



HAL
open science

Simulation, fabrication and electrical characterization of advanced silicon MOS transistors for 3D-monolithic integration

Daphnée Bosch

► **To cite this version:**

Daphnée Bosch. Simulation, fabrication and electrical characterization of advanced silicon MOS transistors for 3D-monolithic integration. Micro and nanotechnologies/Microelectronics. Université Grenoble Alpes [2020-..], 2020. English. NNT : 2020GRALT077 . tel-03219902

HAL Id: tel-03219902

<https://theses.hal.science/tel-03219902>

Submitted on 6 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **Nano Electronique et Nano Technologies**

Arrêté ministériel : 25 mai 2016

Présentée par

Daphnée BOSCH

Thèse dirigée par **Francis BALESTRA**, IMEP-LAHC et
codirigée par **Claire FENOUILLET-BERANGER**, CEA-LETI et
co-encadrée par **Francois ANDRIEU**, CEA-LETI

préparée au sein du **Laboratoire d'électronique et des
technologies de l'information (CEA-LETI)**
dans l'**École Doctorale électronique, électrotechnique,
automatique et traitement du signal (EEATS)**

Simulation, fabrication et caractérisation de transistors MOS avancés pour une intégration 3D monolithique

Thèse soutenue publiquement le **17/12/2020**,
devant le jury composé de :

Mr Guilhem LARRIEU

Directeur de recherche CNRS-LAAS, (Rapporteur)

Mme Cristell MANEUX

Professeur université de Bordeaux-IMS, (Rapporteur)

Mr Pierre Emmanuel GAILLARDON

Professeur associé université UTAH, (Examinateur)

Mme Anne KAMINSKI

Professeur Grenoble INP, (Examinateur, présidente du jury)

Mr Francis BALESTRA

Directeur de recherche CNRS-IMEP-LAHC, (Directeur de thèse)

Mr Francois ANDRIEU

Ingénieur de recherche au CEA-Leti, (Invité)

Mr Franck ARNAUD

Ingénieur STMicroelectronics, (Invité)



Acknowledgements

To whom it may concern,

Whoever is reading these lines, I would like to thank you. Not just because you are a great person (I have no doubts in that), but because if you are here, you had an intent to read this manuscript, or at least the acknowledgements part. I wish you all the best for this colossal task, whatever the option you choose.

I spend most of my time during the third last year between CEA-LETI site, my home and to be honest, mountains. The climbing gym counts into the mountain category. I propose you the evident section subdivision: professional greetings, personal one and mountains. I'm not entirely certain of the content of the last section but no worries: rocks can be fascinating.

Before going further, I would like to thank the members of my jury, for reading my manuscript, attending the defense and for all the feedbacks that improved the quality of the manuscript. The final results is in your hands. It was an honor in having you in my jury.

From the workplace part, my warmest gratitude to the driving force of my PhD (apart myself) my supervisor Francois Andrieu. You succeed in canalizing my energy and my attention towards productivity. I am glad that you pushed me to write for conferences and challenged me several time to instill me a veritable researcher engineer spirit. Thanks again for your guidance, your enthusiasm, your reactivity and the different opportunities you offered me. I would like to thanks my thesis director Francis Balestra, for microelectronics questions related, administrative paperwork, but most important to grab a coffee and discuss about the world. I learned a lot with you about sugar industry, nutrition but also theater. I also had the opportunity to work with Jean-Pierre Colinge. Thanks for the junctionless transistor and to be available when I came to ask the very same question several times in a row. By the way, can you explain me again screening effect? Sorry dat ik ben sneller dan mijn e-mails, maar nogmaals bedankt. Speaking of the people that I could bothered with existential questions about transistors, Gerard Ghibaudo was always available (and willing) to derive equations about mismatch.

Still in the workplace area, the members of the Coolcube meeting helped me a lot when I started to take my marks into this new environment. Among them (I conserve the PhD students for later), Perrine, Didier, Laurent B. and Laurent B. (merci Laurent de m'avoir enseigné la rigueur indispensable pour une utilisation correcte d'EYELIT), Mickael, Hervé and Claire. At the same time, I learned a lot about design/PDK with the members of 3DMUSE: Sébastien, Olivier, Mehdi... And after, the IMC group: Sylvain, Joris (thanks for your support concerning TCAD ©), Sébastien, Mona, Bastien, Jean-Philippe, Jean-Michel: thanks for all the discussion, merging design and technology for the best brainstorm. Speaking of design, my warmest thanks to Lorenzo and Adam who supervised me and gave me all their knowledge about SRAM. Speaking of SRAM, a special thanks to Louis, with who we constructed, deconstructed and constructed again the same story about SRAM security. Between the member of this thematic (I may add Valérie and Heimann), "la confiance règne". And last, but not the least, thanks to all the person who helped me, either in the characterisation laboratory, in the cleanroom, at the office, at the tea corner... Here, a non-exhaustive list: Bernard, Jean-Michel, Yves, Christophe, Alexandre, Zdenek, Olivier, Benoit, Yves, Cyrille, Claude, Cyril, Thomas, Nils, Maud, Xavier, Denis, Jacques, Jean, Giovanni, Rabah, Arnaud, Christoforos, Jose, Angeliki, Jean-Luc, Artemisia, Alain, Fabienne, Gabriel, Laurent, Mickael, Fred, Jean-Michel, Pablo, Cédric, Ludivine, Christian, Pascal, Virginie, Aurélien, David, Aurélien, Ronald, Francois...

Special dedicace to people who shared my office (and had to manage me all day long): Sotiris, Romane, Simon, Giulia, Rony, Théophile and Thibaud. I wish to the last three names many debates about all kind of topics (I know, it already started). Concerning the PhD student, the "old" generation is composed of

the wise Remy (I'm still using part of your code), the social Jessy (our "spiritual guide", claimed only by itself) and the brave Lina (I always admire you, you are so strong!). For the "current" generation, Giulia and Camila have been the perfect speaker to chit-chat, complain but also low down the stress. Thanks Sylvan for the beautiful drawings in the office! For the "future" generation, good luck Théophile, you will need it.

Passons maintenant à la deuxième partie -qui pour l'occasion est devenue francophone-, un peu plus personnelle. De manière générale, merci pour le soutien de ma petite famille, de ma grande famille et de ma très grande famille qui englobe toutes les rencontres et contacts que j'ai pu nouer au cours de mes 26 ans d'existence. Merci à mes parents ainsi que ma petite sœur et mon petit frère pour mon éducation, votre soutien et votre confiance en moi sans failles. Je ne pourrais jamais vous remercier suffisamment ! Merci à toute ma grande famille, ma grand-mère et grande tante, mes oncles et tantes, neveux et nièces, cousins et cousines, la famille de Catalogne et je manque de vocabulaire pour qualifier le reste... Merci à mes amis d'enfance, Claire, Margot et Alexandra, votre amitié (depuis une vingtaine d'années pour certaines) m'est précieuse. A tous ceux de mes études avec qui j'ai gardé contact et que je revois régulièrement : Lucie, Arnaud, Sylvain, Mickael, Florian, Thomas, Maxime, Pierre, Quentin, Piel, Hélène, Gaëlle, Julie et tous les autres. Merci Julie pour les weekends en Suisses, chacun sont inoubliables à leur manière. Et le plus important à mes yeux, merci à mon partenaire Guillaume d'avoir vécu avec moi cette thèse au jour le jour, d'avoir su être content pour mes succès tout en sachant me soutenir dans les périodes difficiles.

Pour la dernière catégorie, j'aimerais remercier les parcs et massifs suivants pour m'avoir donné une bouffée d'oxygène et des souvenirs inoubliables : la Vanoise, les Dolomites, le Yosemite, le grand canyon, Taroko national park, le Vercors, la Chartreuse (boisson ou massif, à vous de choisir), Belledonne, les Ecrins...

Merci encore,

Sincèrement,

Daphnée Bosch

Contents

Chapter I : Introduction

1- History of Semiconductor industry:	16
a. Dennards' law: happy scaling era (Moore's law):.....	16
b. Physical limit to scaling, apparition of parasitic effects:	17
c. New architectures (FINFET, FDSOI: back bias)	18
i. FDSOI architecture.....	19
ii. Gate all-around Nanowires or nanosheets	20
2- Semi-conductor industry: current challenges, roadmaps and propositions to keep the race to technological node.....	21
a. Picture of 2020 microelectronic ecosystem.....	21
b. Introduction to 3D sequential integration.....	22
i. 3D sequential integration.....	22
ii. 3D sequential integration: More Moore applications	22
iii. 3D sequential integration: More than Moore Applications	23
c. Introduction to In-Memory Computing:	24
3- Thesis objectives:	26

Chapter II: Design Technology Co-Optimization: functionalities provided by 3D monolithic integration

1- VLSI digital design flow	29
a. Overview of a planar digital design flow and EDA tools.....	29
b. Power, performance and area (PPA) design trade-off.....	32
c. From 2D to 3D digital design flow.....	33
i. 3D Design flow	33
ii. Netlist partitioning: examples.....	33
2- State of the art of 3D design performance assessment: Motivation for 3D monolithic integration for digital applications	35
a. Cost analysis.....	35
b. Thermal dissipation issue	35
c. Performances	37
3- 3D design MOSFET environment.....	39
a. 3D tier and intermediate BEOL for CMOS over CMOS integration: Coolcube TM	39
b. SPICE model	40
c. Parasitic element extraction.....	40

d.	Methodology summary.....	41
e.	RO, SRAM benchmark: typical figure of merits.....	42
i.	Ring Oscillator	42
ii.	SRAM.....	42
4-	Routing in 3D designs	45
a.	Buried power rail.....	45
b.	Congestion mitigation and resources sharing between tiers.....	45
c.	Design guidelines for top-tier Back-plane contact	48
i.	Simulated structure.....	48
ii.	Static consideration	49
iii.	Dynamic consideration.....	50
5-	Design-technology co-optimization: top-tier SRAM	51
a.	14nm technology performance	51
i.	Electrical characterization of typical FOM	51
ii.	SRAM: variability issue and impact on FOM.....	53
iii.	Back-bias assist	54
iv.	BTI-induced dynamic variability at the bitcell level	55
a-	BTI mechanism	55
b.	BTI at the bitcell level: experimental results.....	56
c.	Proposition of a novel fine-grain back-bias assist techniques for 3D-monolithic 14nm FDSOI top-tier SRAMs	59
i.	3D monolithic design kit: layout considerations	59
ii.	Fine grain and versatile back-bias assist	60
iii.	Parasitic capacitances reduction	63
6-	Variability as an asset: FDSOI SRAM PUF.....	65
a.	PUF: SRAM based fingerprint	65
b.	Single dopant transport.....	66
c.	Emulation of leaky devices to assist technological choices	66
i.	Impact of channel doping on SRAM devices	67
ii.	Simulation environment	67
iii.	Emulation of resonant transport	71
iv.	Conclusion:.....	73
7-	Conclusion of chapter II	74

Chapter III: Fabrication of junctionless transistor in the scope of 3D monolithic integration

1-	State of-the art	77
a.	3D sequential integration demonstration: review of literature	77
i-	Deposited top-tier channel material.....	78
ii-	Reported top-tier channel material	78
b.	Junctionless transistors	80
i-	Short presentation of the JL transistor (JLT) architectures	80
ii-	Polycrystalline materials	81
iii-	Other materials	82
2-	TCAD simulations	83
a.	Chosen device architectures	83
b.	Physical Model used and justification	85
3-	Junctionless MOSFET operation.....	86
a.	Sub-threshold region: depletion	86
b.	From threshold voltage to flatband voltage: volume conduction	87
c.	Above flatband voltage: accumulation region.....	88
d.	Analytical models.....	88
4-	Characteristics of Junctionless devices	89
a.	Effective channel length modulation.....	89
b.	Mobility.....	89
c.	Capacitances.....	91
d-	Variability	92
d.	Reliability	94
e.	Noise.....	95
f.	Junctionless transistor applications	96
5-	Device sizing	98
a.	Tri-gate junctionless sensitivity to silicon thickness and doping level.....	98
b.	n over p channel	100
i-	PN junction physics.....	100
ii-	CMOS Integration	102
iii-	Sizing of the different layers: TCAD simulations	104
c.	Performances of the different structures compared to IM devices	105
6-	Fabrication process flow.....	108
a.	Gate first integration at high temperature.....	108
b.	Channel material.....	109
c.	Active zone patterning.....	113

d.	Gate stack	114
i-	Gate stack materials.....	114
ii-	Gate stack etching	116
e.	Spacer.....	117
f.	Junction engineering SPER.....	118
g.	Thin silicides	122
7-	Overview of studies related to 3D monolithic integration	123
8-	Electrical results	125
a.	Device fabrication	125
b.	Digital Figure-Of-Merit of Junctionless nMOS	126
i-	Electrical performances.....	126
ii-	Mobility.....	128
iii-	Overlap capacitance.....	129
c.	Analog Figure-Of-Merit of junctionless nMOS	130
i-	Analog gain leveraged by back-bias.....	130
ii-	Reliability and noise.....	131
iii-	RF Figure-Of-Merit of junctionless nMOS.....	133
9-	Conclusion of Chapter 3.....	135

Chapter IV: Assessment of an ultra-dense Non-Volatile Memory cube for In-Memory Computing applications

1-	State of the art of In-Memory-Computing existing solutions.....	138
a.	Existing In-Memory Computing implementations.....	138
i.	Memristors for IMC	138
ii.	Boolean logic.....	139
b.	IMC existing solutions: examples	141
c.	IMC materials/ selectors.....	142
i.	Memristors materials.....	142
ii.	Focus on OxRAM technology.....	143
iii.	Selectors	145
d.	My-Cube project: choices.....	146
i.	Stacked nanowires	147
ii.	Memory element.....	147
iii.	Boolean logic: Scouting logic	147
2-	Sizing simulations	149
a.	Simulated pillar structure	149
b.	Definition of SPICE simulation inputs.....	149

i.	JL performances at W=50nm	150
ii.	Drive current for stacked nanowires at W=75nm.....	151
iii.	OxRAM distribution extraction.....	153
c.	Scouting logic in the pillar	154
d.	MY-CUBE: read and write schemes.....	154
3-	Processing of stacked structures	156
a.	Gate-All-Around stacked nanowires detailed process flow	156
b.	Modification to standard process flow to integrate memory elements.....	157
4-	Variability.....	159
a.	Standard evaluation of the mismatch: Pelgrom plots.....	159
b.	Gate input referred normalized matching parameter.....	161
c.	Drain current local and global variability in all-regimes.....	163
d.	Variability of JAM devices for IMC	165
5-	Conclusion of chapter IV.....	166

General conclusion

1-	Conclusion.....	187
2-	Future Work: short term perspectives	188
3-	Perspectives	189

Glossary

Symbols:

Symbols	Definition	Unit
C_d	Drain capacitance	F
C_g	Gate capacitance	F
C_{gc}	Gate-to-channel capacitance	F
C_{ox}	Oxide capacitance	F
$E_{//}$	Longitudinal field	V/m
E_C	Conduction band energy	eV
E_G	Band gap energy	eV
EV	Valence band energy	eV
E_{eff}	Transverse effective field	V/m
f_{max}	Maximum operating frequency	Hz
g_m	Transconductance	A/V
I_{OFF}	OFF-state current or leakage current of a MOSFET	A (or A/ μ m)
I_{ON}	ON-current or saturation current	A (or μ A/ μ m)
I_{th}	Drain current criterion for threshold voltage extraction	A
k, k_B	Boltzman constant	J/K
L_G	Transistor gate length	m
N_A	Acceptor impurities concentration	Atomes/cm ³
n_i	Intrinsic carriers concentration	cm ⁻³
q	Elementary charge	C
R_{ACC}	Access resistance	Ω (or $\Omega \cdot \mu$ m)
R_{ON}, R_{TOT}	ON-resistance in linear regime and at a given gate overdrive	Ω or Ω/μ m
SS	Subthreshold Slope	mV/dec
t_{ox}	Gate oxide thickness	m
V_B	Back-bias voltage (Body voltage)	V
V_D	Drain voltage	V
v_d	Drift velocity	m/s
V_{DD}	Supply voltage	V
V_{FB}	Flat-band voltage	V
V_G	Gate voltage	V
V_S	Source voltage	V
V_T	Threshold voltage	V
V_{TLIN}	Threshold voltage in linear regime	V
V_{TSAT}	Threshold voltage in saturation regime	V
W	Transistor width	m
ϵ_0	Permittivity of vacuum	F/m
ϵ_{Si}	Permittivity of silicon	F/m
γ	Body Factor	mV/V
θ_i	Mobility attenuation parameters	V ⁻ⁱ
λ_0	Mean free path	m
μ_0	Low-field mobility	m ² /Vs
μ_{eff}	Effective mobility	m ² /Vs
τ	Relaxation time	s
φ_M	Metal work function	eV

φ_s	Semiconductor work function	eV
φ_f	Fermi potential	eV
χ_s	Electron affinity	eV
ψ_s	Surface potential	eV

Acronyms:

Acronym	Definition
3DCO	3D Contact
3DSI	3D Sequential Integration
AFM	Atomic Force Microscopy
AI	Artificial Intelligence
ALD	Atomic Layer Deposition
AMD	Advanced Micro Device
ASIC	Application Specific Integrated Circuit
BEOL	Back End Of Line
BGCO	Back-Gate Contact
BIST	Built-In Self Tests
BL	BitLine
BOX	Buried OXide
BTBT	Band-To-Band Tunneling
BTI	Bias Temperature Instability
CBRAM	Conductive-Bridge RAM
CD	Critical Dimension
CEA	Commissariat à l'Energie Atomique
CMOS	Complementary Metal Oxide Semiconductor
CMP	Chemical Mechanical Polishing
CNT	Carbone NanoTube
DFT	Design For Testability
DIBL	Drain Induced Barrier Lowering
DRAM	Dynamic Random Access Memory
DRC	Design Rule Check
DTCO	Design-Technology Co-Optimisation
DUV	Deep Ultra-Violet
EDA	Electronic Design Automation
ENIAC	Electronic Numerical Integrator Analyser and Computer
EOT	Equivalent Oxide Thickness
EUV	Extreme Ultra Violet
FAST	Field Assisted Superlinear Threshold
FDSOI	Fully Depleted Silicon On Insulator
FEOL	Front-End-Of-Line
FET	Field Effect Transistor
FFT	Fast Fourier Transform
FOM	Figure Of Merit
GAA	Gate All Around
GDS	Graphic Design System
GIDL	gate-induced drain leakage
GNS-LC	Green Nanosecond Laser Crystallization
GP	Ground Plane
HCI	Hot Carrier Injection
HDD	Highly Doped Drain
HDL	Hardware Description Language

HRS	High Resistance State
IC	Integrated Circuit
IM	Inversion Mode
IMC	In-Memory Computing
IMEP-LAHC	Institut de Microélectronique Electromagnétisme et Photonique et le Laboratoire d'Hyperfréquences et de Caractérisation
IMT	Insulator Metal Transition
IOT	Internet Of Things
IRDS	International Roadmap for Devices and Systems
ITO	indium-tin oxide
IZO	indium-zinc-oxide
JAM	Junctionless Accumulation Mode
JLT	JunctionLess Transistor
KMC	Kinetic Monte Carlo
LDD	Lightly Doped Drain
LER	Line-Edge Roughness
LETI	Laboratoire d'Electronique, de Technologie et d'Instrumentation
LFN	Low Frequency Noise
LRS	Low Resistance State
LVS	Layout Versus Schematic
LVT	Low VT
LWR	Line-Width Roughness
MAGIC	Memristor Aided loGIC
MC	Monte Carlo
MIEC	Mixed ionic electronic conductor
MIV	Monolithic 3D Inter Via
MOS	Metal Oxide Semiconductor
MW	Memory Window
NBL	Negative BitLine
NBTI	Negative Bias Temperature Instability
NMC	Near Memory Computing
NMOS	Negative MOS
NW	NanoWire
OPC	Optical Proximity Correction
OTS	Ovonic threshold switch
OxRAM	Oxide-based RAM
PBTI	Positive Bias Temperature Instability
PC	Personal Computer
PCM	Phase-Change Memory
PD	Pull-Down
PDK	Process Design Kit
PEX	Parasitic Element eXtraction
PG	Pass-Gate
PINATUBO	Processing In Nonvolatile memory ArchiTecture for bUlK Bitwise Operations
PMD	Pre-Metal Dielectric
PMOS	Positive MOS
PPA	Power Performance Area
PPAC	Power-Performance-Area-Cost
PPACT	Power-Performance-Area-Cost-Time-To-Market
PU	Pull-Up
PUF	Physical Unclonable Function
RBB	Reverse Back Bias
RC	Resistance Capacitance
RDF	Random Dopant Fluctuations
RFID	Radio Frequency Identification

RO	Ring Oscillator
RRAM	Resistive Memory
RSD	Raised Source and Drain
RTL	Register Transfer Level
RTS	Random Telegraph Signal
RVT	Regular VT
SCE	Short Channel Effect
SCL	Scouting Logic
SEM	Scanning Electron Microscopy
SF	Slow Fast
SIT	Sidewall Image Transfer
SL	Source Line
SNM	Static Noise Margin
SOI	Silicon On Insulator
SOC	System On Chip
SPER	Solid Phase Epitaxy Regrowth
SPICE	Simulation Program With Integrated Circuit Emphasis
SRAM	Static Random Access Memory
SRH	Shockley-Read-Hall
SRRV	Supply Read Retention Voltage
SRS	Surface Roughness Scattering
STT-MRAM	Spin-Torque-Transfer Magnetic Memory
TCAD	Technology Computer-Aided Design
TDDB	Time Dependant Dielectric Breakdown
TEM	Transmission Electron Microscopy
TFT	Thin-Film Transistor
TG	Tri Gate
TRR	Time Resolved Reflectometry
TSMC	Taiwan Semiconductor Manufacturing Company
TSV	Through Silicon Via
TT	Typical (NMOS) Typical (PMOS)
TVS	Threshold Vacuum Switch
VLSI	Very Large Scale Integration
WFV	Work Function Variations
WL	WordLine
WNM	Write Noise Margin
ZnO	ZiNc Oxide

Context:

The first microprocessor was manufactured by Intel in 1971, composed of 2300 transistor on a 10mm² chip (Intel 4004, node 10μm). Its performances were equivalent to the first electronic computer ENIAC (Electronic Numerical Integrator Analyser and Computer) built in 1946 for a total surface of 167m². However, nowadays, the AMD (Advanced Micro Device) ROME detains up to 39.54 billions transistors and is integrated on 1008mm² surface with 7nm node of TSMC (Taiwan Semiconductor Manufacturing Company)[1]. To achieve such a progress, the dimensions have been aggressively reduced (from 10μm to 7nm). At the beginning, during “happy scaling area”, the transistors have been scaled down geometrically. However, at some point, because of physical constraints, some innovation were required to reduce further the dimensions. In this context, several performances booster are introduced, like strain [2] or high-k dielectrics [3]. New transistor architectures such as Fully Depleted Silicon On Insulator (FDSOI) transistors or FinFETs were also developed to mitigate short channel transistor performance degradation. However, as the transistor dimensions are reduced, the density of transistors and interconnection increases, as well as the power consumption per unit area. In fact the performance of a circuit is no longer dictated by transistor performance only but also and mainly by the interconnection delay. The dominant delay for ultra-scaled technological nodes (7nm and bellow) comes from the RC wire delay as indicated by Fig. 1. Furthermore, interconnection congestion limits the area gain when shrinking further the transistor dimensions.

