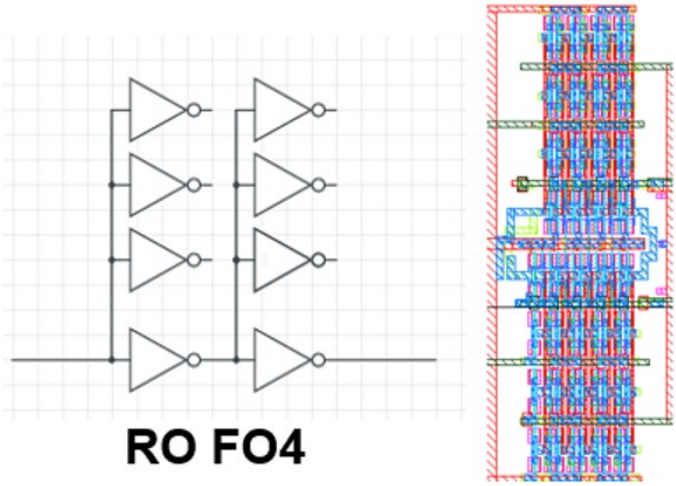


Figure 2.4.1.3 Ring Oscillator schematic piece representing a fan-out four configuration. On the right, a layout for a RO with increased fan-out. The main gates are in the center, while additional dummy gates are in the extremities.

The planar ring oscillator was simulated to assess its properties and as a form of design environment sanity check. Several RO were simulated in different conditions, such as the FO1 or FO3 and different number of inverters in the chain as illustrated in Figure 2.4.1.4. The output frequency formula is shown as the function of the inverter delay, which is composed by the raising and failing delay. The FO3 additional parasitic capacitances slow down the RO, i.e. it increases the delay. In Chapter Three, the different back-end solutions for the 3D sequential integration will be discussed (other than Cu Metals and Low-k dielectrics), and their contribution to the RO performances will be evaluated.

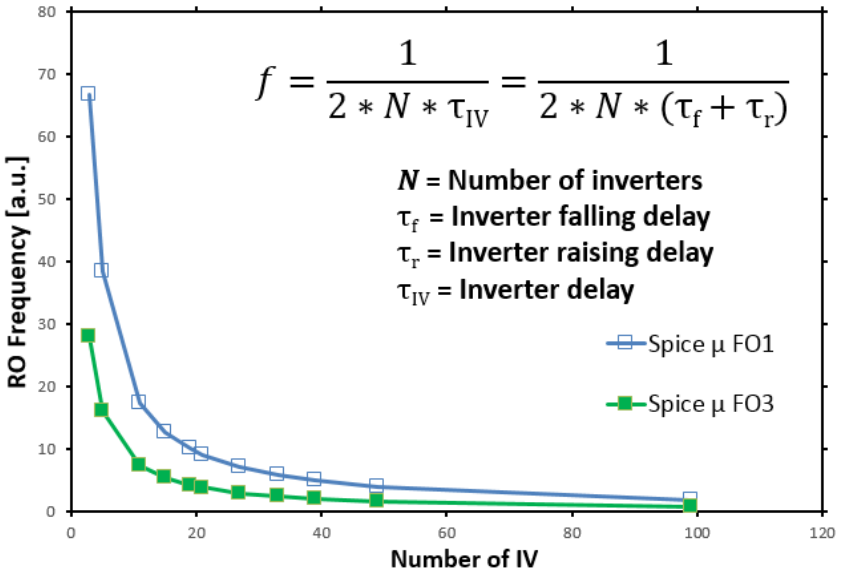


Figure 2.4.1.4 Ring Oscillator output frequency for different number of inverters in a chain. Blue curve represents the FO1 and green curve represents FO3. [Ayres 2016]

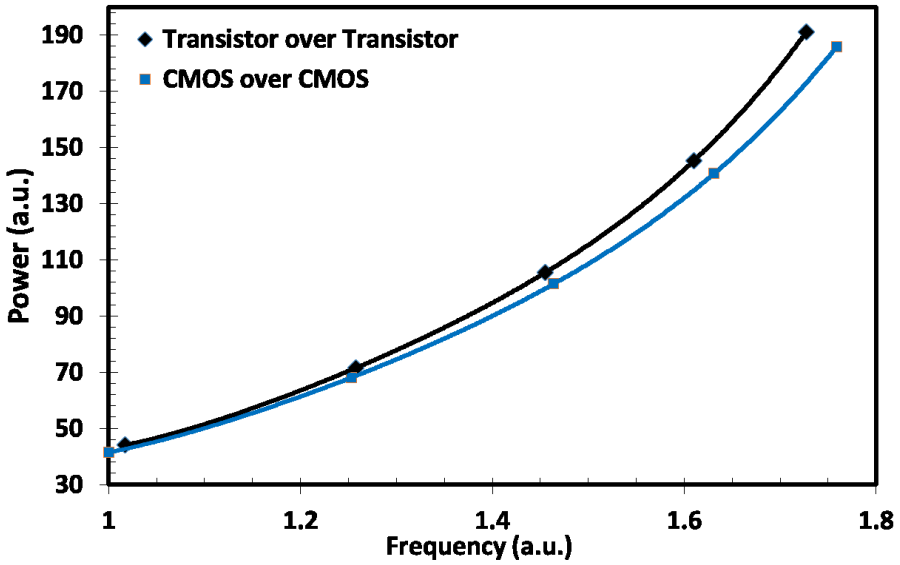


Figure 2.4.1.5 Ring Oscillator output frequency versus the total power comparing two different integration approaches: CMOS over CMOS and Transistor over Transistor.

The ring oscillators are very useful to do a first order technology benchmark, as it has a simple layout and operate the inverter gates at a high switching frequency. For example, a figure of merit regarding the output frequency versus the power consumed by the circuit can be used to determine the best design approach. The power exponentially increases with the frequency, because the transistor current drivability exponentially grows with the gate voltage, raising the power losses in the parasitic impedances. The energy stored in parasitic capacitances also increase due to higher supply voltage. Higher frequency also impacts the power due to the increased number of cycles for a given period. Therefore, this leads to an exponential power dependency over the circuit frequency. The circuit total power is the sum of the leakage power (power consumed during no switching) plus the dynamic power (power under switching operation), and are described respectively by (2.1) and (2.2).

$$P_{STATIC} = V_{DD} I_{LEAK} \tag{2.1}$$

$$P_{DYNAMIC} = \frac{1}{T} \int_{0}^{T} i_{SUPPLY}(t)\, V_{SUPPLY}(t)\, dt \tag{2.2}$$

In Figure 2.4.1.5 the **Transistor over Transistor is compared to CMOS over CMOS integration** for ring oscillators using the same number of stages and the same transistor characteristics. Interestingly, despite of a huge layout difference, the two integrations achieve the same result with a **marginal difference of 3%** in higher frequencies (which are achieved by increasing the V_{SUPPLY}). The result proves the quality of 3DCOs, even using one 3DCO for each inverter gate in the transistor over transistor case. The small parasitic elements from such vias are enough to match the CMOS over CMOS integration.

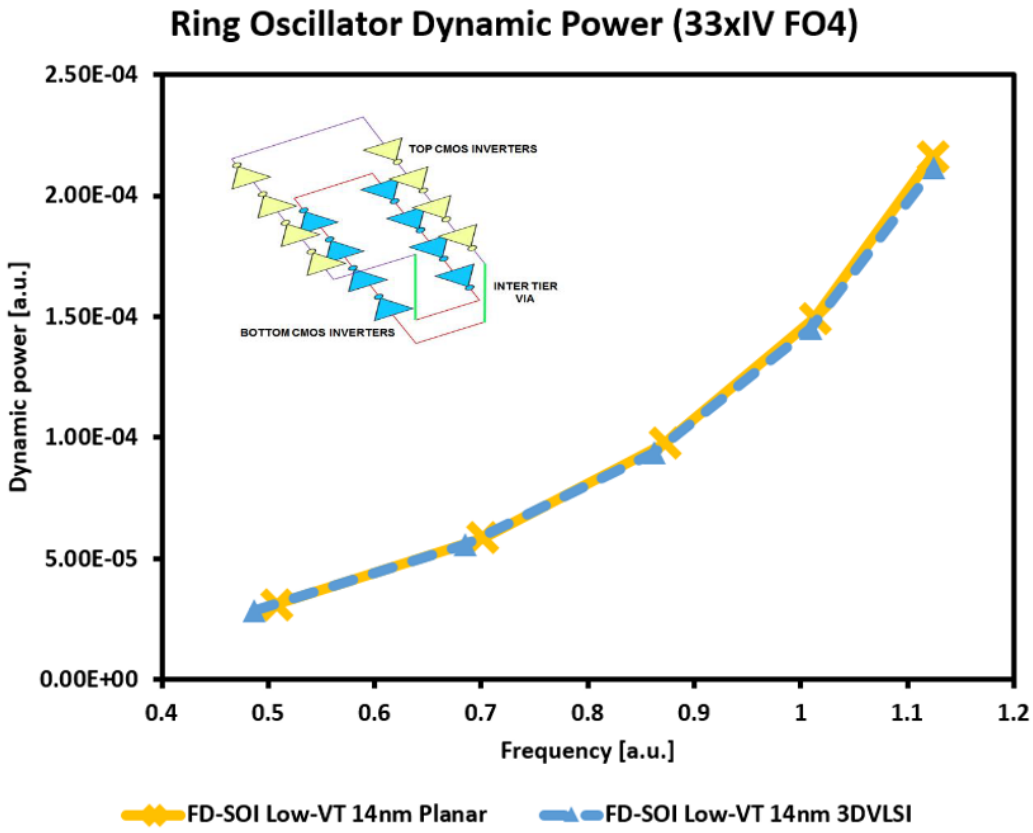


Figure 2.4.1.6 Ring Oscillator benchmark comparing 14nm planar to 14nm 3DVLSI CMOS over CMOS, for the dynamic power versus the ring oscillator output frequency.

The planar RO was also compared to a 3DVLSI CMOS over CMOS ring, using only two 3DCO in order to reduce the area overhead, as depicted on the top of Figure 2.4.1.6. Both RO are simulated after a layout parasitic elements extraction. The 3DVLSI matches the planar performance. From the frequency versus dynamic power figure of merit, two important insights are observed. The **3DCO does not degrade the signal performance**, or does not add a significant RC delay in the circuit. The second insight is about the 3D environment: as the layout was designed with CMOS gates aligned in top and bottom, **the capacitance between tiers, and the coupling among then is negligible**. This outcome holds true for top transistors using back-gate (providing shielding from bottom top metals), and assumes iBEOL with similar characteristics of planar circuits.

2.4.1.2 STR

Besides the ring oscillators, another circuit based in the ring topology has been used to benchmark the 3DVLSI environment. The Self Timed Rings (STR) are circuits that the output frequency does not only depend on the number of stages composing the ring, opposed to the conventional inverter ROs. Indeed, the output frequency of STR can also be controlled by the initial state, making it programmable. Moreover, if each stage is used as an output, the STR can be exploited as a multi-phase oscillator, making it useful for system clock generators. The STR is a ring composed by Muller C-elements, which have been implemented

in CMOS using the Van-Berkel topology as in Figure 2.4.1.7. With this implementation, the PMOS and NMOS can have the same size, except the transistors N6 and P6, which can have larger width to increase the element drivability. In this topology, the gate is composed by 14 transistors, including an inverter for one input. The diagram of the STR for L stages is shown in Figure 2.4.1.8. The STR oscillations can have different modes, including an evenly spaced mode. This mode is due the propagation delay of the Muller C-element, which behaves as follow: the smaller separation time of the two inputs causes a higher delay [Fesquet 2014]. This is known as the Charlie effect (this has been discovered by Charles Molnar in the 70's), and it is plotted in Figure 2.4.1.9. Decreasing the separation between the input signals of the C-element, increases the effective delay (D_{EFF}).

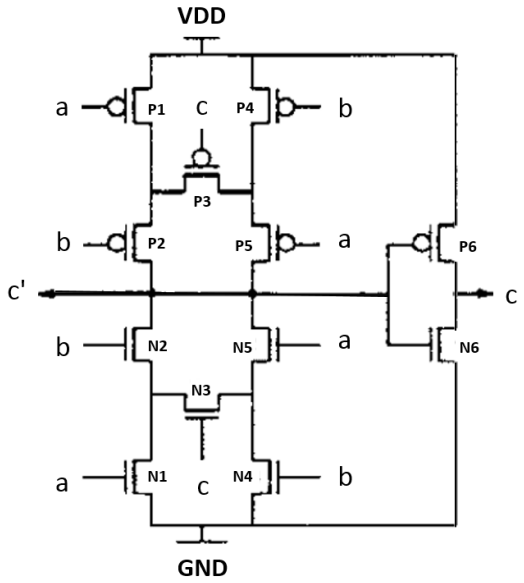
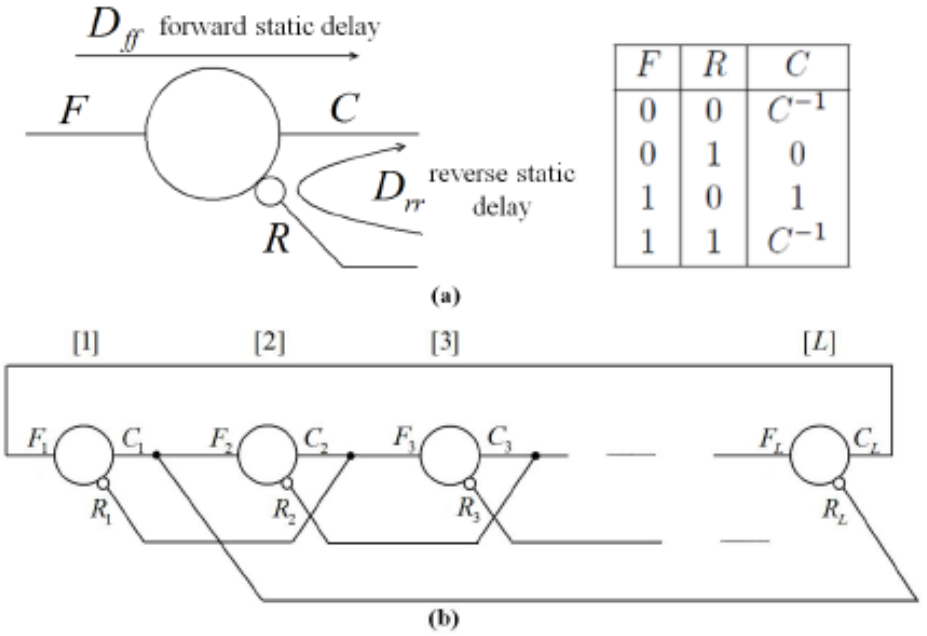


Figure 2.4.1.7 C-Element using Van-Berkel Topology [Shams 1996]

The evenly spaced oscillation in the STR, confirms that the Charlie Effect is operating in the circuit. This fact can be used to benchmark the quality of the interconnections, as the interconnections parasitic elements can degrade the signal, perturbing this analog effect.



F	R	C
0	0	C^{-1}
0	1	0
1	0	1
1	1	C^{-1}

Figure 2.4.1.8 Self Timed Ring composed by Muller C-element and its truth table. [Fesquet 2014].

To verify the behavior of the STR in the 3D environment, the circuit has been partitioned into two tiers as in Figure 2.4.1.10. All the C-elements have the same footprint, only changing the specific layers for bottom or top tier description. The circuit has four 3D contacts, and the top gates are positioned aligned with the bottom gates. This circuit has a total of 256 transistors, including the isolation gates.

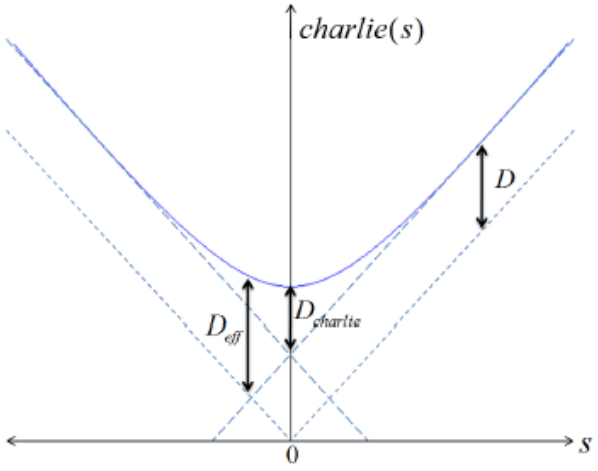


Figure 2.4.1.9 Charlie diagram representing the Charlie Effect. The x axis represents the separation time between two inputs of the C-element. D_{eff} represents the C-element delay. [Fesquet 2014]

The SPICE simulations were done using the extracted parasitic elements layout. This extraction takes into account all front-end and back-end parasitic elements, including the 3DCOs and the capacitance between the top tier and bottom tier. The initial conditions are set in the netlist, in order to an evenly-spaced oscillation mode in the STR.

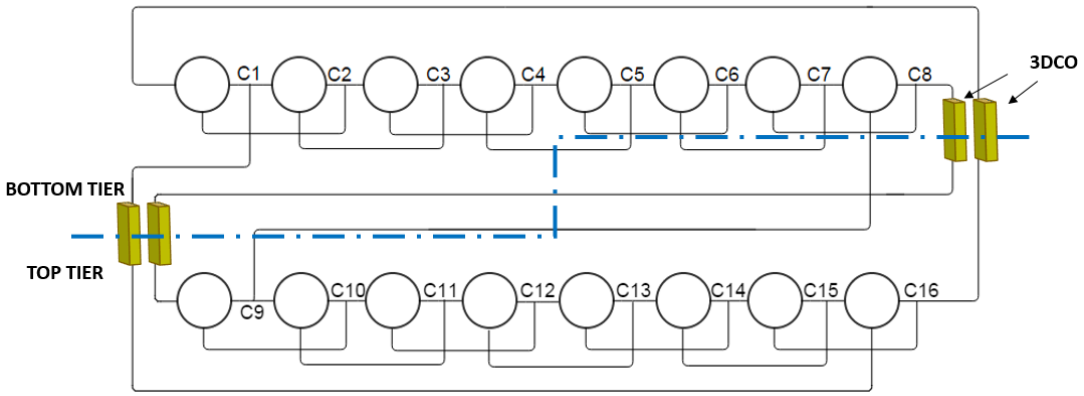


Figure 2.4.1.10 3DVLSI STR schematic. A total of four 3D contacts are used to make connection between the tiers.

The 3D partitioned STR oscillation frequency is illustrated in Figure 2.4.1.11. The output frequency is measured at three different nodes: C16, C13 and C5 defined in Figure 2.4.1.10. The first node is the 3DCO connection between top and bottom tier, and the other two nodes are in the middle of top and bottom chain respectively. The oscillation takes several periods to stabilize, as illustrated in the first frequency measurement. After some periods, the oscillations are measured again, and all the nodes are stabilized at the same frequency. The wave in all nodes exhibits the evenly-spaced mode, or in other words, the waveforms have a duty cycle of 50% and the phase distances between the outputs are equally-spaced. It is important to notice that the evenly spaced-mode is not obtained, if there is a performance mismatch between the connected C-elements on the same tier and the connected ones through a 3DCO. This leads to a burst oscillation mode. In this case, the duty cycle is no longer 50% and the transitions on the outputs are no more evenly-spaced! This result confirms the capability of the 3D environment and 3DCO to operate at high frequency, above 10GHz, and more importantly to preserve the Charlie effect. This gives us a really good indication that 3DCO connections do not strongly affect the performances. Indeed, the Charlie effect is sensitive to a degradation of the routing quality.

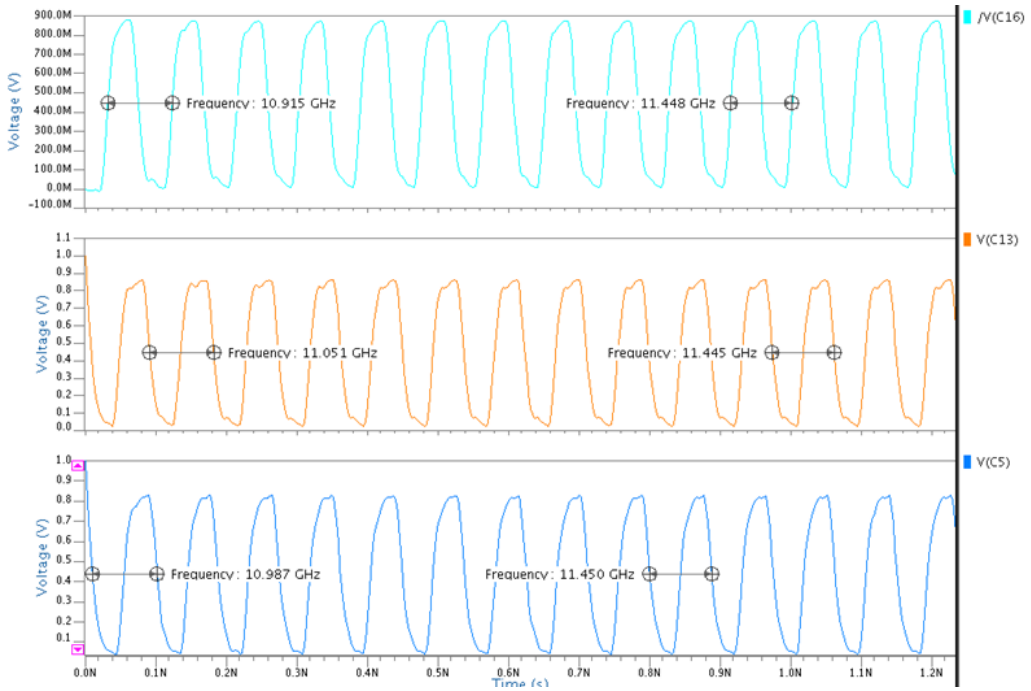
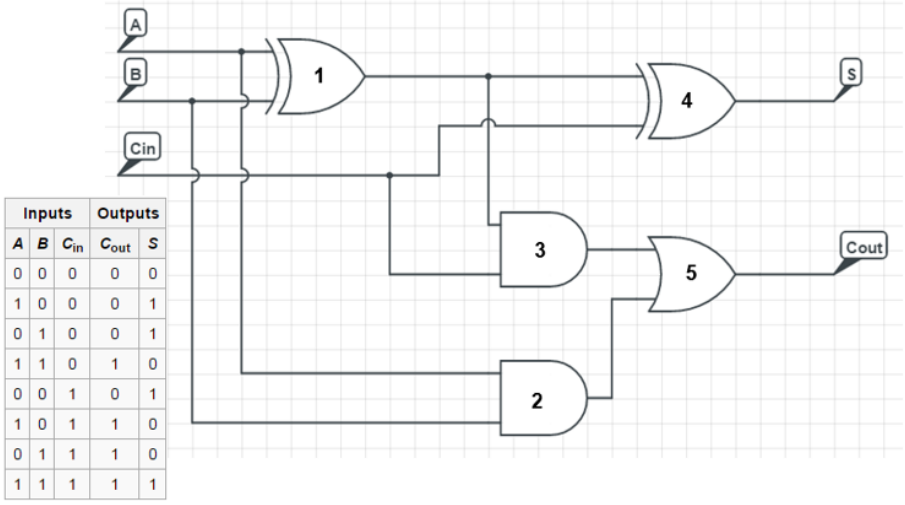


Figure 2.4.1.11 3DVLSI STR output frequency according the C-Element position in the chain.

2.4.1.3 Full Adders

A full adder was designed as a full custom circuit in order to check the potential gains in the 3DVLSI environment. The full adder can sum up two one-bit inputs and consider carry in and carry out signals. One typical implementation of a full adder schematic using combinational logic was used as illustrated in Figure 2.4.1.12. The circuit uses five standard logic gates, such as: ANDs, excusive ORs (XOR), and OR ports.



Inputs			Outputs	
A	*B*	C_{in}	C_{out}	*S*
0	0	0	0	0
1	0	0	0	1
0	1	0	0	1
1	1	0	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0
1	1	1	1	1

Figure 2.4.1.12 Full adder using combinational logic with five standard gates. The truth table with the 3 inputs and 2 outputs is presented on the left.

The layout was implemented in the virtuoso environment following the 14nm design rules, and considering the isolation gates, this circuit has about 76 transistors. The 2D layout was done focusing the best performance and shortest wirelength possible. The final 2D placement is shown in Figure 2.4.1.13a. In this planar implementation, the longest wirelength connects the gates #3 and #5 measuring 1.7µm (approximately 19 times the CPP). The carry out signal delay is directly impacted by the parasitic elements of this connection. To reduce this wirelength, a 3DVLSI implementation is proposed as in Figure 2.4.1.13b. The OR gate #5 is changed to the top tier, and the previous 1.7µm interconnection becomes a 3DCO. The total area is unbalanced between tiers, but this can provide the best performance, as the short connections stay on the same tier.

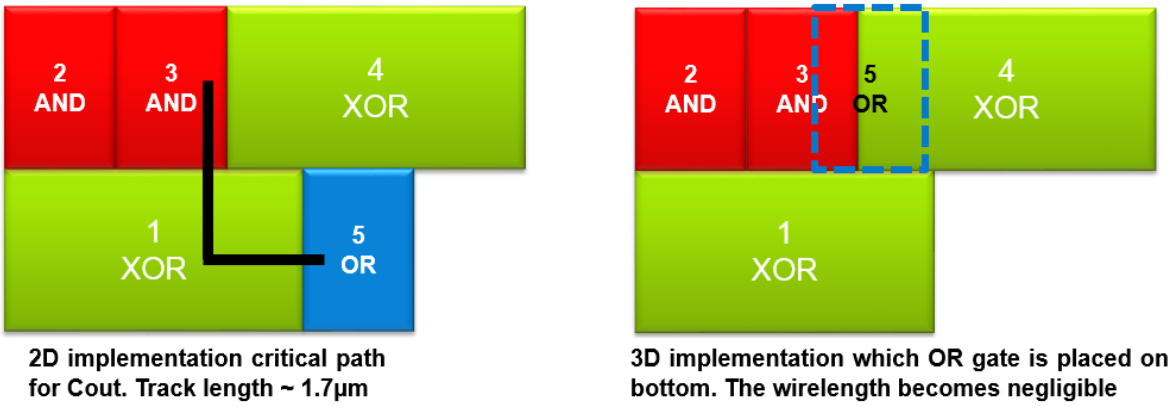


Figure 2.4.1.13 On the left, (a) The planar full custom implementation of the full adder. The maximum wirelength is this setup is 1.7µm. On the right, (b) a 3DVLSI implementation using two-tier. The limiting wirelength of previous case is eliminated, by placing the OR gate in the top level (#5 as dashed line).

A test case was designed to evaluate the outcome of the shorter interconnections. Note that the optimized wirelength only affects the carry out signal, as the rest of the circuit stays routed like the planar case. The test setup is shown in Figure 2.4.1.14a; the inputs B and C_{IN} are changed to high, causing the carry out transition from low to high. The carry out transition time is measured, along the full adder average power during the period. The results for planar and 3DVLSI are illustrated in Figure 2.4.1.14b. The reduced interconnection makes the carry out signal transition low to high 2% faster than the planar circuit. The circuit power consumption is also reduced by 2% compared to the planar case, due to lower parasitic elements. In this fashion, the optimization of very small blocks had been emphasized, as in circuits like two input full adder. The 3D sequential integration partitioning and reduction of wirelength can bring marginal gains to very small logic blocks, considering this design close to the transistor level.

Initial State				
Inputs			Outputs	
A	B	C_{in}	C_{out}	S
1	0	0	0	1

Final State				
Inputs			Outputs	
A	B	C_{in}	C_{out}	S
1	1	1	1	1

2D Integration	
Cout t_{plh}	16,9ps
Average power	301,6µW

3D Sequential Integration	
Cout t_{plh}	16,6ps
Average power	295,2µW

Figure 2.4.1.14 On the left, (a) The transition of the B and C_{IN} inputs on the Full Adder; this test setup was done in planar and 3DVLSI layouts. On the right, (b) The comparison between planar and 3DVLSI considering the carry out propagation time and the average power consumed by the Full Adder.

2.4.2 Conclusion

In this section, the bottom-up design was analyzed to provide insights and guidelines for EDA development. The bottom-up means the strategy to start designing the layout with transistor basic blocks, and assess the performance by doing full custom-layouts. The Ring Oscillators (RO) provides an excellent first order analysis, as the fan-out capacitances plays with the dynamic switching operation. Using ROs full custom layouts in 3D environment the performance is assessed comparing transistor over transistor approach to CMOS over CMOS. The result shows a marginal advantage for CMOS over CMOS. This is the first indication that the 3DCO does not degrade the performances, as the transistor over transistor integration uses one 3DCO per inverter gate. Besides the integration flavor, a planar RO was also compared to a 3DVLSI CMOS over CMOS "folded" RO. The 3DVLSI matches the planar performance, proving that the 3D environment does not degrade the circuit performance, even considering coupling between top and bottom tier. The result was strengthened by benchmarking 3DVLSI STRs, which showed a good environment performance by operating in the evenly-spaced mode, meaning that the analog Charlie effect was retained in the environment. Finally, a Full Adder optimization in 3DVLSI was done, by cutting the longest wire present in a full custom planar implementation. By doing so, a marginal performance and power gains were observed. This outcome shows that optimizations close to the standard cell implementation are limited, and large circuits with long critical paths should be the target of 3DVLSI in order to extract more performance, and reduce power and area; effectively increasing the PPA. The implementation suggests an optimal 3DCO count per number of gates to grant x0.5 area increase, and a 3DCO placement following standard cells directives.

2.5 Chapter Conclusion

In this chapter, the 3DVLSI environment design was explored in order to provide guidelines to process development and the construction of specific EDA tools for 3D sequential integration, such as partitioning. The work has been done using SPICE simulations and layouts inside a 3D monolithic PDK based on the 14nm technology node. Therefore, the circuits are done in a full custom fashion, limited to simple circuits but expressively useful for benchmarking.

The 3D layout drawing guidelines are:

- **3D Contacts:** The vias connecting different tiers. They are built in the sequential process and are aligned to previous layers, allowing the via size and pitch compared to Metal 1 vias, in other words, enabling very high density via placement. Although a principal limitation imposed by the Design Rules Manual (DRM) is the impossibility to pass through active regions. This work suggests a creation of standard cell like structure for 3DCO placement, which can easily be manageable by an automatic EDA tool.
- **Integration Granularity:** The 3DCO high density allows a small integration granularity such as transistor over transistor. Despite of potential process gains by doing only unipolar transistors in a tier (either N or P), the issues in the design level firmly opposes such integration. Density loss and routing blockage caused by the high 3DCO via count are demonstrated. The solution used in the layouts of this work, is the CMOS over CMOS integration. A huge advantage of this scheme is the possibility to reuse the planar standard cells, scripts, parametric cells, etc.
- **3D Design Overhead:** Some partitioning tools can define and limit the number of 3DCO in the physical layout. By using a standard cell approach in CMOS over CMOS integration the number of gates per 3DCO has to be above one hundred in order to achieve 50% area gain with 2 tiers. The worst-case scenario is when only 3DCOs are placed inside a standard cell.