A solution considered by the International Roadmap for Devices and Systems (IRDS) is to stack transistors of top of the other sequentially (called 3D monolithic integration) to achieve an equivalent node in terms of performance and density without scaling further the devices. Connections at the transistor level between two tiers can de-congestion the interconnections, improve the RC delay and increase the performance compared to planar circuit with the same silicon footprint. In this context, chapter II focuses on the performance advantages of such an integration for SRAM and chapter III presents the fabrication and electrical characterization of devices in the scope of 3D monolithic integration.

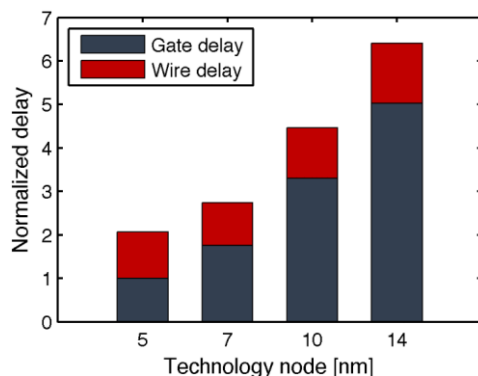


Fig. 1: Gate and Wire delay for advanced technologies nodes, taken from [4].

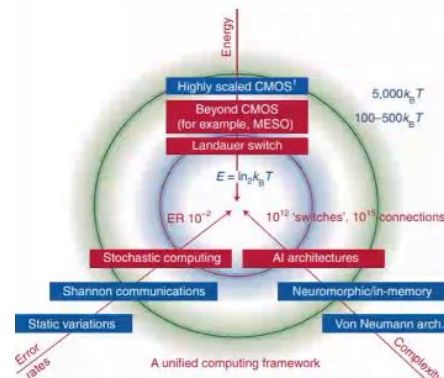


Fig. 2: Schematic presenting the improvements axis for computing: energy, error rates and complexity. Figure taken from [5].

In a similar way, past years were dedicated to lower down the energy required for a computation. However, there are additional levers to increase the overall performance of a circuit, such as playing on the complexity or error rates, improving reliability or lifetime instead of focusing only on energy (Fig. 2). For instance, futures technologies can provide high level functionalized circuits and add value by differentiating. In this PhD manuscript, we propose to reduce the energy of the computation system by reducing the data transmission between memories and computing part. In fact, most of the bandwidth

and the power of nowadays circuits is used to access the memory. To break this memory-wall a possibility is to perform computation (or to pre-process) directly in the memory. In this scope, chapter 4 proposes a 1T-1R cube for in-memory computing (IMC).

Manuscript organization:

This thesis was conducted between CEA-LETI (Laboratory of Electronics, Technology and Instrumentation -French Atomic Energy Agency) and IMEP-LAHC (Institut de Microélectronique Electromagnétisme et Photonique et le Laboratoire d'Hyperfréquences et de Caractérisation) both located in Grenoble, France.

The manuscript is organized as followed:

Chapter 1 is dedicated to the presentation of semi-conductor industry and its current challenges. The history of semi-conductor industry is discussed and the major technological changes to overcome industrial problems are highlighted. New architectures are also proposed to increase electrostatic control and enable to scale down further the transistors. In particular, 3D monolithic integration is discussed as an alternative to traditional scaling in the context of More Moore applications. Also an emphasis is done on in-memory computing, which by gathering memory and computational part promises energy savings.

Chapter 2 consists in the design-technology co-optimization of 3D monolithic SRAM devices. A back-bias assist using specific features of 3D monolithic integration is proposed. SPICE simulations are done using a FDSOI 14nm model card. A performance/area gain is seen with 3D monolithic architecture, making such a technology interesting for more than Moore applications. Also, a SRAM based Physical Unclonable Function for security applications is analysed in depth.

Chapter 3 explains the choice of junctionless devices for 3D monolithic integration, its fabrication in CEA-LETI and electrical characterisation. Sizing TCAD studies are exposed. The low-temperature (<400°C) process flow is detailed before electrical characterization. An in-depth characterisation comparative study is done between junctionless, accumulation and inversion mode devices targeting mixed digital-analog applications.

Chapter 4 is about in-memory computing to reduce the interactions (data transfers) between memory and computation parts. A 3D structure composed of stacked junction transistors co-integrated with memory devices is proposed. Simulations based on junctionless electrical measurements are performed to explore the feasibility of scouting logic. An emphasis is put on junctionless mismatch. Planar JL-RRAM are fabricated to demonstrate the working operation.

Chapter 5 ends the thesis manuscript with a general conclusion, and the perspectives of this work.

Additional details are given in appendix.

Chapter I: Introduction

This chapter presents the thesis work overall context. First a summary of the history of the semiconductor industry is presented, focusing on CMOS technology scaling and nowadays energy and performance challenges. Finally, the last section highlights 3D monolithic integration interest for More Moore applications and In-Memory Computing, which will be explored in this thesis work.

<u>1-</u>	<u>History of Semiconductor industry:</u>	16
a.	<u>Dennards'law: happy scaling era (Moore's law):</u>	16
b.	<u>Physical limit to scaling, apparition of parasitic effects:</u>	17
c.	<u>New architectures (FINFET, FDSOI: back bias)</u>	18
i.	<u>FDSOI architecture</u>	19
ii.	<u>Gate all-around Nanowires or nanosheets</u>	20
<u>2-</u>	<u>Semi-conductor industry: current challenges, roadmaps and propositions to keep the race to technological node</u>	21
a.	<u>Picture of 2020 microelectronic ecosystem</u>	21
b.	<u>Introduction to 3D sequential integration</u>	22
i.	<u>3D sequential integration</u>	22
ii.	<u>3D sequential integration: More Moore applications</u>	22
iii.	<u>3D sequential integration: More than Moore Applications</u>	23
c.	<u>Introduction to In-Memory Computing:</u>	24
<u>3-</u>	<u>Thesis objectives:.....</u>	26

1- History of Semiconductor industry:

a. Dennards' law: happy scaling era (Moore's law):

In the beginning of 20th century, electronics was based on vacuum tubes and permitted the first electronic computer in 1945 (weight: 30 000kg, surface: 167m², power consumption: 150kW and performances: 38 divisions per seconds [6]). Previous computers, like Z3 in 1941 were based on mechanical switches using binary algebra to perform operations. However, the vacuum tube technology became obsolete and is replaced by the emergence of transistor devices. In fact, invented by William Shockley, John Bardeen and Walter Brattain in 1947 (bipolar transistor in 1948), the transistor were more reliable, produced less heat and consumed less power. But before 1958, the discrete transistors were manufactured independently and Jack Kilby suggested that transistors could be integrated on a same substrate and connected together, making the integrated circuit manufacturing closer to nowadays one. And since this time, the microelectronics industry has evolved to provide Personal Computer (PC) in the 90's, democratisation of internet (cable or Wi-Fi in 1998), phones and smartphones in beginning of 21th century, connected objects (Internet Of things) in the past ten years. With the promised of 5G and an ever more connected world for customers, the semi-conductor technologies had to evolved (and will) to provide cheaper and smaller components with more performances and functionalities.

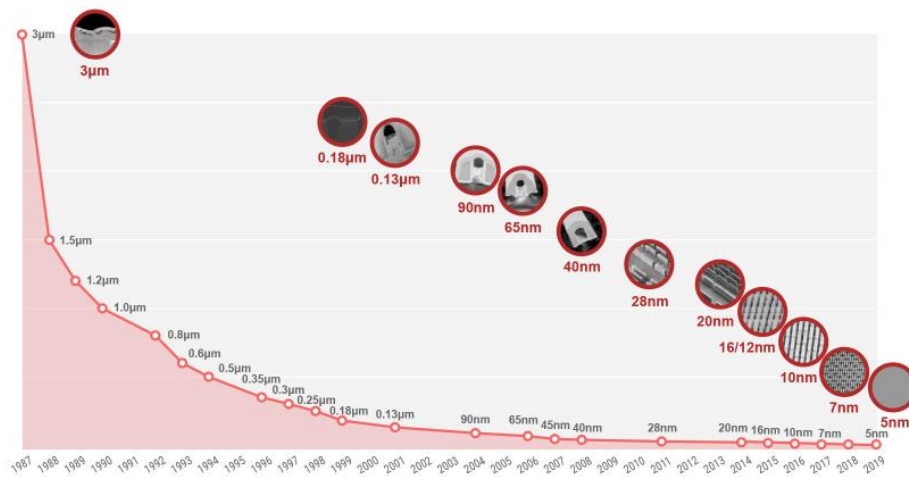


Fig. 3: Evolution of TSMC technology node from 1987 to today taken from [7]. Nowadays, the technology nodes no longer correspond to the smaller dimension but are artificially reduced by a 0.7 factor from one generation to the next one.

In fact, if we have a look on TSMC technology node evolution the past years (Fig. 3), we can observe that in only 30 years, the transistor technology has evolved from 3 μ m to 5nm. This drastically shrinking of dimensions comes along with a price per transistor reduction, mainly induced by a higher transistor density. In fact, in 1965, Gordon E. Moore, co-founder of Intel, predicted that the number of transistors on a chip would double every two years at least for a decade. This declaration, based on six years data (from 1959 to 1965), became the “Moore’s Law” and drove the semi-conductor industry for several years. This transistor miniaturization is also announced by a cost reduction of integrated circuits (IC) as highlighted in Fig. 4. So the interest of shrinking transistor dimensions is mainly a price reduction and a performance increase. The main flavor of transistors used is the Metal Oxide Semiconductor Field Effect Transistor (MOSFET) due to requirements for energy consumption reduction. From this, Dennard *et al.* set straightforward scaling rules based on the “constant-field scaling method” in 1974 to reduce the MOSFET dimensions without additional technological development [8]. In fact, he noticed that from one technological node to the next one, if we maintain the same power density, the transistor dimensions must be scaled by 30% (x0.7) to reduce circuit delay (x0.7) and thus increase operating frequency (x1.4). The supply voltage is also reduced by 30% and the area by 50%. For informative purpose, the scaling factors of the device or circuit parameter are given in Fig. 5. We might observe that by scaling the device

dimensions (t_{ox} , L , W), the current, supply voltage, capacitance or delay are scaled down by the same factor. So, the power density stays identical from one node to the next one, while increasing transistor density and performance. This time is referred today as the “happy scaling” era since there were no trade-off between cost, functionality and performance.

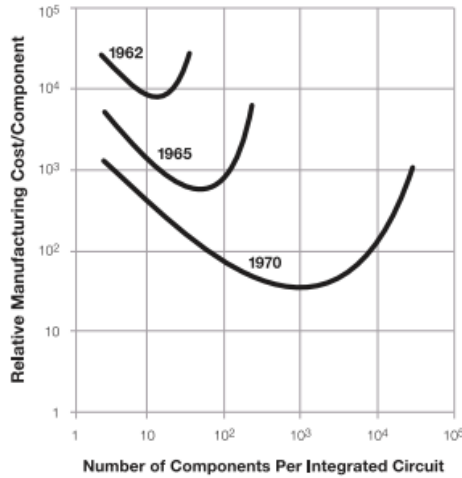


Fig. 4: Manufacturing cost as a function of transistor density, taken from [9].

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

Fig. 5: Dennard's scaling rules: presentation of the scaling factor of device or circuit parameters, figure from [8].

However, for small dimensions the scaling is no longer straightforward (for instance, t_{ox} cannot be reduced anymore) and parasitic effects due to physical limits tend to appear. In fact, beyond 90nm node (around 2005), some adaptations were needed to stick to Moore's law induced industrial roadmap.

b. Physical limit to scaling, apparition of parasitic effects:

The transistor device consists in a three terminal device named gate, source and drain. The substrate can also be biased but is usually kept to ground. The current flow between the source and drain is controlled by the voltage applied on the gate electrode. The ideal MOSFET must be closed (*i.e.* OFF state and no current flow) if the voltage applied on the gate (V_G) is below a threshold voltage noted V_T . If V_G is above this value, the MOSFET is in ON state and a current flows from the source towards the drain. In reality the dissociation of these two states is not abrupt and the invert of the slope corresponding to the transition is called Subthreshold Slope SS (or Subthreshold Swing). In fact thermodynamics' laws impose the limit of $SS = \ln(10) \cdot kT/q = 60\text{mV/dec}$ at ambient temperature below threshold for MOSFET. As far as the threshold voltage is concerned, its values is usually extracted for a drain current $I_{th} = 100\text{nA}W/L$. However, for sub-90nm nodes, in addition to this deviation from ideal working operation, unwanted effects for small gate length, called short channel effect (SCE) has risen. In fact, for small gate lengths the electrostatic control by the gate on the channel is degraded and might not be longer efficient to dissociate ON and OFF state. Among these limitations (including SCE) we can notice:

- Electron/hole mobility degradation.
- Subthreshold slope: the transistor does not switch from ON to OFF abruptly.
- Gate-induced drain leakage current.
- Gate leakage.
- Threshold voltage roll-off: the threshold voltage tends to decreased for smaller gate length.
- Parasitic resistances: the shorter the channel length, more important (in relative) are the source and drain access and contact resistances w.r.t the channel resistance.
- Drain Induced Lowering Barrier (DIBL). In fact for short channel devices, the threshold voltage is no longer independent of the drain voltage since physically, the drain is close enough to the source. It induces a negative threshold shift and a degradation of the subthreshold slope. A measure of DIBL is given by:

$$DIBL = \frac{V_T^{DD} - V_T^{low}}{V_{DD} - V_D^{low}} \quad Eq. 1$$

- Source to drain tunneling: for ultra scaled gate dimensions (physical gate length below 10nm [10]), electrons in the source can directly tunnel to the drain, the probability of transmission being determine by the barrier width/height and silicon effective mass.
- Punchthrough: if the physical gate length is small enough, the source and drain depletion regions can merge, leading to a large undesirable current flow between source and drain.

These degradations results in an increase of the leakage current, limiting further the MOSFET scaling.

Due to these effects, the pure geometrical scaling (Dennard's rules) couldn't be applied anymore. For instance, let us consider the oxide thickness scaling. According to Dennard's scaling rules, the oxide thickness is scaled down for each node to maintain a constant vertical field (together with a V_{DD} reduction) at the expense of an increase of the gate leakage due to tunneling currents. To overcome this leakage, high-k dielectrics have been introduced to increase the gate oxide capacitance C_{ox} without reducing the oxide physical thickness t_{ox} . The Equivalent Oxide Thickness (EOT) is defined as the equivalent SiO_2 thickness of the capacitance made of high-k materials. The formula is expressed as:

$$EOT = t_{high-k} \frac{\epsilon_{SiO_2}}{\epsilon_{high-k}} \quad Eq. 2$$

For these reasons, hafnium-based dielectrics detaining an high permittivity ($k-HfO_2=25$) have been introduced in the gate stack along with metal gate satisfying the 45nm node requirements [3].

In addition, Dennard's scaling imposes a reduction of the supply voltage while the threshold voltage should be maintained not to degrade the leakage current. As a result, the gate overdrive $V_{DD}-V_T$ decreases and $C_{ox} \cdot (V_{DD}-V_T)$ as well. To compensate for the drive current loss, mechanical stress is introduced in Intel 90nm technology [2]. In fact a compressive stress from SiGe S/D for PMOS and a tensile stress from a stressed SiN layer for NMOS will boost the carrier mobility and thus improve performances without impacting the leakage current.

c. New architectures (FINFET, FDSOI: back bias)

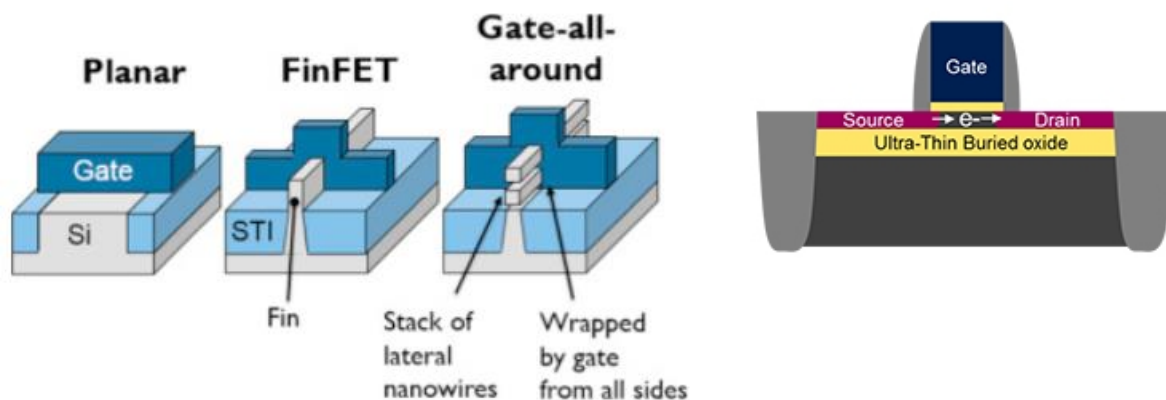


Fig. 6: Presentation of Bulk architecture, FinFET architecture FDSOI and GAA-NW architecture. Figures from [11] and adapted from [12].

To continue Moore's law, the transistors density was required to increase. To counteract SCE, new architectures have risen to improve the electrostatic control of the gate on the channel. Among them we can cite Fin-shaped Field Effect Transistors (FinFET), Fully-Depleted SOI (FDSOI) and stacked Gate-All-Around Nanowires (GAA-NW) or stacked nanosheets. The geometry differences between these devices are outlined in Fig. 6. The main idea is to create smaller channel dimension (either silicon thickness or width) with a higher gate electrode surface to strengthen the electrostatic control. Compared

to bulk technologies, where the silicon used for the transistors channel was thick and wide, the FINFETs architecture proposes to reduce the device width and increases the transistor height, forming a device in a FIN shape. This device electrostatics is controlled by top-gate but also by lateral gates since the gate surrounds the channel. It was first manufactured by Intel at the 22nm node [13]. At the opposite, FDSOI architecture enables a better electrostatic control thanks to the insertion of a buried oxide, which depletes entirely the thin silicon film, preventing leakage currents between the S/D and the bulk. It is then possible to bias the region below the buried oxide and use it as a back-bias to modulate V_T to achieve the best trade-off between performance and power consumption. To go further, devices with a gate wrapped around the channel (GAA) are created to have a full gate control. In this PhD manuscript we will focus on FDSOI devices and GAA-NW ones, which will be the object of next parts. We will present more in details FDSOI and NW architecture in the next sub-sections.

i. FDSOI architecture

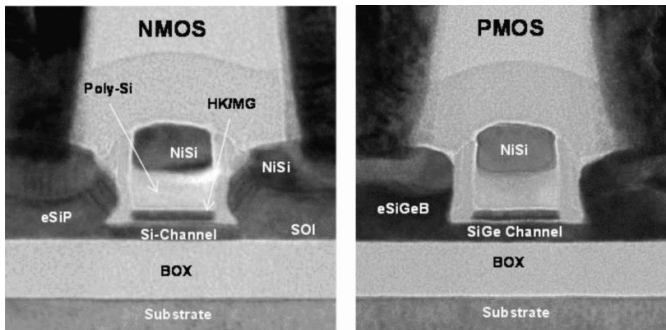


Fig. 7: FDSOI transistor TEM cross-section developed for the 22nm node. Figure from [14].

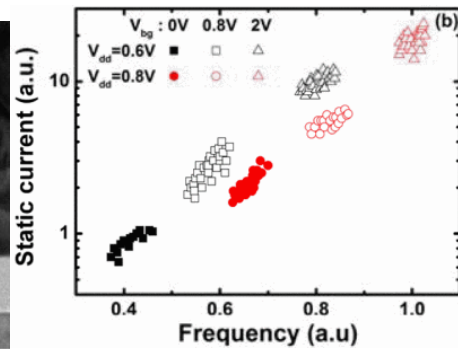


Fig. 8: Static current as a function of frequency for different back biasing. A positive back-bias increases the performance as well as static current. Figure from [14].

First introduced at the 28nm nodes [15] and developed for 22nm [14] and 14nm nodes [16], FDSOI architecture details a thin isolated channel which is well controlled by the gate. Fig. 7 presents a TEM cross-section of a FDSOI device for the 22nm node for both NMOS and PMOS. Note that some of the previously discussed boosters to produce strain in the channel are integrated. Unlike Bulk devices, the channel is on top of a Buried OXide, called BOX, which isolates the device from the substrate. This particular kind of devices uses SOI substrates, which are fabricated with the Smart Cut technique. If the silicon channel is thin enough, the channel can be entirely depleted and in this case the depletion depth is equal to the silicon film thickness. Thus the electrostatics is enhanced compared to planar bulk technologies.

Additionally, there is a coupling between the channel and the body, only separated by the BOX. In fact, the threshold voltage can be modulated by back-bias and ground plane (GP) doping [17]. This modulation is expressed by the body factor $\gamma = \Delta V_T / \Delta V_B$ and is higher for lower values of BOX. Thanks to this modulation, it is possible to switch from a low power state (high V_T) to a high performance one (low V_T). In fact, two back-bias regimes can be distinguished depending of back-bias polarity:

- **Reverse back-bias:** a negative (respectively positive) voltage is applied on the NMOS (PMOS) body, which increases the transistor absolute threshold values and lower the leakage current. Low leakage devices are obtained at the expense of performances.
- **Forward back-bias:** a positive (respectively negative) voltage is applied on the NMOS (PMOS) body, which decreases the transistor absolute threshold values and increases the leakage and drive current. High performance devices are obtained at the expense of power consumption.