Full custom layouts have been presented for Ring Oscillators, Self-Timed Rings and Full Adders. Those circuits are representative for benchmarking the logical circuits, especially the analog signal behavior in the 3D environment. The ROs shows an ability to work at high speed through 3DCOs without degradation, and no coupling between tiers. Further, the STR can oscillate in an evenly-spaced mode using the 3D sequential integration, proving the good quality of 3DCO connections; as the analog Charlie effect is preserved. Finally, the full adder modification to a full custom 3D layout shows that only marginal gains are possible by cutting the length of back-end interconnections in such a small scale.

REFERENCES

Ayres, A., O. Rozeau, B. Borot, L. Fesquet, G. Cibrario, P. Batude, and M. Vinet. 2015. "Guidelines on 3DVLSI Design Regarding the Intermediate BEOL Process Influence." In *2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 1–2. doi:10.1109/S3S.2015.7333540.

Ayres, A., O. Rozeau, B. Borot, L. Fesquet, and M. Vinet. 2016. "Delay Partitioning Helps Reducing Variability in 3DVLSI." In *2016 46th European Solid-State Device Research Conference (ESSDERC)*, 67–70. doi:10.1109/ESSDERC.2016.7599590.

Batude, P., C. Fenouillet-Beranger, L. Pasini, V. Lu, F. Deprat, L. Brunet, B. Sklenard, et al. 2015. "3DVLSI with CoolCube Process: An Alternative Path to Scaling." In *2015 Symposium on VLSI Technology (VLSI Technology)*, T48–49. doi:10.1109/VLSIT.2015.7223698.

Billoint, O., H. Sarhan, I. Rayane, M. Vinet, P. Batude, C. Fenouillet-Beranger, O. Rozeau, et al. 2015. "A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool." In *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, 1192–96. doi:10.7873/DATE.2015.1110.

Fesquet, L., A. Cherkaoui, and O. Elissati. 2014. "Self-Timed Rings as Low-Phase Noise Programmable Oscillators." In *2014 IEEE 12th International New Circuits and Systems Conference (NEWCAS)*, 409–12. doi:10.1109/NEWCAS.2014.6934069.

Panth, S., K. Samadi, Y. Du, and S. K. Lim. 2015. "Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34 (4): 540–53. doi:10.1109/TCAD.2014.2387827.

Panth, S., S. Samal, Y. S. Yu, and S. K. Lim. 2014. "Design Challenges and Solutions for Ultra-High-Density Monolithic 3D ICs." In *2014 SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 1–2. doi:10.1109/S3S.2014.7028195.

Poiroux, T., O. Rozeau, S. Martinie, P. Scheer, S. Puget, M. A. Jaud, S. E. Ghouli, J. C. Barbé, A. Juge, and O. Faynot. 2013. "UTSOI2: A Complete Physical Compact Model for UTBB and Independent Double Gate MOSFETs." In *2013 IEEE International Electron Devices Meeting*, 12.4.1-12.4.4. doi:10.1109/IEDM.2013.6724616.

Poiroux, T., O. Rozeau, P. Scheer, S. Martinie, M. A. Jaud, M. Minondo, A. Juge, J. C. Barbé, and M. Vinet. 2015a. "Leti-UTSOI2.1: A Compact Model for UTBB-FDSOI Technologies #x2014;Part I: Interface Potentials Analytical Model." *IEEE Transactions on Electron Devices* 62 (9): 2751–59. doi:10.1109/TED.2015.2458339.

———. 2015b. "Leti-UTSOI2.1: A Compact Model for UTBB-FDSOI Technologies #x2014;Part II: DC and AC Model Description." *IEEE Transactions on Electron Devices* 62 (9): 2760–68. doi:10.1109/TED.2015.2458336.

Sarhan, H., S. Thuries, O. Billoint, and F. Clermidy. 2015. "An Unbalanced Area Ratio Study for High Performance Monolithic 3D Integrated Circuits." In *2015 IEEE Computer Society Annual Symposium on VLSI*, 350–55. doi:10.1109/ISVLSI.2015.102.

Sawicki, S., G. Wilke, M. Johann, and R. Reis. 2009. "A Cells and I/O Pins Partitioning Refinement Algorithm for 3D VLSI Circuits." In *2009 16th IEEE International Conference on Electronics, Circuits and Systems - (ICECS 2009)*, 852–55. doi:10.1109/ICECS.2009.5410761.

Shams, M., J. C. Ebergen, and M. I. Elmasry. 1996. "A Comparison of CMOS Implementations of an Asynchronous Circuits Primitive: The C-Element." In *Proceedings of 1996 International Symposium on Low Power Electronics and Design*, 93–96. doi:10.1109/LPE.1996.542737.

Singh, K. J., and A. Sangiovanni-Vincentelli. 1990. “A Heuristic Algorithm for the Fanout Problem.” In *27th ACM/IEEE Design Automation Conference*, 357–60. doi:10.1109/DAC.1990.114882.

Weber, O., E. Josse, F. Andrieu, A. Cros, E. Richard, P. Perreau, E. Baylac, et al. 2014. “14nm FDSOI Technology for High Speed and Energy Efficient Applications.” In *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, 1–2. doi:10.1109/VLSIT.2014.6894343.

Weber, O., E. Josse, J. Mazurier, N. Degors, S. Chhun, P. Maury, S. Lagrasta, D. Barge, J. P. Manceau, and M. Haond. 2015. “14nm FDSOI Upgraded Device Performance for Ultra-Low Voltage Operation.” In *2015 Symposium on VLSI Technology (VLSI Technology)*, T168–69. doi:10.1109/VLSIT.2015.7223664.

INTRODUCTION TO CHAPTER THREE

Previous chapter considers the back-end process identical for each tier back-end. The interconnections are assumed to be made of Copper and Low-k dielectric isolation. In this chapter, an evaluation of different process for intermediate back-end is done, in the case where standard planar BEOL is not feasible for 3D integration.

The 3D circuits are benchmarked as function of the process assumptions, especially for the back-end process flavor. For example, if 3DVLSI wafer breaks in front-end machines, the tiers already built can contaminate the tools with BEOL metals, thus a different process approach is needed in order to reduce contamination severity.

Finally, a comparison of 3DVLSI integration to planar scaling trend is done using advanced nodes compact models, such FDSOI, FinFETs and Nanowires, up to the 5nm node. As the BEOL scales, the interconnections have its parasitic elements increased per normalized length, decreasing the circuit performance. This illustrates the possible insertion of 3D sequential integration in the scaling roadmap.

Chapter Three – BEOL Process Influence on 3D Design

3.1 Guidelines on 3DVLSI BEOL process development

3.1.1 IBEOL Limitations

The process for a monolithic 3D integration has a limited temperature to preserve the FET already built [Fenouillet-Beranger 2014]. This leads to **difficulties of using standard ultra low-k** (ULK) dielectrics on the intermediate BEOL; in a worst-case scenario SiO_2 with a permittivity of 3.9 could be used as an intermediate dielectric. Also, due to **wafer break risk**, the use of copper can be **an issue for contamination** reasons. A solution can be the use of tungsten for the filling lines as it has already been integrated in the FEOL of several products.

3.1.1.1 W and Cu as metal

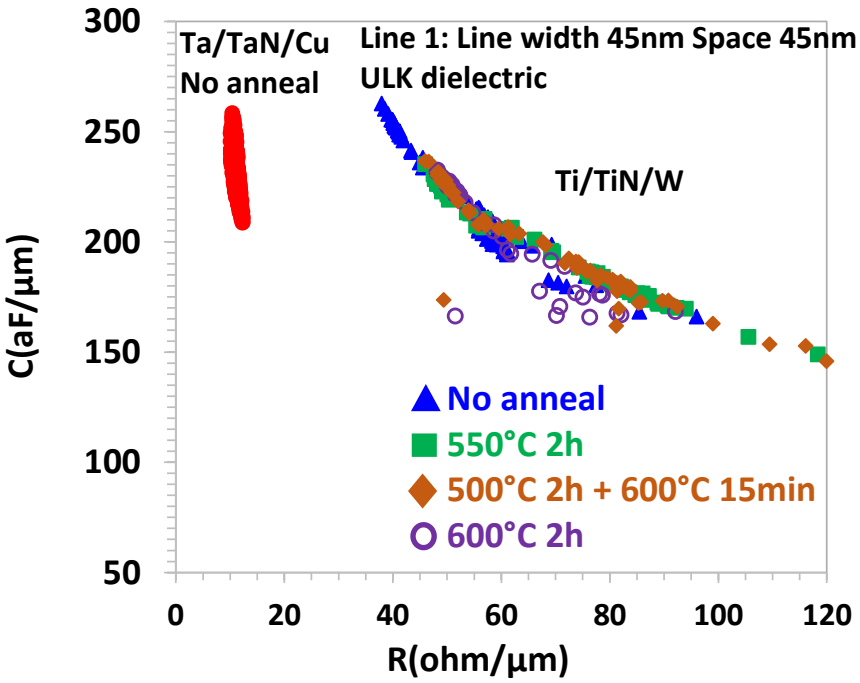


Figure 3.1.1.1 Lateral capacitance versus resistance for line 1 W/ULK interconnection with Ti/TiN barrier before and after annealing (line width and space=45nm). No anneal Cu line is plotted as reference. [Fenouillet-Beranger 2017]

Figure 3.1.1.1 shows one line capacitance versus resistance measured on a specific multi-fingers/serpentine test structure for Tungsten (W) interconnections before and after annealing with the Ti/TiN barrier in Ultra Low-K dielectric (ULK). This figure confirms a factor 6 times higher resistance value for the W as compared to the copper one and the W/ULK stability up to 550°C during 5 hours. In addition, dies are functional at 600°C during 2 hours but the measurements dispersion increases as the temperature increases beyond 550°C [Fenouillet-Beranger 2017].

3.1.1.2 Dielectric Stability

The ultra low-k materials are currently under research for a low temperature process. Also from (Fenouillet-Beranger et al. 2017), the ULK material is shown as stable for a thermal budget of 500ºC up to two hours, and no defects were found. However, the evaluation of extrapolated lifetime is still necessary to determine the dielectric endurance over the years.

Metal	Barrier	Permittivity	C, R	Stability	Reliability
W	Ti/TiN	ULK	R x6 /Cu, C ok	Ok up to 550°C 5h	Ok up to 550°C 5h
	F-free W	ULK	R x4 /Cu, C ok	Ok up to 550°C 5h	Ok up to 550°C 5h
	Ti/TiN	TEOS/SiN	R x6 /Cu, C NOK	Ok up to 600°C 2h	N/A
Cu	Ta/TaN	ULK	R, C Ok	Ok up to 500°C 2h	In progress

Figure 3.1.1.2 Summary of the different interconnections stability behavior versus thermal budgets regarding R, C and reliability performances.

3.1.2 IBEOL flavors and Ring Oscillators

3.1.2.1 Parasitic Elements Extractions

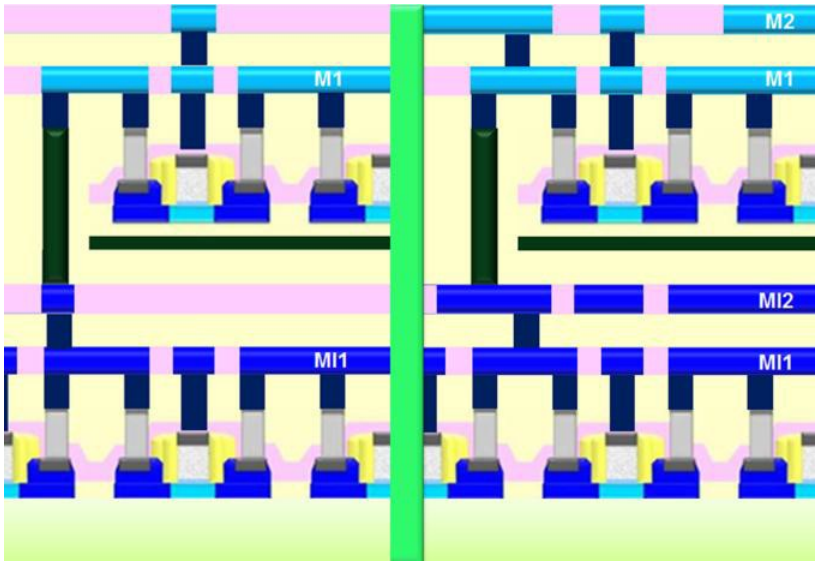


Figure 3.1.2.1 3D sequential stack with 2 IBEOL metals, which is the process setup for 3DVLSI parasitic extraction. [Ayres 2015].

Ring oscillators in FO1 and FO4 configuration are simulated after the PEX, using the 3D process stack as in Figure 2.3.3.1. The evaluation methodology consists **in changing the characteristics of intermediate BEOL from copper metals and low-k dielectric to tungsten and SiO2 with 3.9 permittivity** as illustrated in Figure 3.1.2.2, then the performance impact due to limited thermal budget process is accounted. The simulation focuses in the RO frequency output, which is directly affected by the interconnections. Higher resistance or capacitances in the interconnections will degrade the signal timing for the same drive current. The signal transition t_{PHL} or t_{PLH} have a higher delay R.C product, thus a lower frequency is expected in the ring

oscillator. In this section, the goal is to provide guidelines to process development, by evaluating the outcome of different BEOL processes in the final circuit performance. Finally, those guidelines can be used by the integration development team to evaluate the **tradeoff between the performance impact and the process manufacturing constraints**.

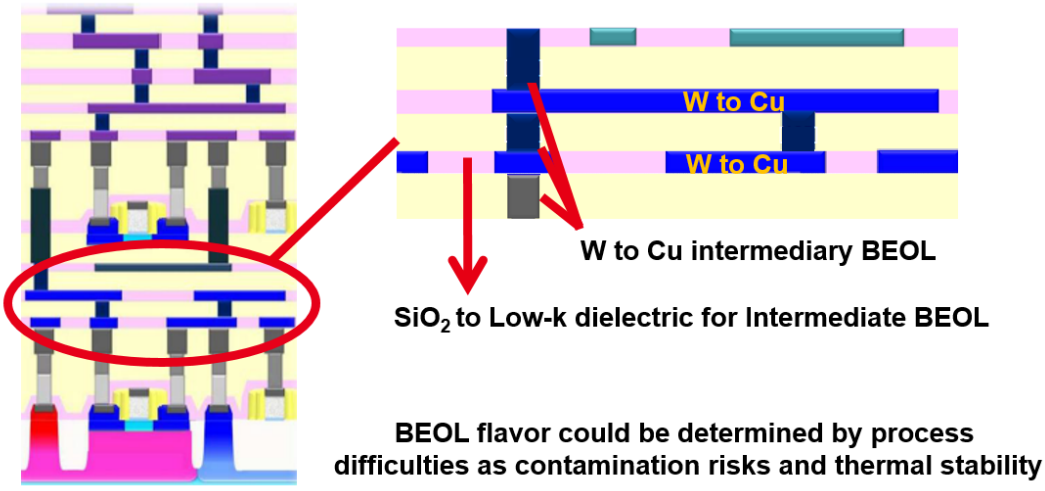


Figure 3.1.2.2 Back-end flavors setup for parasitic extractions. The metals can have their resistivity changed as well the dielectric have several permittivity.

3.1.2.2 Sources of performance impact

Employing iBEOL cases W/SiO_2 or Cu/ULK depict the worst/best case scenario. By measuring the 3D CMOS gate-level ring oscillator FO4 output frequency, the worst case has a 2% degradation compared to the best case for FO1 and 4% for FO4. Another strategy was employed in order to evaluate the sources of this degradation. In further simulations, only a specific iBEOL element was modified from the best case, for instance some simulations have only intermediate vias in tungsten, while some simulations only employ SiO_2 dielectric. The sum of the performance impact from each element reaches the total impact, reinforcing the accuracy of the evaluation. Those simulations show that the main constraining factor is the dielectric permittivity, due to the aggressive metal spacing. The via contact resistance has also a minor influence in the delay as shown in Figure 3.1.2.3. Finally, no influence of the metal line resistivity is observed, because in the RO layout the length of intermetal lines is short (<2µm).

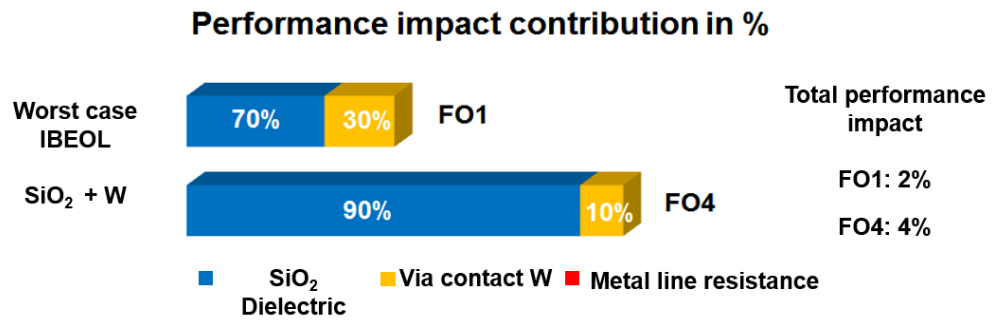


Figure 3.1.2.3 Performance impact contributing factors when using SiO2 and Tungsten. The SiO_2 dielectric is the main limiting factor in intermediate BEOL. No impact from tungsten in intermediary metal lines with this setup. [Ayres 2015]

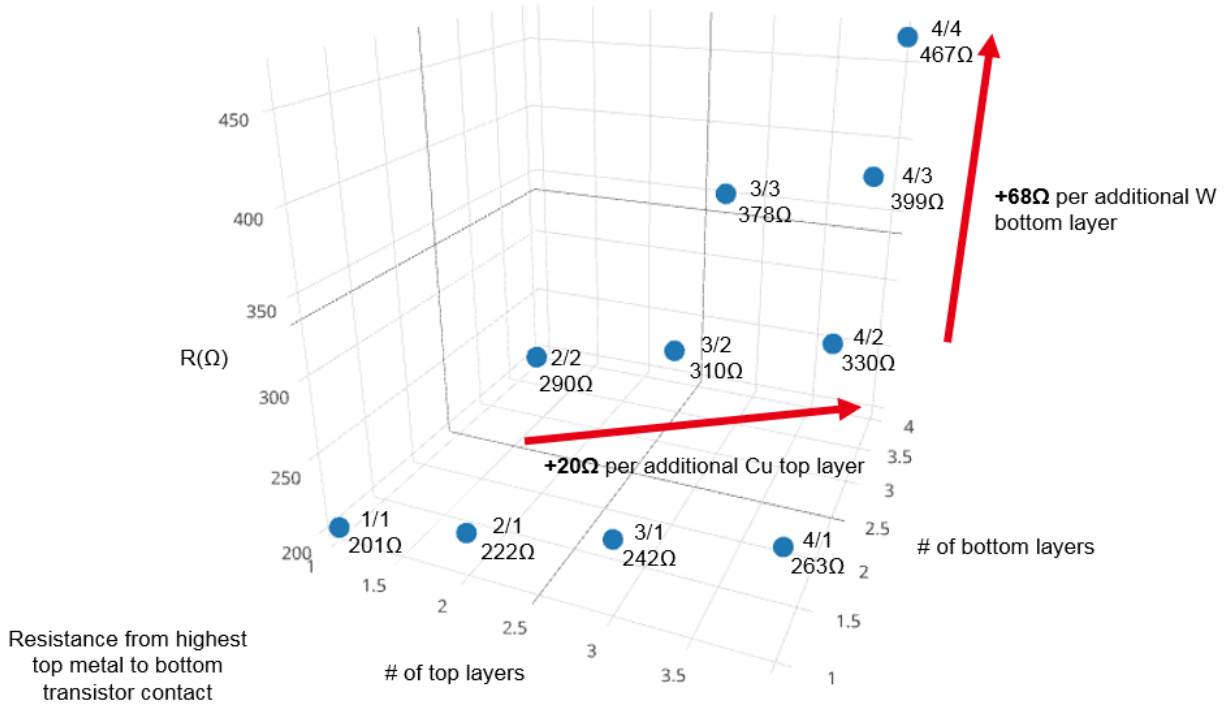


Figure 3.1.2.4 Total resistance from the highest top metal to the bottom transistor contact. The resistance increases with the total number of metal layers. The iBEOL is assumed as tungsten and top BEOL is made of copper.

The 3DCO can have a huge role in the circuit performance, especially for the **power delivery network (PDN)**. A study was performed using the 14nm design rules, and considering tungsten as metal for iBEOL and copper for top BEOL. Increasing the number of metal layers in the back-end augments the resistance from the highest top metal to the bottom transistor contact, mainly due to additional vias between the metals. In Figure 3.1.2.4 the total resistance is plotted as a function of the number of top and bottom back-end layers; from one layer at minimum up to four layers in each tier. The number above the resistance is

in the format T/B, where T is the number of top metal layers while B is the number of bottom metal layers. The resistance increases by 20Ω per additional copper metal in top tier, and the additional bottom metal layer increases the resistance by 68Ω. In this study, the vias were considered at minimum size possible, thus the resistance is quite high. Also, in upper layers above Metal 4, the density is usually lower than previous layers, and the resistance penalty is lower. The main message and guideline from this study, is that **by increasing the iBEOL number of layers it degrades the vertical path for device biasing**. A practical example is illustrated in Figure 3.1.2.5, showing the supply voltage close to the transistor in a ring oscillator under operation. As the current flows into the circuit, **the parasitic resistances cause a voltage drop seen by the transistor**. This effect decreases the circuit performance, as the gates operate lower than the nominal voltage. In Figure 3.1.2.5a, the supply voltage is measured close to the bottom transistor contact. In this case, the minimal via size is used, including the 3DCO with tungsten metal. This results in a maximal voltage drop of 10mV, or 1% considering the supply voltage at 1V. Another layout setup was done, improving the PDN vias resistance by enlarging their size. The result is shown in Figure 3.1.2.5b, where the maximum voltage drop is in the order of 1mV or 0.1%, thus the ring oscillator can operate at the nominal voltage.

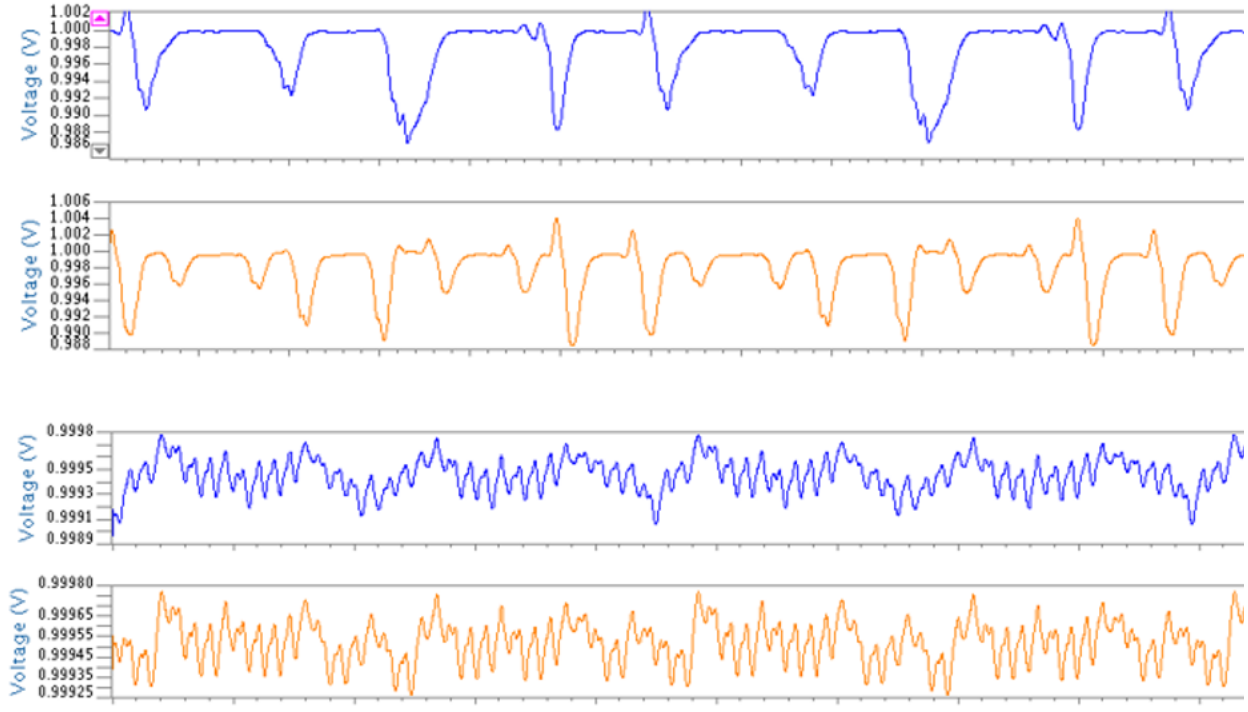


Figure 3.1.2.5 Ring oscillator supply voltage measured in two different points (yellow and blue) close to the transistor contact. In the top, (a) a layout using minimum size tungsten vias for PDN. In the bottom, (b) the same circuit using relaxed density PDN vias. A voltage drop of 10mV is seen in the first case, while less than 1mV is observed in the second case.

3.1.2.3 Wirelength Study

The vast contribution in performance increases from 3D integration comes from the shorter wirelength paths [Billoint 2015] and, in this fashion, 3D integration permits to reduce critical path length. The wire length in 2D circuits is studied employing a planar ring oscillator FO4 split into two parts with extended wire length as in Figure 3.1.2.6.

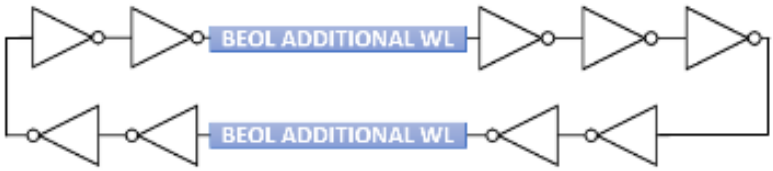


Figure 3.1.2.6 Planar ring oscillator using increased wire length. Two connections have longer interconnections, then the total additional wirelength is the sum from those wires.

The planar layout setup was incremented with several metal lines in parallel to the signal path to simulate a real circuit capacitance. The impact of tungsten interconnections is compared to copper in Figure 3.1.2.7. The RO frequency at 14nm design rules have been measured and reveals no major change of propagation delay with respect to the two types of interconnections for wirelength lower than 5µm. Increasing the wirelength, the copper back-end performs better, due to reduced parasitic resistance compared to tungsten. However, for a 39 µm interconnection, the output frequency is degraded by 20% for copper, while tungsten loses the same amount of performance for 35µm interconnections. Therefore, another test case with ROs was crafted to show the potential 3DVLSI gain, by cutting long interconnections.

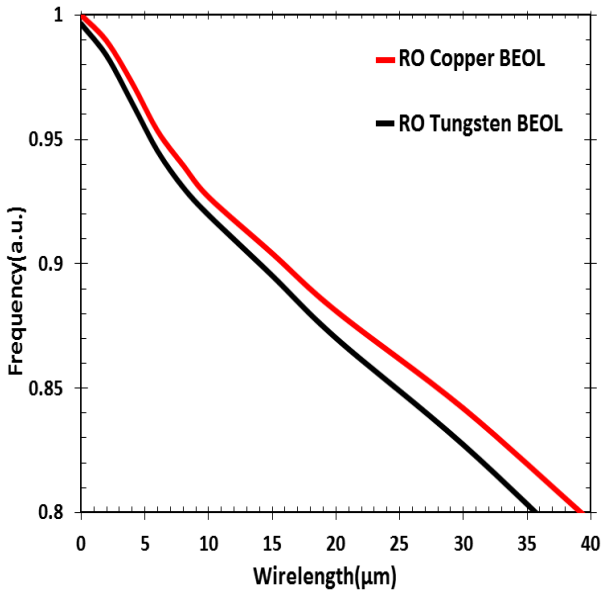


Figure 3.1.2.7 Output frequency for RO using different back-end metals for increased interconnection length.

The 3D ring oscillator FO4 was made in the CMOS gate-level integration due to better routability compared to transistor over transistor. The 3DVLSI RO is finally compared to the planar equivalent as in Figure 3.1.2.8. The 3D circuit is evaluated at fixed wirelength, while the planar wirelength is changing. This approximation can be done because for each planar wire cut the routing tool has to minimize the 3D path, especially if it is taken into consideration that the bottom tier should be used for local routing only. The critical planar wirelength WL_{IBEOL} (purple vertical dashed line) is defined as the **length upon which a gain in performance is obtained by cutting the planar wire and stacking in a 3D configuration**. For an intermediate BEOL using a standard process, with low-k and copper, the critical wirelength named $WL_{IBEOLstd}$ (green vertical dashed line) is zero, meaning that **3D performance matches planar**, and it allows the maximum number of planar tracks to be cut. At this point, the **frequency is increased for each track cut**. In fact, the circuit can be folded in any part without any loss, which is especially interesting for memory applications. In the worst-case scenario, the critical wirelength $WL_{IBEOLworstcase}$ increases. i.e. This scenario represents a technological solution not found in order to obtain a lower k dielectric than 3.9 and no other metal with resistivity lower than W reducing the contamination risks. In order to match the planar performance in this configuration, 3D connection should replace a 5µm planar track. This means that shorter metals tracks should stay in planar integration if the designer is only looking for performance, although it is possible to pay a small penalty in frequency to gain density. On the other hand, when the critical WL_{BEOL} is achieved, the 3D integration gains in both density and performance. The frequency performance enhancement depends on the length of the cut wires. For a 40µm planar track, the frequency gain is 21%. Thus, the opportunity for performance gain with 3DVLSI will be higher for complex circuits containing long wire length distribution.

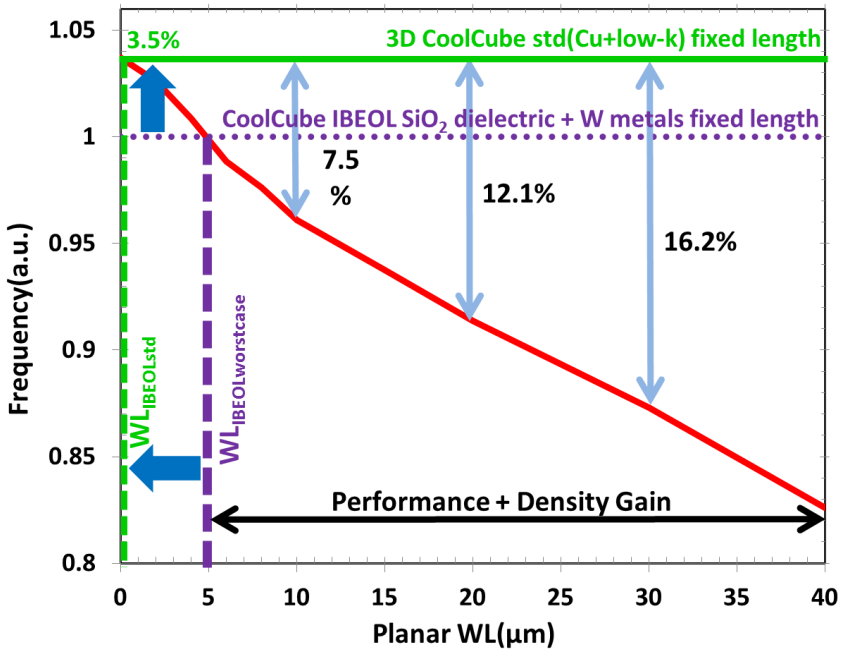


Figure 3.1.2.8 Planar ring oscillator FO4 increased WL compared to 3D ring oscillator with fixed length. The horizontal dotted line represents a worst-case process in 3D intermediate BEOL. The process influences on planar WL to 3D tradeoff is shown in the vertical dashed line. [Ayres 2015]

3.2 BEOL Limitations in Advanced Nodes

The 3DVLSI can increase the circuit performance by optimizing the BEOL interconnections RC parasitic elements using the vertical direction. In this section, the planar BEOL scaling is discussed in order to provide insights about physical limits, and consequently a good opportunity window to 3D sequential integration for digital circuits.