As an example, Fig. 8 presents the modulation of a ring oscillator frequency and static current figure of merit with forward back-biasing. The higher the back-bias, the higher the operating frequency is but the higher the static current is.

We can also think of this feature to compensate process variability between dies: a forward back-bias can be applied on slower dies. Also, this modulation is not limited to static compensation but can rather be used in a dynamic way. The use of back-bias will be discussed more in details into chapter II. Additionally, details about FDSOI structure fabrication will be given in chapter III.

ii. Gate all-around Nanowires or nanosheets

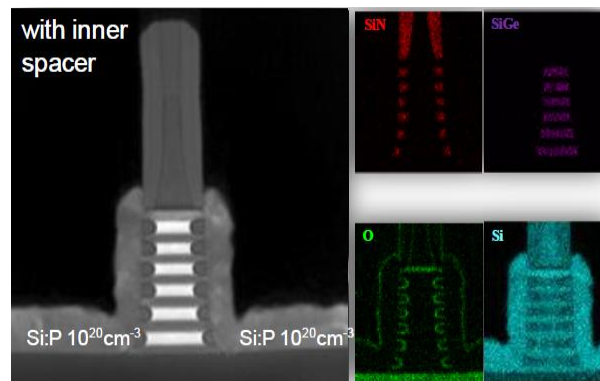


Fig. 9: TEM cross-section of seven stacked nanowires. Figure taken from [18].

The ultimate CMOS device consists in wrapping entirely the channel by the gate to have the best electrostatic control. This structure is called nanowire. However, even if this architecture is relevant to counteract SCE, their small width, due to mechanical constraints, delivers a low drive current. That is why, stacking vertically nanowires to increase the equivalent device width (and thus the drive current) appears as a viable solution. Up to seven stacked nanowires (Fig. 9) have been demonstrated in [18] with excellent electrostatic control. It is also possible to enlarge the transistor width to create nanosheets, which are promising devices for sub 5nm nodes [19]. This structure will be investigated in Chapter IV and details about fabrications will be given.

In this part, the semi-conductor industry history from Moore's law, Dennard's scaling rules to SCE limitations have been presented. Some technological boosters such as the introduction of high-k dielectrics/metal gate to reduce EOT and strain into the channel have been discussed. Later on, some new architecture have emerged to ensure a better electrostatic control of the gate on the channel. In the next part, we will dress an overview of today semi-conductor industry, highlighting challenges and roadmaps.

2- Semi-conductor industry: current challenges, roadmaps and propositions to keep the race to technological node

a. Picture of 2020 microelectronic ecosystem

In France in 2019, 99% (77%) of the 18-24 population (whole population) detains a smartphone [20]. Combined with IOT, which requires back and forth communication between the device and the “cloud”, around 463 exabytes (1000^6) of data which will be generated each day in 2025 [21] and some of them need to be stocked in clusters of servers and memory banks called data centers. A veritable data deluge is predicted, especially with 5G development, deep learning and the democratisation of Artificial Intelligence (AI) and big data. That is why for data centers, there is a need of performance, while mastering power issues.

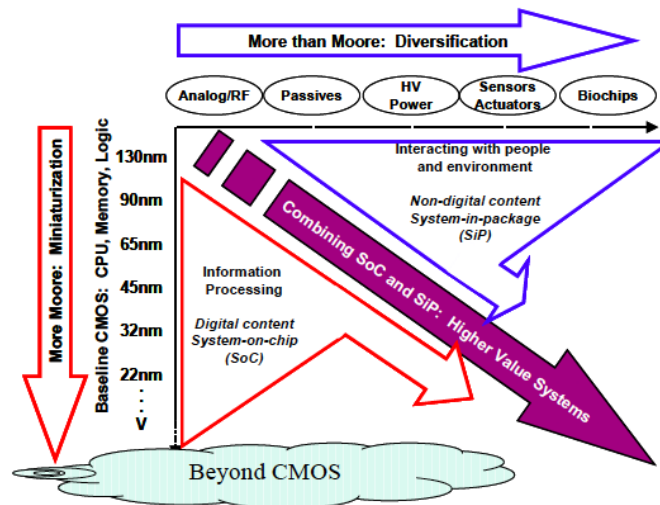


Fig. 10: Presentation of the IRDS roadmap. Even if the miniaturization still drives the semi-conductor industry (More Moore: miniaturization), the diversification towards several applications is desired (More than Moore). Combined together, this paves the way to higher value systems. Schematic from [22].

Targeting these future applications, the International Roadmap of Devices and Systems (IRDS) provides requirements for logic and memory technologies over a 15 years horizon. The main considered points are power, performance, area and cost (PPAC metric). According to their 2020 report, the main applications of nowadays logic technologies is high-performance and low-power/high density logic. Even with the improvement of lithographic tools, such as the extreme-ultraviolet (EUV) tools, the ground scaling is forecast to slow down and saturate around 2028. This traditional scaling must go with design-technology-co-optimization (DTCO) to reduce further the area limited by the design rules. Additionally, the standard transistor miniaturization is limited by parasitic elements but also by the prevalence of interconnections, which dictate nowadays the circuit delay. At the same time, power density poses a serious challenge, which when combined with the scaling of gate drive, could limit clock frequency at 0.8GHz in 2034. From this statement, a new paradigm has emerged, where microelectronics is no longer driven by PPAC but tends to diversify to propose added functionalities to standard devices (Fig. 10). This is called More than Moore applications and is not an alternative to Moore’s law but rather a complement to digital signal and data processing. From one hand, 3D monolithic integration (or sequential integration) by stacking transistors on top of the other, appears as an alternative to standard miniaturization to decrease further the delay between transistors, but also as a lever to add functionalities. This technology will be presented in the next sub-section. From the other hand, the limiting factor for performances is no longer the number of operations per second but rather the speed of communication between logic and memory chips and the associated energy. To break this memory

wall, In-Memory computing proposes to gather memory and logic (computation) to reduce data movement and thus power consumption. It will be discussed in the second sub-section.

b. Introduction to 3D sequential integration

As previously stated, 3D sequential integration is interesting for More Moore and More than Moore applications. After a brief explanation of the technologies characteristics (which will be more detailed in Chapter II and III), we will see how 3D monolithic can be part of industrial roadmaps.

i. 3D sequential integration

3D sequential integration (3DSI), also called 3D monolithic integration consists in stacking active device layers on top of each other in a sequential manner. The sequential term is in opposition with parallel which described an integration (3D parallel integration or 3D packaging) where different chips are processed independently before being stacked and connected vertically. The connections between substrates can be done Through Silicon Via (TSV).

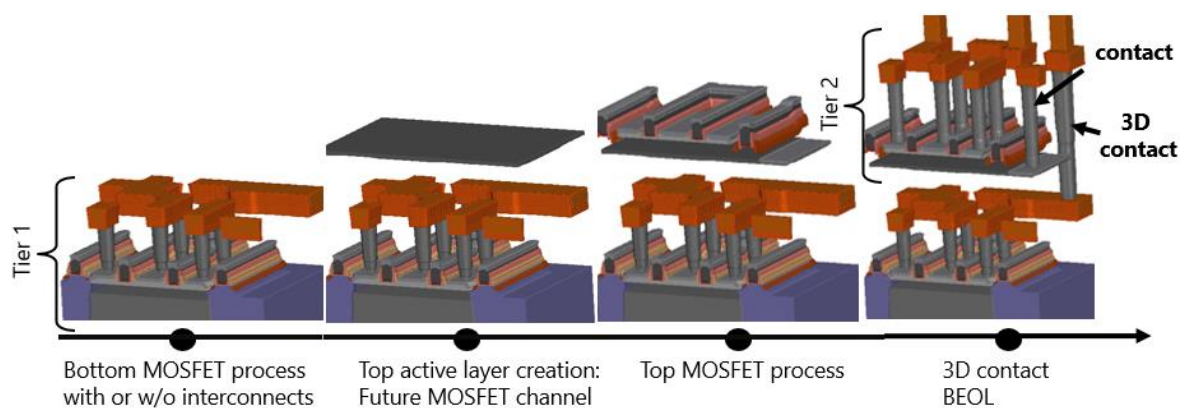


Fig. 11: Presentation of 3D Sequential Integration process flow.

Fig. 11 presents a typical 3D sequential integration process flow, where first the bottom MOSFET tier (bottom tier or tier 1) is processed and followed by the top active layer creation. It can be done either by direct deposition or wafer bonding. This step is detailed in chapter III. The final top active layer is thin enough to align the top transistor with the bottom level. Then the top layer is processed at low thermal budget ($<500^{\circ}\text{C}$, 2hours) to avoid bottom tier degradation. Finally interconnections (3D contact) are done between the two tiers. Unlike TSV (diameter $\sim 1.7\mu\text{m}$), 3D contact retains dimensions similar to traditional ones, offering unique inter-tier connectivity opportunities thanks to precise alignment between tiers. In fact, the alignment accuracy is only limited by stepper resolutions [23]. As far as parallel integration is concerned the interconnection density is limited by the bonding alignment (around 200nm). Additionally, regarding the 3D contact dimensions, a high via density can be reached: over 100 million/ mm^2 is projected with 14nm ground rules in [24].

To conclude this part, 3D monolithic integration enables the formation of multi-tier devices with a high interconnection density between the tiers that is not feasible in TSV technology. Next part will present why a dense interconnection network is required for More Moore applications.

ii. 3D sequential integration: More Moore applications

The first justification of 3DSI was to pursue Moore's law and create an equivalent node by staking instead of shrinking transistors to improve circuit performances. Fig. 12 presents the granularity scale of stacking devices, 3DSI enabling a fine grain interconnection network between tiers that is not achieved with 3D parallel integration. In fact, for an identical silicon footprint (or die size), more devices will be integrated with shorter connections, improving the RC delay which dictates the circuit speed for advanced nodes (see Fig. 13). The gain of performances compared to planar devices are detailed in

chapter II but in a nutshell Shi *et al.* [25] show that a transistor level partitioning in a 14nm technology node yields 20% improved performances among with 30% power saving compared to 2D IC [25].

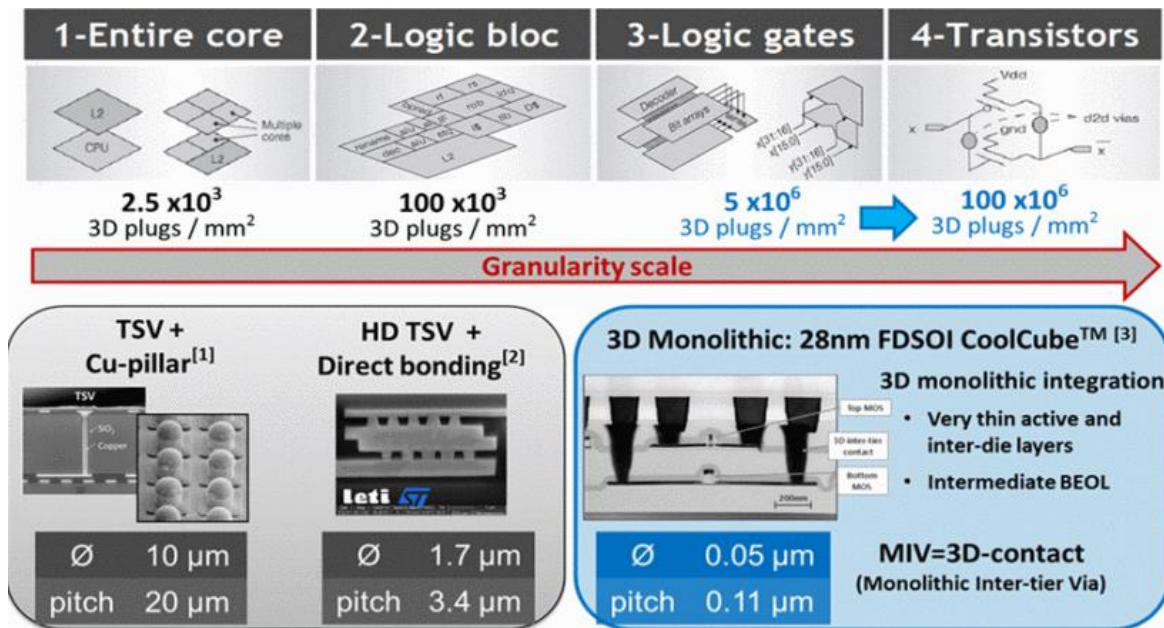


Fig. 12: Definition of the different granularity scale for 3D integration, which are entire core, logic bloc, logic gates and transistor level. Due to the size of the contact, 3D parallel integration is limited to the first two level of abstractions while 3D sequential integration can cover the whole levels. This figure is taken from [26].

iii. 3D sequential integration: More than Moore Applications

In the scope of More than Moore applications, the idea is to integrate different layers types (analog layer, sensors and actuators, memory...) with 3D monolithic integration according to the targeted application. There is already done in other co-integration solutions like System-On-Chip but they are costly (large die size) and the process is not necessarily optimized for all signal domain (analog, digital...). Fig. 14 presents the advantages of heterogeneous integration, in particular 3D stacks to reduce system size, increase performances and reduce cost. Please note that 3D stack is not limited to 3D monolithic integration but can also comprise TSV technology, which, depending of the application, can be more relevant. For instance, we can think of a digital layer with an advanced CMOS node with high performances on top of an analog one with a relaxed node, which is less costly. The connections between analog and digital layers can be either fine grain or between entire blocks. Both technologies could be optimized for each applications (digital and analog in this example).

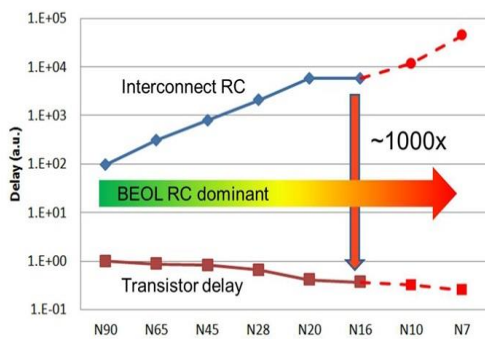


Fig. 13: Transistor and interconnection delay for sub 100nm nodes. Even if the transistor delay is reducing, the overall delay is dominated by back-End of Line (BEOL) RC. Reproduction from [27].

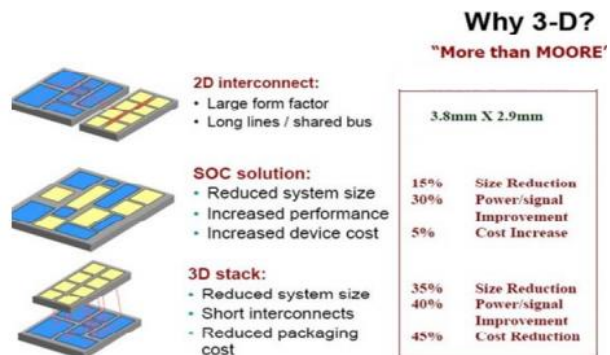


Fig. 14: Comparison between planar, System-On-Chip (SOC) and 3D stack to highlight advantages of heterogeneous integration. This figure is taken from [28].

In this PhD manuscript we will tackle the topic of 3D monolithic technology from both design point of view and fabrication one. In fact, chapter II will describe Design-Technology Co-Optimization required to take all the benefits, in terms of performances, area and power from such a technology. Later, chapter III will propose the physical and electrical analysis of JL transistors and their low-temperature integration ($<500^{\circ}\text{C}$), making them compatible with 3D monolithic integration. However, before going further, we will introduce In-Memory Computing applications, which do not rely on transistor miniaturization to gain performances but rather on gathering the memory block from computational one to reduce data transfer delay.

c. Introduction to In-Memory Computing:

The aim of this part is to provide general knowledge about In-Memory Computing, the context and their application field. For detailed information about the working principle, please refer to chapter IV. First, to present the pre-dominance of data access, various applications are presented in Fig. 15 according to data needs, computational complexity and computational precision. We do observe that either for security, deep learning or scientific applications, the data transfer between memory and computation part is primordial. However, this data exchange is translated into additional latency and power consumption for the well-known Von-Neumann architecture. In fact, due to this computing centric architecture -and not data-centric-, data movements in the memory hierarchy result in 50% energy waste [19] and is the main factor, limiting further improvements in computing performances. This limit is generally referred as the “memory wall”.

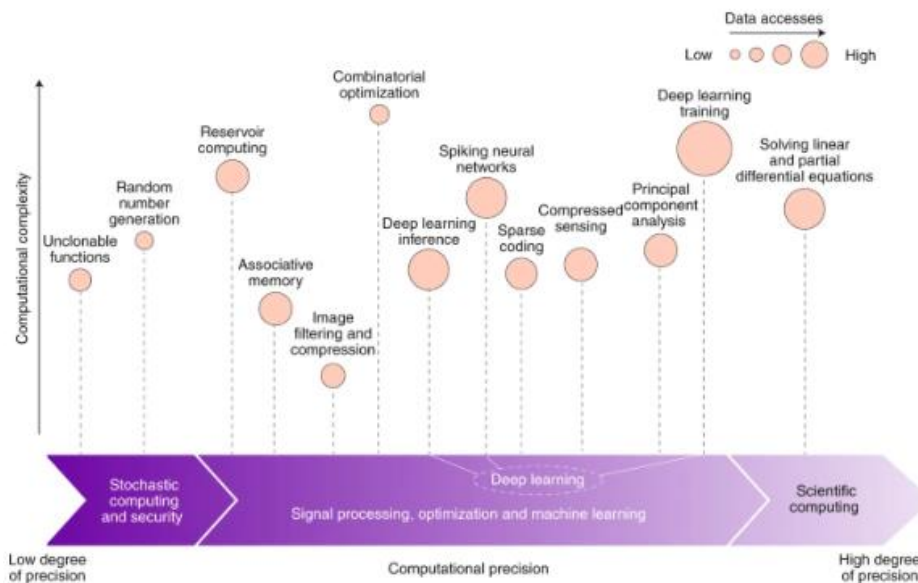


Fig. 15: Data access for various type of applications organised by computational precision and complexity. This figure is taken from [29].

To overcome this limitation, In/Near-Memory Computing (IMC/NMC) rises to be a solution with the co-location of data and logic operations, reducing drastically data movements. The idea is straightforward and illustrated in Fig. 16. In a Von-Neumann architecture, the processing unit will ask the memory block for the data, compute it and transfer again the result into the conventional memory. In an IMC system the processing unit will ask the computational memory block to perform the operations, whose results will be stocked directly into the memory array. For this, they exist several approaches based on charge or resistance memory devices. Several IMC approaches can be found in literature, shared between volatile (DRAM or SRAM) and non-volatile memory (Resistive memories as well as charge storage) with promising energy efficiency.

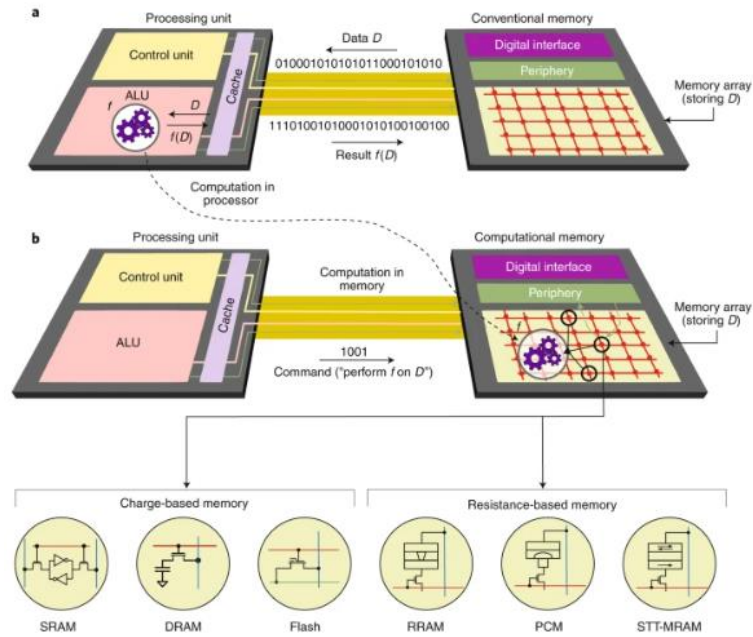


Fig. 16: Illustration of Von-Neumann architecture and In-Memory Computing (IMC) one, taken from [29].

Chapter IV will explain what kind of computation operation can be performed in IMC architectures and propose an implementation of so-called “scouting logic” into a low-power high-density 3D cube.

3- Thesis objectives:

This chapter presented the history of semiconductor industry as well as the current challenges. From the transistor miniaturisation trend enounced by Gordon E. Moore in 1965, a happy scaling era (with constant scaling factor between nodes) lasted until the 21 century. With the shrinking of dimensions, short channel effects limiting the device operation appeared. To mitigate them, boosters have been introduced and new device architectures rose. Nevertheless, digital circuit performance are no longer dictated by intrinsic transistor delay but rather by interconnections. At the same time, with the increase of transistor (and interconnection network) density, power consumption and dissipation is now an issue. From both aspects, 3D monolithic integration by staking transistors on top on the other can solve these issues by enabling shorter interconnections and lower silicon footprint. Chapter II will explore 3D monolithic designs to analyze the PPA gain from planar to 3D designs. For the manufacturing point of view, chapter III describes the fabrication of low-temperature junctionless transistors and their electrical characterization. Additionally, it is also possible to merge memory and computational part to avoid data transfers (*i.e* save energy) through separated blocks. In-Memory computing is foreseen as an alternative to Von-Neumann architecture for efficient and low power computation. In this scope, Chapter IV proposes a low-power high-density 3D cube. Simulations based on experimental data demonstrates Boolean operation feasibility.

The main topics tackled in this manuscript are:

Chapter II:

- Proposition of a 3D VLSI design flow.
- How to share resources between different tiers? How efficient is the partitioning?
- Can we take benefit from the 3D architecture to integrate back-planes for top-tier transistors?
- SRAM as physically unclonable functions.

Chapter III:

- TCAD comparison of JL devices, n/p devices and inversion-mode one.
- Description of junctionless devices process flow at low-temperature.
- Electrical characterization of Junctionless and Inversion-Mode devices (analog applications, digital FOM and variability).

Chapter IV:

- Introduction of a 3D cube co-integrating junctionless nanowires and memory elements for IMC through “Scouting Logic”.
- Choice and sizing of the materials.
- Presentation of the process flow.

*Chapter II: Design Technology Co-Optimization:
functionalities provided by 3D monolithic
integration*

3D monolithic integration is foreseen as an alternative to traditional transistor scaling to pursue Moore's law. Stacking devices with a fine grain contact grid between tiers allows the reduction of the wire length and could leverage new architectures improving both performance, power and silicon footprint. The aim of this chapter is to optimize 3D structure design with such a technology and quantify the gain provided. In the first part, the VLSI digital planar design flow is presented with insights and modifications required to create a 3D one. In the second part, the state of the art of 3D design assessment is done in terms of performance, power consumption and area. In the third part, the 3D environment used in this PhD work is presented. Then, 3D monolithic routing, wire decongestion and design guidelines of back gate contact are discussed. Afterwards, a specific assist technic for 3D monolithic SRAM is proposed to compensate SRAM deviation from reference one. This technic is enabled by a specific feature of this technology: the back gate integration. To finish with, variability in SRAM is used as an asset to generate physical unclonable function for security purposes.

1-	VLSI digital design flow	29
a.	Overview of a planar digital design flow and EDA tools	29
b.	Power, performance and area (PPA) design trade-off	32
c.	From 2D to 3D digital design flow	33
i.	3D Design flow	33
ii.	Netlist partitioning: examples	33
2-	State of the art of 3D design performance assessment: Motivation for 3D monolithic integration for digital applications	35
a.	Cost analysis	35
b.	Thermal dissipation issue	35
c.	Performances	37
3-	3D design MOSFET environment	39
a.	3D tier and intermediate BEOL for CMOS over CMOS integration: Coolcube™	39
b.	SPICE model	40
c.	Parasitic element extraction	40
d.	Methodology summary	41
e.	RO, SRAM benchmark: typical figure of merits	42
i.	Ring Oscillator	42
ii.	SRAM	42
4-	Routing in 3D designs	45
a.	Buried power rail	45

b.	Congestion mitigation and resources sharing between tiers	45
c.	Design guidelines for top-tier Back-plane contact	48
i.	Simulated structure	48
ii.	Static consideration	49
iii.	Dynamic consideration	50
5-	Design-technology co-optimization: top-tier SRAM	51
a.	14nm technology performance	51
i.	Electrical characterization of typical FOM	51
ii.	SRAM: variability issue and impact on FOM	53
iii.	Back-bias assist	54
iv.	BTI-induced dynamic variability at the bitcell level	55
a-	BTI mechanism	55
b.	BTI at the bitcell level: experimental results	56
c.	Proposition of a novel fine-grain back-bias assist techniques for 3D-monolithic 14nm FDSOI top-tier SRAMs	59
i.	3D monolithic design kit: layout considerations	59
ii.	Fine grain and versatile back-bias assist	60
iii.	Parasitic capacitances reduction	63
6-	Variability as an asset: FDSOI SRAM PUF	65
a.	PUF: SRAM based fingerprint	65
b.	Single dopant transport	66
c.	Emulation of leaky devices to assist technological choices	66
i.	Impact of channel doping on SRAM devices	67
ii.	Simulation environment	67
iii.	Emulation of resonant transport	71
iv.	Conclusion:	73
7-	Conclusion of chapter II	74

1- VLSI digital design flow

First, this part presents the VLSI planar design flow commonly used to design complex circuits. Then, an emphasis is done on power, performance and area (PPA) metrics. To finish, the adaptation of the planar design flow for 3D monolithic technology is presented.

a. Overview of a planar digital design flow and EDA tools

With an increasing number of transistors to manage, the Very Large Scale Integration (VLSI) design flow has become automated. It is composed of various sequential stages with a high level of abstraction to build complex circuit up to billion of transistors. Fig. 17 presents the sequential steps to generate the layout. Let's consider the example of a ring oscillator to explain the different building blocks.

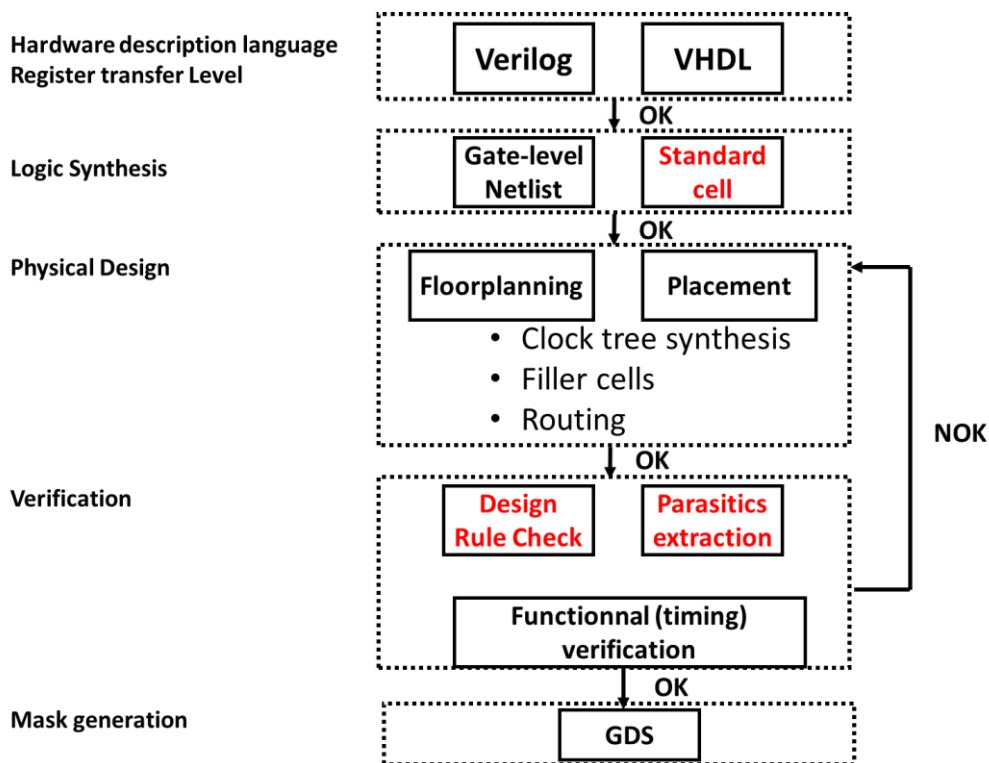


Fig. 17: Usual planar design flow. The part tackled in this work are highlighted in red.

A ring oscillator (RO) is a device composed of an odd number of NOT gates (inverters) in a ring. The output oscillates between two voltage levels, representing true (noted 0) and false (noted 1). The NOT gates, or inverters, are attached in a chain and the output of the last inverter is fed back into the first. A three stage RO is presented in Fig. 18 and its ideal output in Fig. 19.

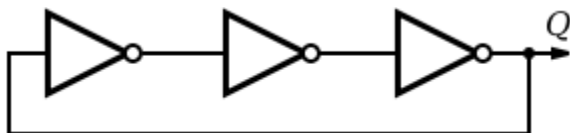


Fig. 18: Example of a three ring inverter. The output frequency depends on the inverter delay τ and is $1/6\tau$.

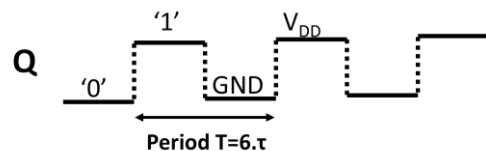


Fig. 19: Schematic of the desired waveform output. An oscillation is expected from a low state (gnd, '0') to a high state (V_{DD} , '1').

One practical way to represent this ring oscillator is to code it using a hardware descriptive language (HDL) like in Fig. 20. For instance, Verilog or VHDL can be used to model a synchronous digital circuit in terms of the flow of digital signals (data) between hardware registers, and the logical operations

performed on those signals. The described circuit is usually synchronous, *i.e.* the change of state of each memory element is regulated by a clock signal.

```
library ieee;
use ieee.std_logic_1164.all;
entity ring_oscillator is
    port (ro_en : in std_logic;
          delay : in time;
          ro_out : out std_logic);
end ring_oscillator;

architecture behavioral of ring_osc is
    signal gate_out : std_logic_vector(2 downto 0) := (others => '0');

begin
    process
    begin
        process
        begin
            gate_out(0) <= ro_en and gate_out(2);
            wait for delay;
            gate_out(1) <= not(gate_out(0));
            wait for delay;
            gate_out(2) <= not(gate_out(1));
            wait for delay;
            ro_out <= gate_out(2);
        end process;
    end process;

end behavioral;
```

Fig. 20: This three inverter ring oscillator code is given as an example. The input/output ports are highlighted in red. The input delay have been added to be able to simulate the RO at this stage. Also, when the logic is synthesized without specific constraints, the redundant logic cell are suppressed and the ring oscillator described above will be replaced by a single inverter.

Then, the synthesis tool considers the combinational and sequential logic described by the HDL at the RTL level and synthesises the logic. It means that the RTL blocks are associated to the smallest level constructs called standard cells. The standard cells come from a library and perform specific operation. For instance, an inverter (Boolean function NOT) with input I and output O can be a standard cell. More complex structure such as 2-bit full adder are also available in the standard cell library. The layout of standard cell are fixed height (but variable width) to ease their future placement in rows. For instance for the 14nm, the standard cell height is 880nm, delimited by power rails. They are optimized full custom layout, minimizing delay and area. Usually they are designed by the Application Specific Integrated Circuit (ASIC) manufacturer and are presented under several views such as symbols or electrical schematics (see Fig. 21). The final collection of standard cells and the required electrical connections between them is called a gate-level netlist. A timing analysis can be done at this stage to ensure the proper operation of the circuit.

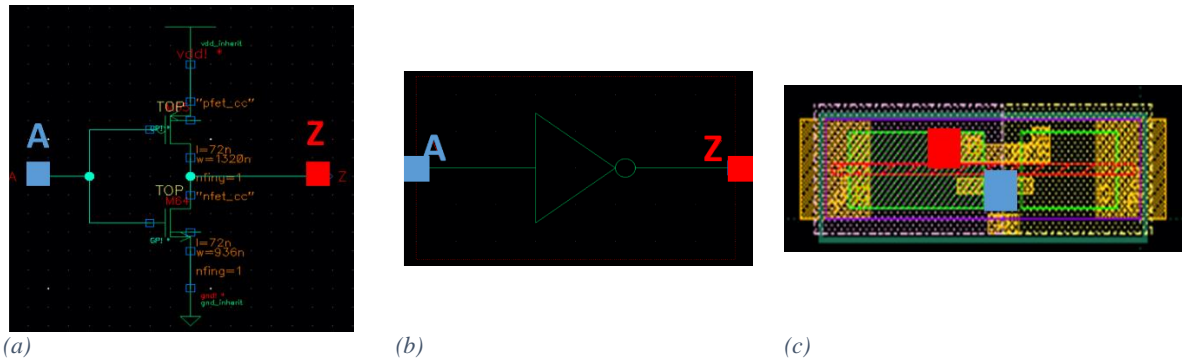


Fig. 21: Different representation for the same entity (a) Electrical schematics of an inverter taken from [30]. The NMOS and PMOS are represented and the pin in blue materialized the input port (A) and in red the output port (Z). (b) Symbolic representation, the inverter is seen as a black box with input (A) and output (Z). Behind this representation, the circuit in (a) is implemented. That is why this entity can be directly used in more complex circuit. (c) Associated layout of the inverter.

After, the physical design consists in placing and optimizing the gate position of the netlist on a floorplan. It is possible to define a specific partitioning to separate some blocks from the others. Once the gates are physically placed, the clock tree is synthesised to drive correctly the flip-flops and minimize the skew and insertion delay. Filler cells complete the unused space to ensure performance and reliability. Then, the root tool will make physical connections between the standard cells with back-end metal rails and via. Usually, the wire length is minimized to avoid additional delay but should not lead to a wire congestion. An example of the obtained layout is presented in Fig. 22. From a general point of view, all the tools search to reduce area, timing (increase performance) and power consumption. Some specific requirement can be done on a constraint (maximum power consumption for instance) at the expense of the others. However, if the constraints are too restricted, the place and root tool cannot find a solution and a trade-off between power, performance and area must be figured out.

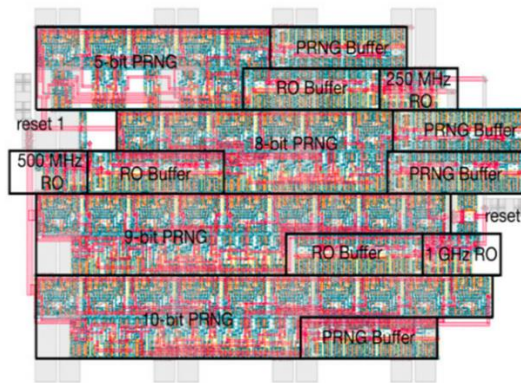


Fig. 22: Example of a layout combining several Ring-oscillators, physical random number generation taken from [31].

Final physical verifications are done prior mask generation. For instance, a Design Rule Check (DRC) ensure that the generated layout respect the design restrictions for device processing. As an example, the DRC contains spacing rules between metallic layers to make sure that they are electrically independent. A specific DRC is done for each technology. Also, the circuit timing is verified (and thus proper circuit operation is ensured) considering all the parasitic elements (capacitance, RC wire delay...). Waveforms function can be generated for timing analysis. Note that similar verifications are done for each step of the design process flow but are not detailed.

To finish, the Graphic Design System (GDS) is generated. It is a binary file format which represents planar geometric shapes, text labels, and other information about the layout. It can be directly used to generate masks for future device processing.

The general planar VLSI design flow have been presented, but before going further and propose a 3D alternative, the trade-off between power, performance and area will be explained to give insights of design optimization.

b. Power, performance and area (PPA) design trade-off

When introducing a new technology node, the progress compared to the previous one are usually shown in terms of gain on power consumption, performance and area. For instance TSMC 7nm node provides a 20% speed improvement at iso-power, a 40% power decreased at iso-speed and a density multiplied by 1.6 with respect to TSMC 10nm node [7]. From this marketing announcement, three important criteria can be figured out: power, performance and area. Some variants of this metric (not used here) are Power-Performance-Area-Cost (PPAC) and Power-Performance-Area-Cost-Time-To-Market (PPACT). In fact, increasing the transistor density from an N-1 node to an N node means increasing the integration capability. It also implies shorter connections between devices and less silicon used to perform similar operation. Thus, with a lower silicon footprint, the same operation should be cheaper to perform from N-1 to N node. For the performance aspect, speed (or frequency) is a good indicator to see if the N node is better than the N-1. However, nowadays, power is a major concern. The first reason concerns the power density, which increases drastically when the dimensions shrunk and can lead to device overheating and prematurely aged components. A second reason is the need for low energy devices, such as for Internet Of Things (IOT).

As far as power is concerned, it is possible to reduce the overall chip consumption by optimizing the circuit-level power at the expense of area or performance. Static power must be differentiated from dynamic power and can be express as:

$$P_{tot} = \sum_n \left(\frac{1}{2} \cdot \alpha \cdot f \cdot C \cdot V_{DD} + I_{leak} \cdot V_{DD} \right) \quad Eq. 3$$

With n the number of gates, α the activity factor for each gate, f being the transistor frequency, C the charging capacitance, V_{DD} the operating voltage and I_{leak} the leakage current. The switching activity factor is a number between 0 and 1 representing during a clock cycle how often the transistor will be ON.

When considering Eq. 3, an efficient way to lower both dynamic and static power will be to reduce the supply voltage V_{DD} [32]. For instance, wider transistors can be designed to deliver the same amount of drain current but at a lower operating voltage (area penalty). The V_{DD} can be directly lowered down at the expense of speed circuit (performance penalty). Also, part of the circuit can be shut down ($V_{DD}=0$) when unused with power gating technics to lower leakage current [33]. In fact a high V_T sleep transistor is added to shut off power supply of part of the design. Similarly, clock gating technics can be used to prevent the clock input to idle modules [34]. The granularity of power gating (or also clock gating) can be adapted to the circuit but increases both area and time delay. The switching energy can also be reduced by carefully designing different frequency domains or using techniques such as dynamic voltage and frequency scaling [35]. Some optimisations can be also done during logic synthesis such as path balancing [36] or state encoding [37].

To conclude this part, power, performance and area are part of a trade-off and is design dependent. Each circuit should be designed with specific constraints in mind. Next part will present the modifications done to the planar design flow to create 3D designs.

c. From 2D to 3D digital design flow

i. 3D Design flow

For 3D monolithic design, the first two steps (hardware description and logic synthesis) are unchanged. Basically, all the steps will remain the same, except that the place and route tool must consider a floorplan with several tiers (here two) instead of one. At this time, there is no commercial dedicated 3D floorplanning and routing tool. The practical way is to separate the 2D netlist into netlist 1 and netlist 2 using a given separation strategy prior placement. Then each part is partitioned to a specific floorplan accounting for tier 1 and 2. The tool will map netlist 1 to floorplan 1 and netlist 2 to floorplan 2. However, the tiers are not independent and are connected by 3D contacts (3DCO). That is why a step of 3DCO placement is inserted in the standard planar design flow (see Fig. 23).

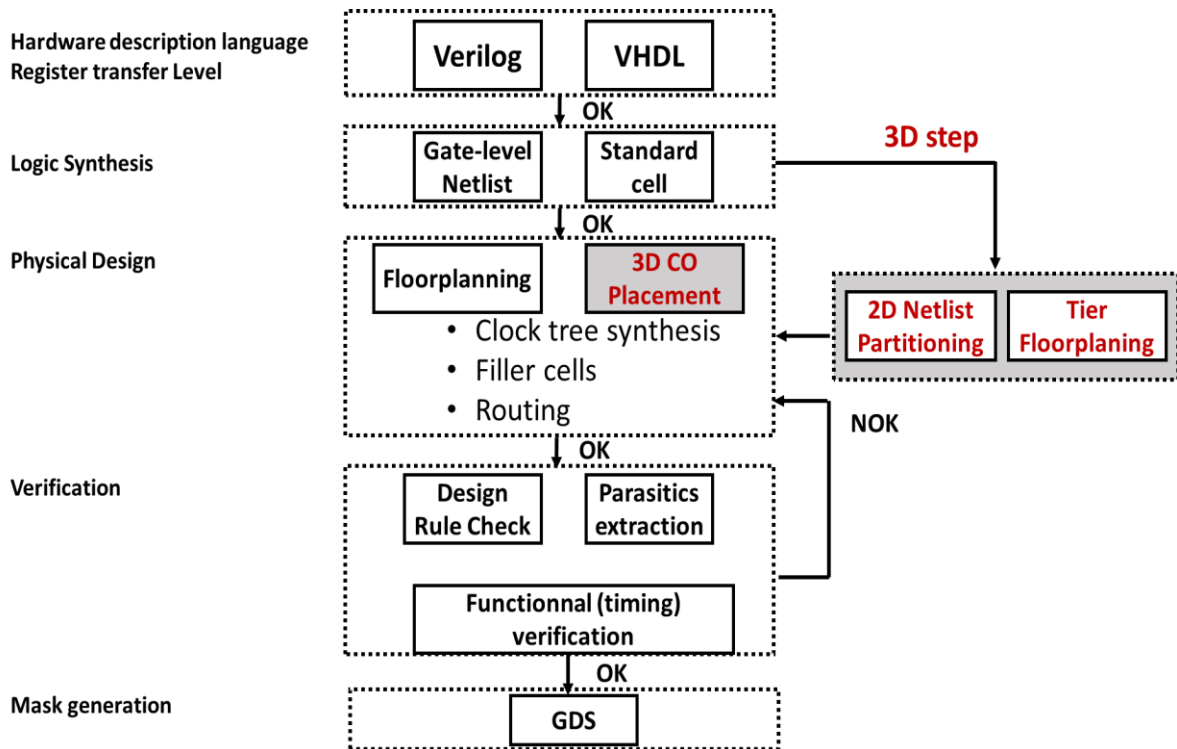


Fig. 23: 3D design flow. The modifications compared to planar one are highlighted in red. A floorplan partitioning step of the 2D netlist generated by logic synthesis is added. Also, 3DCO are placed.

ii. Netlist partitioning: examples

Several partitioning technics have been proposed to separate the netlist into two parts [38], [39], [40] accounting for different optimization strategies. For instance, Sarhan *et al.* [39] propose to sort the wire lengths after a 2D placement. Then, wires longer than a defined threshold will be cut, *i.e.* partitioned into two tiers to reduce the wire length. For instance, the length cut-off threshold can consider the maximum number of 3DCO needed. Also, some specific interconnections can be constrained to a specific tier for optimization. However, these technics tend to limit the number of 3DCO or Monolithic 3D Inter Via (MIV) and do not take fully advantage of the 3D architecture. In [41], the mathematical formulation of MIV placement is presented and a new partitioning tool based on simulated annealing algorithm coupled with a dedicated cost function is presented. The iterative algorithm minimizes the wire cost and balances the area between both tiers without limiting the number of MIVs. Compared to min-cut algorithm, the total wire cost is reduced by up to 44% [26].

With such an approach, the overall wire length will decrease and less delay and parasitic elements will be associated to wires. It can be intuited an overall performance gain. Also, the addition of a third dimension could enable the reduction of critical paths and buffers and repeaters to achieve a gain on power consumption. If the gain of area from 2D to 3D is straightforward, the advantages of a 3D technology for both power and performance must be analyzed. The next part presents a literature review on 3D monolithic gain assessment.