3.2.1 Scaling Expectations

Moore's scaling is roughly reducing transistor dimension by two each eighteen month. To keep the trend alive, new transistors architectures were introduced in the 28nm/22nm nodes; such as FD-SOI and FinFET [Jan 2012],[Boeuf 2008]. Those new transistors were designed to overcome limitations of the traditional bulk devices, which suffers, for example, of leakage when the gate length is scaled too aggressively. The FD-SOI and FinFET have excellent performance characteristics, allowing a reduction in the supply voltage, decreasing the power consumption and were adopted by the industry as solution for scaling [Skotnicki 2008]. In 5nm node, the stacked gate all around is expected to be introduced, increasing further the electrostatic control by the gate. The back-end of the line (BEOL) has to follow the transistor shrink, or in other words, the interconnections have to be scaled in the same ratio of the transistor. The BEOL scaling tendency is shown in Table 3-I as minimum metal pitch. The metal pitch is composed by the metal line width plus the dielectric width separating two conductors. In this chapter, we simulate the BEOL for future nodes using a common setup layout, and then by parasitic extractions, we evaluate the BEOL impact in performance via SPICE simulations.

TABLE 3-I
SCALING IN ADVANCED NODES

Node	CPP [nm]	Metal Pitch [nm]	Supply Voltage [V]
14	90	64	0.8
10	64	48	0.7
7	46	36	0.64
5	32	24	0.6

3.2.2 Wirelength Delay in Advanced nodes

The BEOL shrinking augments the connection resistance and capacitance for a given length, considering the same BEOL integration for different nodes. This problem has a limited impact in the circuit performance, as the scaling reduces the CPP. This translates in a short interconnection length; thus compensating, or even increasing the interconnection performance despite the increased normalized parasitic elements per length. A layout has been designed in a PDK as follows in Figure 3.2.2.1, in order to simulate interconnection parasitic elements. The back-end connects an inverter logic gate in a FO3 configuration. Metal 0 is placed under the interconnection connected to supply source, as well M2 above connected to ground. Parallel to M1 signal, two lines are connected to supply and ground respectively.

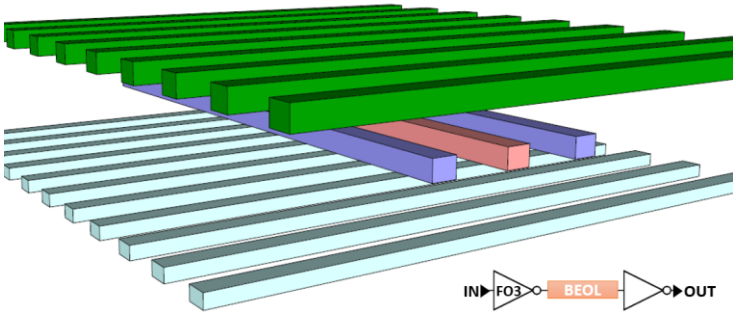


Figure 3.2.2.1 Layout for delay benchmarking. M2 in green, M1 in dark blue and M0 in cyan. The connection under evaluation is highlighted in reddish.

The connection length is normalized to node, as hundred times the CPP for typical interconnections and ten thousand the CPP for hypothetical critical cases. In real circuits, the 10KCPP connections would be done in less dense metal and have buffers to not degrade the delay timing. This condition is used to illustrate a case where the BEOL delay is dominant, and does not depend on the transistor capabilities. The metal width and spacing follows the Table I for each node, along with the metal thickness adjusted x0.7 of previous node. After the layout parasitic extraction, the circuit was simulated for several nodes and the back-end delay evaluated as in Figure 3.2.2.2.

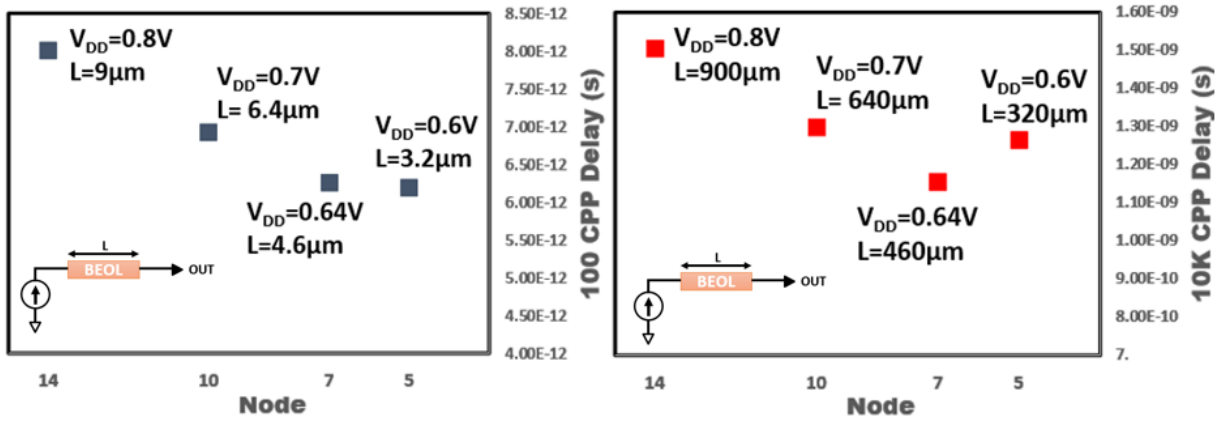


Figure 3.2.2.2 Back-end delay for at fixed current for node given voltage. On the left, (a) evaluation for a connection of 100 CPP in length. On the right, (b) same illustration for ten thousand CPP in length.

In this simulation, all cases use the same fixed current for the node given voltage, hence only the BEOL performance is evaluated. With the same BEOL composition (Cu/ULK), a trend reversal is observed in the 5nm node. Due to BEOL geometries, the interconnection length reduction is not anymore enough to compensate the parasitic elements. In Figure 3.2.2.2 the same netlist with the parasitic extraction is used, and simulated using SPICE compact models for FD-SOI, FinFET and GAA. The total delay illustrated being composed by transistor delay and BEOL delay. For 100CPP the BEOL delay represents approximately 60% of the total delay. In Figure 3.2.2.3, the delay reduction still occurs for the transition from 7nm to 5nm node. This outcome is explained by the 5nm transistor better current drive, compensating the degraded BEOL delay described in Figure 3.2.2.2a. However, for very long wires, in the range of 10K CPP, the BEOL delay is dominant as seen in Figure 3.2.2.5; and for 5nm, the performance is extremely impacted, confirming the results from Figure 3.2.2.2b. As the BEOL performance does not scale from 7nm to 5nm, some solutions process solutions can be employed. An additional case for 5nm BEOL was extracted; employing air-gaps in the M1/M2, which are already implemented in upper metal levels of devices in mass production [Natarajan 2014] as illustrated in Figure 3.2.2.4. The air-gaps are implemented in the PEX files, simulating a relative permittivity (εr=2.2) for M1/M2 dielectrics. For 100CPP, the benefits are minimal, nonetheless a major improvement is seen for very long connections in 5nm, placing the node back in the general trend of Figure 3.2.2.5.

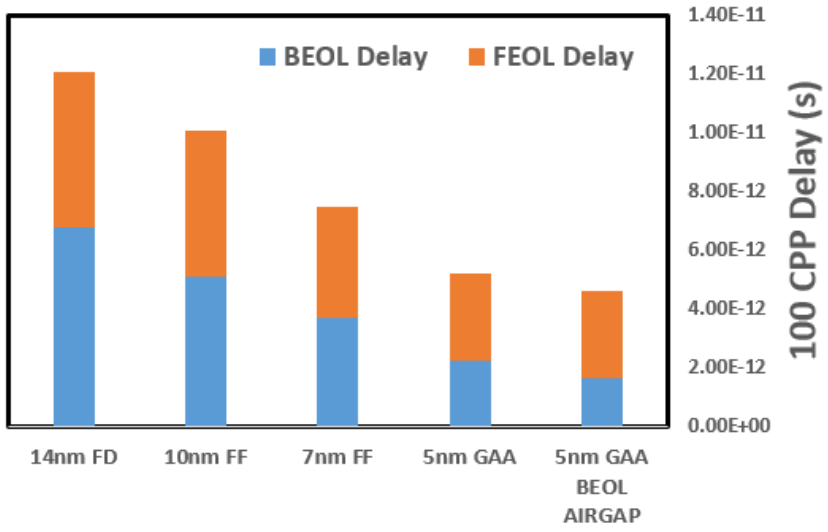


Figure 3.2.2.3 Total delay considering previous layout setup and using SPICE compact model for each node. Evaluation for a connection of 100 CPP in length while.

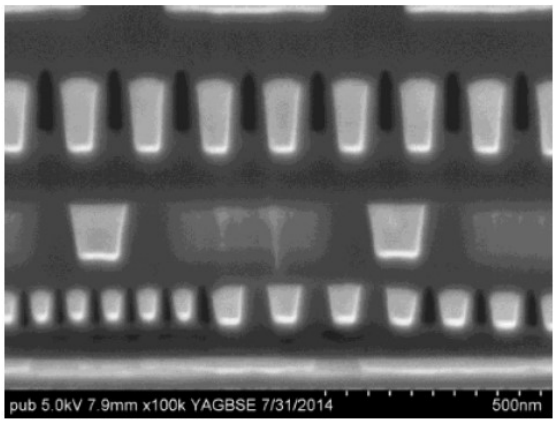


Figure 3.2.2.4 Back-end showing air-gaps in the dielectric material to reduce the relative permittivity. [Natarajan 2014]

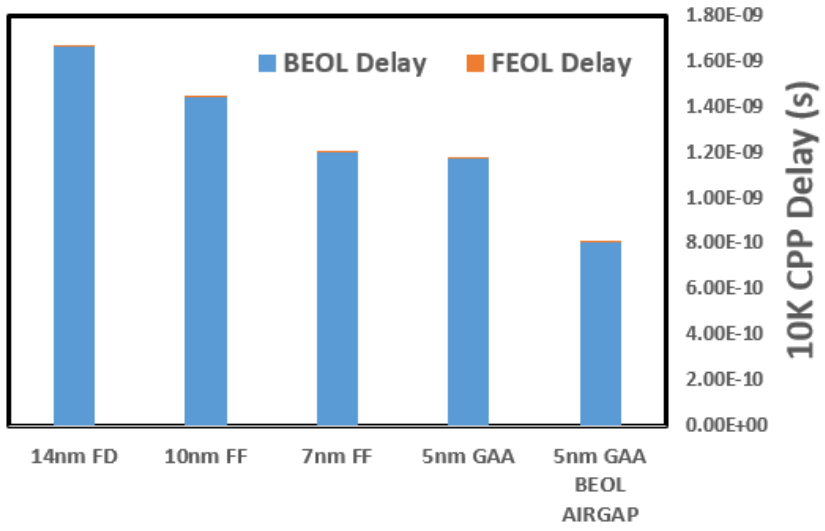


Figure 3.2.2.5 Total delay considering previous layout setup and using SPICE compact model for each node. Evaluation for a connection of 10 thousand CPP in length while.

The simulations consider the metal resistivity (ρ) of 4μΩ.cm for all nodes. Nonetheless, copper wire resistivity for minimum width increases in each node due to electron surface scattering and grain-boundary scattering. As the dimensions scale down, the metal widths are in the order of electron mean-free path, augmenting those effects and increasing the resistivity [Chen 1998]. The setup was redone, at this time, considering the resistivity increase for advanced nodes. The cooper resistivity is extracted from the evaluations of [Huynh-Bao 2017]. For example, the minimum metal width in 7nm node has a resistivity of 5.5μΩ.cm, while in the 5nm node 8μΩ.cm. The simulations are illustrated in Figure 3.2.2.6.

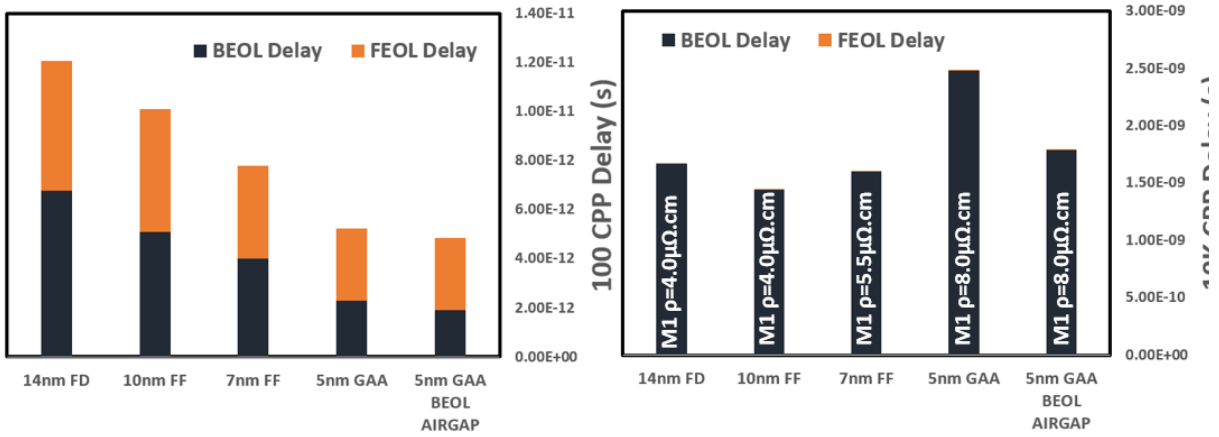


Figure 3.2.2.6 Total delay considering previous layout setup and using SPICE compact model for each node and adjusted metal resistivity. (a) Evaluation for a connection of 100 CPP in length while (b) for ten thousand CPP in length.

Another possible way of BEOL scaling, is the use of 3D sequential integration. In this integration, the circuits are positioned in different levels, namely tiers. A 3D contact connects the bottom and top tier. In Figure 3.2.2.7 a simulation compares the delay of 1K and 100 CPP in 5nm BEOL with air-gap to the delay of ten 3D contacts. Here, it is shown that if technological solutions are not found to reduce the BEOL resistivity induced by the scaling in advanced nodes, the 3D sequential integration can be a suitable candidate, as expanding to the vertical direction can further reduce the BEOL delay.

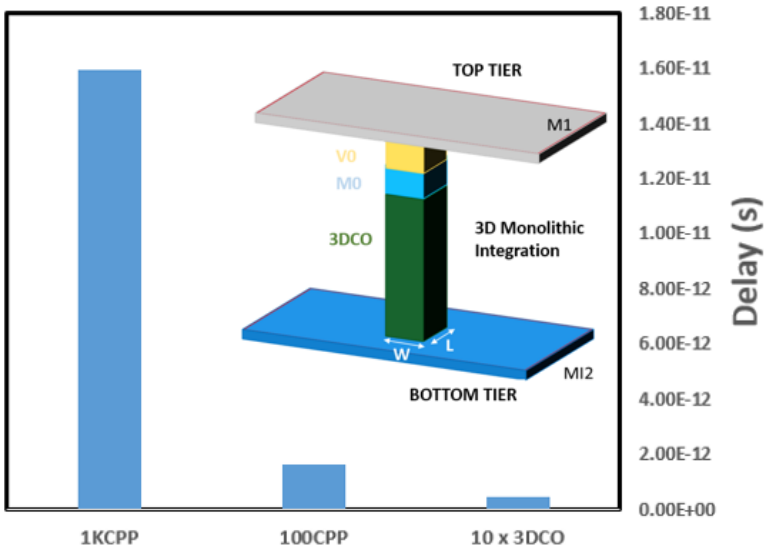


Figure 3.2.2.7 Comparison of BEOL delay in 5nm node with Air-gap to 3DCO delay. The 3D sequential integration 3D via (3DCO) is shown as an alternative to 2D BEOL scaling.

3.3 Chapter Conclusion

Due to some process thermal budget and to the risk management, the intermediate back-end (iBEOL), namely the back-end between two tiers can be done in tungsten and SiO_2 rather than with the usual copper and low-k dielectrics. A study using ring oscillators shows the higher permittivity is crucial for the RO performance. Despite of tungsten being six times more resistive than copper, no impact due to metal routing was observed, mainly because ring oscillator interconnections are very short. However, when using tungsten vias, the designer should be aware of the voltage drop in bottom tier, as the higher resistivity requires a larger via. The supply voltage drop is very important in the PDN design, otherwise the bottom transistors will have a different operating point than the top transistors.

3D integration can increase the performance by transforming long wires of planar circuits (critical paths) into 3DCOs reducing the net parasitic elements. A planar RO was modified in order to create such long wires and compare to 3D ROs. The results show a performance increase of 16% for a 30µm planar wire cut into 3D using 14nm design rules. This reconfirms that the performance increase opportunity comes from large scale digital circuits were a complexity is needed in order to present gains; otherwise for small wirelength cut, the possible gains will be lower and less attractive.

Finally, a benchmark was done using advanced nodes from 14nm up to 5nm. The back-end scaling will become a significant problem limiting the performance, if no process change is done. The scaling reduces the metal and dielectrics width and thickness, increasing the parasitic resistance and capacitance. Until the 7nm this effect is compensated by the lower wirelength, because the gates scaling makes shorter the distances. However, in 5nm, a trend reversal appears because the back-end performance can limit the circuit performance. The 3D integration is shown as a strong contender for Moore's planar scaling, especially because of these back-end limitations.

REFERENCES

Ayres, A., O. Rozeau, B. Borot, L. Fesquet, G. Cibrario, P. Batude, and M. Vinet. 2015. "Guidelines on 3DVLSI Design Regarding the Intermediate BEOL Process Influence." In *2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 1–2. doi:10.1109/S3S.2015.7333540.

Billoint, O., H. Sarhan, I. Rayane, M. Vinet, P. Batude, C. Fenouillet-Beranger, O. Rozeau, et al. 2015. "A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool." In *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, 1192–96. doi:10.7873/DATE.2015.1110.

Boeuf, F., M. Sellier, A. Farcy, and T. Skotnicki. 2008. "An Evaluation of the CMOS Technology Roadmap From the Point of View of Variability, Interconnects, and Power Dissipation." *IEEE Transactions on Electron Devices* 55 (6): 1433–40. doi:10.1109/TED.2008.921274.

Chen, Fen, and D. Gardner. 1998. "Influence of Line Dimensions on the Resistance of Cu Interconnections." *IEEE Electron Device Letters* 19 (12): 508–10. doi:10.1109/55.735762.

Fenouillet-Beranger, S Beaurepaire, F. Deprat, A. Ayres, L. Brunet, P. Batude, P. Besombes, et al. 2017. "Guidelines for Intermediate Back End Of Line (BEOL) for 3D Sequential Integration." In *2017 47th European Solid-State Device Research Conference (ESSDERC)*.

Fenouillet-Beranger, C., B. Mathieu, B. Previtali, M. P. Samson, N. Rambal, V. Benevent, S. Kerdiles, et al. 2014. "New Insights on Bottom Layer Thermal Stability and Laser Annealing Promises for High Performance 3D VLSI." In *2014 IEEE International Electron Devices Meeting*, 27.5.1-27.5.4. doi:10.1109/IEDM.2014.7047121.

Huynh-Bao, T., J. Ryckaert, Z. Tökei, A. Mercha, D. Verkest, A. V. Y. Thean, and P. Wambacq. 2017. "Statistical Timing Analysis Considering Device and Interconnect Variability for BEOL Requirements in the 5-Nm Node and Beyond." *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* PP (99): 1–12. doi:10.1109/TVLSI.2017.2647853.

Jan, C. H., U. Bhattacharya, R. Brain, S. J. Choi, G. Curello, G. Gupta, W. Hafez, et al. 2012. "A 22nm SoC Platform Technology Featuring 3-D Tri-Gate and High-K/Metal Gate, Optimized for Ultra Low Power, High Performance and High Density SoC Applications." In *2012 International Electron Devices Meeting*, 3.1.1-3.1.4. doi:10.1109/IEDM.2012.6478969.

Natarajan, S., M. Agostinelli, S. Akbar, M. Bost, A. Bowonder, V. Chikarmane, S. Chouksey, et al. 2014. "A 14nm Logic Technology Featuring 2nd-Generation FinFET, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588 #x00B5;m2 SRAM Cell Size." In *2014 IEEE International Electron Devices Meeting*, 3.7.1-3.7.3. doi:10.1109/IEDM.2014.7046976.

Skotnicki, T., C. Fenouillet-Beranger, C. Gallon, F. Boeuf, S. Monfray, F. Payet, A. Pouydebasque, et al. 2008. "Innovative Materials, Devices, and CMOS Technologies for Low-Power Mobile Multimedia." *IEEE Transactions on Electron Devices* 55 (1): 96–130. doi:10.1109/TED.2007.911338.

PART TWO: VARIABILITY

INTRODUCTION TO CHAPTER 4

In this chapter, the typical variability sources for planar circuits are introduced. A discussion of the usual way to analyze those variations is done, such as the Monte Carlo method and the clever solution of using process corners. Those tools will be later reused for 3DVLSI circuits in Chapter Five.

The main goal of this chapter is to discuss the notions of planar variability; and to illustrate how it is treated and managed. As the circuit performance is shifted depending on process characteristics, the intrinsic process variability causes a performance distribution among produced chips. In this chapter, the Monte Carlo method is depicted as the main tool to evaluate the circuit performance, and then the process corners are described as a powerful solution to design phase, granting a statistical perspective of circuit performance and yield.

Chapter Four – Variability in VLSI

4.1 Variability in VLSI Circuits

4.1.1 Sources of Process Variability

Process variability arises during the fabrication process of the transistor. **Some transistors attributes differ from the desired nominal value by a certain amount of variation**. For example, thickness, length, width and roughness are parameters that may have a deviation during the fabrication process. Process variations happen in all technologies as the **machines and processes have a certain amount of uncertainty due to intrinsic stochastic behavior**. These variations between transistors, which are supposed to be identical, are called mismatch. Beyond random variations, another source of variations is the systematic variation, caused by the circuit layout implementation. Effects like mechanical stress, orientation in the silicon crystalline structure and layout induced lithography proximity can occur due to layout design. A classification of sources of variability based on their root causes is illustrated in Figure 4.1.1.1.

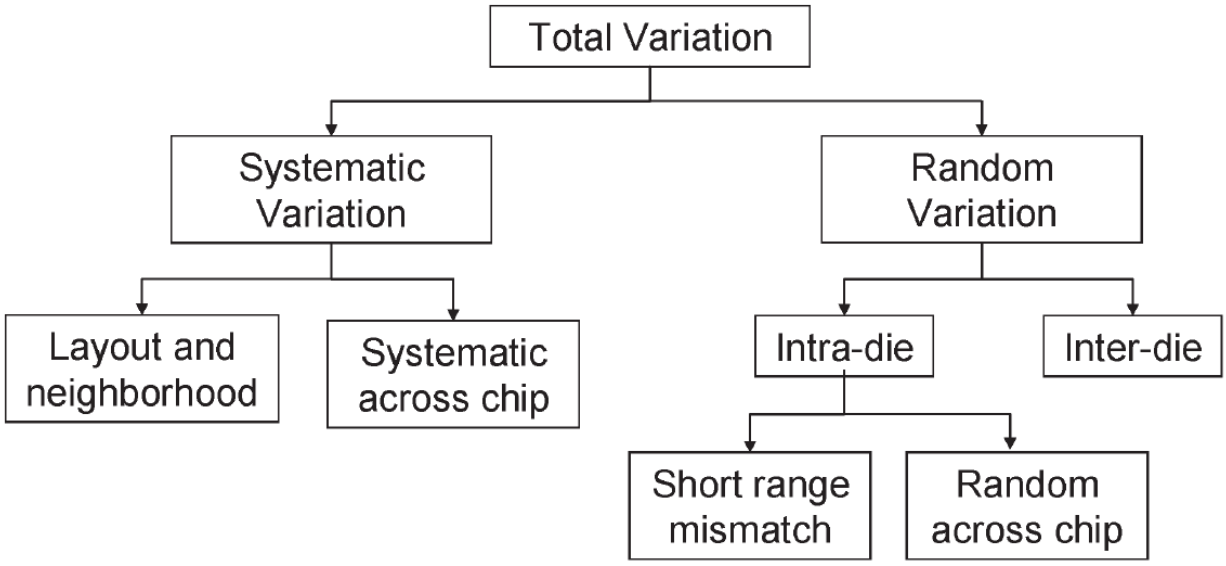


Figure 4.1.1.1 Transistor variability sources divided in two classes: layout dependent and process random variations.[Saxena 2008]

This thesis is focused on the **random variations branch**. The systematic variations are highly dependent on the circuit layout and schematic and are usually managed during the LVS step. In order to treat the general case in 3DVLSI, the circuits were simulated at a schematic level, considering some pre-layout effects, meaning that parasitic elements of source and drain contacts, as well the M0 (Metal 0) interconnections are taken into account. The random variations are separated into **intra-die** (variations in the same wafer) and **inter-die** (variation across all produced wafers). Further, the intra-die variations can be classified as **local variations** (or Pelgrom's variation) and **across-chip variations** (distance dependent variations).

4.1.2 Pelgrom's Variability – Local Variations

Local variations affect transistors individually. The transistors on the same chip can have different performance compared to each other. Physically, the process variations in one dimension or in two dimensions are for example results from edge roughness, thickness variations, channel doping, etc. This

introduces the notion of fluctuation dependency on the device length (L) and width (W), or precisely area [Drennan 2003]. The dependency between the variance of a parameter P and the device size can be described as in (4.1).

$$\sigma_P \propto \frac{1}{WL} \tag{4.1}$$

Intuitively the variations **depend on the scale of dimensions**, as device size increases, the "averaging" of variations is reached; in the other sense, for smaller devices, the variation becomes a significant part compared to the desired nominal value. The effect is represented in Figure 4.1.2.1.

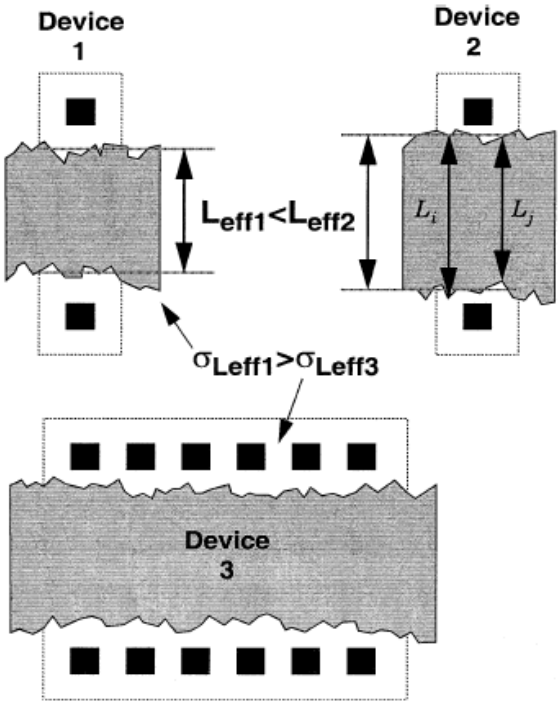


Figure 4.1.2.1 Local variability dependence on size illustrated as gate length. For longer widths, the final length is "averaged". [Drennan 2003]

The relation between the device size and local variations was deduced by [Pelgrom 1989] highlighting the **mismatch variance ΔP** for a parameter pair P depending on the device size (4.2). The original work deduces the formula from spatial frequencies of fluctuations and by using Fourier transform on two rectangular devices with the same size.

$$\sigma^2_{\Delta P} = \frac{A_P^2}{WL} + Device\ Spacing \tag{4.2}$$

The parameter A_P is an area proportionality constant for a given process. The second term in the original equation refers to the device separation, thus the work addresses the total intra-wafer variability. However, for accuracy reasons, the variability originating from the **device spacing** is treated in this thesis as **across-chip variations**. The mismatch becomes a linear function with angular coefficient A_P when evaluating the variance only considering the first term in 4.2. Plotting the standard deviation versus the

inverse of the area square root: $1/\sqrt{WL}$; the linear relation becomes evident, and this function plot is usually referenced as Pelgrom's plot. Increasing the X axis direction means that the device area is lowering.

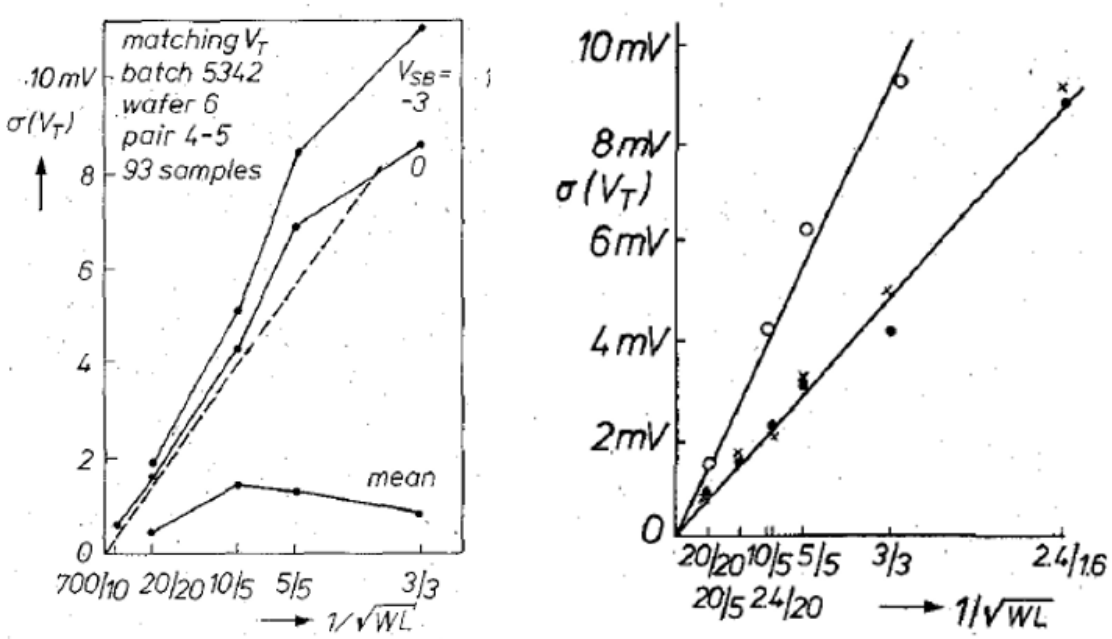


Figure 4.1.2.2 Pelgrom's work showing the linear relation between the sigma V_T and the inverse of the device size. [Pelgrom 1989]

Interestingly, the observation made by Pelgrom circa 1989 holds true for some parameters in advanced nodes, as for example in the standard V_T deviation in 14nm node depending on the transistor size. The extracted A_{VT} (the area proportionality constant for the V_T) can be used as benchmark parameter to directly compare the process variation.