2- State of the art of 3D design performance assessment: Motivation for 3D monolithic integration for digital applications

a. Cost analysis

The first evaluation of 3D monolithic technology concerns its cost. In fact, to be industrially envisioned, this technology must be cheaper to produce chips achieving similar performance than planar technology. Two types of cost intervene: the design cost (additional EDA tools, 3D design engineers...) and the manufacturer cost (processing cost such as die, metal, bonding and cooling cost [42], [43]). In this part, only the cost to fabricate 3D wafers is analysed since the design cost will tend to the planar one when the technology will be mature. Cost analysis of 3D monolithic chip is not straightforward. In fact, stacking devices increases device performance but add complexity for the process. As presented in [43], a 3D cost model expressed for TSV technology must consider wafer/die yield, wafer test cost, stacked die test cost, die area, I/O count, package yield, number of TSV, die temperature and bonding yield. For 3D monolithic, some of these indicators are lower but others are higher. For instance the 3D fault free dies are the combination of a bottom tier fault free, interconnections fault free and top-tier fault free, and intuitively 3D defectivity should be lower. Furthermore, the processing time for 3D monolithic wafers is longer since more steps are required to fabricate the intermediate metal lines and top-tier but results in a larger amount of dies [44]. Gitlin *et al.* [45] consider 3D yield (composition of Bose-Einstein yield) to provide a 3D cost model and investigate different scenario such as CMOS over CMOS, nMOS over pMOS. Up to 50% lower cost is seen for large die (~250mm²) with a CMOS over CMOS integration (28nm 12ML and 4 intermediate BEOL). This range of benefits is found back for more advanced node. In fact -50% cost is seen compared to planar devices for a 400mm² die with a transistor-level partitioning design for 14nm technology node [25]. As far as the 7nm node is concerned, 33% die cost reduction is seen compared to standard planar for 125mm² die area for heterogeneous (memory and logic part are separated from analog and IO which are manufactured in the N28 technology at top-tier) 3D monolithic integration [46].

To conclude, cost analysis of 3D monolithic integration indicates an opportunity and a motivation to develop dedicated process flow and explore 3D designs. Nevertheless, thermal dissipation is an issue in nowadays TSV technologies and before going further, we have to ensure that this technology can efficiently dissipate heat.

b. Thermal dissipation issue

Thermal dissipation is a widely known drawback of bonding technologies, it could even be a potential show stopper [47]. In the general category of bounding technologies, which provides 3D solutions, we can cite TSV (Though Silicon Via), Face-to face Copper-to-Copper (F2F Cu-Cu) or hybrid bonding. In this part, we will compare mainly TSV and 3D monolithic but most of the argument are also valid for 3D technologies in general. In fact, the reduction in footprint area increases the power density by the same factor. The heat is generated by Joule effect in the MOS transistor and wires and can propagates though the network of dielectrics and metal lines. Silicon and copper detain high thermal conductivities (150 and 390 W/m.K). However, with the circuit miniaturization, the interconnections shrinks, having a higher resistivity. That is why, the heat density in nowadays circuits reaches high values. The impact of this temperature rise for circuit can be divided in two categories. The first one is about the physical impact on the single MOS transistor. The electron and hole mobility in silicon is reduced with temperature [48], decreasing the transistor drain current. At the same time, the bandgap decreases, which leads to higher leakage current and thus, increases the Joule effect. The reliability is also limited and the electromigration issues decoupled. Usually, a temperature limit of 125°C is fixed for CMOS devices. The second category concerns the discrepancies of this temperature rises. In fact, high computational

systems, such as computing cores, will present locally high heat density flux, called hot-spot. Due to the presence of hot spots, the temperature of a chip can vary by up to 30°C [49] and impacts the variability and reliability of the circuit. Significant intra-die performance difference are seen due to the temperature difference. Dissipation technology can use natural (heat sink, localized or not [50]) or forced (fans) convection to get rid of the heat and manage hot spots. Furthermore, a power-driven design optimization can mitigate hot spots. For instance, parallel processing (two spaced cores instead of one) will decrease the local rise of temperature. That is why it is important to understand and monitor the dissipation paths for device performances.

As far as the 3D integration is concerned (both TSV and monolithic), the dissipation paths becomes 3D and thermal coupling between tiers appears. However, not the same materials and sizes of via and intermediate layers are used between TSV technology and monolithic one. In fact Santos *et al.* [51] show that the copper pillars for TSV standard technologies [52] has poor thermal hot-spot dissipation results since the underfill layers necessary for stress issues have poor thermal conductivity. However, Coolcube™ technology (3D monolithic technology from CEA-LETI) has a good thermal coupling between tiers, mainly due to the thin dielectric layers and the absence of bulk silicon [53]. For instance, the peak temperature of an 8 stacked dies is below 100°C for Coolcube™ and around 140°C for TSV. The corresponding thermal maps are given in Fig. 24, highlighted that Coolcube™ technology dissipates efficiently hot spots. In the case of a uniformly distributed power density, the heat flow is vertical and depends mainly on packaging. Brocard *et al.* [54] analyzed the thermal difference between top and bottom tier for a 3D buffer. In fact, for some applications (like analog), it is important than the top and bottom device performances, depending on temperature, matches. This deviation increases with load capacitance and decreases with routing capacity. The worst case is 7°C difference between top and bottom devices.

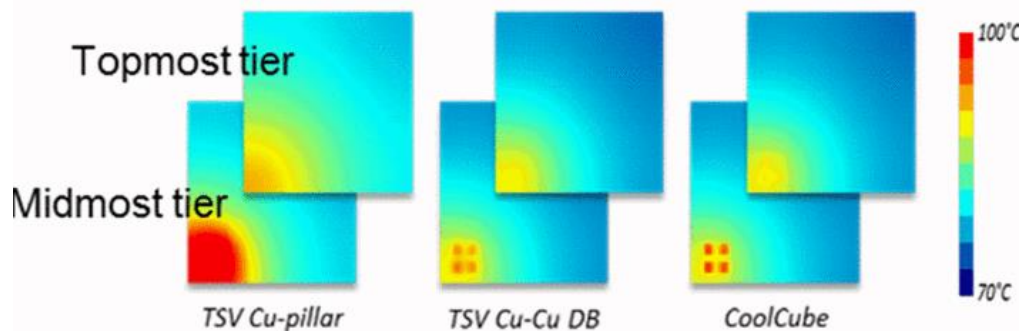


Fig. 24: Thermal maps of the middle and topmost tiers in the case of hot spot dissipation in a 8-die stack: a) TSV-based with cu-pillars; b) TSV-based with hybrid direct bonding; c) Coolcube™ taken from [51].

Similar results with thermal maps are expressed in [53] for the comparison between TSV and 3D monolithic technologies. The impact of TSV on thermal dissipation, unlike 3DCO, due to their large dimensions is emphasized. A fast thermal model is proposed to accurately analyze 3D monolithic designs. Based on this model, a thermal aware floorplanning algorithm is proposed. The floorplanner is run a first time with a wire length cost function and area constraint. Then, floorplaning is done again to minimize the temperature without impacting the area constraint (5% area slack). The 3DCO are not minimized.

Similarly, Hung *et al.* [55] proposed a 3D evolution of the 2D tool hot spot developed in [56] to estimate the chip by thermal-electrical duality. Based on, the 3D thermal-aware floorplanner shows the importance of taking into account the interconnections power consumption to estimate the peak temperature. Up to 15°C peak temperature difference is seen when the interconnections are not considered for an Alpha microprocessor. Also, a maximum on chip temperature reduction by 56% is demonstrated in [57]. The thermal aware floorplanning algorithm combines a resistive model

representation (accurate but long) and a closed-form model (faster but less accurate) to measure the thermal effect.

To finish with, Falkenstern *et al.* [58] propose to develop concurrently the 3D floorplan and the Power/Ground network to minimize IR (ohmic) drop due to the introduction of the Power/Ground network. The addition of a power delivery network to a 3D OpenSPARC T2 processor core design reduces by 48°C the maximum temperature [59] but without considering the increased congestion during wire routing especially because of 3DCO placement. When this additional constraint is considered, the power delivery advantage is more mitigated but up to 13.9% signal wirelength and 17.6% total power reduction is obtained in [60] for a 7nm advanced encryption standard (AES) design.

To conclude this part, unlike TSV technologies, 3D monolithic integration detains a good thermal coupling between tiers and can efficiently dissipate hot spots. Furthermore, several thermal driven floorplan algorithms are proposed in the literature to lower the peak temperature. Next, we will investigate the performance gain by going to the third dimension.

c. Performances

We explained the interest of a 3D monolithic integration to reduce the die cost and how to alleviate thermal dissipation issues. Now we will tackle the 3D circuit performances and investigate the speed, the power and area gain compared to 2D circuits or TSV-based 3D ICs. For a fair comparison, the gain (for example of area) is done with the other metrics fixed (iso-speed and iso-power). The general assets of this technology will be detailed and some specific design cases taken in the literature will be described.

Like for TSV-based circuits [61], the motivation of stacking transistors are miniaturization, reduction of interconnects delay, increase of the memory bandwidth and the possibility of heterogeneous integration. However, unlike TSV, 3D monolithic integration achieves a higher via density [24] thanks to the excellent alignment between tiers limited only by lithographic tools. In fact, over 20 million/mm² have been demonstrated [62] and up to 100 million/mm² is envisioned for 14nm rules [63]. This high contact density enables connections between tiers at the gate level without adding wire complexity and congestion. Also, the wire lengths are even shorten between blocks, reducing its capacitances and lowering the Energy Delay Product (EDP). Shi *et al.* [25] show that a transistor level partitioning in a 14nm technology node yields 20% improved performances among with 30% power saving compared to 2D IC. In addition, 3D designs can take advantage of the coupling between tiers or the dynamic threshold voltage modulation thanks to back-gate integration for top devices [64].

As far as the heterogeneous integration is concerned, several groups in the literature propose original stacks for a specific application. For instance, the next generation of 5G devices combined with the increasing connectivity of IoT devices will induce a real “data deluge”. To manage all this information, high speed systems gathering separate chips, each optimized for a specific application (RF chip, radio chip, digital chip...) are used [65]. In this case, 3D monolithic integration can be an asset to gather different optimized technologies to create a hybrid chip considering all the technology boosters. Similarly a smart pixel is proposed in [66] combining memory, computing and sensing layers for image processing. In [67] the logic and the memory are split into two layers to form a 3D FPGA. This configuration yields a 55% area reduction compared to 2D FPGA and a 47% improvement on EDP thanks to lower routing congestions.

The optimization of a basic cell (SRAM) in 3D design will be discussed to have insights of what 3D monolithic integration can achieve. SRAM blocks represent more than 60% of the total chip area and could be monolithically integrated to reduce it. Usually the SRAM bitcell is designed with six transistors (6T) to achieve a good stability during read and write operation. Thomas *et al.* [68] propose to partition the transistors between tiers to take benefits from the back-gate of top transistors to modulate

dynamically part of the transistors V_T . Thanks to this feature, the static noise margin (the indication of read operation stability) can be improved by 10% and the area is reduced by 20% with 45nm design rules. In the same spirit, the use of dynamic back-biasing enables a stable 3D 4T SRAM with a low power consumption (6 times reduction for write operation, 28nm node) [69]. It is also possible to split NMOS and PMOS between tiers to reduce the area by 33% in 22nm node (6T SRAM) [70] while maintaining the same read/write stability. Even more, the superposition of two 6T SRAM cells (20nm technology) with connection between the internal nodes enables in-memory computing and increases the write ability of 17%, the read stability ($\times 2.2$) and the access time by 6.6% [71] without changing the silicon footprint. For larger circuits, a 3D RISC-V in 28nm rules shows a 23.61% area reduction compared to a 2D RISC-V at iso-performance and power [72].

In this work, the CMOS over CMOS integration will be studied focusing on how to share resources between tiers and how to take benefit from top-tier back-gate to propose enhanced functionalities. The aim of the following 3D monolithic design study is to evaluate the potential of this technology as an alternative to transistor scaling. That is why we will focus on standard cell or small circuit full custom design such as SRAMs.

3- 3D design MOSFET environment

This part will explain the choices we made to analyse 3D monolithic performance for technology, methodology and benchmark. The first part will present the Coolcube™ Design Kit, the second one the SPICE model used, the third the extraction of parasitic elements and the last one the design circuit chosen for benchmark and the associated figures of merit.

a. 3D tier and intermediate BEOL for CMOS over CMOS integration: Coolcube™

The CMOS over CMOS integration, also known as 3D gate-level integration, uses both PMOS and NMOS transistors for each tier. Compared to NMOS over PMOS integration, less 3DCO are needed since the CMOS structure can be done into a given tier. The main advantage of CMOS over CMOS integration is that the planar standard cell can be directly imported into the 3D environment with small modifications. Layers and connectivity associated to 3DCO and intermediate BEOL must be added. A choice could be to define the 3DCO as a standard cell in order to place and route it automatically before the filler cell placement. The number of 3D monolithic vias in a standard cell depends on the 3DCO pitch. Ayres *et al.* shows that a 49.6% area gain can be obtain for large circuits (1200 transistors) [73] with such an integration. For smaller circuit (33 stage RO), the area gain is of the order of 30%, limited by the area overhead due to 3DCO.

As stated in [74], merging environment for different technologies is a major challenge but required for 3D monolithic optimisation. In fact, due to the high contact density, separate design environments for each tier can no longer be representative of reality. That is why a unified design environment for 3D sequential technology by merging Process Design Kits (PDKs) of different technologies related to different tiers is used. In this PhD manuscript, two distinct 3D environments are used for two different reasons. The first one, called here CoolCube™ consists in 14nm CMOS over 14nm CMOS and is used in a prospective way for Design-Technology Co-Optimisation (Fig. 25). The second one consists in 65nm-like CMOS over 28nm CMOS heterogeneous integration for mixed digital-analog applications (Fig. 26). This 3D environment is made to design, fabricate and test chips and simple demonstrators. The associated process flow is presented in Fig. 27.

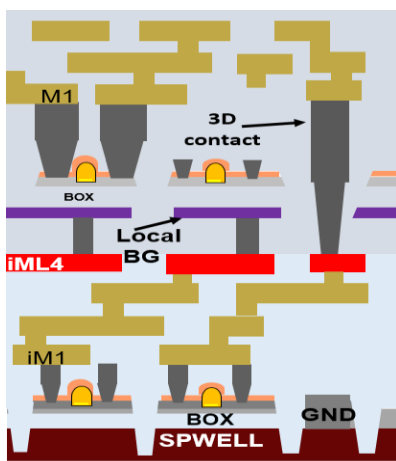


Fig. 25: Schematic stack of CoolCube™ 14nm Design Kit with intermediate vias between the back plane and the upper intermediate metal line.

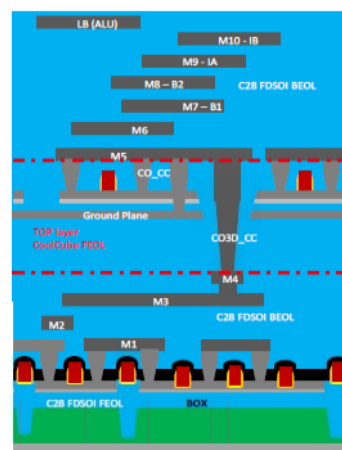


Fig. 26: Schematic stack of the proposed heterogeneous integration. The bottom tier is done in 28nm technology and the top-tier is done with an adapted 65nm low-temperature process flow. Top tier devices minimum dimension is $L=67\text{nm}$ and $W=89\text{nm}$.

The main work is done on the CoolCube™ design kit (see Fig. 25). Transistors of bottom tier and top tier are adapted from 14nm FDSOI CMOS technology. Intermediate backend of line metal lines, noted iMLX are required to take all the benefits of 3D sequential integration. Four intermediate metal lines are chosen for this design kit. In Fig. 25 iML4 is highlighted in red and iML4 parameters and sizing are equivalent to M4. 3D contacts are feasible to connect bottom to top tier. Furthermore, this technology allows a local back-plane underneath each transistor with an associate via. To finish, metals lines are integrated. Similarly, heterogeneous integration DK allows the integration of ground planes and connections between the two tiers with 3D contacts (or MIV).

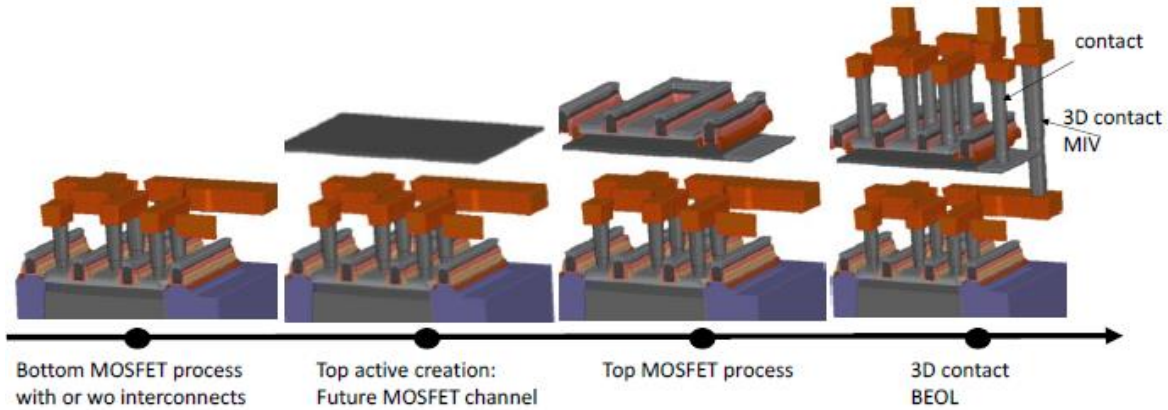


Fig. 27: process flow to create 65nm-like top devices on top of 28nm devices. 4 intermediate metals lines are available to route bottom tier and a 3D contact (MIV) between tiers is available.

b. SPICE model

The SPICE model used in the simulation is LETI-UTSOI2 model [75], [76] declined for 28nm and 14nm node. All model cards included in our PDK are based on the performance of the 14nm FDSOI CMOS and fits with the performance reported in [77], [78]. The assumption that top tier transistors performance are equivalent to bottom tier is made. The state of the art 3D sequential process is in agreement with this hypothesis. Batude *et al.* experimentally demonstrate that the low temperature process performance matches the planar one [79].

As far as the SPICE simulations are concerned, it is possible to define the functionality of a small unit such as the inverter. Then this small element can be duplicated to form bigger circuit such as ring oscillators or an array to study for instance environment effects or leakage issues.

c. Parasitic element extraction

The Parasitic Element eXtraction (PEX) consists in the computation of parasitic effects from device interconnections such as resistance and capacitances. Parasitic elements must be considered since they affect timing performance (RC delay), signal integrity and also power consumption. They are related to technology and design. The Coolcube™ technology stack is described in Fig. 28 from iM4 to BEOL and enables the integration of local back-gate and 3DCO between the two tiers. Some modifications will be done to this stack to evaluate other aspects of 3D monolithic technology. For instance, the advantages brought by the via integration between iM4 and Top-tier back-gate will be explained in subsection 5-c.

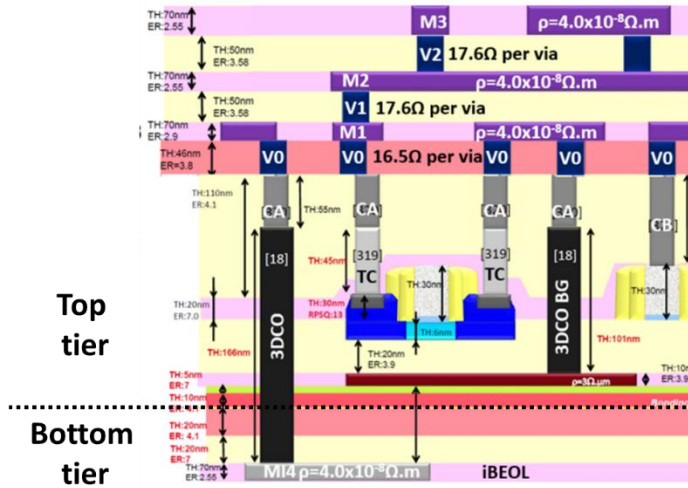


Fig. 28: DRM showing upper level tier, with M14 iBEOL level, back-gate, top tier FEOL and BEOL, taken from [80].

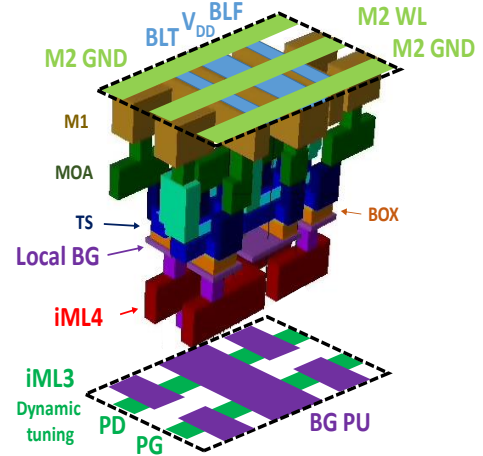


Fig. 29: Representation of the parasitic elements produced by CLEVER SILVACO tool. The capacitances and resistances between pre-defined nets are computed.

That is why, for more flexibility, the CLEVER tool from SILVACO is mainly used in this work to compute the parasitic elements which are layout dependent. The approach is layout driven and each layer is emulated from a layout file to construct a 3D representation of the structure. An example of top-tier SRAM CLEVER output is given in Fig. 29. The tool can handle lithographic effect, linewidth variations, corner rounding and non-uniform etch rates but in this prospective study, manhattan structures (ideal rectangular block shapes) have been chosen. The Coolcube™ process specificities (material, sizing and resistivity) are taken into account. From this 3D representation, resistances and capacitances between user-defined electrodes are computed and compiled into a netlist. This netlist can be directly injected in SPICE simulations. The SPICE simulation itself (without the parasitic element netlist) considers also some local parasitic elements related to the transistor (without BEOL elements) such as the gate to source capacitance. That is why it is important to include both descriptions to model the full parasitic network without overlaps. Design configuration (such as plain back-plane or individual back-gates) can now be directly compared in terms of capacitances, resistances but also on typical figures of merit such as drive current.

d. Methodology summary

Once the 3D environment is set and the reference layout is designed, some layout variants are done (see Fig. 30). In order to compare it with respect to the reference, the parasitic element (which highly depend on layout) are computed and considered in the SPICE simulation. Then, the circuit waveforms are generated and timing analysis can be performed and the circuit functionality verified. However, to be able to compared two different circuits or technologies, some metric must be defined which are representative of the performance, power or area. As far as area is concerned, the area overhead in % with respect to the reference layout is representative of a gain or a penalty. As far as power is concerned, the leakage current or total power consumption can be analysed to compare two designs, depending on the application. For instance, if a circuit is meant to be idle, the static power is more critical than the dynamic one. For the performance, the maximum operating frequency of the system can be a good indicator. However, depending of the application (for instance low-power device), one can attribute more importance on one criteria (power) and defined a custom figure of merit through a formula. In this case, where performance are not required but where power is an issue, a possible FOM is $FOM = P_0 / (P + A_0 / Area)$. Furthermore, a design can be high-performance and an other-one power-efficient for the same silicon footprint and thus, it is important to define which is the best for the tackled application.

Next paragraph will present the FOM of the chosen benchmark circuit.

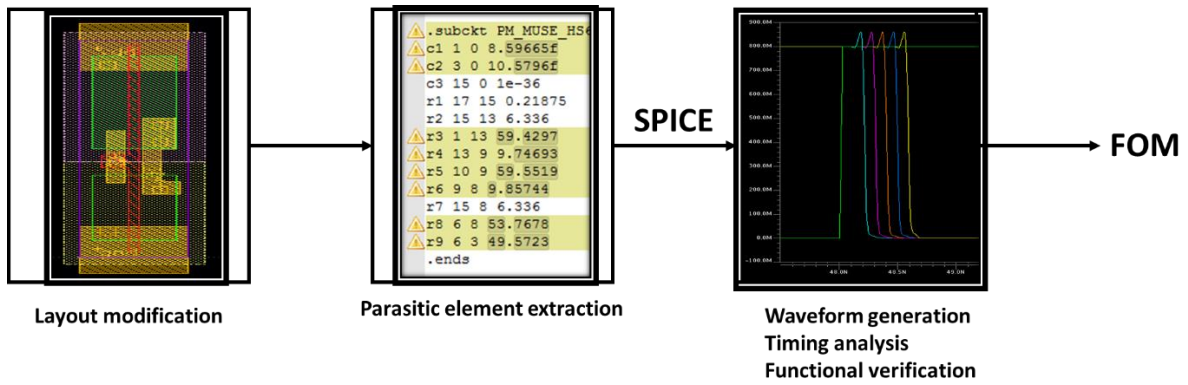


Fig. 30: Methodology used in this PhD work. To compare two layout variants (or technology), the PEX is done and injected into SPICE simulations. From this, the FOM are extracted and a comparison is done between the two designs.

e. RO, SRAM benchmark: typical figure of merits

Full custom circuits were done in the 3D sequential PDK, complying with DRM. The chosen benchmarked circuits are Ring Oscillators and SRAM. Those circuits are simple enough to be full-custom designed without synthesis tools, nevertheless they are a representative benchmark for digital designs. They are both composed of inverter gates which is a basic building block. The idea is to define some criteria systematically used to evaluate the pertinence of a particular design or a new technological approach. Such criteria are called figure of merit (FOM) and are discussed in the next paragraph for ring oscillator and SRAM.

i. Ring Oscillator

As explained previously, the ring oscillator consists in an odd number of inverters connected in series. It will generate an oscillating signal with a specific frequency, depending of the number of inverter and the inverter delay. Each inverter is called a stage. Also, each inverter can drive more than one inverter, this number is called the fan-out (FO). For instance, a fan-out three inverter has three inverters connected to its output. The RO can be done in FO3 or FO4 to be closer to a real circuit implementation. The frequency is higher for a lower FO since there is less parasitic elements. Also, the higher the number N of stages, the lower the frequency is.

This kind of circuit is very useful to evaluate technology processes, because it is simple to design and to check the logic functionality. It is also directly linked to performance (switching frequency).

To determine the best design approach, the output frequency versus the power consumption can be considered. Thus the performance and the power consumption can be easily compared.

ii. SRAM

Static Random Access Memory (SRAM) is a volatile type of memory in the sense than information is lost when unpowered. However, unlike Dynamic RAM (DRAM), no periodical refreshment of the memory element is needed for proper operation. Usually, six transistors are used to create and access a memory point (6T-SRAM). 4T –SRAM are also proposed for density reasons and 8T or 10T SRAM for stability one. In this PhD manuscript, only 6T-SRAM will be presented and analysed. In this introductory part, 6T-SRAM operation will be presented before explaining the SRAM typical figures of merit.

a- SRAM operation

The schematic of the 6T-SRAM bitcell is given in Fig. 31. The bitcell is composed of two cross-coupled inverters to store the information. Two additional transistors (named pass gates, PG) are needed to access the memory point. The stored state can be either a '0' (GND on internal node BLLI) or a '1' (VDD on internal node BLRI). It can be read, written and maintained to its current state using two bitlines (left or right BL) and a wordline (WL).

In the data retention mode, the Pass Gate (PG) transistors are biased in the OFF state ($WL=0$). Bitlines BLL and BLR can be either be precharged to VDD or GND or left floating. The two cross-coupled inverters are able to maintain the state if the supply voltage is sufficiently high.

To perform the read operation, the WL is biased at VDD ($WL=VDD$) and both BLs are precharged to VDD. In this configuration, on the side where a '0' is stored, a read current flows through the PG and the PD, discharging the bitline from VDD towards GND. On the other side, where a '1' is stored, no current is seen since the potential of the source and the drain are both equal to '1'. The read operation consist in sensing the difference between the two sides. However, to ensure that the internal node is maintained to '0' during read operation, the PD must be stronger (lower resistance) than the PG (noted here $PD > PG$). If not, for extreme cases, the read operation can change the '0' stored into a '1', writing the cell instead of reading it. This strength ratio is usually achieved by designing the appropriate width ratio between the PD and PG transistors.

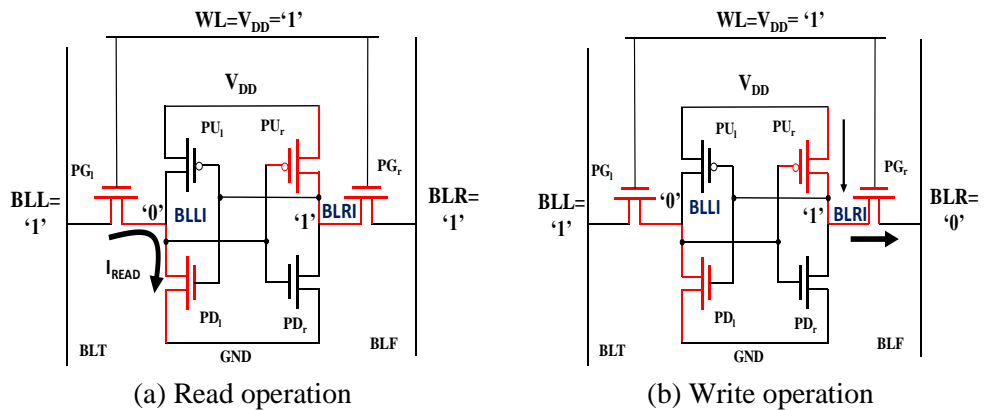


Fig. 31: 6T-SRAM schematics. WL and BLs bias are indicated in read (a) and write (b) operation. Taken from [78].

During the Write operation, PG are biased in ON state ($WL=VDD$) and both bitlines are biased according to the value of the bit to be written. As depicted in Fig. 31-b the write mechanism consists in pulling the internal node storing '1' to GND through the bitline. In order for this to be possible the PG has to be stronger than PU ($PG > PU$). Combining this criterion with the one for read, the general strength relations between the transistors in the SRAM cell can be defined as $PD > PG > PU$.

b- SRAM Figures of Merit (FOM)

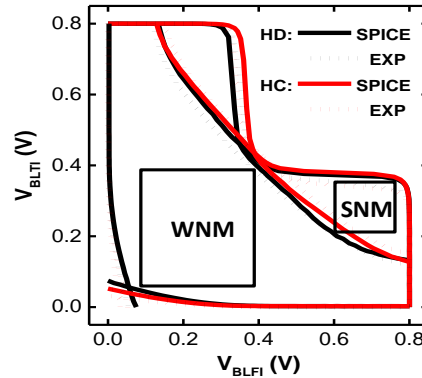


Fig. 32: Experimental vs. simulated butterfly curve at $V_{DD}=0.8V$. The spice simulation is done using the 14nm model card and is taken from [78]. SNM and WNM can be extracted from the curves and are defined as the smallest square, which can be inserted into the curves.

To characterize the SRAM bitcell, different metrics are defined, such as the Static Noise Margin (SNM) [81] and the Write Noise Margin (WNM) [82]. Experimental and simulated metrics extraction for two 14nm 6T-SRAM cells (High Density HD, $0.078\mu m^2$ and High Current HC $0.098\mu m^2$) is illustrated in Fig. 32.

SNM: During Read operation, the voltage of the internal node depends on the PD>PG ratio (voltage divider). If the internal node voltage is higher than the trip point of the other inverter, the data stored in the cell will flip, leading to a read failure or a destructive read. This condition corresponds to SNM violation. To measure the SNM experimentally, an internal node voltage sweep is carried out while monitoring the voltage on the other internal node. Doing this process for both SRAM cell inverters, one can plot the so-called butterfly curve. The side of the largest square embedded between the two characteristics gives the partial SNM (one for each lobe). The SNM is the minimum of the two partial SNMs. The lower the supply voltage, the lower the SNM.

WNM: in a similar way, two voltage transfer characteristics are measured in write conditions and the partial WNM is defined as the side of the smallest square embedded between the curves. The lower the WNM, the more likely the write operation will fail.

Read and write currents give a fair approximation of the read/write operation speed. In addition, the leakage current ($WL=0$ and $BL=0$) represents the static power consumption and has to be kept in mind. From the experimental curves (Fig. 32), a trade-off can be derived: an increase of the SNM can decrease the WNM. Furthermore, as far as the design is concerned, the width of each transistor is limited to retain a small silicon footprint.

Leakage current represents the stand-by power of the bitcell and thus the power consumption.

To conclude this part, the benchmark designs chosen to evaluate 3D monolithic technology are RO and 6T-SRAM. For the RO, power versus frequency or delay is considered to compare different designs or technology approaches. As far as the SRAMs are concerned, a major criteria is the cell stability when reading (SNM) or writing (WNM). Read and write current are also considered to give insights about read and write operation speed.

The 3D design MOSFET environment (DK, SPICE, PEX and benchmark structures) used in this work has been defined. Next, we will address the routing in 3D designs.

4- Routing in 3D designs

One of the main advantage of 3D monolithic integration concerns the reduction of overall wire length, thanks to tier partitioning, in order to increase the circuit performance by reducing wire delays. In this part, tier-partitioning strategies won't be addressed but we will rather investigate the benefits of 3D monolithic integration (fine grain connection between tiers capability) at the cell level. First, the additional routing resources can be used to contact the top-tier by behind to avoid a longer top connection. Secondly, we can think of sharing some resources between the two tiers such as clock signal or power rail. To finish with, back planes can be dynamically accessed and a technological sizing study is carried out to define back-plane design guidelines.

a. Buried power rail

Buried power rail is envisioned for planar devices to scale down the circuit and limit the IR drop of low voltage technologies. A ruthenium lines have been proposed in [83], Fig. 33, detaining lower resistance than W and super via trench [84] to efficiently deliver power and reduce the wire resistivity. Prasad *et al.* [85] analyse back-side and front side power delivery network to reduce the IR drop. These options are benchmarked using the Arm Cortex-A53 CPU at IMEC 3nm technology node and the front side power delivery reduces the worst IR drop by 1.7 while the back side by 7. Salahuddin *et al.* [86] show that the use of buried power rail can improve the write margin (340mV) and read speed (30%) of a 3nm SRAM, because the power rail can be enlarged without impacting the SRAM footprint. The second one consists in burring interconnections to reduce the area of routing limited standard cells. Zhu *et al.* [87] demonstrate a 9-13% chip area reduction thanks to a buried interconnect layer for a 7nm node with FPU and MIPS from Open Cores as well as a Cortex M0 testcases. We propose to investigate in the next paragraph if a buried power rail in between the two tiers is feasible in 3D monolithic design. The benefits would be to share the rail between the tiers and reduce the sizing constraint to provide larger power rail for both tiers.

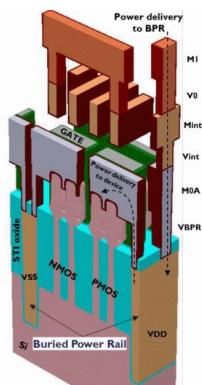


Fig. 33: A buried power rail (in shallow trench isolation STI and Si substrate) runs parallel to the fins. The power grid of VDD and VSS lines is designed at Mint level. Shifting the grid to the FEOL, reduces standard cell height. Taken from [83].

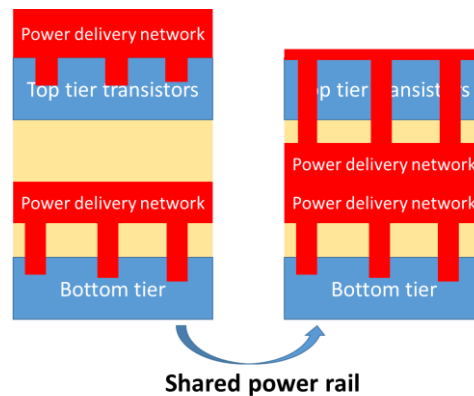


Fig. 34: Schematics presenting how a share power rail can be used. In fact the bottom network on Mi1, Mi2, Mi3 and Mi4 can be duplicated and directly enforced on upper intermediate lines.

b. Congestion mitigation and resources sharing between tiers

Auth *et al.* [88] evaluate the density benefit between technologies with a metric combining NAND (60%) and scan flip-flop (40%) density. That is why, to evaluate the benefits of 3D monolithic integration, a 14nm NAND2 is designed with the 14nm 3D PDK and two scenarios of connections are envisioned in Fig. 35. The layout succeeds the LVS check. The first one consist in contacting the VDD and GND power rail by intermediate metal lines to share it with bottom devices and enlarge the power rail width. The second one investigates further the freedom of 3D monolithic integration by deporting

the output of the cell to bottom tier to save area on top-tier. Without modifications, the 14nm NAND2 height is $H=813\text{nm}$ (presented in Fig. 35-b) and performs the operation $Z=\text{NAND}(A,B)$. Only the M1 layer and PC are presented to highlight the interconnection scheme. In our case, two PMOS are in parallel and connected to two NMOS in series. If the power is delivered by intermediate lines (Fig. 35-c), the cell size does not evolved, since the 3DCO can be directly connected to GND/VDD. In our case, we chose to let the GND/VDD lines at M1 (and not fully buried the line with iML4 because in this case, no area gain was seen due to the distance constraint between 3DCO and PC). However it is possible to contact GND/VDD to iML4 with different scenarios for instance to facilitate different voltage domains. Also, the width of the power rail can be enlarged without impacting the top and bottom cell area. Here the study was about feasibility, impact on area and automatic adaptation of a 2D design. Thus here, a shared power rail between tiers seems to be feasible without impacting the top-level standard cell and is straightforward to implement. In fact, 3DCO from the bottom-tier power grid to GND/VDD top-tier can deliver top-tier power by duplicating reversely the integration scheme as explained in Fig. 34.

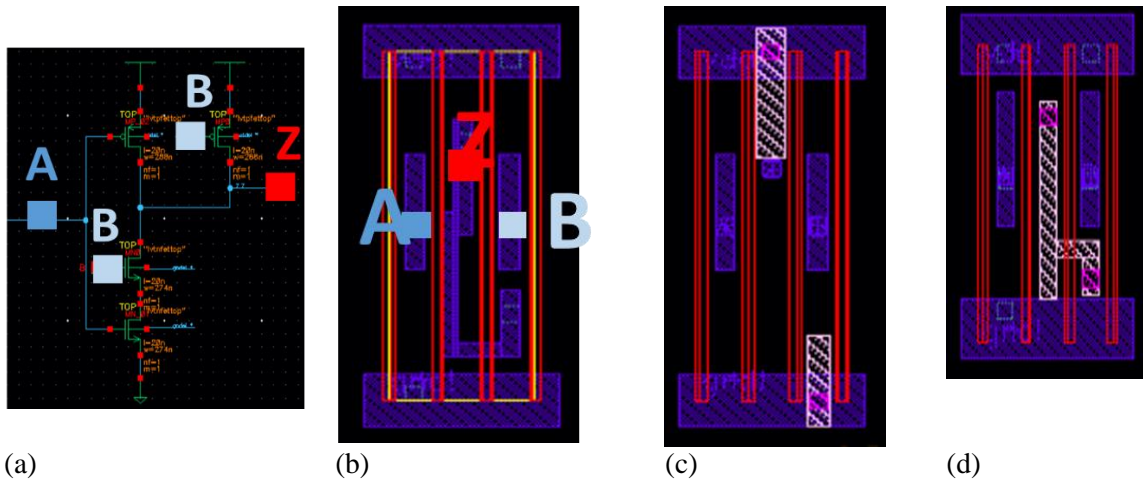


Fig. 35: (a) Electrical schematic of a NAND2, $Z=\text{NAND}(A,B)$. (b) Associated top-tier layout where the purple lines are M1 and the red one PC, to highlight the connections between transistors. The height without optimization is $H_{REF}=813\text{nm}$ and with optimization $H_{op}=0.94 \times H_{REF}=767\text{nm}$. (c) VDD and GND are connected with intermediate metal lines instead of M2, the height is the same as (b) without optimisation. (d) In this case, the output Z is made with intermediate metal lines and vias and connect directly the source and drain of upper transistors. The final height is $615\text{nm} = 0.75 \cdot H_{REF}$.

To optimize further, Fig. 35-d presents a layout where the output Z have been deported to iML and the contact to the active area is done by beneath thank to a via, called iV5 between iML4 and top tier. Without the output routing at top tier, the height of the cell is reduced (x0.75). Also, since the output is at iML4, if the next stage is in the bottom tier, the interconnections between this NAND output and the next input can be reduced. However, such a via cannot cross the back gate and must be integrated without damaging the top-tier process and remained an integration challenge. Some technological drawbacks or advantages are advanced in the next lines, but it consists in considerations rather than real studies. The main issue is the metallic contamination. That is why, this iML4 to top active area contact cannot be done before top-device processing, since no naked metal is wanted in front-end tools. Furthermore, with nowadays technologies, the wafer bonding process consists in an oxide-oxide bond and not an oxide/metal lines to silicon. For these reasons, the contact must be done during back-end of line device processing. If so, the contact will punch though the active until to reach the iML4. This configuration, as far as the transistor electrical characteristic is concerned will be similar to borderless contacts proposed in [89]. However, a borderless contact increases significantly the leakage current. Nevertheless, at the same time, this contact could reduce the parasitic capacitance and must be investigated to verify if the capacitances and routing reduction gain compensate the leakage current as well as the additional integration complexity.

The design modifications steps are the following:

- The desired M1/M2 routing path is changed to iML4/iML3 and the via V0 is replaced by iV5.
- A hole is done into the back gate to let the via
- The top cell is routed again to decrease its area and make sure than the iV5 is not in contact with V0.

The two top steps can be done automatically but the last one should be done by hands and inherently is more costly.

To verify the interest of the methodology, the same exercise is done with a OAI22 (Fig. 36), where the implemented function is $Z = \text{NOT}(A \& B) \text{ OR } (C \& D)$. Similar results are seen: a small area gain when optimizing the bitcell and using a shared power rail and a final height structure of $0.79H_{\text{REF}}$ when intermediate interconnection are used.

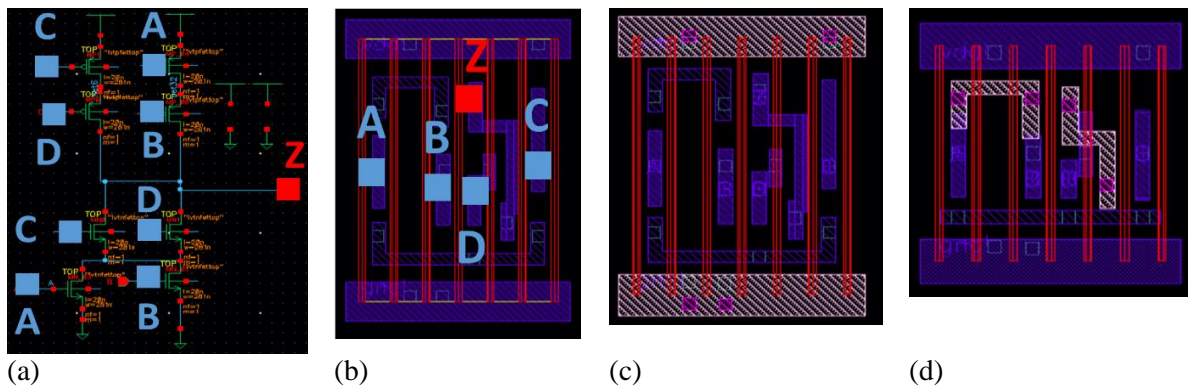


Fig. 36: (a) electrical schematic of an OAI22, $Z = \text{NOT}(A \& B) \text{ OR } (C \& D)$. (b) associated top-tier layout where the purple lines are M1 and the red one PC, to highlight the connections between transistors. The height without optimization is $H_{\text{REF}} = 882\text{nm}$. (c) V_{DD} and G_{ND} are connected with intermediate metal lines instead of M2 and the height cell is optimized $H_{\text{pwr}} = 836\text{nm}$. (d) In this case, the output Z is made with intermediate metal lines and via and connect directly the source and drain of upper transistors. The final height is $660\text{nm} = 0.79H_{\text{REF}}$.

However, such a via is complicated to integrate so we made the decision to perform the same study only with 3DCO and 3D buried power rail. A more complex design is chosen, a D flip-flop with a multiplexer to allow scan chains for testability. It detains a clock input as well as enable input. A balanced clock tree should be designed to deliver a synchronous signal to every sequential cell with small skew and under skew constraint. The clock can be gated to deliver the signal only to operating cell to lower the power consumption. Several algorithms propose a thermal and slew aware clock routing for TSV circuits [90], [91], [92]. In [93], the 3D stacked clock distribution/generation network achieves a 2.29 times energy-efficiency improvement compared to the H-tree structures in 45nm CMOS technology.

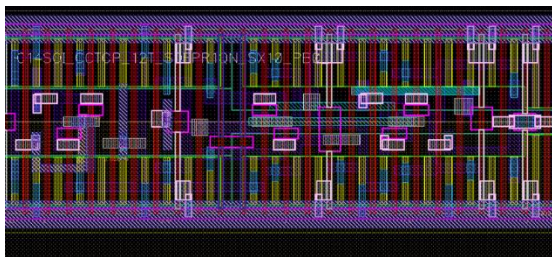


Fig. 37: D-flip-flop with scan chain for testability layout.