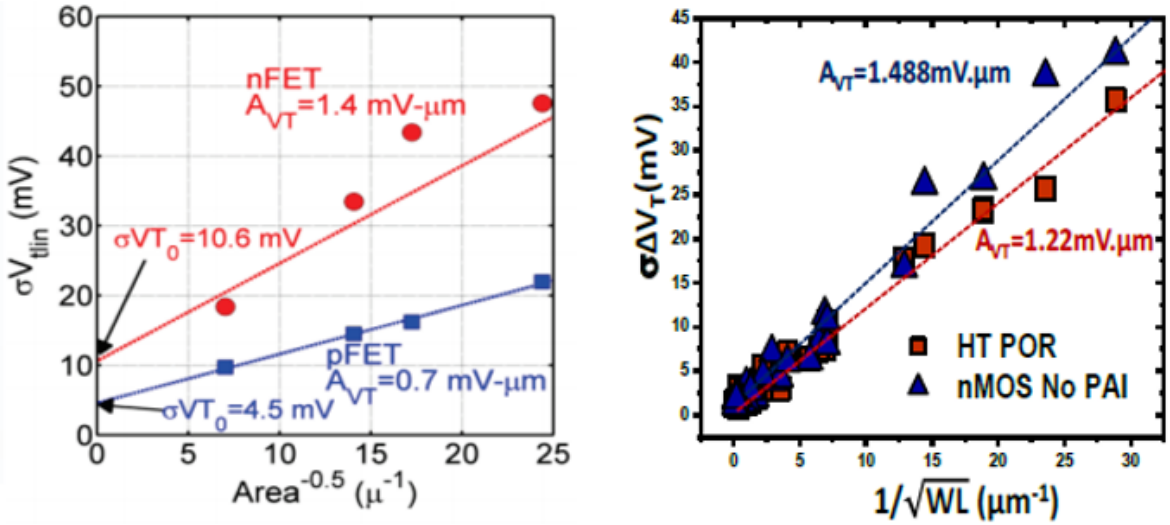


Figure 4.1.2.3 Pelgrom Plot for 14nm nodes. (a) SOI-FinFET sigma V_T. [Paul 2013] (b) Low-Temperature process for 3D monolithic 14nm FD-SOI. [Pasini 2016]

The local variations mismatch can be measured directly for the parameter or the Δ_P difference in the parameter for close devices and later evaluating the standard deviation for the different $\sigma\Delta_P$ [Kuhn 2011]. The relation between the $\sigma\Delta_P$ and σ_P measurements is $\sqrt{2}$ as shown in (4.3) and (4.4) for the V_T.

$$\sigma_{random-one-device} = \frac{\sigma(V_{TA} - V_{TB})}{\sqrt{2}} = \frac{\sigma(\Delta V_T)}{\sqrt{2}} \quad (4.3)$$

$$\sigma_{random-pair} = \sigma(V_{TA} - V_{TB}) = \sigma(\Delta V_T) \quad (4.4)$$

4.1.3 Global Variability

This variability **equally affects all transistors in a chip or a wafer**. Therefore, the statistical model used in simulation is the same for all the devices. Considering a Gaussian model, each parameter is defined as a distribution $N\sim(\mu_{GLOBAL},\sigma_{GLOBAL})$ which is shared by all the devices in the chip. **The notion of global variation is extended further than a single chip, all the devices that belong to a population can have its global variation extracted**. The global variation arises from small alterations in the environment or in the machines during wafers production over months or years. **Despite the rigid control of a clean room environment standard, wafers produced in different dates can have a variation.** For example, if a machine has a component repaired, it may not have the exact same characteristics as before [Castaneda 2012]. In addition, the global variability is often referenced as worst or best-case scenario for some circuits due to the immunity to the local variations. This property will be discussed in ring oscillator variance analysis. An important remark is that the **global variation is not correlated with the local variability**, e.g. a parameter has independent statistical models for local and global variations.

4.1.4 ACV

Across-Chip Variations, namely **ACV**, are variations **dependent on distance separation between two devices**. Beyond the mismatch, the concept of **correlation** and device variance needs to be described to provide the full variability figure. Usually the ACV correlations can be visualized in wafer contour plots for a certain parameter. The color changes with the amplitude of the parameter measured as represented in Figure 4.1.4.1 for SOI thickness.

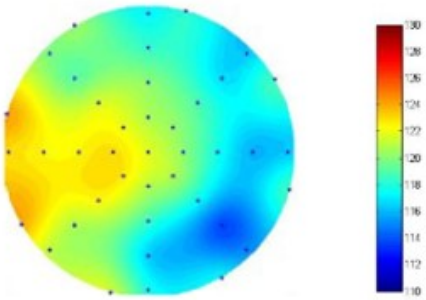


Figure 4.1.4.1 300mm SOI wafer scale Across-Chip Variation for silicon thickness. The contour plot evidence the correlations between close regions, as they remain in the same color. [Schwarzenbach 2011]

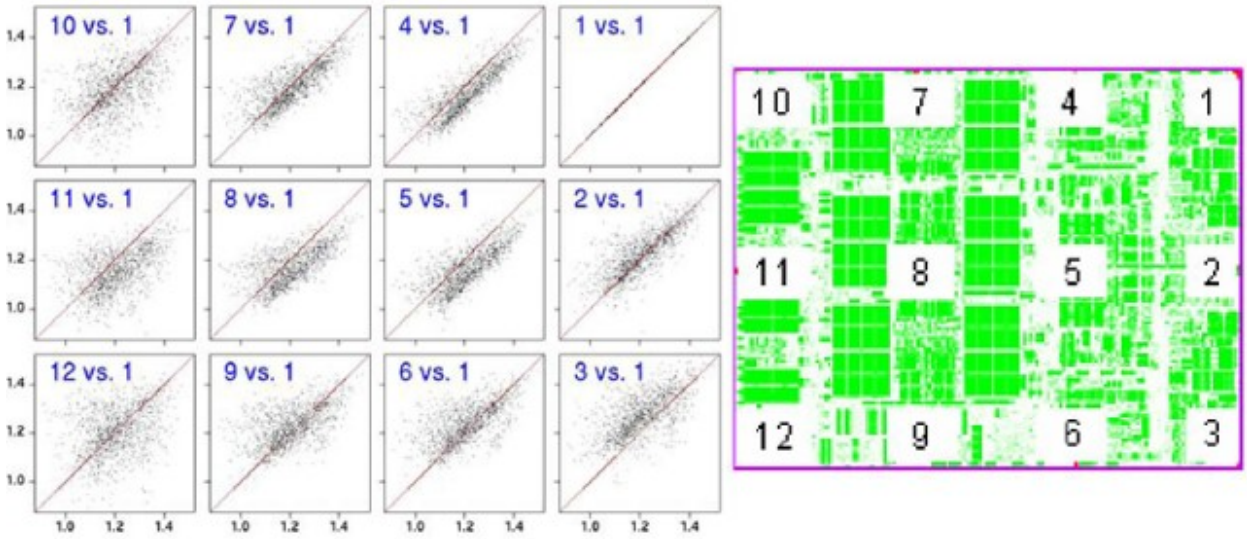


Figure 4.1.4.2 Across-Chip Variation analysis for Ring Oscillators inside a die of commercial chip. Comparing the reference RO #1 with others, the correlation dependence on the separation distance becomes noticeable. [Gattiker 2006]

The measured variability along this 300mm wafer, is at the maximum range of ±5Å [Schwarzenbach 2011]. This translates in a **different device performance depending on the position**. Another important aspect, the correlations, can be seen as the colors gradient. If two devices have **almost identical value for a given parameter, the colors will be similar**. In the illustration, it is possible to observe that for a **given point in the wafer the colors in the neighborhood tend to stay the same**. This represents the correlation between

close devices. **The length, for which the devices remains correlated, is called the correlation length**. A clever way to evaluate the correlation lengths in actual wafer under production is reported by [Gattiker 2006]. ROs were added to a real die circuit in different positions as shown in the right of Figure 4.1.4.2

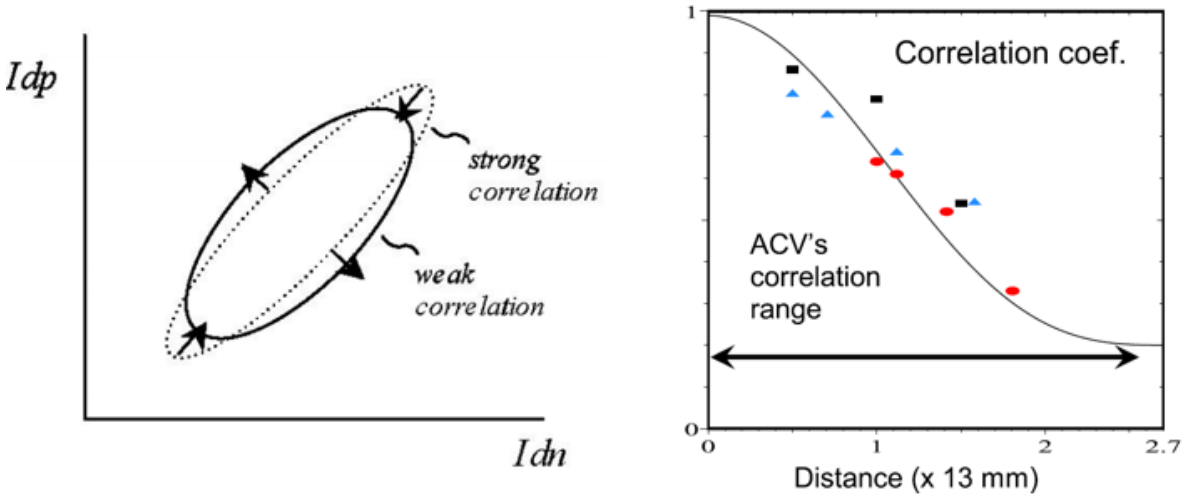


Figure 4.1.4.3 (a) XY plot to illustrate correlations [Eikyu 2006] (b) Across-Chip Variation analysis for Ring Oscillators inside a die of commercial chip. Evaluation of correlation depending on distance and the correlation length. [Lu 2014]

By measuring the RO frequency in one position and comparing it to the other RO frequencies at different locations allows to calculate the correlation between them as a function of distance. In the left side of the illustration, ROs in position #1 are compared to ROs in position #2 by plotting each position output frequency in an axis. The plot has a certain dispersion in both X and Y direction, and the **Pearson correlation** can be evaluated as illustrated in Figure 4.1.4.3a. In the big picture comparing all positions to the reference #1, it is possible to see the **correlation decreasing as the distance increase**, as illustrated in Figure 4.1.4.3b.

4.1.5 Monte Carlo Analysis

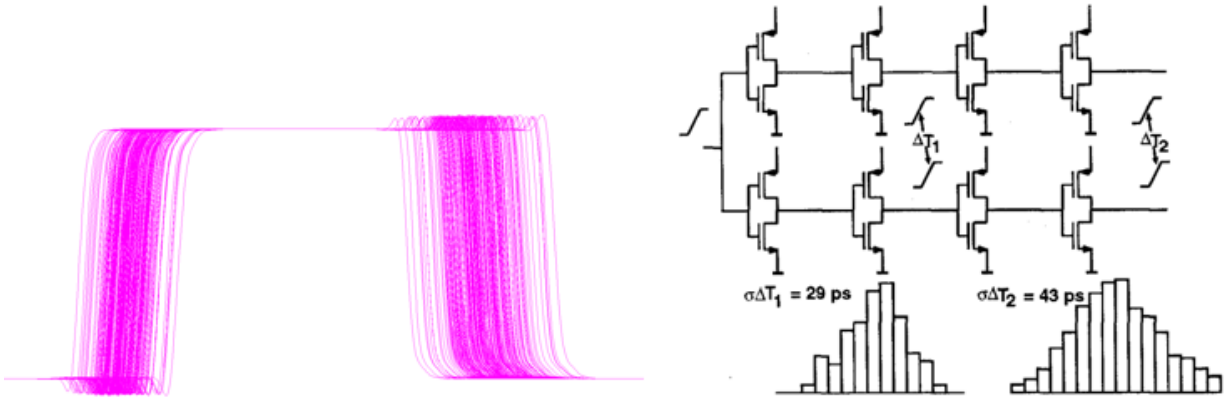


Figure 4.1.5.1 (a) Pulse variability simulation with Monte Carlo SPICE simulation. In the right, (b) Clock skew simulation for between two branches. [Pelgrom 1998]

The Monte Carlo method (MC) is a powerful tool that uses **random inputs** limited by a specific distribution, giving multiple results in order to solve problems that are difficult to answer when using a deterministic approach. In semiconductor simulations, the method is mainly utilized to analyze a probability distribution figure of merit. As **the transistor performance depends on many parameters**, computing each parameter contribution to the final performance, and the cross-parameter interactions for a circuit with more than one device can be extremely **computationally costly**. The simulations of intra-die and inter-die variations taking into account correlations are possible thanks to the Monte Carlo method. **In digital design, the variations affect the circuit timing**, and should be precisely accounted in synchronous circuits, as the **functionality depends on the clock signal synchronization**. The MC simulations can provide for example the pulse width, rise and fall time as illustrated in Figure 4.1.5.1a. The method also provides the probability function of the desired figure of merit, as the Clock skew in Figure 4.1.5.1b. This brings a powerful guidance for circuit designers, as the circuit architecture needs to cover not only the expected nominal attributed value, but also its variation.

4.1.6 Process Corners Management

In VLSI, the transistor count can surpass the 10^9 mark. Even though the MC method is suitable for a large number of input variables, **the simulation time increases with the number of devices to analyze**, becoming impractical to evaluate a circuit with billions of transistors. To overcome this limitation some technics can be employed, as simulation partitioning and **process corners**. As the multiple MC simulations have to be avoided, a **definition of extremes from MC simulations** can be used and then replicated across the entire circuit. For an example, an inverter logic gate can be simulated using the MC and have its **best and worst** delay extracted. Then for each inverter in the circuit, the results of the worst and best case can be utilized. As the corners cover almost or all possibilities between them, the method grants the **robust circuit operation** independently of the variability. The process corner can be applied for all MC simulations figure of merit, e.g. I_{ON} vs I_{OFF}, I_{NMOS} vs I_{PMOS}, frequency vs static power, etc. Several methodologies can be applied to define the corners, as for example the method in Figure 4.1.6.1. In this case, the MC simulation consists in two distributions giving dispersion from the mean; they are plotted for inverter delay versus the average energy in the switching. The corners are defined as one sigma from the V_T mean value and, in this case, this definition is enough to **englobe 95% of the points**. In some scenarios, the corner fitting may require larger corner definition, as two or three sigma corners. The importance of fitting many points as possible is due the final process yield; a robust design has the ability to handle all possible variations.

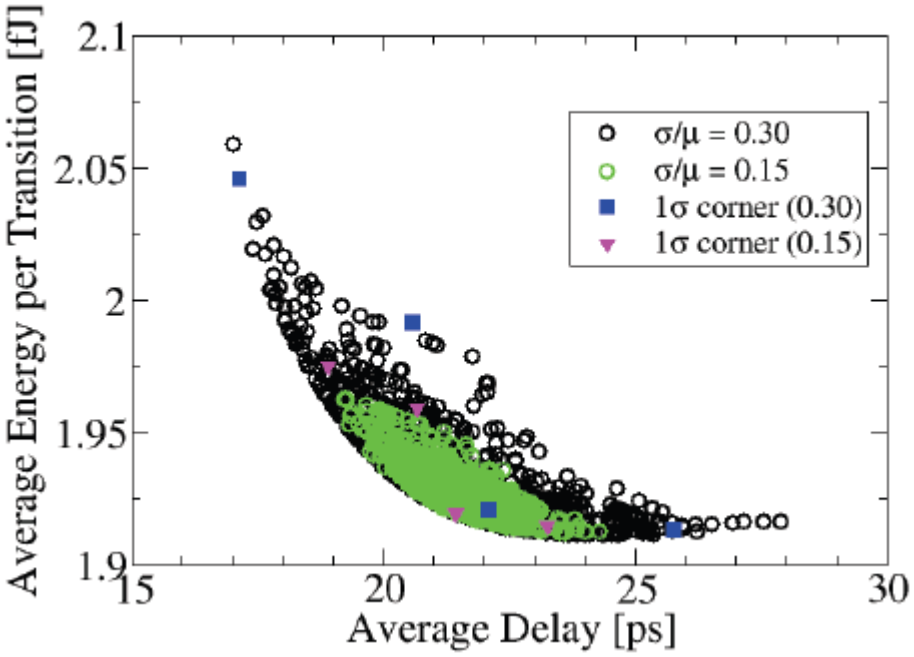


Figure 4.1.6.1 Statistical simulations for an inverter assuming two different process dispersions. In both cases, the one sigma corner edges from the mean value are enough to cover 95% of the points. [Asenov 2010]

4.2 SPICE Model Statistical Evaluation

4.2.1 Statistical Inputs

The SPICE model has **statistical parameters** that are used during a MC simulation. In simulations, where the MC analysis is not used, the simulator uses the nominal value for those parameters. The ELDO simulator from Mentor Graphics used for the simulations in this thesis, support several statistical distributions for a given parameter e.g. normal, uniform, truncated normal, weighted, and user defined distributions. A normal distribution is defined by its mean value μ and by its standard deviation σ as in (4.5), hence it can be represented as N~(μ,σ). The probability density function of a normal distribution has a bell curve shape, and the standard deviation can be understood as the percentage of values within the sigma bands as illustrated in Figure 4.2.1.1a. The one sigma band or ±σ represents 68.27% of the total distribution, ±2σ and ±3σ represents 95.45% and 99.73% respectively.

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt[2]{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (4.5)$$

The parameter variations can usually be approximated by a normal distribution as illustrated in Figure 4.2.1.1b for the measurements of the silicon thickness in a wafer. The 3σ value is around 5 Angstroms, meaning that 99.73% of the points measured are inside this tolerance value.

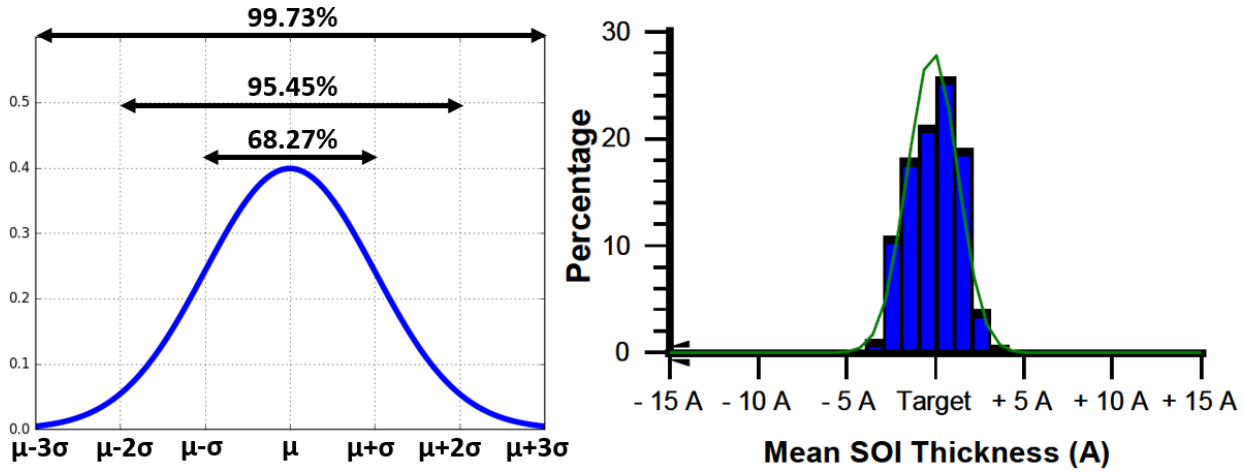


Figure 4.2.1.1 (a) Normal distribution with the sigma bands. (b) Silicon thickness distribution in a SOI wafer with a normal Gaussian fitted in green. [Schwarzenbach 2011]

The LETI-UTSOI2 SPICE model [Poiroux 2013] used in the simulations has the statistical parameters illustrated in Figure 4.2.1.2. Uncorrelated normal distributions N~(μ_{LOCAL},σ_{LOCAL}) are applied for the local variations, meaning that each transistor has a different value for a given parameter. The Table 4-I show the variations used for the flat band voltage and channel low field mobility. The variations are expressed in term of the Pelgrom variability, or A_P^2 area constant proportionality constant considering a fixed area of 1μm².

TABLE 4-I
SPICE PARAMETERS FOR LOCAL VARIATIONS IN LETI-UTSOI2 MODEL

Symbol	Parameter	α^2 Constant
V_{FB}	Flat Band Voltage	1.2 V/µm²
μ_0	Channel Low Field Mobility	57 %/µm²

Global variations are applied to the statistical parameters using normal distributions for each parameter $N\sim(\mu_{GLOBAL},\sigma_{GLOBAL})$. For the global variations, all the transistors have the same value for a given parameter during the same MC simulation run. The global variation values are shown in Table 4-II. **The local and global variations are not correlated**, for example the V_{FB} parameter has a statistical part due to the global variations and another part caused by the local variations. Changing the value of one part should not affect the other one, as they are independent**. During a simulation, the statistical input can be chosen as only global, only local or both**. This important feature allows to determine the circuit sensibility to different variation sources. The circuit simulator usually employs global and local variations in the simulations. The Chapter Five discusses the effects of local and global variations in circuits, and later the need to also consider the across-chip variations.

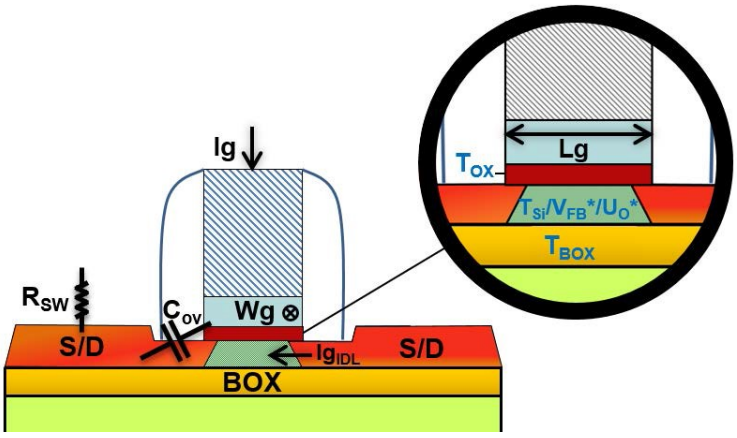


Figure 4.2.1.2 Monte Carlo simulation parameters that have a statistical model based on process analysis. Some parameters as V_{FB} and μ_0 are evaluated for local and global variations. [Ayres 2016]

TABLE 4-II
SPICE PARAMETERS FOR GLOBAL VARIATIONS IN LETI-UTSOI2 MODEL

Symbol	Parameter	Variation
T_{SI}	Silicon Thickness	$1{,}7 \times 10^{-10}$ m
T_{BOX}	Buried Oxide Thickness	$3{,}3 \times 10^{-10}$ m
L_g	Gate Length	$7{,}0 \times 10^{-10}$ m
W_g	Gate Width	$1{,}0 \times 10^{-9}$ m
V_{FB}	Flat Band Voltage	$5{,}0 \times 10^{-3}$ V
T_{OX}	Gate Oxide Thickness	$3{,}0 \times 10^{-11}$ m
μ_0	Channel Low Field Mobility	1,8 %

4.2.2 Parameter Sensitivity

The circuit sensitivity to certain parameters can be evaluated to check the **model trends**, the accuracy, and check how the statistical inputs affect a determined figure of merit. A sensitivity test is done for a ring oscillator with 3 inverters. In Table 4-III the output frequency is checked versus the nominal frequency, where the parameter under sensitivity test has it value changed to μ±3σ. In this way, the RO output frequency represents the worst and best case with a 99.73% confidence. The result shows the flat band voltage higher impact in the output frequency.

TABLE 4-III
RING OSCILLATOR FREQUENCY SENSITIVITY TO PARAMETER VARIATIONS

Symbol	Parameter	Frequency sensitivity to parameter -3σ [%]	Frequency sensitivity to parameter +3σ [%]
T_{SI}	Silicon Thickness	96.92	102.94
T_{BOX}	Buried Oxide Thickness	98.51	101.39
T_{OX}	Gate Oxide Thickness	104.73	95.42
V_{FB}	Flat Band Voltage	105.66	94.42

The static power figure of merit is also evaluated in the same simulation setup as shown in Table 4-IV. For this given analysis, the flat band voltage is shown as critical, as the worst circuit can consume 50% more power. The 3σ for the silicon thickness affects the static power by 13%, evidencing the need of rigorous management of SOI thickness during the wafer production [Schwarzenbach 2011]. Another interesting

TABLE 4-IV
RING OSCILLATOR STATIC POWER SENSITIVITY TO PARAMETER VARIATIONS

Symbol	Parameter	Static power sensitivity to parameter -3σ [%]	Static power sensitivity to parameter +3σ [%]
T_{SI}	Silicon Thickness	86.34	112.96
T_{BOX}	Buried Oxide Thickness	94.00	105.98
T_{OX}	Gate Oxide Thickness	105.67	95.27
V_{FB}	Flat Band Voltage	148.05	69.06

factor from the sensitivity analysis, is **the possibility to check the figure of merit non-linearity to a given parameter**. Comparing the symmetry of the -3σ to +3σ result, one can easily determine the linearity and quickly inspect the model. For instance, the absolute static power changes are 48% with -3σ for $V_{FB,}$ but only 31% with 3σ. This represents the I_{OFF} exponential dependency on V_{FB}, and permits to conclude that the model tendency works as expected. A limitation of this technique is the need of simulation for each circuit, since different designs can have distinct parameter sensitivity.

4.3 Chapter Conclusion

In this chapter, an introduction to variability handling in nanoelectronics is done, and the planar variability analysis tools are described. The goal of this chapter is to introduce the planar variability, whose concepts will be used for 3DVLSI variability analysis in the next chapter.

The sources of variability in planar and 3DVLSI circuits were discussed, which can be divided in systematic and random variations. In this work, in order to consider a general case, the focus was on the random variations, which depends on the process parameters and not on the physical layout.

The random variations sources can be divided in three parts:

- **Global**: The variations that equally affects all devices in the die.
- **Local**: The local variations independently affect the transistors and the amount of variability is tied to the device size.
- **Across-Chip Variations**: This component depends on the device spatial separation. It also introduces the notion of correlation distance, where close devices are correlated, and devices outside the correlation range are not anymore correlated.

In SPICE level simulations, the Monte Carlo (MC) method can be employed to verify the circuit behavior due to variations and provide a sensitivity analysis, to determine which process parameters are critical for the circuit design. Although Monte Carlo is a very powerful tool, the computational cost of the analysis is not feasible for VLSI, because of the large number of transistors. Therefore, the process corners are usually utilized in the industry.

REFERENCES

Asenov, P., N. A. Kamsani, D. Reid, C. Millar, S. Roy, and A. Asenov. 2010. "Combining Process and Statistical Variability in the Evaluation of the Effectiveness of Corners in Digital Circuit Parametric Yield Analysis." In *2010 Proceedings of the European Solid State Device Research Conference*, 130–33. doi:10.1109/ESSDERC.2010.5618458.

Ayres, A., O. Rozeau, B. Borot, L. Fesquet, and M. Vinet. 2016. "Delay Partitioning Helps Reducing Variability in 3DVLSI." In *2016 46th European Solid-State Device Research Conference (ESSDERC)*, 67–70. doi:10.1109/ESSDERC.2016.7599590.

Castaneda, G., A. Juge, G. Ghibaudo, D. Golanski, D. Hoguet, J. M. Portal, and B. Borot. 2012. "Test Structures for Interdie Variations Monitoring in Presence of Statistical Random Variability." In *2012 IEEE International Conference on Microelectronic Test Structures*, 36–42. doi:10.1109/ICMTS.2012.6190609.

Drennan, P. G., and C. C. McAndrew. 2003. "Understanding MOSFET Mismatch for Analog Design." *IEEE Journal of Solid-State Circuits* 38 (3): 450–56. doi:10.1109/JSSC.2002.808305.

Eikyu, K., T. Okagaki, M. Tanizawa, K. Ishikawa, T. Eimori, and O. Tsuchiya. 2006. "Global Identification of Variability Factors and Its Application to the Statistical Worst-Case Model Generation." In *2006 International Conference on Simulation of Semiconductor Processes and Devices*, 154–57. doi:10.1109/SISPAD.2006.282861.

Gattiker, A., M. Bhushan, and M. B. Ketchen. 2006. "Data Analysis Techniques for CMOS Technology Characterization and Product Impact Assessment." In *2006 IEEE International Test Conference*, 1–10. doi:10.1109/TEST.2006.297743.

Kuhn, K. J., M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai. 2011. "Process Technology Variation." *IEEE Transactions on Electron Devices* 58 (8): 2197–2208. doi:10.1109/TED.2011.2121913.

Lu, N. 2014. "Modeling of Distance-Dependent Mismatch and Across-Chip Variations in Semiconductor Devices." *IEEE Transactions on Electron Devices* 61 (2): 342–50. doi:10.1109/TED.2013.2283076.

Pasini, L., P. Batude, J. Lacord, M. Casse, B. Mathieu, B. Sklenard, F. P. Luce, et al. 2016. "High Performance CMOS FDSOI Devices Activated at Low Temperature." In *2016 IEEE Symposium on VLSI Technology*, 1–2. doi:10.1109/VLSIT.2016.7573407.

Paul, A., A. Bryant, T. B. Hook, C. C. Yeh, V. Kamineni, J. B. Johnson, N. Tripathi, et al. 2013. "Comprehensive Study of Effective Current Variability and MOSFET Parameter Correlations in 14nm Multi-Fin SOI FINFETs." In *2013 IEEE International Electron Devices Meeting*, 13.5.1-13.5.4. doi:10.1109/IEDM.2013.6724625.

Pelgrom, M. J. M., A. C. J. Duinmaijer, and A. P. G. Welbers. 1989. "Matching Properties of MOS Transistors." *IEEE Journal of Solid-State Circuits* 24 (5): 1433–39. doi:10.1109/JSSC.1989.572629.

Pelgrom, M. J. M., H. P. Tuinhout, and M. Vertregt. 1998. "Transistor Matching in Analog CMOS Applications." In *International Electron Devices Meeting 1998. Technical Digest (Cat. No.98CH36217)*, 915–18. doi:10.1109/IEDM.1998.746503.

Poiroux, T., O. Rozeau, S. Martinie, P. Scheer, S. Puget, M. A. Jaud, S. E. Ghouli, J. C. Barbé, A. Juge, and O. Faynot. 2013. "UTSOI2: A Complete Physical Compact Model for UTBB and Independent Double Gate MOSFETs." In *2013 IEEE International Electron Devices Meeting*, 12.4.1-12.4.4. doi:10.1109/IEDM.2013.6724616.

Saxena, S., C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky, and M. Quarantelli. 2008. "Variation in Transistor Performance and Leakage in Nanometer-Scale Technologies." *IEEE Transactions on Electron Devices* 55 (1): 131–44. doi:10.1109/TED.2007.911351.

Schwarzenbach, W., X. Cauchy, F. Boedt, O. Bonnin, E. Butaud, C. Girard, B. Y. Nguyen, C. Mazure, and C. Maleville. 2011. “Excellent Silicon Thickness Uniformity on Ultra-Thin SOI for Controlling Vt Variation of FDSOI.” In *2011 IEEE International Conference on IC Design Technology*, 1–3. doi:10.1109/ICICDT.2011.5783188.