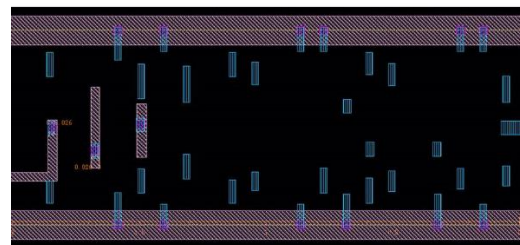


Fig. 38: Modified layout with 3D buried power rails, clock and testability enable routed with iML4 and contacted to the top tier with 3DCO. For visibility reasons, only the modifications are represented. The area is increases by 4.4%.

That is why, there is an interest in having shorter clock wire between two tiers to balance the skew. Thus in the SDFPR1QN design, the clock signal and the Design For Testability (DFT) signal are routed using 3DCO. In the non-modified design, these nets are not above active area or back-gate. That is why only slight modifications to the existing designs are done and most of them can be automatized:

- Via0 => 3DCO except for the 3D buried power rail where a 17nm shift is done to respect the 3DCO to backgate distance constraint. The overall area is increased by $2 \cdot 17\text{nm} \cdot L$.
- M1 => Mi4
- M1pin => Mi4pin
- MOA_UP to be enlarged to contact 3DCO

The final area is increased by 4.4% and the clock tree can be shared between tiers. It is also possible to let the power rail at the top level to avoid IR drop and a bonding on thick metallic lines.

To conclude, sharing resources between tiers such as power rail or clock tree can be done with a slight area cost but allows the reduction of wire lengths. A prospective work would be to place and route cells with and without this punch-through via to analyse the impact on latency and performances. The next part tackle another topic: back-plane contact for top-tier.

c. Design guidelines for top-tier Back-plane contact

I contributed to a prospective study to elaborate design guidelines for the back-gate contact. In fact 3D monolithic integration enables the creation of individual back-gates which can be dynamically controlled. The main question is what should be the distance between back-gate contact and the device back-gate to ensure a correct signal propagation when dynamically biased. To answer this question, a RO is simulated and the back-gate are statically and dynamically biased. In this part, unlike the previous and the following one, the design kit used is the one of heterogeneous integration (65nm over 28nm).

i. Simulated structure

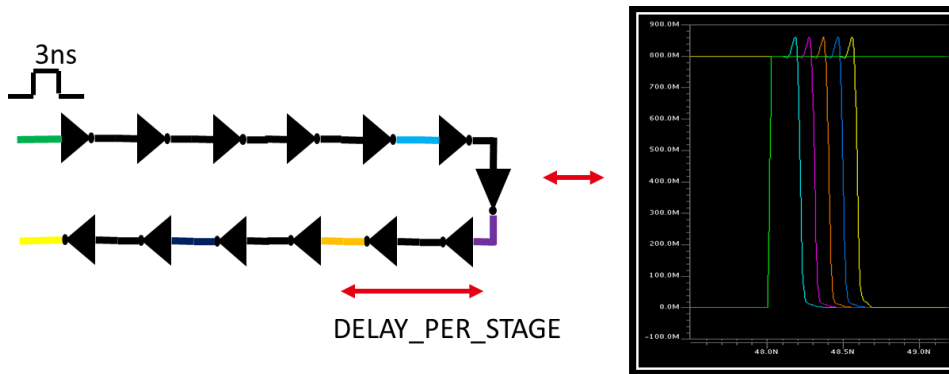


Fig. 39: Schematics of the simulated circuit. The propagation of a 3ns square signal through 13 inverters in series is presented in the waveform.

The simulated circuit consists in 13 inverters in series (13 stage ring oscillator). A 3ns squared impulsion is given in input and the delay per stage or the frequency per stage is considered. It is defined as the time between the seventh inverter equals to $V_{DD}/2$ and the ninth equals to $V_{DD}/2$. In Fig. 39 the propagation through the inverter is seen. The supply voltage will vary between 0.6 to 1V, a fanout of 1 and 5 will be considered, a $\alpha_{\text{dynamique}}=0.1$ and $T=25^{\circ}\text{C}$. The distance between the back-gate contact and the inverter, noted X will vary from 0 to $10\mu\text{m}$ (Fig. 40).

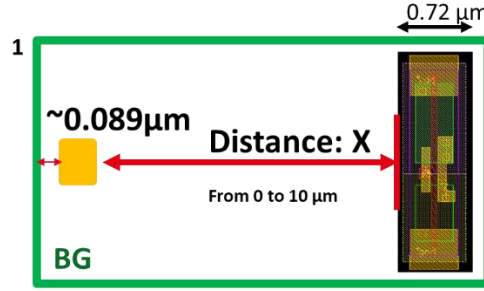


Fig. 40: Presentation of the simulated situations: the back-gate contact is at the minimum distance to the BG and the distance between plug and inverter varies from 0 to 10 μm .

ii. Static consideration

In a first time, we will consider that a static signal is applied on the back-plane contact. The first step is to verify if the back-gate bias has an impact on the considered FOM. Fig. 41 presents the delay through the 13 inverters (ns) as a function of back plane polarization (from -1V to 1V). This short modulation range translates into a 4ns delay difference on the inverter. Monitoring the delay or conversely the maximum operating frequency f_{max} can indicate the voltage applied on the back-plane. A 1V back plane voltage is statically applied before the propagation of the 3ns squared signal through the inverter and the maximum frequency is extracted for various plug distances (Fig. 42). When static, the distance between the contact and the standard cell do not matter up to 5 μm but the type of extraction (RCC or C+CC) does. RCC type of extraction means that both intrinsic and coupling capacitance as well as distributed elements are considered. C+CC considers only capacitances. In fact, when the resistances (RCC mode) are considered, in addition to capacitances, the delay is larger and f_{max} lower. Later on, RCC extraction is done. Please note that in 3D monolithic integration, if the back plane is of the same polarity no latch-up (failure due to excessive current typically between p and n junctions for wells [94]) can occur since the back plane is isolated by the oxide. To mitigate this, usually a distance constraint on well contacts is applied, which won't be the case for 3D monolithic integration.

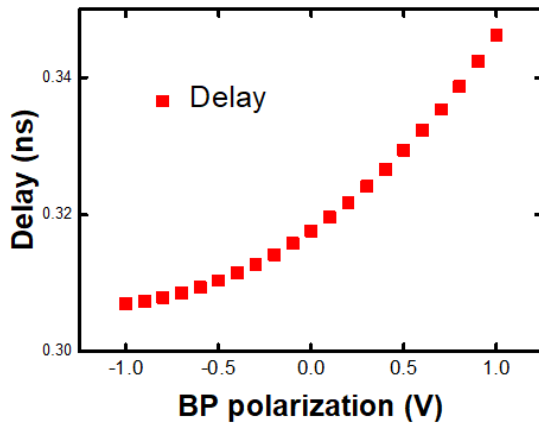


Fig. 41: delay through the 13 inverters (ns) as a function of back plane polarization (from -1V to 1V).

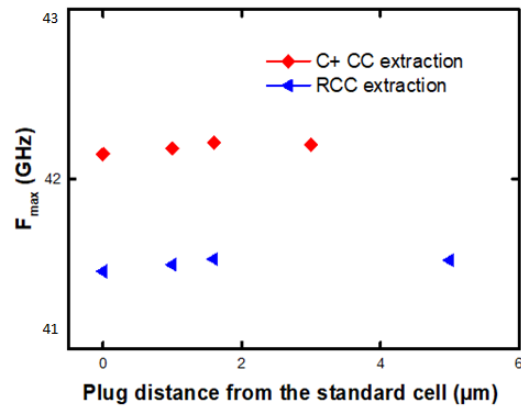


Fig. 42: Maximum operating frequency in GHz as a function of the plug distance. In one case the capacitances are considered only and in the other case, resistance + capacitances lower the maximum operating frequency.

iii. Dynamic consideration

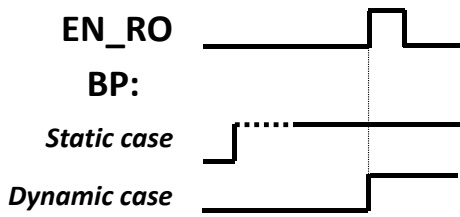


Fig. 43: Polarization scheme between static and dynamic case.

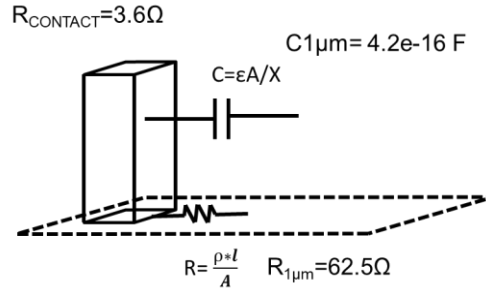


Fig. 44: Representation of the capacitance and the resistance network of the back-gate contact. The values are given for $1\mu m$.

In this dynamic back plane access case, the back plane is biased at the same time that the RO enable signal to see the propagation of the signal as schemed in Fig. 43. A 1V polarization is chosen, increasing the RO delay if the node is correctly biased. Fig. 44 presents the parasitic elements considered in the PeX file. Similarly to the previous static case, up to $10\mu m$, no difference were seen for the RO delay, indicating that the 1V signal had the time to propagate on the back-plane. However, for an extreme case (the contact at $1000\mu m$ from the RO cell) the observed delay was the one of $V_B = 0V$. In fact, as presented in Fig. 45 the delay per stage is modulated up to $200\mu m$ indicating that the ground plane is polarized positively up to this distance in dynamic. It sets an upper bound for the ground plane contacts: $200\mu m$.

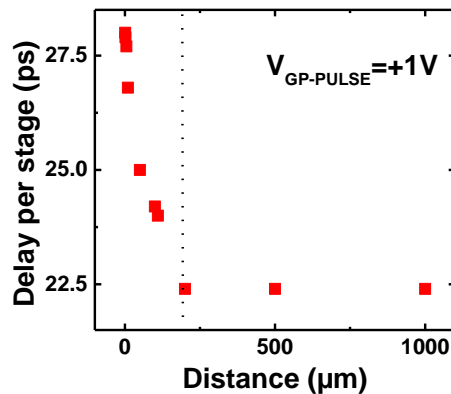


Fig. 45: Delay per stage as a function of the distance. A +1V pulse is applied on the back-plane, increasing the delay of the RO. The modulation is effective up to $200\mu m$.

In this part we first proposed a “hand” design methodology to share resources between tiers by exploring a shared power rail, shared clock signal and test one. Then we defined sizing guidelines to take fully advantage of a dynamic back plane modulation. However, to have a fine grain back-plane or back gate connections at the transistors level implies several back-gate contact leading to a significant area overhead. That is why, in the next part, we will go further and analyse the advantages of a contact directly between the back-gate and intermediate metal lines.

5- Design-technology co-optimization: top-tier SRAM

The intermediate back-end-of-line can be used to route bottom tier and top tier from below. The idea here is to use this additional connectivity to contact the top-tier back-gate from below thanks to an intermediate via iV4. Previously, the back-gate contact BGCO (metal resistivity $2.8 \times 10^{-7} \Omega \cdot m$ and width $W=32nm$) was between the back-gate and M1 (top-tier), leading to additional area to respect the minimum distance to the active zone (see Fig. 46). In fact in Fig. 46, only the design rules of the BGCO with respect to BG or active zone are given but there are additional rules concerning the distance of BGCO with metal0, trench contact and poly. However, if the connection is from below, like in Fig. 47, no additional space is needed for the BGCO and for the back-gate additional extension. Thus, it will enable the use of local back-gate without any area overhead, since the connections are below in the same silicon footprint. That is why an iV4 is defined following the same design rules as iV3 (Width 32nm and distance to BG min of 16nm) and having the metal resistivity of $2.8 \times 10^{-7} \Omega \cdot m$. An SRAM bitcell is taken to evaluate the interest of a fine grain back-gate network which can be dynamically accessible.

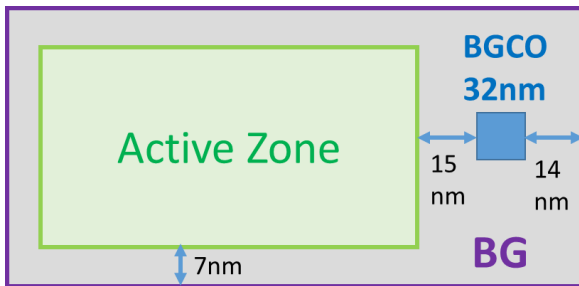


Fig. 46: Schematics of the design rule for back-gate contact (BGCO) in the Coolcube™ integration.

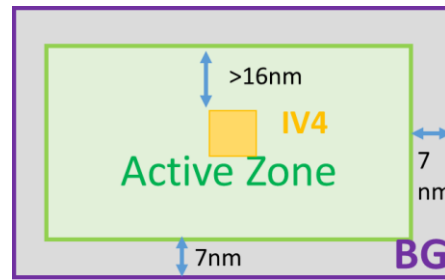


Fig. 47: Schematics of the design rule for the proposed back-gate - iV4 connection. An area gain can be intuited.

The methodology of this work consists in demonstrating the interest of a back gate for 14nm FDSOI SRAM before analysing a 3D architecture. For this, we will first describe the electrical characterisation results of 14nm FDSOI planar SRAMs, in terms of typical FOM, back-bias sensitivity and reliability. Then, we will propose a back-bias assist for 3D monolithic SRAM, based on layout studies and simulations.

a. 14nm technology performance

i. Electrical characterization of typical FOM

14nm planar CMOS devices were fabricated at STMicroelectronics featuring 6nm-thick silicon channels, 20nm gate length (L), SiGeB/SiP in-situ doped sources/drains, 90nm Contacted Poly Pitch and 64nm Metal Pitch [95]. SRAM cells were fabricated down to $0.078 \mu m^2$ and are declined into two flavors: high-density (HD) and high-current (HC). The bitcell device dimensions are summarized in Fig. 50. The strength criteria $PU < PG < PD$ is done by modulating the gate width, the gate length being constant, equals to 30nm for all the devices. For instance for the high-density cell, $W_{PU}=45nm < W_{PG}=66nm < W_{PD}=68nm$. A SEM top view of the HD bitcell observed at the gate level is presented in Fig. 48. All the transistors of both High-Density ($0.078 \mu m^2$) and High-Current ($0.098 \mu m^2$) cells, *i.e.* the Pull-Up (PU) pMOS as well as the Pass-Gate (PG) and Pull-Down (PD) nMOS are built on silicon channel and with a single p-type metal gate and single p-doped well (SPWELL) (Fig. 49), which can be biased at V_{well} .

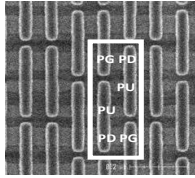


Fig. 48: 14nm High Density SEM observed at the gate level.

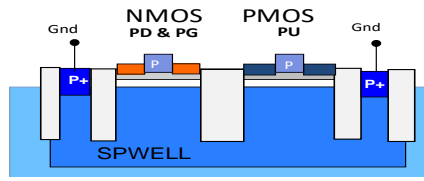


Fig. 49: Schematic device cross section.

Cell	Surface (μm^2)	L_{ALL} (nm)	W_{PU} Pull-Up (nm)	W_{PD} Pull-Down (nm)	W_{PG} Pass Gate (nm)
High Density (HD)	0.078	30	45	68	66
High Current (HC)	0.098	30	52	114	112

Fig. 50: Key dimensions of 14nm FDSOI 6T-SRAM.

The measurements are done on 30 cells and without further indication, the median value is given. To compare HC and HD cells and see what is the limiting operation, the FOM explained in part 3-e are considered. Excellent experimental static performance is obtained in nominal conditions ($V_{\text{DD}}=0.8\text{V}$, $V_{\text{well}}=0$) for both cells. The HD cell features: $\text{SNM}=139\text{mV}$; $\text{WNM}=320\text{mV}$, $I_{\text{read}}=10\mu\text{A}$, $I_{\text{write}}=15\mu\text{A}$ and static leakage in the retention mode $I_{\text{leak}}=6.3\text{pA}$. The margins are even higher for the HC cell: $\text{SNM}=148\text{mV}$ and $\text{WNM}=351\text{mV}$, the stability is improved thanks to the SRAM transistor sizing.

However, starting from these nominal references, a lower V_{DD} reduces the SNM and WNM margins, as presented in Fig. 52. As far as the temperature is concerned, when the temperature increases, the threshold voltage decreases and therefore the SNM decreases [96]. That is why the worst case in our measurements is at $V_{\text{DD}}=0.55\text{V}$ and $T=125^\circ\text{C}$ (see Fig. 51). For this case, the median WNM value is around 230mV for both HC and HD cell and the SNM ranks between 60mV (HD) and 90mV. The SNM and WNM dependence on V_{DD} and SNM degradation with temperature is seen on Fig. 51. Considering global variability, one finds that 14nm FDSOI SRAM stability is read-limited, especially for the HD cell. For instance, the HD cell SNM is 68mV at $V_{\text{DD}}=0.55\text{V}$ and $T=125^\circ\text{C}$, to be compared with 139mV in nominal conditions ($V_{\text{DD}}=0.8\text{V}$ and $T=25^\circ\text{C}$). That is why we will now focus mainly on improving the read operation stability (SNM) for the HD cell.

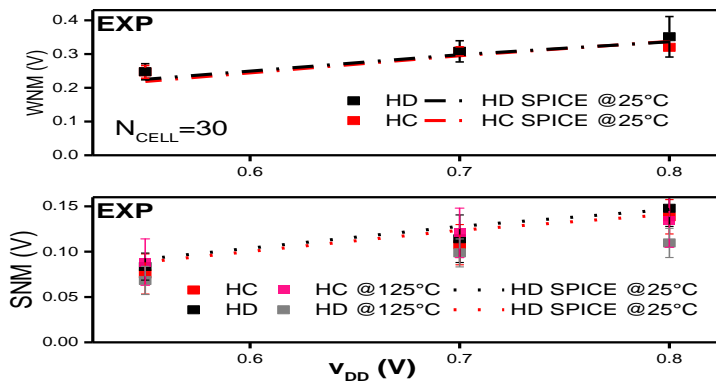


Fig. 51: Read and write stability. 14nm FDSOI SPICE model vs. experiment for different V_{DD} and temperature values. Both high-density and high-current cells are read limited. Taken from [78].

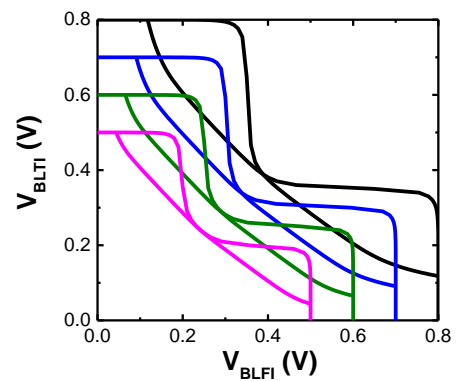


Fig. 52: Butterfly curve at various V_{DD} with SNM representation. The lower the supply voltage, the lower the SNM is.

In fact, to enhance SNM, several standard assist techniques are available [97] to modulate the $\text{PD}>\text{PG}$ strength criteria. The idea consists in dynamically (temporary) modifying the PG (or PD, PU) resistance during the read operation. To increase the read operation stability, either the PG resistance can be increased or the PD resistance can be decreased. It can be done by using a negative ground (larger V_{GS} for the PD and thus lower PD resistance), using a V_{DD} boost or a partial Bit-Line Precharge or Word Line underdrive (WL). These techniques will increase the read margin only during read operation, so the write operation won't be done with such adjustment and WNM do not have to be considered. More specifically, Fig. 53 shows that a Word-line (WL) underdrive (lower $V_{\text{GS-PG}}$ implies larger PG resistance) by 20% improves the SNM by 37% at $V_{\text{DD}}=0.8\text{V}$.

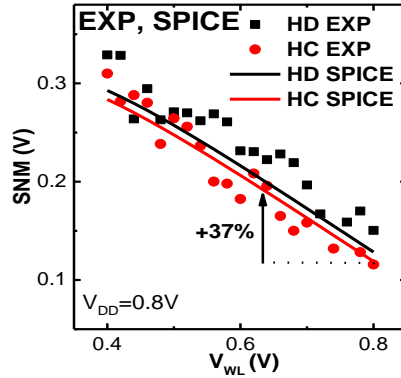


Fig. 53: Experimental vs. SPICE SNM as a function of WL voltage (worldline underdrive read assist). A 20% reduction on WL supply voltage leads to a 37% SNM gain taken from [78].

The 14nm FDSOI SPICE model and a design kit introduced in part 3-b were adjusted using these devices. It should be highlighted in Fig. 51 and Fig. 53 that this model reproduces well the behavior of planar SRAM SNM and WNM even at low V_{DD} and under WL underdrive.

This part presented the typical read and write stability for 14nm FDSOI SRAM high-density and high-current cells. The read stability for the HD cell is the main limitation for voltage scaling. Several read-operation assists are feasible with a dynamic control of voltages. However, a control of transistor threshold voltage through back-biasing is also feasible and could be used to increase SNM.

ii. SRAM: variability issue and impact on FOM

In theory, a defined design (W and L for each transistor) delivers a known amount of current and leads to specific value of SNM and WNM (or other FOM). This theoretical case is called typical case and without layout or SPICE modification, its simulation will give a repeatable output. However, in reality, the cell is not perfectly symmetrical due to process-related or dynamic variability. For more information on process variability, see chapter III, part 4-d. In fact, the strength criteria $PD > PG > PU$ relies on equivalent resistance and is impacted by width/length variations and V_T variations. For instance, when we consider the corner SF (Slow for NMOS and Fast for PMOS) the SNM reduces drastically just because of a higher V_T for NMOS and a lower V_T for PMOS (see Fig. 54). In this SF corner, the obtained SNM and WNM values are far from the SNM and WNM in the typical case. The variation between typical case and extreme cases (corners SF, FS, SS, FF) are inherent from a technology. It means that for a specific process, a parameter called matching parameter A_{vt} can be defined to quantify the threshold voltage variability. This parameter is expressed in $mV \cdot \mu m$ and when dividing by $1/\sqrt{W \cdot L}$ indicates the threshold voltage variability. Two variabilities must be considered. The first one considers the overall variability between one device and another one in another die. It is called global variability. The second one, considers the V_T shift between adjacent devices and is called local variability or mismatch. It is experimentally measured with a pair of transistors as close as the technology allows.