INTRODUCTION TO CHAPTER 5

The variability of transistors, the sources of variability and the methodology to handle this issue is well understood in planar circuits, however it must be explored in 3DVLSI.

This chapter was developed by using circuit simulations, thus the consistency of the transistor statistical models is briefly checked, along the analysis of parameter sensitivity for ring oscillators.

The main goal is to show the variability behavior in 3D partitioned circuits. The analysis of partitioned RO and SRAMs was done, and the outcome is not similar to planar circuits. This work discusses how, and why the variance for determined figures of merit are different in the 3DVLSI case.

Later, the importance of treating correlation between devices is pictured. A unified statistical model was created to handle all sources of variations in a 3DVLSI environment, directly treating the correlations in the SPICE model.

In such a way, by understanding how variability affects the circuit, the 3D partitioning is shown as a feature in order to reduce the circuit variance for a given figure of merit.

Chapter Five – Variability Effects in 3DVLSI Design

5.1 Global and Local Effects in Ring Oscillators and SRAMs

This section simulates RO and SRAMs, in order to verify the statistical circuit behavior and sanity check of design environment. Then the planar circuit partitioning effects are discussed. The classical approach, using global and local variations is used. The ACV component is embedded inside the local statistical variance, however the correlation effects due distances is not evaluated, and will be later discussed in section 5.2. The circuit simulations are based on the 14nm FD-SOI, and the variability model-cards are the same as discussed in Chapter Four.

5.1.1 Planar Behavior

A ring oscillator with 33 inverters has been designed to evaluate the static power by output frequency figure of merit; as an usual way to evaluate process variation [Kuhn 2011]. In Figure 5.1.1.1, the simulation results are shown for planar RO using only local variations for the statistical parameters. The distribution has a low dispersion, both for frequency and static power. All the points are concentrated near the mean value. **The local variability causes some gates to be faster than the others. However, the particular situation of gates connected in series results in an overall averaging, in the sense that slower gates are compensated by the fast ones, converging to the process average.**

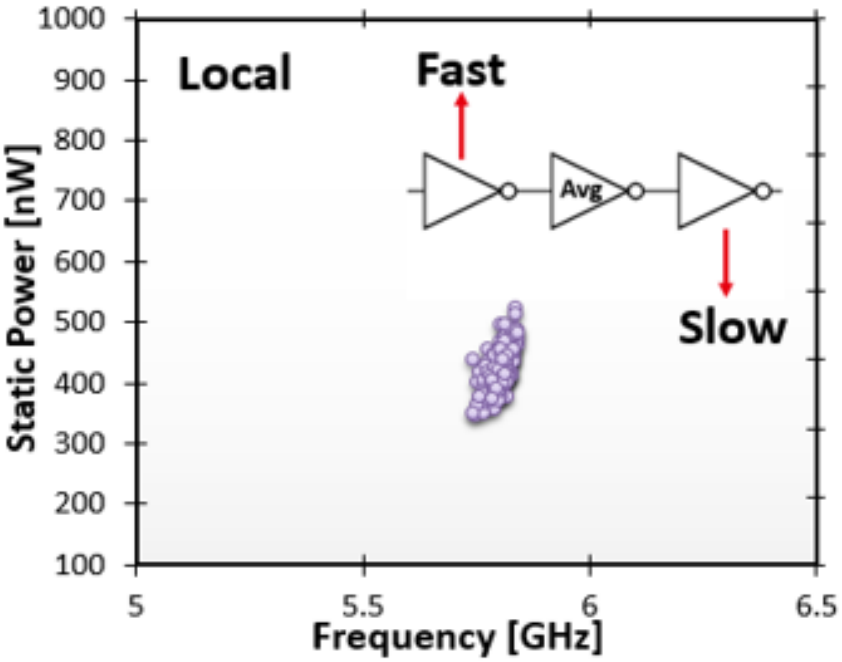


Figure 5.1.1.1 Monte Carlo simulations for planar RO with 33 inverters: Frequency vs Static Power distribution for local variations [Ayres 2016]

Using the same circuit setup, the global variations are added to the simulation, and now, the total dispersion in both static power and frequency becomes visible in Figure 5.1.1.2. **The global fluctuations cause a parameter change that is shared by all the devices.** In the circuits, where gates are connected in series, this means that all the gate characteristics are shifting in the same direction on a given MC run. Therefore, the whole performance will be impacted in the same sense, increasing the circuit sensitivity to those process variations. Another standard circuit used to benchmark the processes is the SRAM

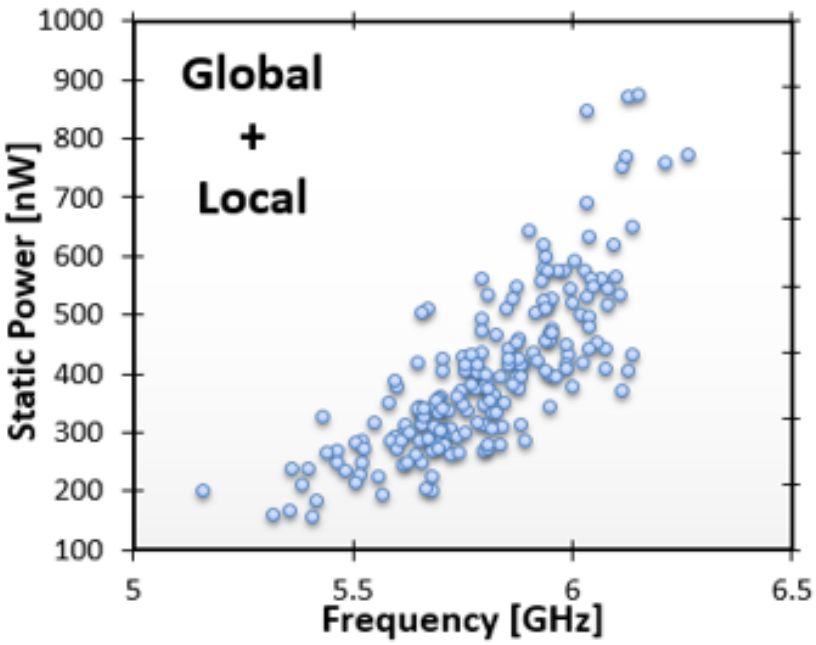


Figure 5.1.1.2 Monte Carlo simulations for planar RO with 33 inverters: Frequency vs Static Power distribution for global plus local variations [Ayres 2016]

memory, where density and speed are classical tradeoffs. A 6-transistor SRAM cell (6T-SRAM) illustrated in Figure 5.1.1.3, based on [Weber 2014] has been simulated. In addition, the variability is also an important aspect in SRAM cells, because it affects the timing and power performances, the stability and can lead to defective cells. The usual way to evaluate the **SRAM robustness** through the variability is the **design metrics** described in detail by [Guo 2009]. For example, the **Read Static Noise Margin (RSNM)** is a figure of merit that exhibits the cell capacity to hold information during a read operation. The bitcell inverter Voltage Transfer Characteristics (VTC) are plotted along each other in a way that the both inverters flipping points can be seen. This plot is also known as the butterfly curves. The RSNM is characterized by the side of the largest square confined inside the VTC of the same bitcell. **Appendix B** shows the Python code and SPICE netlist to extract the RSNM. The RSNM of an array or several MC runs can be defined by the lowest value, representing the worst-case scenario. **Opposite to ROs, the SRAM bitcells do not have an averaging effect**, and for a given process their characteristics are mainly defined by the transistors sizing (length and width). In this fashion, **the local variability plays a huge role compared to the global variations** because the random local variations individually affect each transistor forming the bitcell. Moreover, the small transistor size also contributes to make predominant the local variations. **The global variations affect all transistors in the bitcell, keeping the important design ratios for RSNM almost unchanged.** For example, the pull-down to pass transistor width ratio will remain the same, as all NMOS have the same value for global variations. Figure 5.1.1.4 shows the RSNM calculated for the same cell using a SPICE model with global and local parameter variations using one hundred MC simulations for each VTC. Two main consequences are observed for local variations during a bitcell simulation. It is required to simultaneously simulate both inverters from the bitcell and the correlation between them has to be taken into account. Otherwise, the VTC curves symmetry towards each other will not be accounted, and this important effect is discussed in section 5.2. The final consideration about the RSNM is the high sensitivity to the parameter variance, especially V_T as reported in [Kurude 2016].

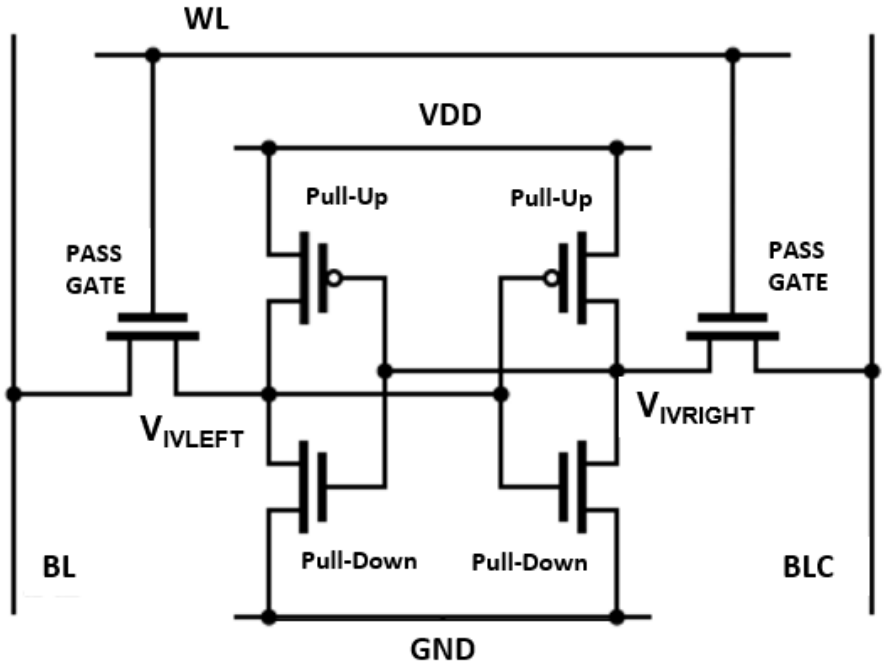


Figure 5.1.1.3 SRAM bitcell with 6 Transistors. The VTC curves are extracted from $V_{IVRIGHT}$ and V_{IVLEFT} nodes. Write-line is referenced as WL, Bit-line as BL and Complementary Bit-line as BLC.

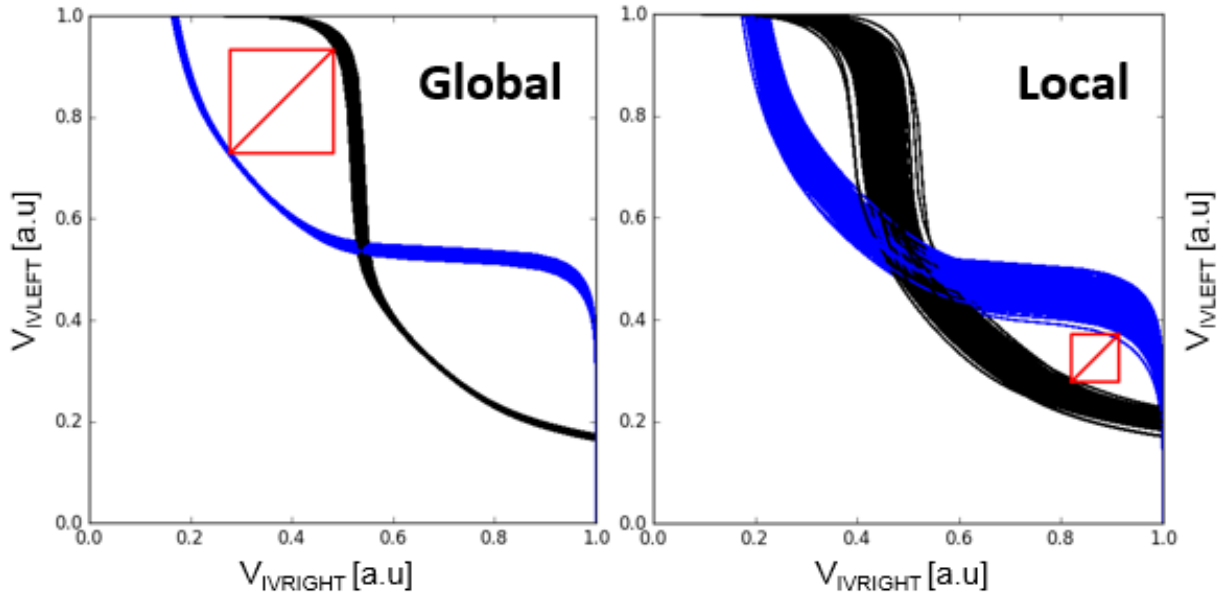


Figure 5.1.1.4 Monte Carlo SNM simulations for the same planar SRAM bitcell. On the left only using global variations. On the right, a simulation for local variations

The first order variability evaluation of planar ROs and SRAMs are done in this section. The next section will reassess the variability for those circuits partitioned using two-tiers 3DVLSI.

5.1.2 3D Partitioning Effects

Partitioning is peculiarity of 3D design, which means that the netlist has to be separated into tiers. The area ratio between tiers does not necessary needs to be equal, for example, in a two-tier integration the top tier can have 60% of the gates while the bottom tier only has 40%. Several RO have been created with different partitioning ratios. For each tier, a parameter P has a Gaussian N~(μ,σ) distribution model for global and local variations and considering no correlation from the top to the bottom tier ($\boldsymbol{\rho_{t,b}}$ =0). The simulations have been performed using the partitioning ratio given by the number of inverters on the top tier over the number of inverters on the bottom tier. Notice that, in this particular case of ROs, the area partitioning ratio between the top and bottom tiers is the same of the delay partitioning ratio because inverters have identical performance on top and bottom tiers [Batude 2015] and the same area. Hence in (5.1), the final average is not affected by partitioning. The output frequency average and standard deviation (σF) are observed on each partitioning ratio as seen in Figure 5.1.2.1 (each point is a MC run for a different partitioning ratio).

$$\mu = \sum p_i \, \mu_i \qquad (5.1)$$

Adopting $\boldsymbol{p_t}$ as the weight ratio of delays in the top tier and $\boldsymbol{p_b}$ = 1-$\boldsymbol{p_t}$ as the delay ratio in the bottom tier as defined in (5.2) for two-tier integration. The total $\boldsymbol{\sigma_{F3D}}$ dispersion is reduced when the circuit delay is equally partitioned between the two tiers. An analytical model has been proposed in [Ayres 2016] which describes the frequency variability as a weighted sum of normal variables as in (5.1), with the total variance described as in (5.3). The particular example of frequency was shown; however, the general formula works for all global parameters variations, which are used in a partitioning in a 3D partitioned circuit.

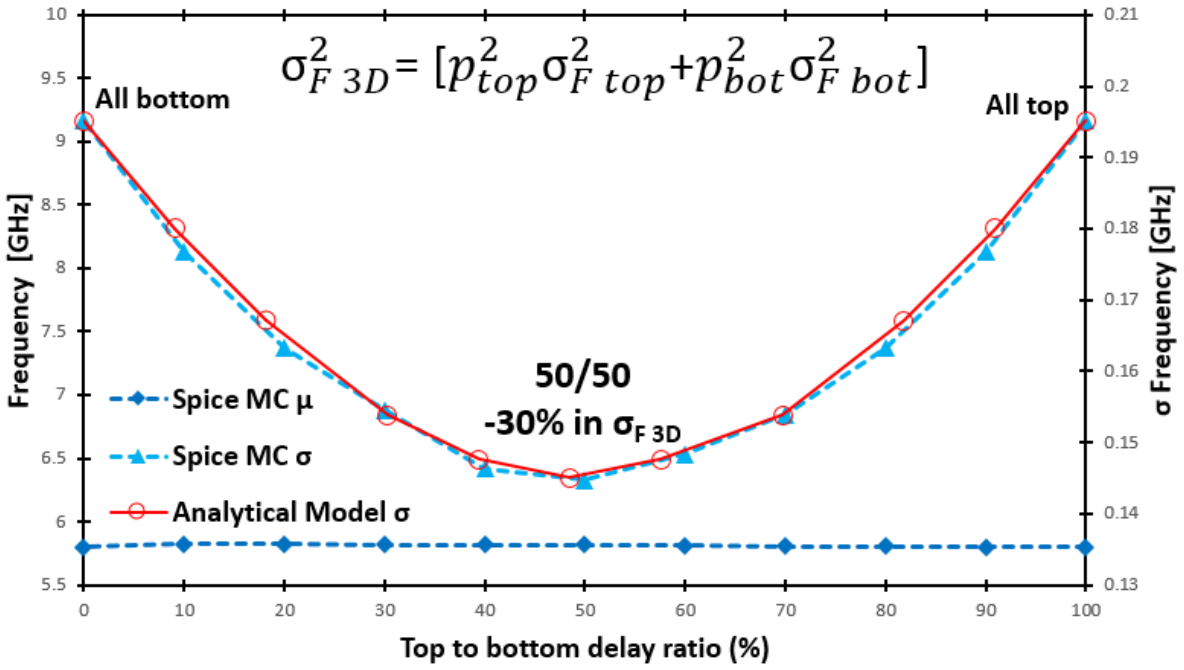


Figure 5.1.2.1 Partitioning effect on RO's output frequency. The dashed line is the mean and stay constant for all partitioning ratios. The variance is at minimum for 50/50 partitioning.

$$\sum p_i = 1 \tag{5.2}$$

$$\sigma^2_{3D} = p^2_{top}\sigma^2_{top} + p^2_{bot}\sigma^2_{bot} + 2\rho_{top,bot}p_{top}p_{bot}\sigma_{top}\sigma_{bot} \tag{5.3}$$

The physical interpretation of this effect is that each global tier variation is tangled to another one, **reducing the probability of the worst/best case at the same time** as illustrated in Figure 5.1.2.2. By contrast a planar circuit is always tied to the same global variations. **The effect is maximized for uncorrelated processes** for which one-tier statistical variance is not related to another tier. Otherwise the correlation $\rho_{t,b}$ has to be taken into account in the third term in (5.3). Such data can be a guideline for 3DVLSI process development, as the tiers process variations should stay uncorrelated. A further simulation was done in the same setup, however using the hypothetical case of coupled statistical models for the tiers. This case means that the processes are correlated and the process dispersion in top tier is exactly the same as in bottom tier as illustrated in Figure 5.1.2.3. The frequency distribution remains unaltered whatever the partitioning ratio as shown in Figure 5.1.2.4. This outcome expresses the need of uncorrelated processes variations in order to deliver unconnected statistical parameters. The probability to obtain uncorrelated processes in the same path is harder to achieve in planar circuits, even impossible in advanced nodes due the nanoscales, hence a unique opportunity to exploit this effect in 3DVLSI. The frequency dispersion reduction can be viewed as a reduction of the design corners as shown in Figure 5.1.2.5.

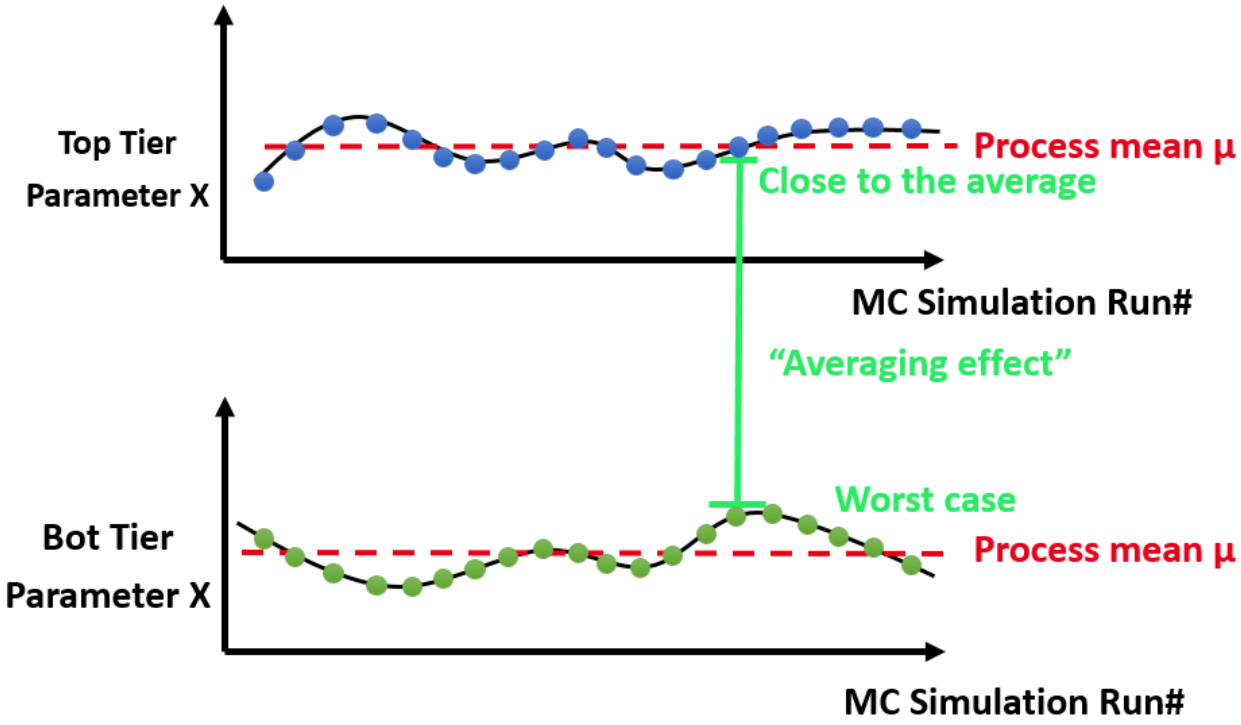


Figure 5.1.2.2 Planar circuit parameter variation depending on the MC run. If the tiers are uncorrelated, the global variation hardly will cause one parameter be at the worst/best situation at the same time. Thus, there is an averaging effect.

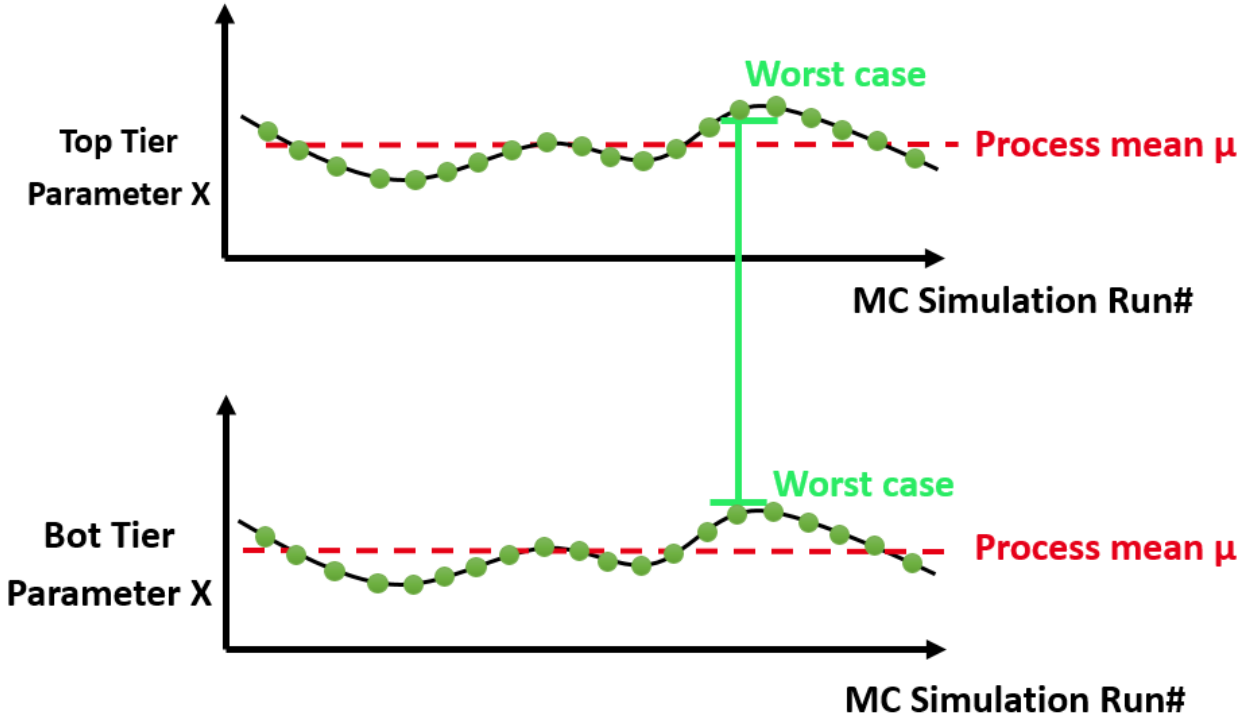


Figure 5.1.2.3 Planar circuit parameter variation depending on the MC run. If the both tiers are totally correlated, there is no averaging effect.

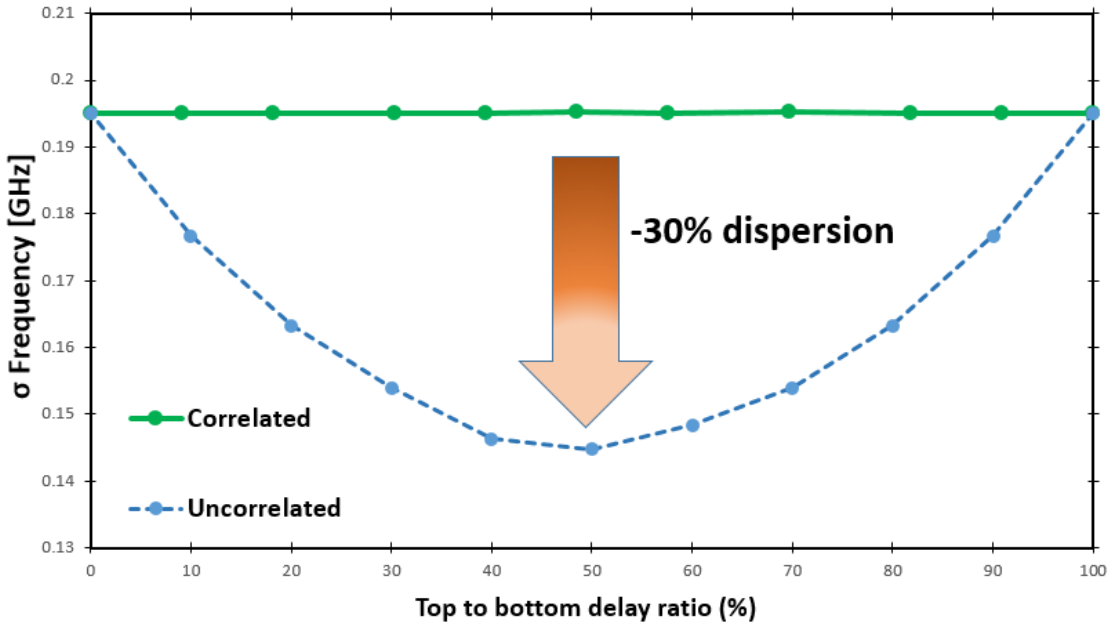


Figure 5.1.2.4 A fully correlated process compared to an uncorrelated process in a Monte Carlo Spice simulation. The frequency dispersion mitigation can only be achieved if the $\sigma_{F\ TOP}$ and $\sigma_{F\ BOT}$ are not correlated.[Ayres 2016]

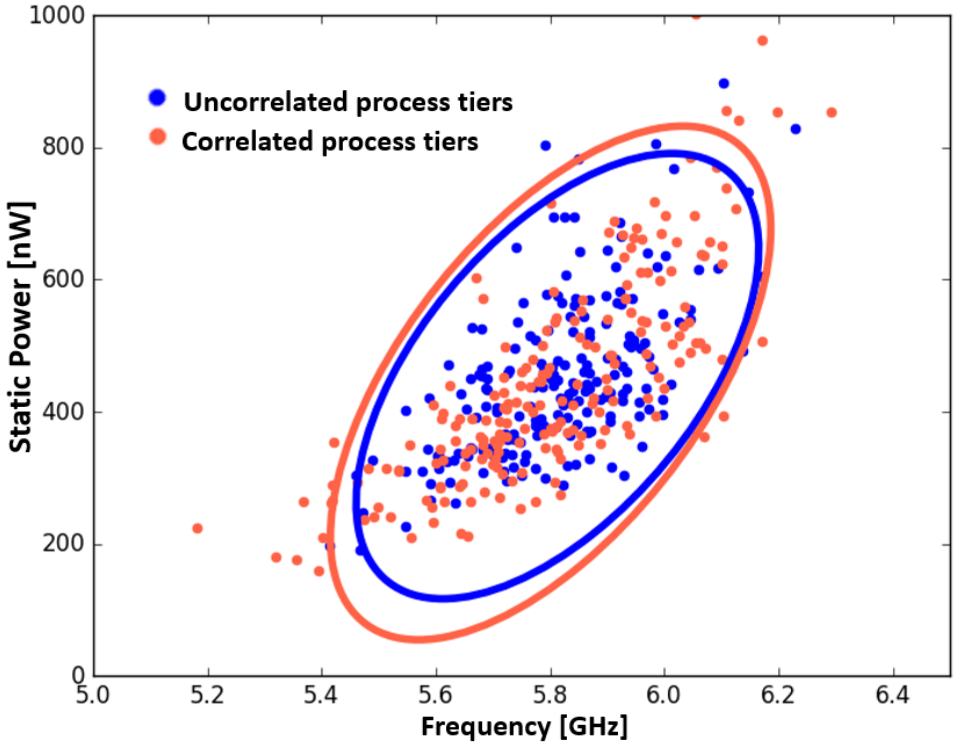


Figure 5.1.2.5 Monte Carlo simulation plot for uncorrelated case in blue and fully correlated process in orange. Confidence ellipses are plotted to fit 95% of the points. Lower dispersion decreases the ellipse area and consequently the corners, enhancing the circuit design. [Ayres 2016]

Expanding the analytic model (5.3) to a general form for N-tiers is given as (5.4). For an example, a 3-tier circuit, which has parameter correlations between tiers, has a non-null covariance term in (5.4). Considering the weights (p_a, p_b, p_c) and a parameter random variable (e.g. t_{si} thickness) for each tier as X, Y and Z, the second term in (5.4) can be written as shown in (5.5).

$$\sigma_I^2 = \sum p_i^2 \sigma_i^2 + 2 \sum_{i,j:\, i<j} p_i\, p_j \sigma(X,Y) \tag{5.4}$$

$$\sigma_I^2 = 2p_a^2\sigma_b^2 COV(X,Y) + 2p_b^2\sigma_c^2 COV(Y,Z) + 2p_c^2\sigma_a^2 COV(Z,X) \tag{5.5}$$

The partitioning effect can produce an additional boost predicted by the model. Indeed, increasing the number of uncorrelated process tiers and designing the delay path allocated equally in N-tier as shown in Figure 5.1.2.6 would reduce the frequency dispersion.

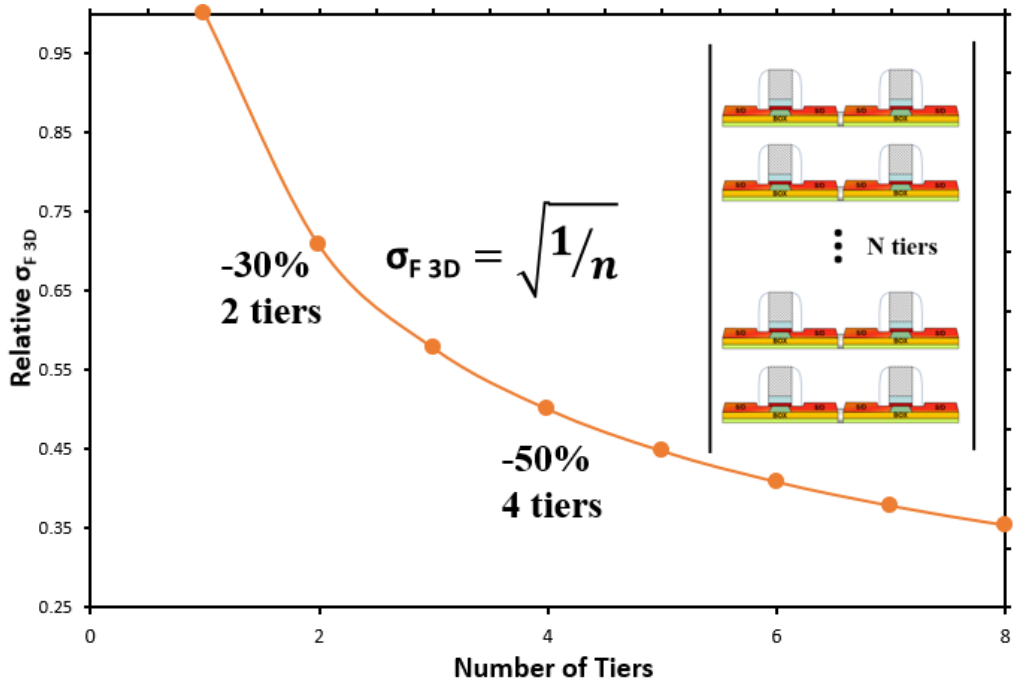
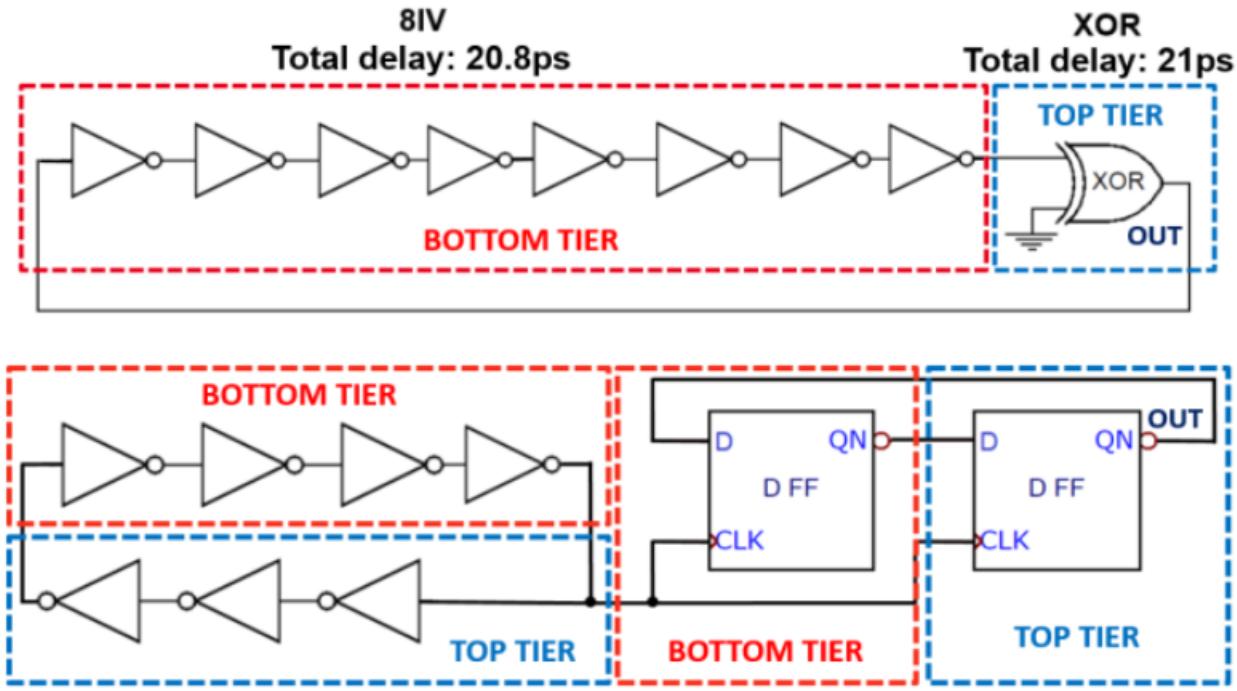


Figure 5.1.2.6 The analytical model for 3D uncorrelated processes considering N tiers can reduce further the relative frequency dispersion compared to a planar circuit. Maximum theoretical gains are shown [Ayres 2016]

Circuits based on RO were tested to verify the impact of the delay ratio partioning between tiers as shown in Figure 5.1.2.7. In those circuits, the area ratio between the top and bottom tiers are different as well as the number of transistors; however, by characterizing the gates, it is possible to determine the delays, and then design a balanced delay circuit equally splitting the delays between tiers. The first circuit uses a XOR gate in series with inverters. As the XOR delay is equivalent to eight inverter delay, thus the circuit delay partitioning is balanced. Comparing the frequency output dispersion of the planar case to the 3D balanced delay one, **a 30% reduction in the frequency dispersion is observed**. Another circuit example which

benefits from similar reduction in dispersion are the D Flip-Flops (FF) connected in series with a clock generated by a two-tier ring oscillator. Contrarily to the RO presented before, the inverter chain with the XOR gate does not reduce the static power variance by 30% because the area and the number of transistors is not balanced between tiers.



CIRCUIT	TOP TO BOTTOM SURFACE%	TOP TO BOTTOM DELAY %	OUTPUT FREQUENCY REDUCTION
RO	51,51	51,51	-30%
RO with XOR	66,66	49,76	-30%
RO + FF chain	50	50	-30%

Figure 5.1.2.7 Circuits with different area partitioning and delay partitioning. The main factor to reduce the frequency dispersion is the delay partitioning.

The 3DVLSI partitioned ROs have their variability accessed in this section, mainly considering the global variations, which is the main source of impact for this case. The local variability affects both tiers, and can't be reduced by employing partitioning. This section considers global and local variations (with ACV dispersion included, but no correlation evaluation). However, the ACV source of variation has to be properly accounted for circuits which global variations are not the main factor in variability, and then the device correlation due distances has to be accessed for an accurate evaluation. The next section will discuss a unified model considering all sources of variation with ACV correlations and will evaluate ROs and SRAMs variability.

5.2 Statistical Unified Model

5.2.1 Model Definitions

As shown in the last section the 3D partitioning can reduce a parameter dispersion, however for circuits where local variability is the main component on the final performance, the correlations integrating distances in their *formulae* should be evaluated. Therefore **a 3D unified statistical model has been developed, where global, local and ACV sources can be accounted** as in (5.6). In this section, the simulations are done explicitly separating the local component from ACV, and considering the ACV correlations. For the same tier, the ACV behaves like a planar circuit. However, considering a 3D circuit, the top tier is not correlated anymore to the bottom tier. In this chapter, the ACV modeling is based on [Lu 2014], which defines six properties for an accurate mismatch model:

1) Mismatch cannot be approximated by a series of step functions (excluding grid approaches, as the non-continuity brings errors in small scales);

2) The mean value of a given transistor parameter is constant;

3) The standard deviation of a parameter is constant (the parameter μ and σ are always the same, independent of the wafer position);

4) Pelgrom's variability still holds true for every device and device pair;

5) For distant devices, when the pair distance increases, the rate of mismatch increase should be reduced (even for very large distances, the amount of mismatch is finite);

6) The solution has to be compact to be implemented in SPICE models.

$$\sigma^2 \equiv \sigma_{GLOBAL}^2 + \frac{1}{2}\sigma_{LOCAL}^2 + \sigma_{ACV}^2 \tag{5.6}$$

The dependence of the ACV correlation and the distance is derived by [Lu 2014] which is compliant with a Monte Carlo implementation in SPICE modeling. The 2D random spatial frequency models for a parameter P mismatch is given in (5.7) and (5.8).

$$P_i = \mu + G_0\sigma_{GLOBAL} + \frac{g_i}{\sqrt{2}}\sigma_{LOCAL} + \sigma_{ACV}\sum_{n=1}^{M} a_n[g_{n,1}cos\varphi(z_i) + g_{n,2}sin\varphi(z_i)] \tag{5.7}$$

$$\varphi(z_i) = \frac{x_i g_{n,3}}{d_n} + \frac{y_i g_{n,4}}{s_n}, i = 1,2,\dots N \tag{5.8}$$

Where G_0, g_i, and $g_{n,1}$, $g_{n,2}$, $g_{n,3}$, $g_{n,4}$ are independent normalized random variables and a_ns are normalized weights. M is the total number of independent random variables used. The model is powerful because it **describes the correlations using a sum of normalized Gaussian contributions from different spatial frequencies**, which are implemented in a circuit simulation environment. This model is also translational invariant, meaning it **does not depend on the absolute value of the device position** $z_i=(x_i,y_i)$ and more

importantly**, it is a continuous model**. The model has been developed further by [Poiroux 2015], creating a unified statistical model compatible with SPICE modeling for planar circuits. The unified model has been expanded for 3D environment, describing the device correlations in (5.9), which contains two statements:

- If the devices g and h are on the same tier, then the pair correlation deduced for planar is used.
- Else, if the devices are in different tiers the correlation is zero.

$$\begin{cases} \rho_{G,H} = e^{-\pi P_x^2/(D_G^2+\pi\Lambda_{xi}^2)} e^{-\pi P_y^2/(D_H^2+\pi\Lambda_{yi}^2)} & if\ T_G = T_H \\ \rho_{G,H} = 0 & if\ T_G \neq T_H \end{cases} \quad (5.9)$$

In the formula above, P is the pair distance, D is the device size and Λ is the physical correlation length. As previously discussed in (4.6) the model keeps independent the local variance from the correlation. The Poiroux's model is used for devices in the same tier, describing all variations sources and capable to generate Pelgrom's plot; accurately modeling the local variations. The model uses a FD-SOI technology, but this unified model is easily extendable to other technologies. Extracted from [Poiroux 2015], a correlation length component of 2.8 µm is used. **The correlation between devices is evaluated in the circuit netlist** using (5.9). For the devices on the same tier, the correlation varying with the distance is shown in Figure 5.2.1.1. The correlation range is treated as isotropic, meaning that the correlation range in the x direction is the same as in the y position for devices situated in the same tier. This plot confirms the veracity of the model, as the curve shape describes the ACV variation in planar circuits as discussed in Chapter Four Figure 4.1.4.2. Notice that the pair correlation tends to be zero as the distance increases. The correlation for the pair is used as input for the transistor model and is handled directly in the circuit simulator as described in Figure 5.2.1.2.

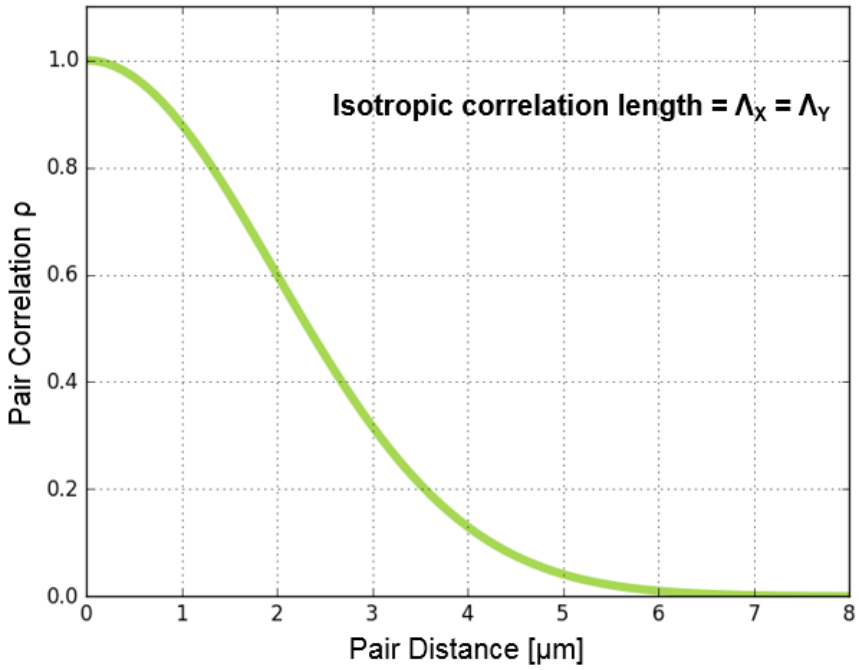


Figure 5.2.1.1 Device pair correlation varying with the distance for a fixed transistor size and isotropic correlation length.

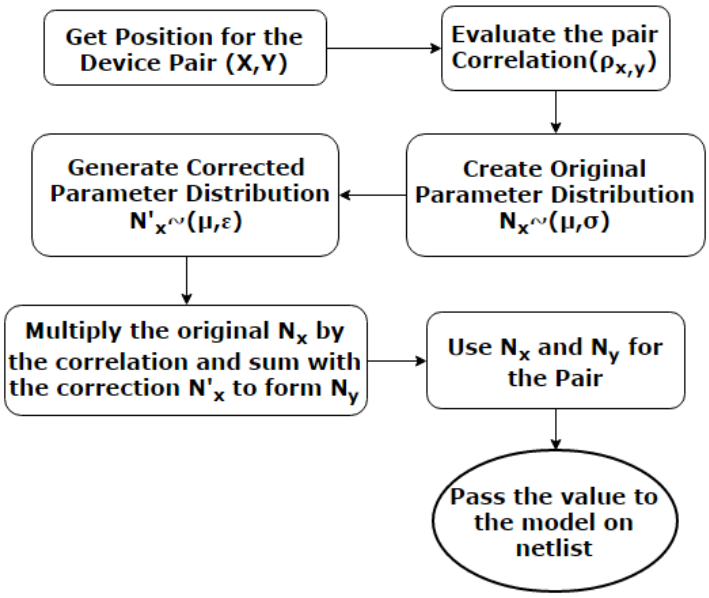


Figure 5.2.1.2 Circuit simulation flowchart treating ACV correlations. The handling is entirely done inside the circuit netlist and SPICE model. The correlation is created using a corrected distribution, and finally keeping the original parameter variance.

A procedure to create a correlation between two parameters N_x and N_y **directly in** the **simulator** is used: the initial parameter distribution is created, and then another Gaussian distribution is created using a corrected variance. Finally, the initial and corrected distributions are joined to form the Gaussian distribution with the original sigma and desired correlation. The code for python and SPICE netlist is illustrated in **Appendix B**. For optimization purposes, if the devices are in different tiers, the parameter distribution can skip this procedure and be declared as independent from each other directly in the netlist. In this manner, the model handles all component variations in (5.6) in the SPICE unified model, including the 3D partitioned circuits. The entire process does not require intensive computing, can be scaled up, and avoid the classical way to implement correlations using matrix as described in [Conti 1999], which is hard to implement directly into circuit simulations.

5.2.2 Ring Oscillators Sensibility to Different Sources

Each component in (5.6) can be separated from each other, allowing a component analysis with SPICE simulations. To illustrate this property and the circuit sensitivity to each component, planar Ring Oscillators with 17 inverters have been simulated. Three cases were considered: 1) Only Local variations 2) Local and across-chip variations, accounting the correlations and 3) All considered sources using the SPICE unified model. **All the inverters are not correlated, this hypothesis means that all the inverters in the chain are far from each other**, or at least 7µm distant. The output frequency variance is shown in Figure 5.2.2.1. Monte Carlo simulations were done using 200 random draws. Global parameter variations are the main contributor to the RO frequency variations. The frequency histogram results in a Gaussian distribution as in [Pelgrom 1998].

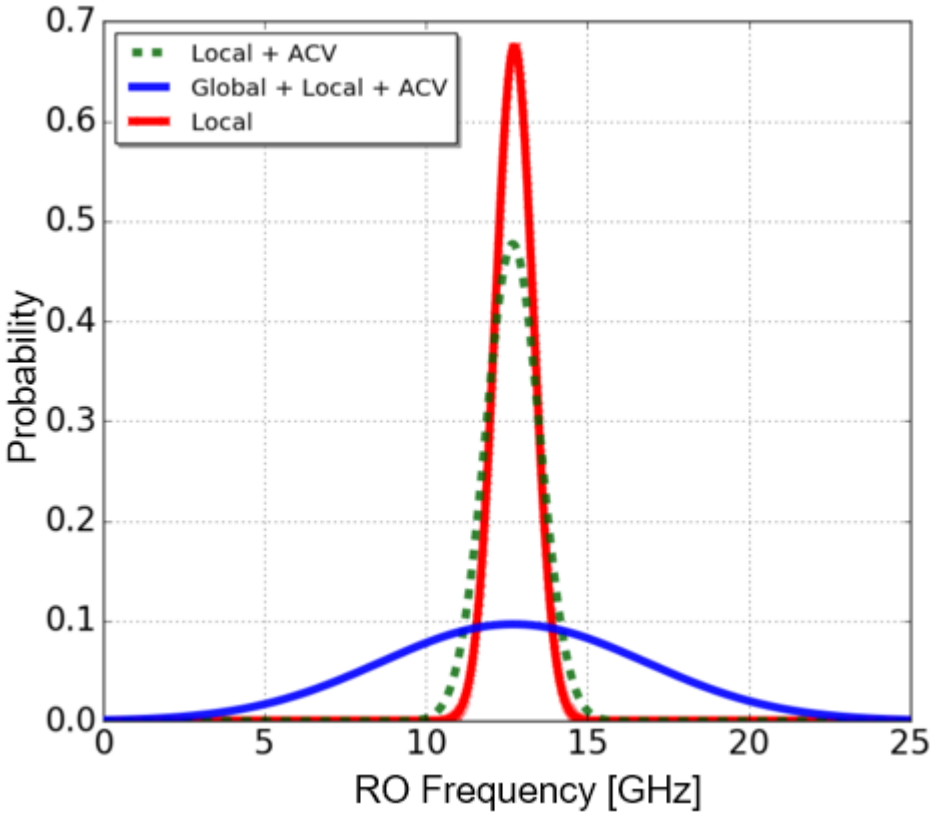


Figure 5.2.2.1 Output frequency normalized distribution for planar RO with 17 inverters. Enabling different variations sources, the global variability is the most sensitive component for ring oscillators.

Another simulation using the same setup was done with ACV correlations plus local variations for the RO. In one case, all the inverters **are not correlated as before**, while in the second case the inverters are all ACV correlated, **depicting a more realistic case, where inverter are close to each other in a RO**. The results are illustrated in Figure 5.2.2.2. The inverter lack of across-chip correlation makes the output frequency less dispersed than the RO with all the inverters correlated. As discussed before, the RO has an "averaging" property. When all **inverters are correlated**, this **averaging effect will no longer work, as the inverter variations are in the same direction**. This outcome is analog to the partitioning effect on the ring oscillator described in Figure 5.1.2.1. With no correlation between inverters, the averaging effect is fully exploited and the output frequency distribution has a lower standard deviation. In planar, the distance required to

have uncorrelated gates are almost unfeasible. However, this can be achieved in a 3DVLSI integration using different tiers. In this case, the output frequency has a lower dispersion from ACV component.

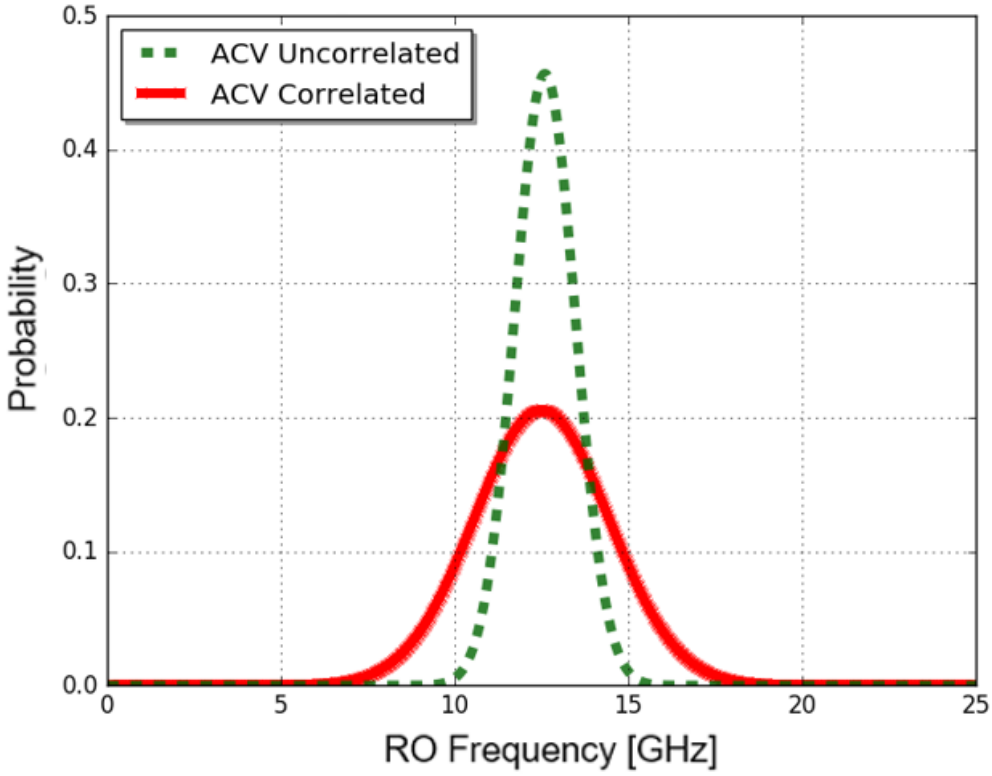


Figure 5.2.2.2 Output frequency normalized distribution for planar RO with 17 inverters. Comparison of correlation effects using ACV statistical parameters.

5.2.3 3D Partitioned SRAM Variability

Considering the FD-SOI technology, we made the assumption that the flat band voltages (V_{FB}) of NFET and PFET from the same tier are also uncorrelated, because the gate stack in this technology can be separately built for the transistors. The unified model can also accept the correlation between parameters as described in [Mazurier 2011] and [Eikyu 2006] where the drain current in different regimes is correlated to transistor V_T and R_{ON} fluctuations. In Figure 5.2.3.1 , we compared a planar SRAM simulation to a 3DVLSI SRAM simulation using the unified model. The SRAM layout and the transistor distances have been extracted from [Weber 2014]. The SRAM layout is shown in Figure 5.2.3.2. On the left upper side, we illustrate a typical planar 6T-SRAM. The V_{FB} parameter is plotted representing the third term in (5.6), namely the ACV. Then each NFET transistor is plotted against another NFET showing the pair correlation. In the planar SRAM, the NFETs pairs M3/M7 and M6/M1 stay correlated with the highest value (r=1) to each other because of the very small distances, while the pairs

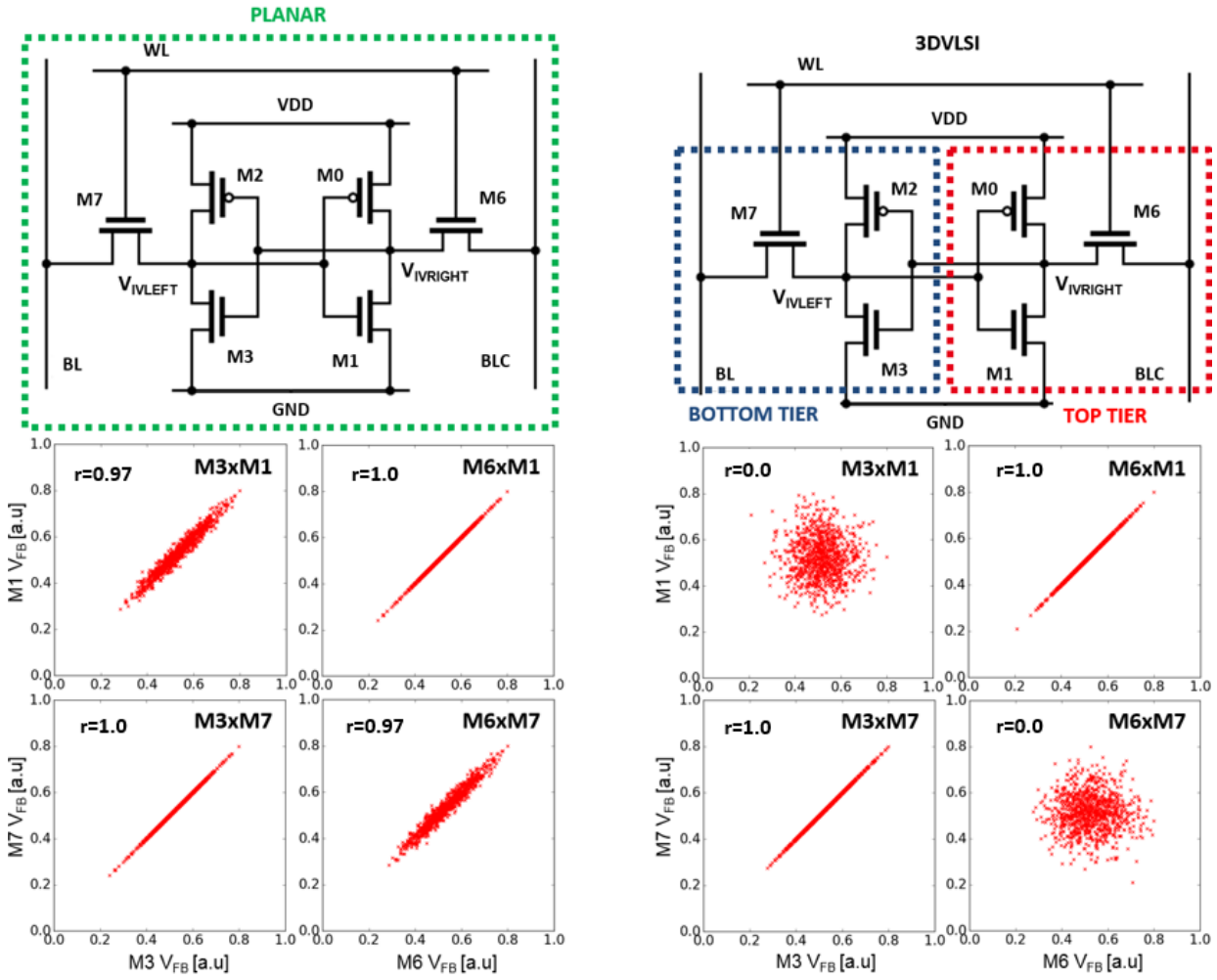


Figure 5.2.3.1 ACV correlations for VFB parameter of NFETs in an SRAM. On the left (9.a) a planar SRAM with all transistors in the same tier. On the right (9.b) the SRAM is partitioned into two tiers.

M3/M1 and M6/M7 presents a lower correlation because of the distance separation, obtained from Figure 5.2.1.1. The PFETs were not shown but also stay correlated (r=1) to each other. Additionally, an

SRAM partitioned into two tiers is simulated. The partitioning is symmetrically splitting the SRAM in equal area and functionality, as one pull-up, pull-down and pass-gate transistors remains in one tier as shown in the right upper side of Figure 5.2.3.1. Again, the V_{FB} parameter is plotted in the same way of planar circuits for the 3D case. **The NFETs in the same tier stay correlated to each other** as before. By comparing the V_{FB} parameter of **NFETs in different tier, the total lack of correlation (r=0) is seen** as defined in (5.9). Also, due this partitioning the PFET parameters are no longer correlated. This outcome has a tremendous impact in the design of the SRAM, meaning that the inverters with the pass-gate transistor are no longer correlated, and the **VTC curves are not anymore symmetrical** (although the local mismatch between NFETs and PFETs is still present in both cases). Further, the ACV physical uncorrelation caused by the tier alteration can be seen as the global uncorrelation presented in ring oscillators. We observe this reproducible effect in a planar circuit by placing the SRAM transistors away from each other, as explained in [Lu 2014], but the required distances shown in Figure 5.2.1.1 to reduce the correlation to zero, will cause a density loss and a degradation in performances due to higher WL routing for a bitcell, hence impractical on real circuits.

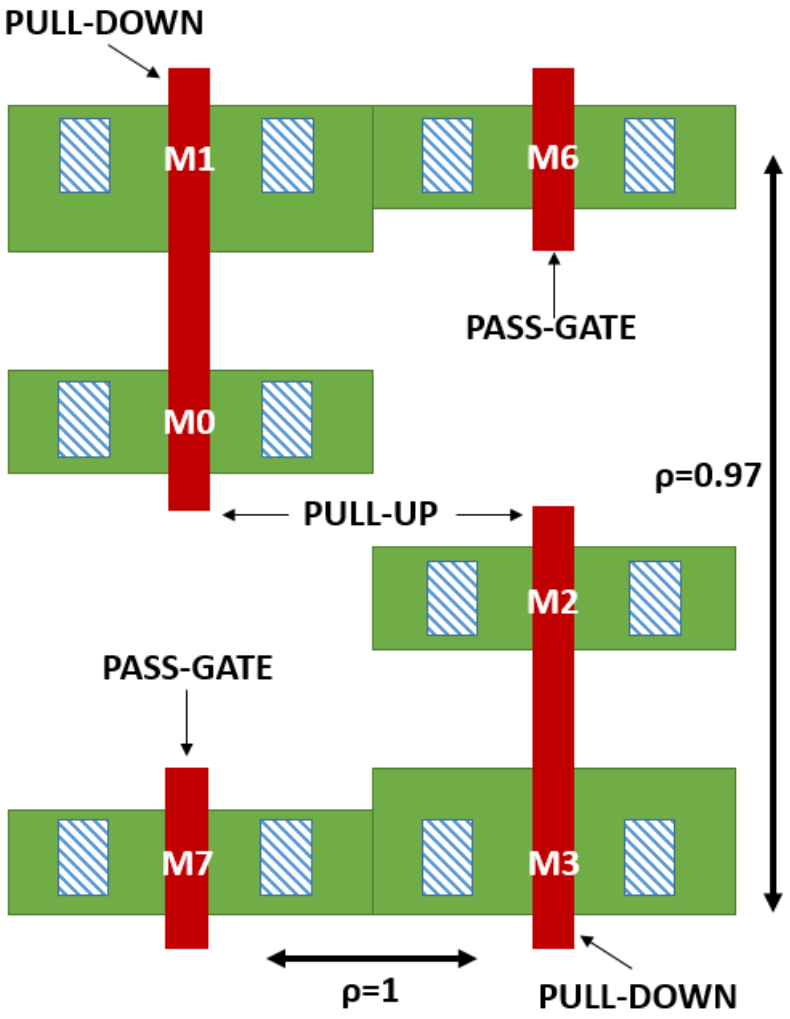


Figure 5.2.3.2 Planar SRAM layout. The NMOS transistors pair are close, thus the Pearson correlation is one. Due to correlation range of 2.8µm, the Pearson correlation between M7/M3 to M1/M6 is 0.97.

5.2.4 SRAM Static Noise Margin

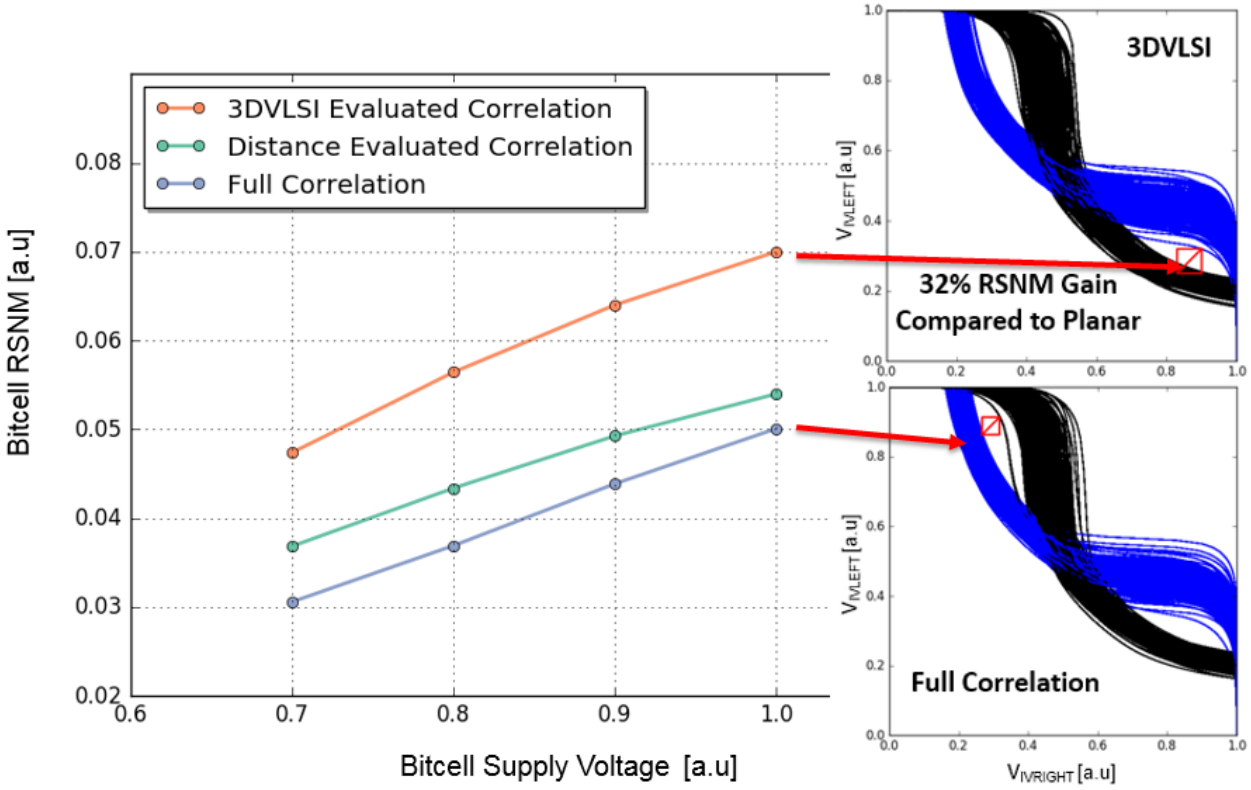


Figure 5.2.4.1 SRAM SNM simulated using the unified model. The SNM increases if the voltage nodes V_{IVLEFT} and $V_{IVRIGHT}$ have lower correlation. The planar case considering correlations has better SNM than a hypothetical planar full correlated case.

The results of the RSNM versus the supply voltage using the unified model considering global, local and ACV variations are shown in Figure 5.2.4.1. Those simulations were done using one thousand MC runs. Three test cases have been simulated, two for a planar SRAM, where one considers the ACV distance, and the other just implements a hypothetical case using fully correlated parameters (r=1). The final case considers the 3D SRAM partitioned and considers the ACV correlations caused by the distance as shown in Figure 5.2.3.2. **The planar case considering the ACV correlations have 8% better RSNM than the fully correlated planar case**. This result shows the need to simulate the ACV correlations for SRAM planar circuits, especially for correlation lengths in the same order of the bitcell size. The need to implement ACV correlations for other planar circuits has already been demonstrated in [Poiroux 2015], for example in digital-to-analog converters. **Partitioning increases the RSNM by 32% for the 3D case compared to the planar case with correlations.** As the uncorrelation between the two SRAM parts decreases, **the chances of the VTC being the worst at the same time are reduced**. In other words, **partitioning lowers the probability to simultaneously have the both worst flipping points.** During the read operation, the voltage node V_{IVLEFT} (defined by Figure 5.2.3.1) is determined by the voltage divider formed between M7 and M3 as explained in [Guo 2009]. The other side voltage node $V_{IVRIGHT}$ (defined by Figure 5.2.3.1) is determined by the transistors M1 and M6 forming the voltage divider as illustrated in Figure 5.2.4.2. As the pair M7/M3 is not correlated to the pair M1/M6 pair, the voltages V_{IVLEFT} and $V_{IVRIGHT}$ are no longer correlated, granting higher lobules in the butterfly curves. The 3DVLSI becomes very attractive for a SRAM design requiring low voltage operation and using technologies with high variability as it can increase further the SNM.

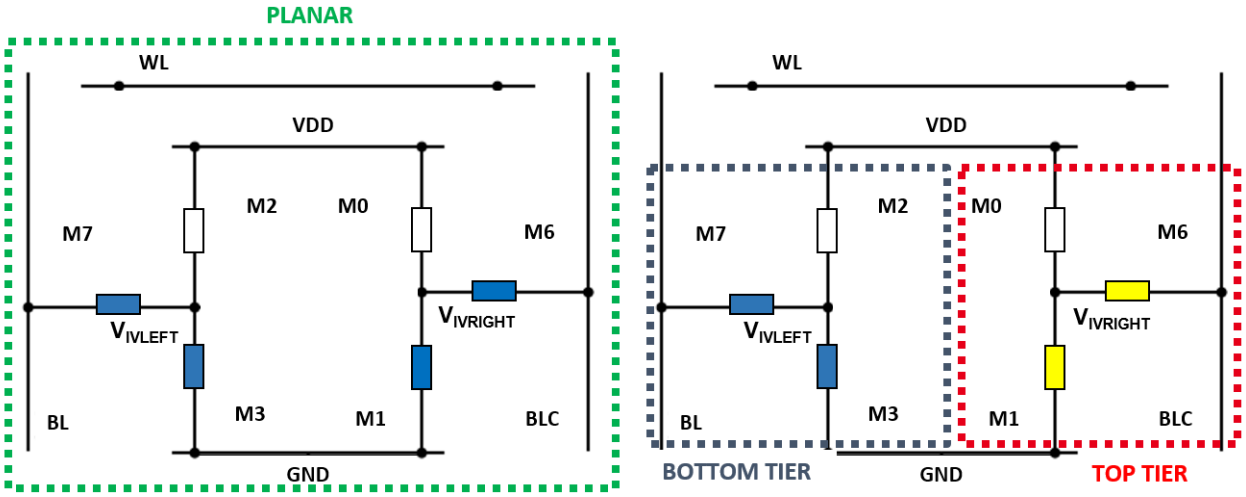


Figure 5.2.4.2 ACV correlations of NFETs in an SRAM during a read operation. On the left, a planar SRAM with all transistors in the same tier. On the right, the SRAM is partitioned into two tiers.

Besides the read static noise margin, another usual figure of merit of SRAMs is the Write Noise Margin (WNM). As commented by [Guo 2009], the WNM indicates the ability to write a SRAM cell. If this figure of merit drops below zero, it becomes impossible to write a bit into the bitcell. **The WNM is quantified by the side of the smallest square inside the read and write VTC at the same bitcell**. The write VTC can be simulated in SPICE by sweeping the voltage in V_{IVLEFT} with BL and WL tied to VDD while BLC is connected to ground. The WNM square should be evaluated at the lower part of the curves, below the half of supply voltage. At the event of a writing, the pass-gate and the pull-up transistor forms a resistive voltage divider. If this voltage divider pulls the storage node ($V_{IVRIGHT}$ or V_{IVLEFT}) below the inverter trip point, the bit is successfully written in the SRAM. The SRAM simulations have taken an initial hypothesis that the NMOS is not correlated to the PMOS in the planar case, neither in the 3DVLSI case. In FD-SOI technology this outcome may be possible because the P and N gate stacks may differ. Even though, **there is a correlation between two PMOS (pull-ups) for the planar case, while in the partitioned case the PMOS are not correlated anymore**. The correlation is picture in Figure 5.2.4.3. This situation is important for the WNM figure of merit, because it depends on the transistors N/P ratio, as cited before, in order to trip the inverter point making a writing successful. A planar SRAM was compared to the 3DVLSI partitioned case for RSNM simulations. The results for WNM using the unified statistical model are illustrated in Figure 5.2.4.4. **The write margins are similar in both cases**. This can be attributed to the fact that in both circuit configurations the PMOS is not correlated to the NMOS, **and the design ratio of the pull-up/pass-gate is uncorrelated in both cases.** The overall conclusion for the SRAM bitcell, is that it can safely benefit from the 3D sequential integration, without degrading important figures of merit, such as RSNM and WNM. **Indeed, the potential lack of correlation between tiers can be a feature, avoiding the worst/best case scenarios and improving the circuit noise margin**, even if those circuits are highly dependent on local variation sources.

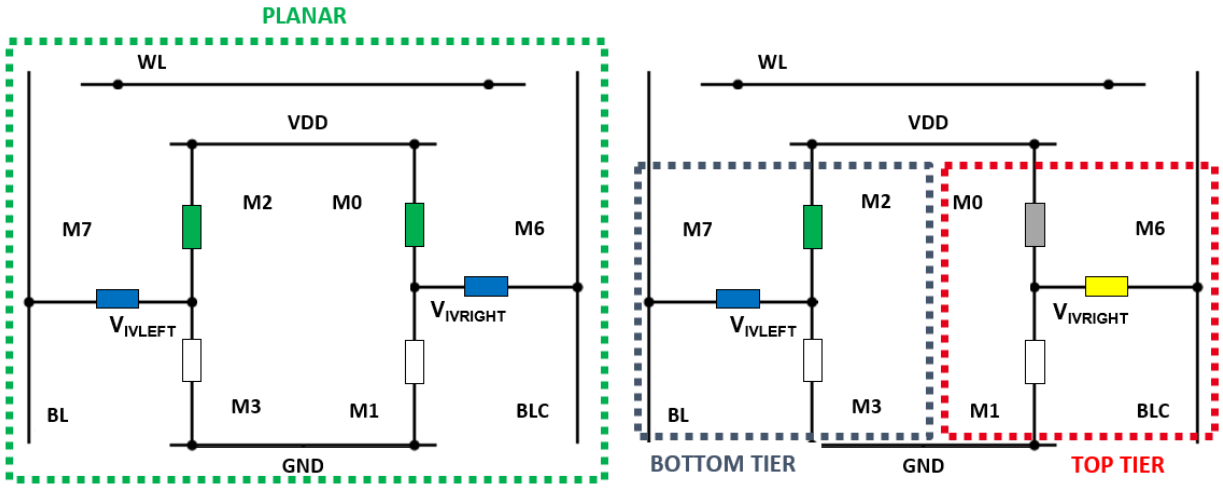


Figure 5.2.4.3 Monte Carlo SRAM simulation for WNM. Due to an initial hypothesis, the NMOS and PMOS are not correlated even in planar case. In 3DVLSI SRAM, they still uncorrelated.

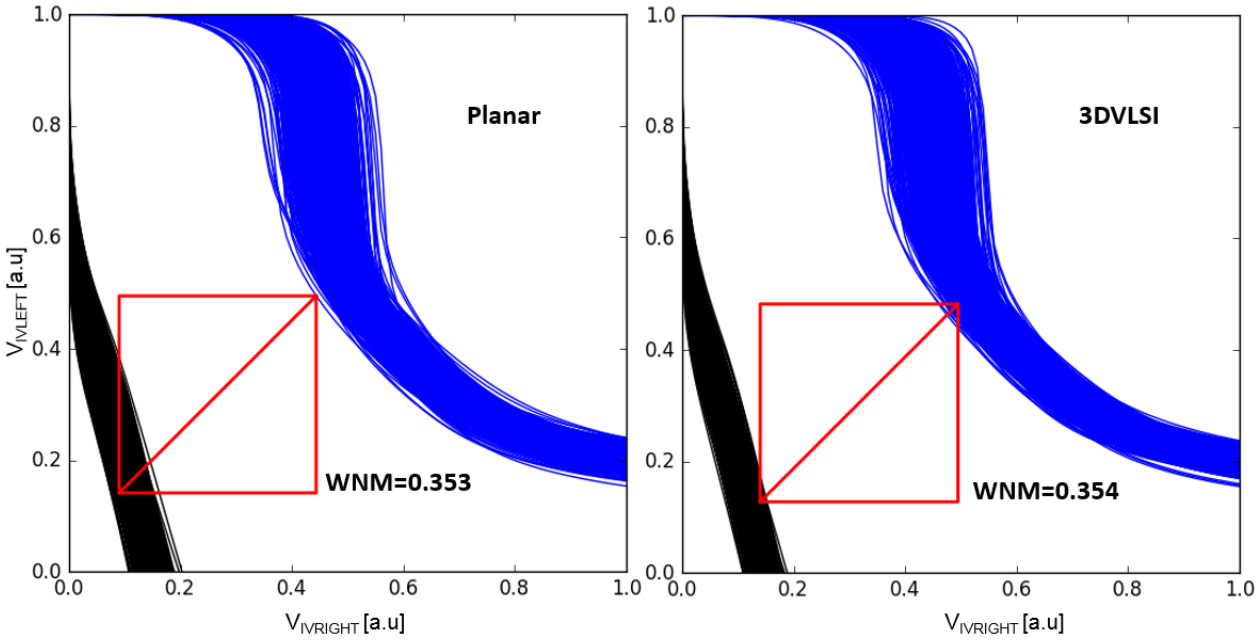


Figure 5.2.4.4 Monte Carlo Write Noise Margin simulations using the unified statistical model. On the left, a planar SRAM; on the right, a 3DVLSI partitioned SRAM.

5.2.5 SRAM Static Power

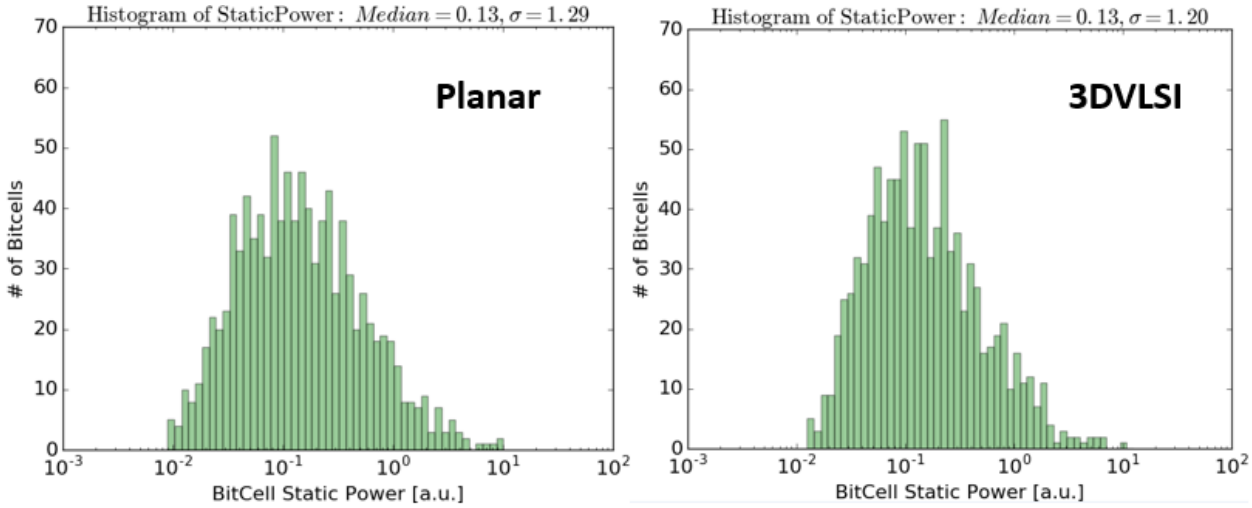


Figure 5.2.5.1 ACV Monte Carlo static power histogram for the same SRAM bitcell. The median and sigma of the fitted lognormal distribution are shown above the plot. On the left (a) a planar correlated SRAM. On the right (b) a 3D partitioned SRAM.

Extending the bitcell analysis, the static power was extracted in both configurations as seen in histograms of Figure 5.2.5.1. The simulation uses the unified model with all the variation sources enabled. The static power is exponentially dependent on the transistor V_T [Gu 1996], thus the MC simulations for SRAM are plotted in a logarithmic scale, and histogram has a lognormal distribution. Comparing the planar full correlated case to the 3DVLSI partitioned case, the median of the lognormal distribution is the same for both cases. This result is expected because the V_T mean and variance are the same for both tiers. Analyzing the static power dispersion, the 3DVLSI case has a slightly lower sigma compared to the planar one. The averaging effect takes place because one part can have a low V_T consuming more static power, while the other part can have a high V_T consuming less. Despite the static power variance being similar to both cases, the results show no degradation in the static power due to the partitioning.

5.3 Chapter Conclusion

In this chapter, the study of variability in 3D circuits was done transposing concepts of the planar integration, and adapting for a monolithic integration environment. After an instruction to the variability handling in nanoelectronics, the planar variability is analyzed. Then the partitioning of circuits in 3DVLSI is discussed.

The planar circuit analysis has been done with ROs and SRAMs, illustrating how the different variability sources affects these circuits. ROs are mostly influenced by global variations, while SRAMs behavior are dominated by local variations, and if a parameter is outside the correlation range for different transistors, the analysis also requires the ACV evaluation.

Partitioning has been described as highly influential element in the circuit variability. In the ROs, applying the partitioning of 50/50, it reduces the output frequency and static power dispersion up to 30% if the process tiers are not correlated. The simulations show that in the worst case, where bottom and top tier are correlated, the final dispersion of the partitioned circuit becomes similar to the planar circuits.

The SRAMs were also evaluated after a partitioning, although in this case it requires the 3D statistical unified model to evaluate the ACV correlation. **The model was developed based on previous statistical unified models and integrated directly in SPICE simulations.** The partitioned SRAM, while considering a correlation length one order higher than the transistor size, shows a better RSNM compared to the planar SRAM. **The lack of correlation between SRAM nodes is described as critical element, as the RSNM is a worst/best case figure of merit**, thus directly influenced by VTC curves correlation.

The main message of this chapter is: 3DVLSI has a unique feature, namely the partitioning, to reduce variance in determined figures of merit.

REFERENCES

Ayres, A., O. Rozeau, B. Borot, L. Fesquet, and M. Vinet. 2016. "Delay Partitioning Helps Reducing Variability in 3DVLSI." In *2016 46th European Solid-State Device Research Conference (ESSDERC)*, 67–70. doi:10.1109/ESSDERC.2016.7599590.

Batude, P., C. Fenouillet-Beranger, L. Pasini, V. Lu, F. Deprat, L. Brunet, B. Sklenard, et al. 2015. "3DVLSI with CoolCube Process: An Alternative Path to Scaling." In *2015 Symposium on VLSI Technology (VLSI Technology)*, T48–49. doi:10.1109/VLSIT.2015.7223698.

Conti, M., P. Crippa, S. Orcioni, and C. Turchetti. 1999. "Parametric Yield Formulation of MOS IC's Affected by Mismatch Effect." *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 18 (5): 582–96. doi:10.1109/43.759074.

Eikyu, K., T. Okagaki, M. Tanizawa, K. Ishikawa, T. Eimori, and O. Tsuchiya. 2006. "Global Identification of Variability Factors and Its Application to the Statistical Worst-Case Model Generation." In *2006 International Conference on Simulation of Semiconductor Processes and Devices*, 154–57. doi:10.1109/SISPAD.2006.282861.

Gu, R. X., and M. I. Elmasry. 1996. "Power Dissipation Analysis and Optimization of Deep Submicron CMOS Digital Circuits." *IEEE Journal of Solid-State Circuits* 31 (5): 707–13. doi:10.1109/4.509853.

Guo, Z., A. Carlson, L. T. Pang, K. T. Duong, T. J. K. Liu, and B. Nikolic. 2009. "Large-Scale SRAM Variability Characterization in 45 Nm CMOS." *IEEE Journal of Solid-State Circuits* 44 (11): 3174–92. doi:10.1109/JSSC.2009.2032698.

Kuhn, K. J., M. D. Giles, D. Becher, P. Kolar, A. Kornfeld, R. Kotlyar, S. T. Ma, A. Maheshwari, and S. Mudanai. 2011. "Process Technology Variation." *IEEE Transactions on Electron Devices* 58 (8): 2197–2208. doi:10.1109/TED.2011.2121913.

Kurude, S., S. Mittal, and U. Ganguly. 2016. "Statistical Variability Analysis of SRAM Cell for Emerging Transistor Technologies." *IEEE Transactions on Electron Devices* 63 (9): 3514–20. doi:10.1109/TED.2016.2590433.

Lu, N. 2014. "Modeling of Distance-Dependent Mismatch and Across-Chip Variations in Semiconductor Devices." *IEEE Transactions on Electron Devices* 61 (2): 342–50. doi:10.1109/TED.2013.2283076.

Mazurier, J., O. Weber, F. Andrieu, F. Allain, L. Tosti, L. Brévard, O. Rozeau, et al. 2011. "Drain Current Variability and MOSFET Parameters Correlations in Planar FDSOI Technology." In *2011 International Electron Devices Meeting*, 25.5.1-25.5.4. doi:10.1109/IEDM.2011.6131613.

Pelgrom, M. J. M., H. P. Tuinhout, and M. Vertregt. 1998. "Transistor Matching in Analog CMOS Applications." In *International Electron Devices Meeting 1998. Technical Digest (Cat. No.98CH36217)*, 915–18. doi:10.1109/IEDM.1998.746503.

Poiroux, T., P. Scheer, A. Juge, and M. Vinet. 2015. "Multiscale Statistically Correlated Variability: A Unified Model for Computer-Aided Design." *IEEE Transactions on Electron Devices* 62 (11): 3605–12. doi:10.1109/TED.2015.2478912.

Weber, O., E. Josse, F. Andrieu, A. Cros, E. Richard, P. Perreau, E. Baylac, et al. 2014. "14nm FDSOI Technology for High Speed and Energy Efficient Applications." In *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, 1–2. doi:10.1109/VLSIT.2014.6894343.

Chapter Six – Conclusion

6.1 Moore's Scaling Perspectives

6.1.1 Limit of Moore's Law

Moore's law has been active during the last five decades. The **device miniaturization is reducing the cost per transistor**. Along with the scaling, the circuit performance (speed) increases at each new node, if benchmarked for a constant power. The quote opening Chapter One: "*There's a basic principle about consumer electronics: it gets more powerful all the time and it gets cheaper all the time*"; it will no longer be true if the scaling cannot continue. Thus, a huge impact in all science fields, economics and human life style will occur as our society is extremely dependent on nanoelectronics.

The scaling of transistor is now reaching atomic dimensions, for example the 10nm Intel FinFETs have only 7nm fin width at half height. Continuing the miniaturization trend is getting harder from a physical point of view. For transistor scaling, the limits are the quantum effects degrading the transistor performances. At the interconnection level, the back-end scaling is increasing the parasitic elements as dielectric isolation and metal routing layers are getting thinner and smaller, increasing resistances and capacitances.

In this context, the future of CMOS logic scaling is uncertainty at long-term. As 2017, the industry and academia researches are focused on:

- **Short Term (5 years):** The miniaturization will continue by employing new transistor architectures and novel process technologies, such as Extreme Ultra-Violet (EUV) lithography. We can notice that nanowires or nanosheets transistors are already in the foundry roadmap.
- **Medium Term (5-20 years):** In this perspective, the research will address both more than Moore approach and alternatives to scaling. The work of this thesis is part of this last category as an option to the ultimate scaling integration. Moreover, the research on new materials to increase the device performance without decreasing dimensions will remain very active.
- **Long Term (20-50 years):** This is more speculative but quantum computing and exotic materials such as carbon nanotubes for mass fabrication seems to have a chance. Even if quantum processing shows a good potential, a hybrid between quantum and traditional CMOS processors is more likely.

The goal of this thesis consists in the evaluation of features, opportunities and issues of 3D sequential integration for logic circuits, which are potential technological solution to extend the Moore's scaling.

6.1.2 The 3D opportunity

3D monolithic, also referenced as sequential integration, is the idea to stack several tiers of transistors, opposed to traditional planar integration, where transistors are built side by side. The great opportunity of this technology is for dense and complex logic circuits, where it can deliver small pitch and size 3D contacts (3DCO) through the tiers. Considering the Performance /Power/Area (PPA) figure of merit, the 3D design must follow the guidelines:

- **Performance**: CoolCube™ integration features back-end in each tier, enabling an optimized 3D circuit routing. By routing in the third dimension with 3DCO, the designer has one more degree of freedom. In order to reduce the interconnections and the parasitic elements, which slow the circuits, the gates should ideally be closer to each other. However, in planar circuits, due to the high density of logic circuits, this ideal case is not always achieved, forming long critical paths

that limit the circuit performance. 3DVLSI design creates the opportunity to eliminate such bottlenecks by placing the gates in different tiers, close to each other. Hence, the circuit will benefit from reduced interconnection wirelength, increasing the overall performance. In this thesis, we show those gains by analyzing Full Adders and Ring Oscillators. At worst case, if no wirelength is cut in 3D integration, it will match the planar performance.

- **Power**: As the interconnections are smaller in 3DVLSI case, the total number of gates in the circuit should be lower compared to the similar circuit in planar, because some signal buffers and repeaters can be avoided. This effect reduces both the circuit static and dynamic power. Another aspect evaluated by this work, is the power delivery network (PDN). The 3DVLSI circuit will be connected to the exterior by the top tier, and then all the power nets will arrive to bottom tiers through the 3DCO. The contacts should be carefully designed to avoid voltage drop in lowermost tiers, otherwise the performance will be degraded in those transistors operating at lower supply voltage.
- **Area**: Stacking wafers increases the total area straightforwardly. The designer job is to avoid area wasting with 3D overhead, for example, with a disproportional number of 3DCO per number of gates. This thesis proses to use the 3DCO placement as standard cells, and evaluates the ideal number of 3DCO per gate in order to achieve an area doubling using two-tier integration. The transistor over transistor integration flavor is discouraged in favor of CMOS over CMOS integration because of the high number of 3DCO, causing a large 3D area overhead and sometimes creating routing congestions.

6.1.3 Advantages of 3D design for variability

Besides the PPA metric, the variability is an important topic in circuit design as it directly impacts the fabrication yield and circuit performance. As the processes are subject to variability, the circuit architecture has to take it into account, otherwise some devices will operate outside their specifications or will not work, decreasing the yield. This is managed by designing the circuits in a way that it can tolerate process variations. However, they will cause some chips to be faster than others, potentially affecting the final user perception of performance. In this context, the variability in 3D sequential circuits has to be measured, furthermore considering a circuit distributed across the tiers, as 3DVLSI granularity can be fine-tuned.

In order to evaluate all the sources of process variations, this thesis proposes a **3D unified model considering global, local and across-chip variations** for circuits implemented in more than one tier. The study presented some insights on 3D variability, and how 3D design can use the netlist partitioning as a feature to reduce a determined figure of merit variability.

The variability in 3DVLSI can be reduced if the design follows the directives based on circuit sensitivity to variation sources:

- **Global variations dependent circuits**: In this thesis, the first order evaluation is done using Ring Oscillators. The results show a partitioning influence on the variability. As the processes for each tier are usually not correlated, the final variability is described as a weighted sum. Thus, for the designers, partitioning is a way to reduce global variability. The main physical reason of this outcome is the avoidance of worst/best case happening on both tiers at the same time, or in

other words, a less performant device can be compensated by the device on another tier, hence reducing variability.

- **Local and Across-Chip variations**: Local variations are intrinsically related to technology and individually affects the transistors. This source of variation cannot be mitigated in 3DVLSI design and should be managed as in planar integration. The across-chip variation is the source of variations dependent on the distance between transistors, and introduces a notion of correlation as function of distances, namely the correlation range. In 3D sequential design, where the netlist can be partitioned with a fine granularity among the tiers, the ACV behavior can be engineered in order to increase a determined circuit figure of merit. As the transistors in different tiers are not correlated due to ACV, the design can exploit this feature. In this thesis, a SRAM partitioned into two-tiers, illustrates this point by showing a better figure of merit than in planar integration. The SRAMs are circuits with a high sensitivity to local and ACV variations. Thus, the designers should carefully partition their SRAMs to make dependent parts uncorrelated. In this fashion, the circuit can avoid the worst/best limiting the ACV correlation impact.

6.2 General Conclusion

3DVLSI integration, also known as monolithic or sequential integration is presented and evaluated in this thesis as a potential contender to continue the scaling for CMOS logic circuits. The main advantages of this technology compared to already existing 3D parallel integration is the **high alignment among tiers**, enabling **small size and pitch 3DCO**. This feature is a must for logic over logic integration, because the circuit needs a refined netlist granularity to optimize performance, while the area overhead caused by those connections remains negligible. Another great 3DVLSI feature is the improved placement and routing compared to planar circuits. Indeed, the **interconnections are shorter** as the design has an additional degree of freedom in the Z direction. Hence long wires in planar circuits can become 3DCO contacts, lowering the interconnection parasitic elements and speeding up the circuit as well as reducing the power usage.

This thesis was partitioned into two parts: the first one analyses and compares the 3DVLSI physical design implementation to the classical planar integration. The fundaments of scaling, and the reasons pushing 3DVLSI to become the mainstream integration for logic advanced nodes is discussed in Chapter One. Chapter Two details the planar and the expected automated design flow for 3D digital circuits. The tools are in a development stage; thus, the bottom-up design methodology is proposed. In this approach, the design is done close to the transistor, in order to answer to open questions during the EDA development. PPA performance is evaluated using simple circuits such as ROs. Small circuits, with few standard cells are depicted as hard to increase performance and reduce power, as the wirelength reduction is not enough to observe remarkable gains. On the other hand, if a planar circuit is stacked into 3D using CMOS over CMOS style, both circuits will match the performances, with no penalties due to the 3D integration. In Chapter Three, some guidelines are proposed in design and process, like the iBEOL composition to reduce contamination risks in front-end machines. An analysis of BEOL scaling in planar nodes depicts a physical barrier that will impact the circuit performance, and 3DVLSI is proposed as a solution for BEOL scaling issue.

A brief introduction to variability in planar circuits is done in Chapter Four. Planar process variations and its management are used as a starting point of 3DVLSI figures of merit variance analysis. The global, local and across-chip variations are discussed in Chapter Five, and evaluated using a 3D SPICE unified statistical model developed in this work. The main conclusion is that the designer can exploit the partitioning as a feature in order to reduce the variance of determined figures of merit.

Finally, this work concludes that **3D integration is a viable option to virtually continue the scaling**. The industry roadmap is focused on scaling for short-term, however miniaturization will face a major physical barrier in future, which may turn Moore's Law uneconomical as happened to 2D NAND memories. A great opportunity to introduce the 3DVLSI lies ahead.

6.3 Prospects

6.3.1 CMOS logic integration and memories – Several Tiers Scaling

3DVLSI is suitable for logic over logic integration, focusing high-performance computing. As described in this work, the **circuit PPA can be enhanced** by employing this integration. Memories are already stacked in sequential processes [Jung 2006; Park 2014], increasing the storage density per area. Logic circuits are still integrated in planar fashion, however as the scaling is becoming more difficult, the 3DVLSI chips integrating logic circuits and memories are the **easiest solution to extend Moore's Law**. Besides the physical limits, the cost of scaling planar technologies can become uneconomical, and then the 3DVLSI is an excellent candidate. This scenario already happened with memories, as scaling of 2D NAND memories has been too costly. Transition to 3D NAND was the path taken by the industry to reduce the costs.

The future of 3DVLSI also depends on the **Z direction scalability**. This thesis mainly worked on an initial case only using two-tiers. However, the possibility of stacking N tiers is needed to guarantee scalability over the years after the introduction of 3DVLSI. An interesting prospect, if 3DVLSI becomes the mainstream, is a Moore's Law like trend for the number of tiers, or in other words, the tier count should double each eighteen months. The present technology is thought to be **N-tier scalable**, as the low temperature process should not degrade another low temperature process. However, a silicon demonstrator need to be done yet, along with more research on 3D design flow for several tiers. Also, with 3DVLSI prototypes, the **thermal validation and several studies** have to be done, in a way that EDA tools can handle the power dissipation in stacked digital circuits.

6.3.2 More than Logic – Functionality Integrated Sequentially

3D parallel integration, or TSV already enables chips packaged with complementary functions, such as sensors, imaging, NEMS, specialized circuitry coupled to logic elements. The main goal is to provide **miniaturization and functionality** in a single package. However, as discussed in Chapter One, the TSV integration has very limited contact density between tiers, restraining the maximum data flow among tiers. The great opportunity of 3DVLSI is to **aggregate a list of functions in a single monolithic chip**, and further extending the capabilities of chips already integrated in parallel 3D.

With new segments driving the nanoelectronics industry, the applications require intensive real-time computing, such as machine learning, self-driving cars, smart grids and connected objects. The self-driving cars requires a huge processing of external world, as the software needs to understand what is happening around the car [Lee 2013]. This typical application requires an extensive data flow from the sensors (imaging, laser, radar and lidar) to the processing unit. An ideal system would have those capabilities integrated in a single monolithic device, in order to reduce costs and increase the system performance. The 3D sequential integration is one of the few candidates able to deliver the needed technical solution, by **integrating dense high-performance logic to fine grain complementary functions**, such as sensor matrix, imaging, and analog circuitry.

REFERENCES

Jung, S. M., J. Jang, W. Cho, H. Cho, J. Jeong, Y. Chang, J. Kim, et al. 2006. “Three Dimensionally Stacked NAND Flash Memory Technology Using Stacking Single Crystal Si Layers on ILD and TANOS Structure for Beyond 30nm Node.” In *2006 International Electron Devices Meeting*, 1–4. doi:10.1109/IEDM.2006.346902.

Lee, G. H., F. Faundorfer, and M. Pollefeys. 2013. “Motion Estimation for Self-Driving Cars with a Generalized Camera.” In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2746–53. doi:10.1109/CVPR.2013.354.

Park, K. T., J. m Han, D. Kim, S. Nam, K. Choi, M. S. Kim, P. Kwak, et al. 2014. “19.5 Three-Dimensional 128Gb MLC Vertical NAND Flash-Memory with 24-WL Stacked Layers and 50MB/S High-Speed Programming.” In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 334–35. doi:10.1109/ISSCC.2014.6757458.

Appendix A

Appendix A

A.1 Thesis Tools Context

A.1.1A.1.1 3D Design Environment

The 3D sequential design environment has been developed **based in the planar environment logic synthesis macroscopic step**. The process design kit (PDK) consists of four design elements based in the Figure A.1.1.1 which are based on technology parameters. The 3DVLSI integration is currently developed in the FDSOI technology, however in future it can be implemented for any transistor architecture. The **simulations are done using the ELDO simulator from Mentor Graphics**, using the proper transistor model. The circuit functionality is tested using the schematic drawings and then electrically simulated using Virtuoso environment from Cadence. Although the focus is digital circuits, the SPICE simulations are analog/mixed-signal simulations which capture all interactions of the transistor compact model. The range of capabilities of the schematic is analyzed through the voltage and current in the circuit nodes, also allowing timing analysis. The layouts are done using the *Cadence Virtuoso* in a full custom environment. After the layout drawing, the **Design Rule Check (DRC)** tool is used to verify if the layout complies with the **Design Rules Manual (DRM)**. Those rules are imposed by the process limitations and are enforced to guarantee the process yield and reliability. For example, avoiding short circuits due to a specific layout routing. After the layout is checked to comply the DRM, it is also checked against the schematic in a tool called **LVS (Layout Versus Schematic)**. This tool confirms that the drawn layout corresponds to the schematic. The LVS is essential, as complex layouts may have a minor mistake that is hard to identify, such as missing via connection. After the DRC and LVS verification steps, the layout parasitic elements can be extracted (PEX). The used tool for PEX includes a complete definition of all layers in the design: including spacing, thickness, dielectric permittivity, metal resistivity, contact resistance, etc. With this information, the tool can calculate parasitic elements from the interconnections, such as resistance and capacitance. The DRC, LVS and PEX are inside the tools suite *Mentor Calibre* from Mentor Graphics. The design flow is then continued by **adding the parasitic elements in the original schematic netlist**. Then the full schematic with realistic parasitic elements can be simulated, in order to verify if the layout design accomplish the design goals, and finally circuit the PPA. If some layout or schematic modification is necessary, then the design flow should restart at the DRC level. This is the **full custom environment proportioned by the 3D sequential PDK**. In future, the expectation is that commercial tools will implement a planar like flow, with standard cells integration and fully automated tools.

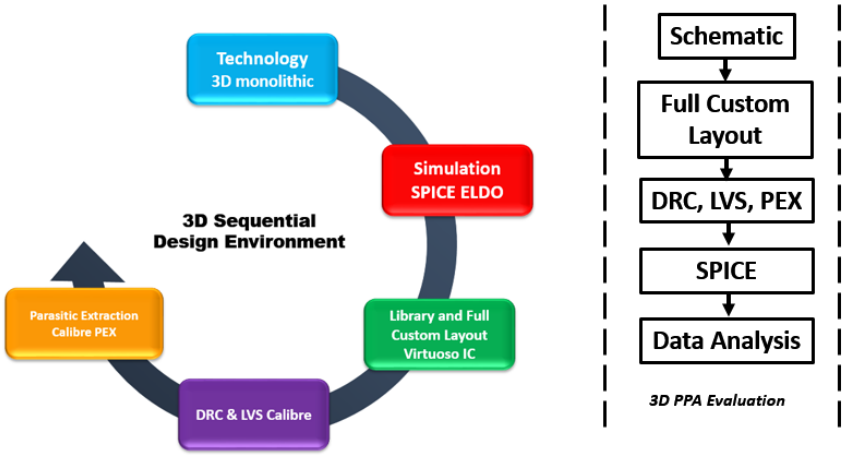


Figure A.1.1.1 3DVLSI Predictive Design Kit (PDK) used on this work.

A.1.2 Full Custom vs Standard Cell Integration

The full custom integration refers to the highly optimized designs, requiring a great amount of man hours. The transistors size, **placement and routing are carefully crafted in order to extract the maximum performance**. The area also is reduced as the optimization can look for short interconnections and avoid blank spaces in the layout. The semi-custom approach refers to layouts which certain circuits can be repeated several times, copying and pasting full-customized blocks. Despite of the existence of automation tools to help the Full-Custom development, it differs from the circuit layout implemented using a full automated flow with Electronic Design Automation (EDA). In this approach, the circuit netlist in the form of gate-level description is used, and each gate correspondent is chosen in a library containing gate layouts. Those gate layouts are named standard cells. The standard cell is a layout of a certain gate, and is drawn to achieve a certain design directive. For example, the standard cell for a given gate can be drawn for best timing performance, or lowest area for high density, lowest power, or high current drive. **It is common that libraries contain more than one flavor of standard cell layout for the same gate**. The usual standard cell metrics are shown in Figure A.1.2.1. The standard cell has to comply with the technology design rules, and its **size is usually defined by the number of metal tracks** that fits inside the cell, and the number of poly in the horizontal axis times the minimum **contacted poly pitch (CPP)**. The cells are done using the CMOS integration and sometimes using predefined spaces for supply rails, such as VDD and GND, in order to facilitate the automated gate placement into the layout. Another common practice is to use grids in the layout, to position the input and output pins, easing the further connection. In this work, **the layouts were developed using the full-custom approach**. The automated partitioning, floor-planning, placement and routing for 3DVLSI are under development and target of numerous publications. The small circuits used to **benchmark the PPA are representative for a first order evaluation**, such as ring oscillators.

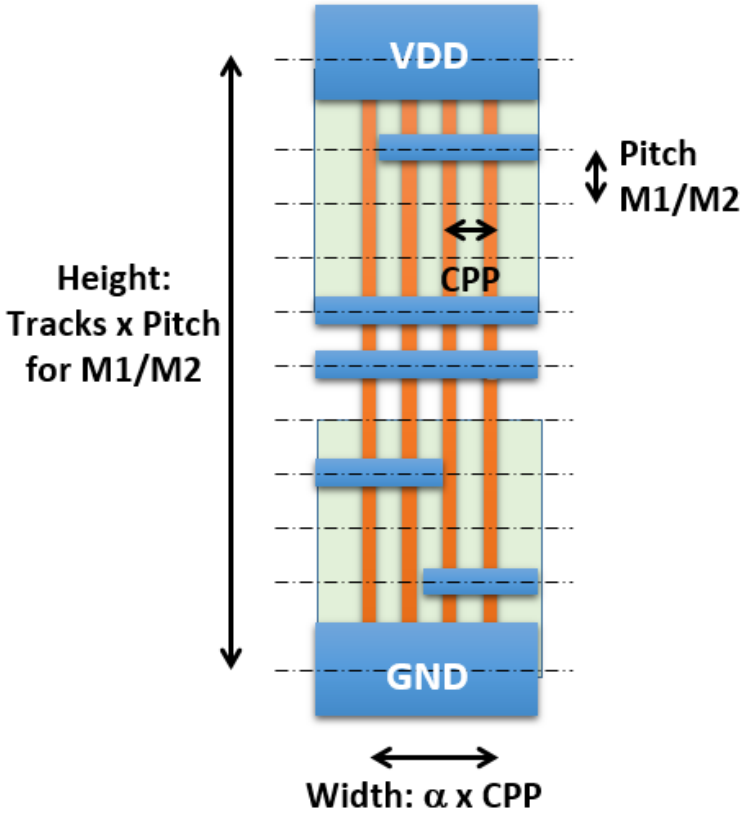


Figure A.1.2.1 Typical standard cell definitions. The cell height is predefined as the number of metal tracks that can fit inside. The width is defined as the number of poly (PC) in the horizontal axis; the CPP (Contacted Poly Pitch) is the minimum distance between two parallel PC (represented in orange).

Appendix B

B.1 SRAM Signal Noise Margin (SNM) Simulations

B.1.1 SRAM SPICE Netlist

The 6 transistors SRAM has been simulated in ELDO SPICE, using the LETI-UTSOI2 model. The netlist is illustrated in Figure B.1.1.1. Each transistor is defined with their respective parameters, such as gate length, transistor width, number of fingers and active region continuous length. For this work, the Pearson correlation is also passed to the parameter. Then, the transistors connections to form the SRAM are declared as well the node voltage and capacitances. To extract the Voltage Transfer Characteristics (VTC) the SRAM inverter output node is swept. A total of a hundred of operation points are evaluated. Finally, the SPICE alter command does the same procedure to the other side of the SRAM, and then the Monte Carlo method is applied, repeating the simulation one thousand times. Each time, the transistors parameters have different values, inside a boundary specified in parameter statistical distribution. The output has a total of 100 sweeps x 1000 MC x 2 VTCs = two hundred thousand points.

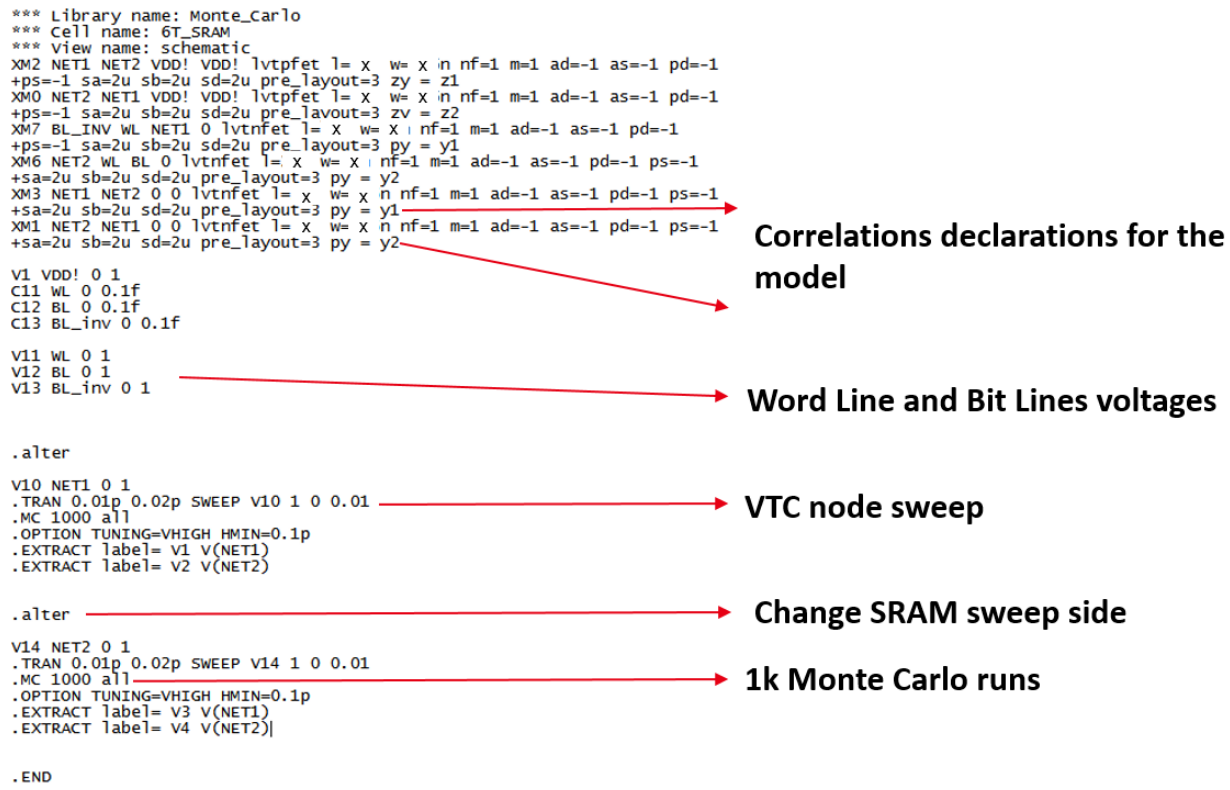


Figure B.1.1.1SPICE netlist for 6T SRAM.

In order to evaluate such amount of data, a python script was done to extract the RSNM. As defined on Chapter Four, the SNM is the size of the biggest square inside the VTC curves. The python script interpolates the data from sweep to increase the precision. Then, it looks for the longest diagonal between the two curves, taking advantage from the fact that the RSNM square is parallel to X-axis. The python script was divided into two parts, to evaluate upper and lower lobules. The diagonal is found by sweeping in X and Y direction at the same time, until it arrives to other VTC curve as illustrated in Figure B.1.1.2. After many steps, the script stops if the diagonal value start decreasing, and the biggest values is returned. The algorithm for upper side search is illustrated in Figure B.1.1.3.

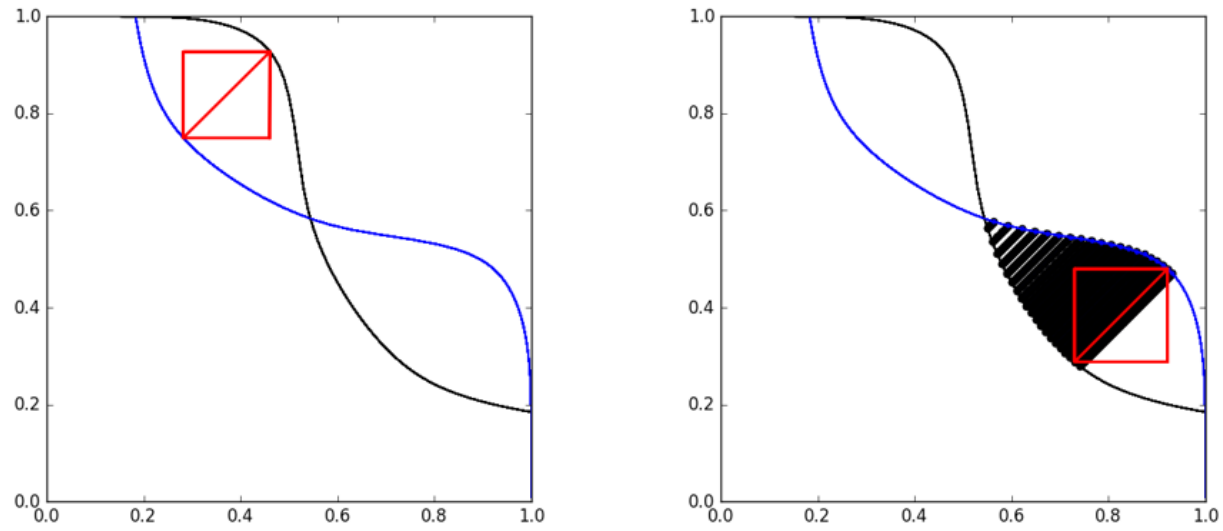


Figure B.1.1.2 SRAM SNM evaluation. On the right, the steps of python code in order to find the longest diagonal.

A better solution can be implemented for diagonal size evaluation by using binary search to find the other VTC curve limit. However, the presented solution was enough to handle the two hundred thousand points for 1mV resolution.

```
f1 = interp1d(x1, y1, kind='linear')                    # VTC1 curve interpolation
f2 = interp1d(x2, y2, kind='linear')                    # VTC2 curve interpolation

for h in np.arange(min(x1)*1.1, intersection*0.9, 0.01):    # X-axis sweep
  for g in np.arange(0, 10, 0.01):                      # Diagonal sweep steps
    x = h + 0.1 * g                                     # x and y are current diagonal position
    y = f1(h) + 0.1 * g
    if y >= f2(x):                                      #Evaltuates if the diagonal arrived in the other VTC curve
      if side < abs(f2(x) - f1(h)):                     #If the current diagonal is higher than stored, save the new one.
        side = abs(f2(x) - f1(h))
        # print(side)
        upper_v_x = x                                   #Store the square vertices for plotting purposes
        upper_v_y = f2(x)
        lower_v_x = h
        lower_v_y = f1(h)
        # plt.scatter([upper_v_x, lower_v_x], [upper_v_y, lower_v_y], color='red', zorder=2)
        break
      break
  if side > abs(f2(x) - f1(h)) and y >= f2(x):           #If the diagonal is not increasing, stops the loop
    break
zz.append(side)                                         #Returns the side of the square
```

Figure B.1.1.3 Python code for SRAM SNM upper lobule evaluation.

The script can be divided in multicore-processing in order to speed-up the processing as shown in Figure B.1.1.4. As the interpolation takes place, the SNM can have higher resolution, but slowing the total search time. Then the multicore processing is a requirement.

```
from multiprocessing import Process, Queue

z = parse()     #Parse all points from SPICE output
a = z[0]        #VTC1 X-axis
b = z[1]        #VTC1 Y-axis
c = z[2]        #VTC2 X-axis
d = z[3]        #VTC2 Y-axis

if __name__ == '__main__':                                          #Main loop

  p1 = Process(target=upper, args=(a, b, c, d, u1, queue1))         #Start 8 processes - Multicore
  p1.start()
  p2 = Process(target=upper, args=(a, b, c, d, u2, queue2))
  p2.start()
  p3 = Process(target=upper, args=(a, b, c, d, u3, queue3))
  p3.start()
  p4 = Process(target=upper, args=(a, b, c, d, u4,  queue4))
  p4.start()
  p5 = Process(target=lower, args=(a, b, c, d, u1, queue5))
  p5.start()
  p6 = Process(target=lower, args=(a, b, c, d, u2, queue6))
  p6.start()
  p7 = Process(target=lower, args=(a, b, c, d, u3, queue7))
  p7.start()
  p8 = Process(target=lower, args=(a, b, c, d, u4,  queue8))
  p8.start()

  p1.join()
  p2.join()
  p3.join()
  p4.join()
  p5.join()
  p6.join()
  p7.join()
  p8.join()

  RSNM1.append(queue1.get())                                        #Gets the multicore results and treat them
  RSNM2.append(queue2.get())
  RSNM3.append(queue3.get())
  RSNM4.append(queue4.get())
  RSNM5.append(queue5.get())
  RSNM6.append(queue6.get())
  RSNM7.append(queue7.get())
  RSNM8.append(queue8.get())
```

Figure B.1.1.4 SRAM SNM evaluation. Main script invoking the upper and lower function to evaluate each SNM lobule.

B.1.2 Mathematical considerations – Correlations

The model presented in section 5.2 is can handle the correlations for N device pairs, as long the correlations for a given parameter are coherent, or in the other words the correlation matrix NxN for a parameter P is positive-definite [Conti 1999]. In Figure B.1.2.1 three devices are illustrated with a certain distance to each other. In order to describe all the correlations simultaneously; the Pearson correlation between pairs (**a**, **b**, **c**) can be treated analog to the cosine angle (B.1). The cosine of z is the correlation a. In addition, the angle z is at the most the sum of the angles x and y. This places a bound limit for the correlation **a** when **b** and **c**

are already defined. By using the cosine addition identity formula, the bound correlation limit is defined by (B.2)

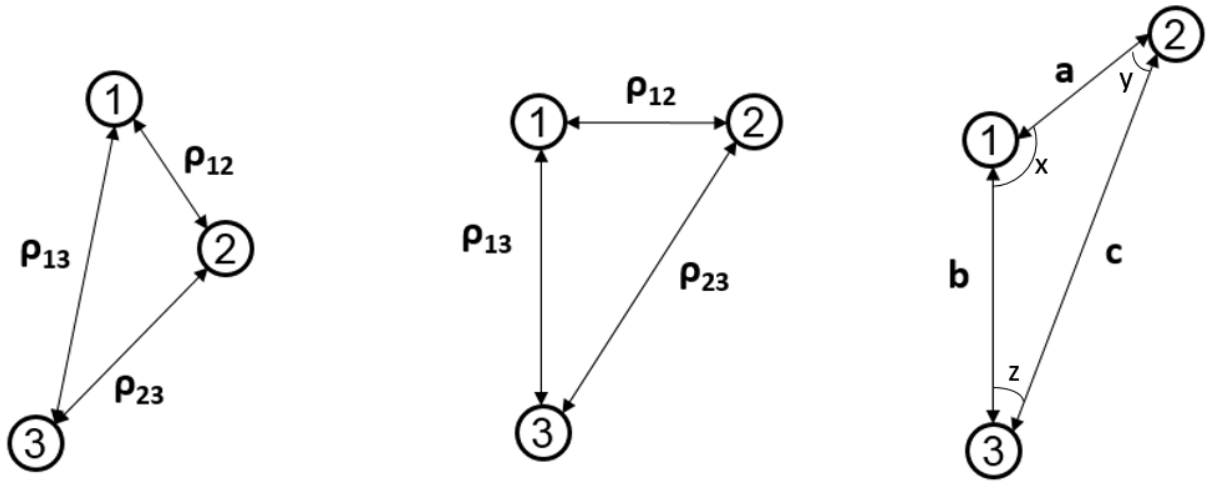


Figure B.1.2.1 Correlation for 3 devices depending on distance. In order to all correlations become coherent, it need follow angle coherence, or in general case, the correlation matrix needs to be positive-definite.

$$\rho_{X,Y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sqrt{E((X-\mu_X)^2)E((Y-\mu_Y)^2)}} \tag{B.1}$$

$$= \frac{\langle X-\mu_X, Y-\mu_Y\rangle}{\|X-\mu_X\|\|Y-\mu_Y\|} = cos\theta_{X,Y}$$

$$a \geq bc - \sqrt{1-b}\sqrt{1-c} \tag{B.2}$$

Another method is to analyze the correlation matrix (B.3), as it needs to be positive-definite, the determinant is non-negative, as in (B.4). This equation can be rewritten, and matches the equation (B.2).

$$\mathrm{C} = \begin{bmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{bmatrix} \tag{B.3}$$

$$1 + 2abc - a^2 - b^2 - c^2 \geq 0 \tag{B.4}$$

This analysis can be done using the exponential elements of the correlations described by the unified statistical model (5.9). The matrix (B.5) becomes positive-definite for the equation (B.6).

$$C = \begin{bmatrix} 1 & e^{-\alpha} & e^{-\beta} \\ e^{-\alpha} & 1 & e^{-\gamma} \\ e^{-\beta} & e^{-\gamma} & 1 \end{bmatrix} \quad (B.5)$$

$$1 + 2e^{-\alpha\beta\gamma} - e^{\alpha^2} - e^{\beta^2} - e^{\gamma^2} \geq 0 \quad (B.6)$$

Where α, β and γ are the terms dependent on the device size, distance and correlation range. **If the input values do not satisfy the equation, then the Cauchy–Schwarz inequality is not respected.** The outcome is at least one pair correlation will not be coherent, or depending on the correlation code implementation, the standard deviation for a given parameter P won't be constant for all devices. The proposed 3D unified statistical model, have a discontinuity when devices are not in the same tier, thus the triangle inequality is not respected in this situation; however, for devices in the same tier, the inequality (B.6) should be satisfied.

B.1.3 Correlation Treatment in the Netlist

As discussed, the correlations are usually treated as correlations matrix. However, this approach is very hard to implement in SPICE due to the lack of matrix processing. A method to overcome the limitation has been presented, and the raw python and circuit netlist are illustrated in Figure . The Pearson correlation is used as input, and for each device the distribution is calculated from a previous normal distribution, inserting an uncorrelated Gaussian part. Then, the desired distribution mean and standard deviation is done by sum and multiplication of the calculated correlated distribution.

- **The python and SPICE code:**

```
def correlatedValue(x, r):
    r2 = r**2
    ve = 1-r2
    SD = math.sqrt(ve)
    e  = random.gauss(0,SD)
    y  = r*x + e
    return(y)

for i in range(100000):
    x = random.gauss(0,1)
    y = correlatedValue(x, r=1)
    z = correlatedValue(y, r=0.1)
    z1.append(x)
    z2.append(y)
    z3.append(z)
```

```
*NFET_BOT*

.param x_nb = '0 LOT/gauss=1'

.param SD1 = '((1-(p1^2))^0.5)'
.param e1 = '0 LOT/gauss=SD1'
.param y1_ = 'p1*x_nb + e1'

.param SD2 = '(1-(p2^2))^0.5'
.param e2 = '0 LOT/gauss=SD2'
.param y2_ = 'p2*y1_ + e2'

.param SD3 = '(1-(p3^2))^0.5'
.param e3 = '0 LOT/gauss=SD3'
.param y3_ = 'p3*y2_ + e3'

.param SD4 = '(1-(p4^2))^0.5'
.param e4 = '0 LOT/gauss=SD4'
.param y4_ = 'p4*y3_ + e4'

.param SD5 = '(1-(p5^2))^0.5'
.param e5 = '0 LOT/gauss=SD5'
.param y5_ = 'p5*y4_ + e5'

.param y1  = (y1_*0.05)+126.3m
.param y2  = (y2_*0.05)+126.3m
.param y3  = (y3_*0.05)+126.3m
.param y4  = (y4_*0.05)+126.3m
.param y5  = (y5_*0.05)+126.3m

*PFET_BOT*

.param x_pb = '0 LOT/gauss=1'

.param zSD1 = '((1-(zp1^2))^0.5)'
.param ze1 = '0 LOT/gauss=zSD1'
.param z1_ = 'zp1*x_pb + ze1'

.param zSD2 = '(1-(zp2^2))^0.5'
.param ze2 = '0 LOT/gauss=zSD2'
.param z2_ = 'zp2*z1_ + ze2'
```

Figure B.1.3.1 Correlations treatment. In this approach, the correlation is built for each device from the previous one. The distribution mean standard evaluation is done by sum and multiplication after the correlation step.

REFERENCES

Conti, M., P. Crippa, S. Orcioni, and C. Turchetti. 1999. “Parametric Yield Formulation of MOS IC’s Affected by Mismatch Effect.” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 18 (5): 582–96. doi:10.1109/43.759074.

Title: 3D Monolithic Integration: Performance, Power and Area Evaluation for 14nm and beyond

Abstract

3DVLSI integration, also known as monolithic or sequential integration is presented and evaluated in this thesis as a potential contender to continue the scaling for CMOS logic circuits. The main advantage of this technology compared to the already existing 3D parallel integration is its high alignment among tiers, enabling small size and pitch with the inter-tier contacts (3DCO). Another great 3DVLSI feature is its improved capability to place and route circuits, compared to the planar approach: the interconnections can be shorter as the design has an additional degree of freedom in the Z direction. For instance, long wires in planar circuits can be cut thanks to 3DCO contacts, lowering the interconnection parasitic elements and speeding up the circuit as well as reducing the power. In this framework, the thesis has been divided into two parts: the first part is dedicated to the evaluation of Performance, Power and Area (PPA) of 3D circuits and gives design guidelines. The second part treats the variability in 3D circuits by using a 3D unified statistical model and propose an approach for the multi-tier variability.

Keywords: 3D Monolithic Integration, PPA, 3D circuit variability, 3D Unified Statistical Model, 3D Design Guidelines, 3D SPICE simulations.

Titre: Intégration monolithique en 3D: étude du potentiel en termes de consommation, performance et surface pour le nœud technologique 14nm et au-delà

Résumé

L'intégration 3DVLSI, également connue sous le nom d'intégration monolithique ou séquentielle, est présentée et évaluée dans cette thèse comme une alternative à la réduction du nœud technologique des circuits logiques CMOS. L'avantage principal de cette technologie par rapport à l'intégration parallèle 3D, déjà existante, est l'alignement précis entre les niveaux, ce qui permet des contacts 3D réduits et plus proches. Un autre avantage, extrêmement favorable à l'approche 3DVLSI, est l'amélioration du placement et du routage par rapport aux circuits planaires, notamment parce qu'elle permet des interconnexions plus courtes et qu'elle offre a un degré de liberté supplémentaire dans la direction Z pour la conception. Par exemple, les fils les plus longs dans les circuits planaires peuvent ainsi être réduits grâce aux contacts 3DCO, en diminuant les éléments parasites d'interconnexion. Il est ainsi possible d'augmenter la vitesse du circuit et de réduire la puissance électrique. Dans ce contexte, la thèse a été divisée en deux parties. La première partie traite de l'évaluation de la Consommation, des Performances et de la Surface (CPS) et donne des recommandations pour la conception des circuits 3D. La deuxième partie traite la variabilité des circuits 3D en utilisant un modèle statistique unifié, et en proposant une approche pour la variabilité des circuits multi-niveaux.

Keywords: Intégration monolithique en 3D, CPS, Variabilité du circuit 3D, Modèle Statistique Unifié pour la 3D, Recommandations pour le Design 3D, Simulations 3D avec SPICE.

ISBN 978-2-11-129229-1