To consider the variability on threshold voltage, statistical simulations called Monte-Carlo are done, each sample considers a new set of V_T for each SRAM transistor. Fig. 55 depicts Monte-Carlo simulation result (WNM vs. SNM, MC=1000) taking into account local, global, local and global variability ($A_{vt}=1.7mV \cdot \mu m$), showing a significant impact both on SNM and WNM. The SF and FS corners are indicated. Global variability can be well represented by typical case TT and SF and FS corners. Moreover the fact that the local variability is by far dominating the cell behavior highlights the importance of running high sigma statistical simulations to ensure a sufficient yield on the bitcell level. To this end, typically a 6σ yield is targeted for all stability metrics. A 6σ process (*i.e.* $SNM - 6\sigma > 0$) ensures that 99.99966% of the cases will be correct (*i.e.* no read failure in the bitcell). Fig. 55 indicates the σ , 3σ and 6σ margins, and in this specific case, even if $SNM_{TT}=85mV$, the $SNM - 6\sigma = 4mV$. In case of SNM and WNM, this can be assessed either by complex importance sampling methods or by

extrapolating into the tail of the partial SNM and WNM distributions as they are Gaussian, contrarily to the full SNM (or WNM) taken as the minimum of two partial ones.

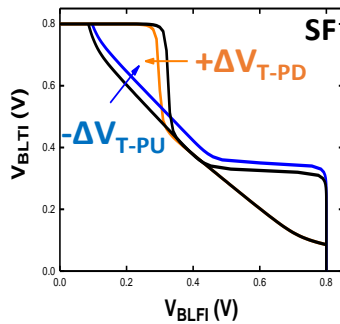


Fig. 54: SF (Slow NMOS and Fast PMOS) corner impact representation on butterfly curve. The SNM is drastically reduced.

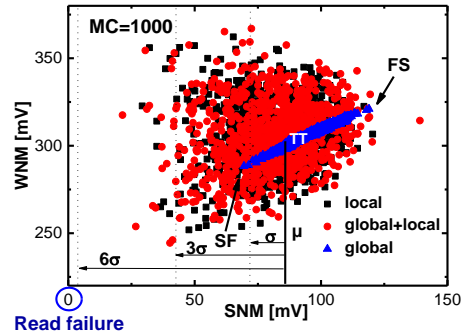


Fig. 55: WNM vs SNM taking into account global and/or local variabilities at $V_{DD}=0.8V$. A read failure is obtained at $SNM=0$. Typical (TT), Slow-Fast (SF) and Fast-Slow (FS) and margins are indicated. Taken from [78].

In this work, we first consider the TT corner (typical values for both NMOS and PMOS) before doing Monte-Carlo simulations (MC=1000), which are time consuming. We can note that changing the corner from SF (Slow for NMOS and Fast for PMOS) to FS allows us to tune the SNM-WNM metrics. In fact controlling independently V_T can modulate the different FOMs.

iii. Back-bias assist

In fact, thanks to the FDSOI structure, the back gate can be electrically biased, changing the threshold voltage of the transistors. Fig. 56 presents the threshold voltage V_T as a function of the well bias V_{well} for NMOS transistors (HD-PG). The back-gate acts as an additional gate, modulating the electron flow in the channel. Thus, a positive V_{well} will lower the V_T and a negative one will increase the V_T . A 63-50 mV/V threshold voltage modulation by well bias (V_{well}) was extracted for the HC-HD Pass-Gate (PG), respectively. This V_T modulation have a consequence on HD-PG I_{ON} - I_{OFF} FOM, presented in Fig. 57. This enables either the PG drive current to be boosted by 44% for $V_{well}=+2V$ or its leakage to be reduced below 0.1pA for $V_{well}=-2V$ on demand. Thus, by tuning the threshold voltage, the well provides an additional degree of freedom for changing the usual SRAM trade-off between read, write and retention operations.

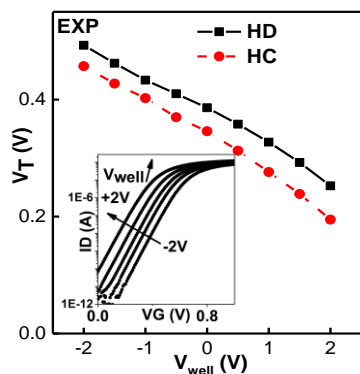


Fig. 56: Experimental Pass Gate threshold voltage modulation with back biasing. $I_D(V_G)$ (inset).

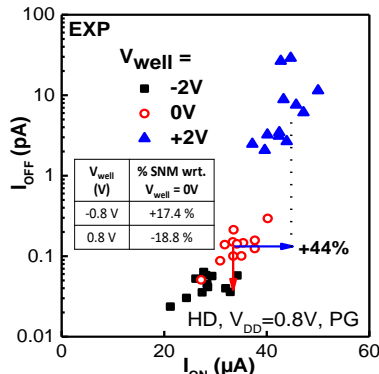


Fig. 57: Experimental High Density Pass Gate $I_{OFF}(I_{ON})$ for different V_{well} . SNM vs. V_{well} (inset).

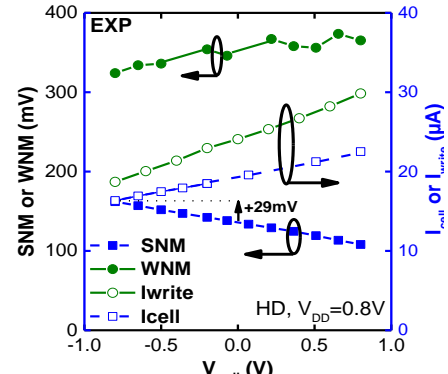


Fig. 58: Experimental Read and Write FOM as a function of V_{well} . A positive V_{well} increases the write margin as well as the current drive in read and write operation. A negative V_{well} increases the read operation margin. Figure reproduced from [78].

Fig. 58 shows the experimental modulation of SNM, WNM and read and write current by the V_{well} . In our case, since the well is shared between all devices in planar FDSOI, $V_{well} < 0$ V strengthens PMOS Pull-Up (PU, $V_T(PU)$ decreases) with respect to NMOS (PG, Pull-Down PD), helping the PU to maintain $BLTI=1$ and $BLFI=0$ during the read operation [98]. In fact, $V_{well}=-1$ V increases the SNM by 29mV. On the contrary, using $V_{well}>0$ V improves the PG/PU strength ratio (and so the WNM). Thus, the sensibility of SRAM to the back bias can be used to assist the read ($V_{well}<0$) and write ($V_{well}>0$) operation.

Furthermore, since variability is an issue for yield, this specific feature can be rather used as a process compensation technique, to narrow die-to-die variations. The SNM have been measured on 24 HD-SRAM cells with $V_{well}=0$ V among the 300mm wafer. The SNM values ranks from 95mV to 175mV as illustrated by the wafer mapping in Fig. 59. Custom back-bias values have been applied on the well to reduce the SNM variability. This technique is called back-bias assist. For instance, a negative (respectively positive) well bias is applied if the SNM value is lower (respectively higher) than the median SNM value. The range of modulation is $[-V_{DD}, +V_{DD}]$, with $V_{DD}=0.8$ V. Fig. 60 presents the 50mV SNM variability gain in HD bitcells across wafer.

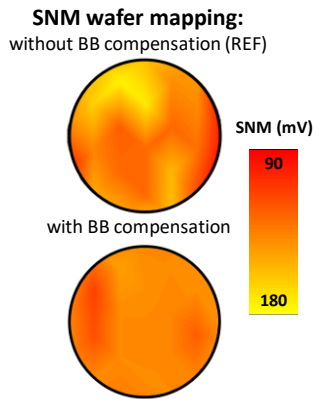


Fig. 59: Exp. SNM wafer mapping before ($V_{well}=0$ V) and after back bias compensation (various V_{well}). The back-bias range of modulation is $[-0.8$ V, $+0.8$ V]. Figure from [99].

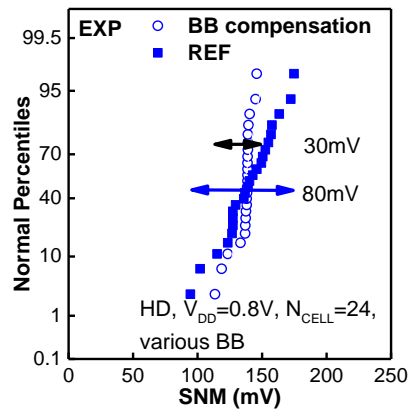


Fig. 60: Exp. SNM distribution before and after back bias compensation for 24 different dies. A 50mV SNM variability improvement is seen. Figure from [99].

To conclude, threshold voltage modulation by well biasing is efficient to improve the SNM of the read-limited HD cell. Back-biasing can be also used as a variability compensation technique rather than read-operation assist. However, if the read operation assist must be dynamic and the back-biasing must occur only during the read cycle, it is not the case for the variability compensation assist. That is why we have to ensure that this static back biasing assist does not cause additional stress (prematurely ageing), decreasing the benefits of this assist.

iv. BTI-induced dynamic variability at the bitcell level

To tackle ageing of the HD SRAM cell with and without the back-bias assist, the ageing process will be explained in a first part before presenting in a second part the electrical measurements.

a- BTI mechanism

Fig. 61 presents the lifecycle of an electronic component, in our case a transistor and is composed of three distinct periods. The first one is called infant mortality and is related to pre-existing defects [100] and considered in the yield. Its duration is of the order of months under normal working condition. The non operating devices can be identified by burn-in electrical tests, such as leakage detection in SRAM to detect the faulty cells. Theses extreme working-condition tests are done in the manufacturing facilities prior commercialisation. After this infant mortality, the failure rate stabilizes to its minimum, some random defects affecting the transistor operation. This region last several years and must be extended

for higher reliability. During the last period, end of product life, called wearout the failure rate increases again.

The goal of reliability tests is to determine the time-to-failure of a product or device, *i.e.* the time to enter in the wearout period. Usually this amount of time is of the order of years. For industrial reasons, one cannot wait years to determine the lifetime of the product before putting a product on the market. That is why the product ageing is accelerate thanks to two parameters: temperature and electrical potential.

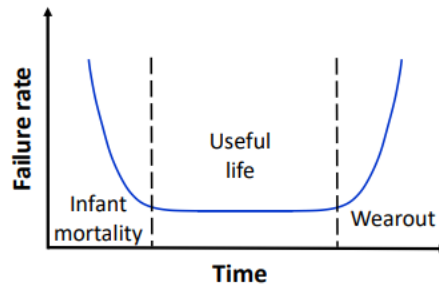


Fig. 61: Bathtub curve showing the lifecycle of a product taken from [101]. Three periods are distinguished: infant mortality where the failure rate is high, useful life where the failure rate is minimum and constant and wearout where the failure rate increases. The useful life duration can be a specification, such as 5 years, of a product.

The time-to-failure is then found back by modelling.

In the case of CMOS technology, the main failure mechanism are Time Dependant Dielectric Breakdown (TDDDB) for sudden failure of transistor, Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI) which are graduals. As far as TDDDB is concerned, a high voltage is applied on the gate and the time for oxide breakdown is measured, giving information about the quality of the gate oxide [102]. For BTI and HCI, the test is similar and consist in applying a stress voltage on the gate, when the device is heated (usually 125°C). For BTI case, the drain electrode is not stressed unlike for the HCI test. The resulting degradation is different but both are linked to oxide quality and oxide-silicon interface quality. More information about oxide defects and failure mechanisms are given in Annex I.

In this work, we will focus on BTI degradation at the SRAM bitcell level and not for individual transistors.

b. BTI at the bitcell level: experimental results

BTI is an issue for SRAM circuits since they are always powered. In particular, under retention mode (*i.e.* information storage, $V_{WL}=0$ V and $V_{BLT}=V_{BLF}=0$ V), half of the transistors are constantly under positive or negative BTI stress (Fig. 62). As explained previously, this phenomenon induces a threshold voltage shift, positive for NMOS and negative for PMOS [103]. As illustrated in Fig. 63, the SNM is reduced due to BTI induced V_T shift. It reduces mainly the read margin [104], which can be critical for the SRAM cell operation. Furthermore, ageing can be detrimental for security. In fact, if the same pattern is stored for a long amount of time, the V_T shift in the SRAM will be representative of the pattern and the previous data can be recovered. Ho *et al.* [105] demonstrates up to 21% data recovery of a 65nm commercial SRAM Lyontek because of ageing effects. This data imprinting effect must be avoided and strategies to relocate the information and minimize the transistor ageing can be done. That is why in this study, we focus only on BTI and not TTBD. The BTI induced threshold voltage shift will be compared between with and without [106] back-biasing. A negative well bias $V_{well}=-0.8$ V is chosen since such a value increases the SNM.

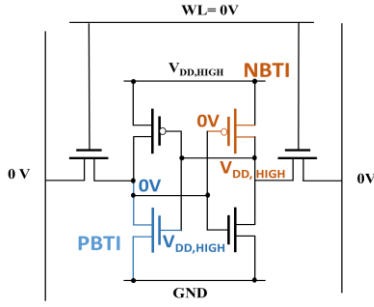


Fig. 62: SRAM schematics. Transistors under Positive and Negative BTI are highlighted when a '0' is stored at left node.

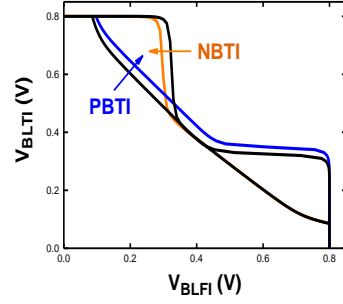


Fig. 63: Illustration of P and N BTI impact. Both mechanisms reduce the SNM.

The idea is to apply a voltage stress for different durations at high temperature and then monitor an SRAM metric to have an idea of the degradation. The standard SNM metric extracted from the well-known butterfly curve is obtained with two measurement steps, which is not fast enough to capture stress and recovery effects. That is why, the Supply Read Retention Voltage metric, representative of the SNM and compatible with fast measurement ($t_{meas} \sim 65 \mu s$ [107]), is chosen. This metric can be extracted from the bitline current measurements. In fact (see Fig. 64), the bitcell is initialized to a known state (for instance '0' in the left node and '1' in the right one) and the bit-lines and wordlines are precharged to V_{DD} . Then the cell voltage V_{cell} , is decreased while the bitline current is monitored. At some point, for V_{flip} , the cell state flips, dropping the current bitline. The SRRV is thus defined as $V_{DD} - V_{flip}$. Fig. 64 shows fourteen measurements on the same bitcell in a 10mV range: the 65 μs SRRV measurement is repeatable.

The measurements are done at 125°C, $V_{DD, stress} = +2V$ and a 1V supply voltage on 40 isolated HD SRAM cells. Two well biasing (20 bitcell for each condition) are chosen to characterize the cell without back-bias assist ($V_{well} = 0V$, reference) and with read assist ($V_{well} = -0.8V$). The stress time at $V_{DD, stress}$ varies between 0s (fresh cell) to 100s.

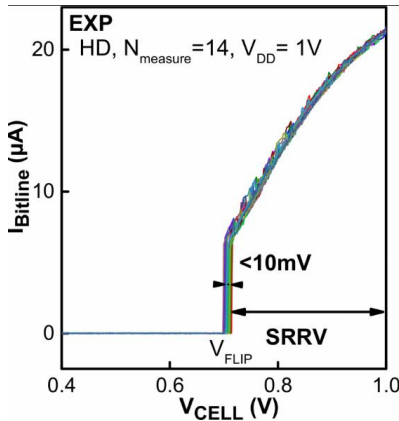


Fig. 64: SRRV measured 14 times on the same bitcell, showing the reproducibility of the measure. SRRV is the voltage difference between V_{DD} and V_{FLIP} (voltage when the information is flipped, extracted from the $I_{bitline} - V_{CELL}$ curve). Taken from [106].

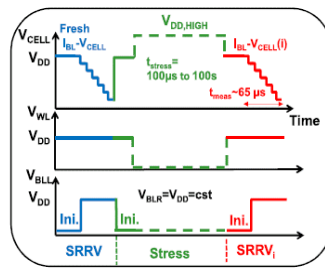


Fig. 65: Fast procedure waveform used for SRAM cell reliability characterization. Taken from [107]. The stress time at $V_{DD, stress} = +2V$ varies between 0s (fresh cell) to 100s.

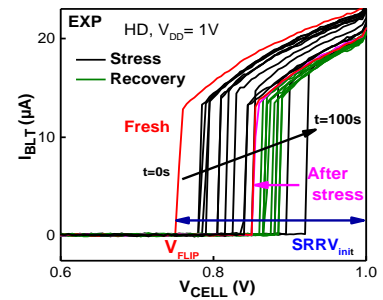


Fig. 66: SRRV measured on a bitcell during stress (black lines) and recovery (green lines). Stress reduces the read stability of the bitcell. Taken from [106].

Our Bit-Line (BL) current fast measurement method (Fig. 65), described in [107] is effective to extract the reduction of the read margin induced by stress. First the SRAM is initialized at a known state and the fresh SRRV is determined (V_{CELL} is decreased while IBL is sensed). Secondly a stress is applied ($V_{DD, stress} = +2V$ or 0V for recovery) during t_{stress} and the SRRV is measured. This sequence is repeated for different stress durations from 0s to 100s. Recovery time is also monitored.

Fig. 66 shows a typical measure of the $I_{\text{BITLINE}}-V_{\text{CELL}}$ curve the different stress times. A SRRV degradation up to 180mV is seen for long stress duration ($t=100\text{s}$). The recovery on SRRV degradation is highlighted in dark green.

The associated ΔSRRV distributions at different stress times for $V_{\text{well}}=0\text{V}$ and -0.8V are reported in Fig. 67. We can note that the mean value $\mu\Delta\text{SRRV}$ increases with stress like the standard deviation. Same ageing is observed (SRRV variation) for both V_{well} biases. It proves that back biasing does not degrade the BTI reliability.

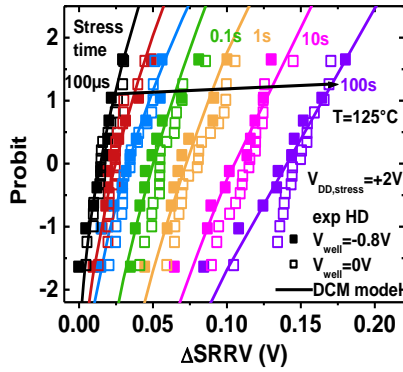


Fig. 67: Experimental ΔSRRV distribution on ~ 20 bitcells for two V_{well} polarizations and fitted using the model [104]. Same ageing is measured for both polarization. Taken from [106].

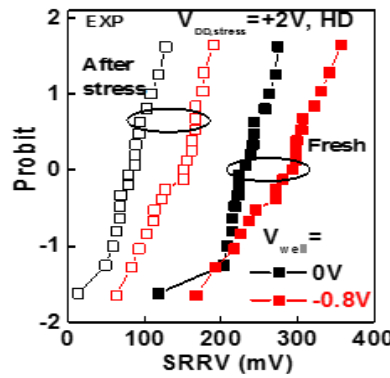


Fig. 68: Fresh and after stress SRRV distributions for different p -well biasing. A higher SRRV metric is seen for a well biasing of -0.8V . Taken from [106].

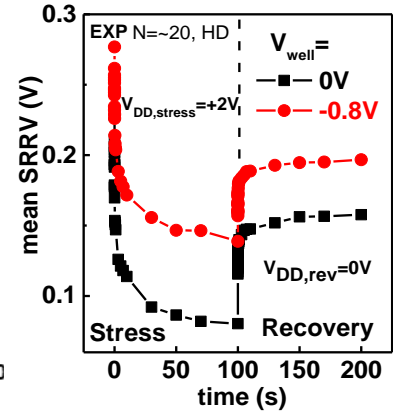


Fig. 69: μSRRV during stress and recovery. A -0.8V V_{well} biasing achieves better stability on fresh, stress and after stress bitcells. Taken from [106].

Finally, Fig. 68 and Fig. 69 shows that $V_{\text{well}}=-0.8\text{V}$ can efficiently boosts SRRV by 65.5mV for fresh bitcells, and that this benefit is preserved after stress. This proves the great interest of back biasing to improve not only performance but also reliability of the SRAM cell.

In this part, we showed that the 14nm high-density FDSOI SRAM cell were read limited. Well biasing have been proposed to increase read stability but also to compensate the SNM variability of the wafer. Such an assist does not degrade the SRAM reliability while increasing the performances. However, in this FDSOI configuration, the measured devices shares the same well limiting the voltage modulation since the PG and PD experience the same V_T shift. Also, the voltage range is limited due to the risk of forwarding the p -well/ n -well junctions [108]. To finish with, the large well capacitance is a drawback for a dynamic assist, limiting the speed. That is why, the use of dedicated back-plane at the transistor level can provide a dynamic assist without limited voltage range, increasing further the performances. In fact, 3D monolithic integration allows to connect the back-gate from behind, offering much greater opportunities for assist techniques application than regular planar FDSOI MOSFETs with diffused back gates. In the next part, such a 3D assist is investigated considering layout effects, parasitic elements and timing.

c. Proposition of a novel fine-grain back-bias assist techniques for 3D-monolithic 14nm FDSOI top-tier SRAMs

First, we will see what design is feasible with the Coolcube™ design kit rules within the area of the bitcell. Secondly, the performance gain with the design assist will be studied in detail. Last, the parasitic element will be computed and timing will be discussed.

i. 3D monolithic design kit: layout considerations

The idea in this part is to integrate a local back plane within the area of the top SRAM bitcell. The top HD SRAM bitcell layout is presented in Fig. 70. The 14nm 3D-monolithic design environment includes four intermediate metal lines iML and a back plane following the same design rules as a back end metal layer as illustrated in Fig. 71. Thus, the design kit allows placing an individual back plane underneath each type of SRAM transistor without area penalty. To connect the back planes, there is no need to differentiate each transistor (PUr from PUI for instance) to maintain the symmetry of the bitcell. So the maximum requirement is to dissociate the back gate of PD (BGPD) from BGPG from BGPU.

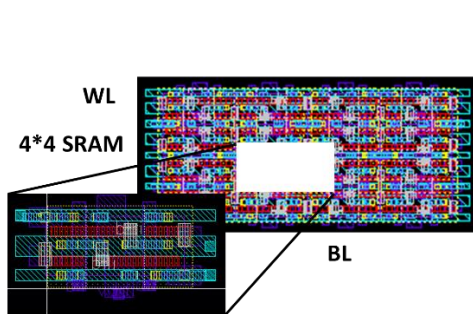


Fig. 70: Top high-density SRAM bitcell layout into a 4 by 4 SRAM matrix. The direction of bitlines and wordlines are indicated.

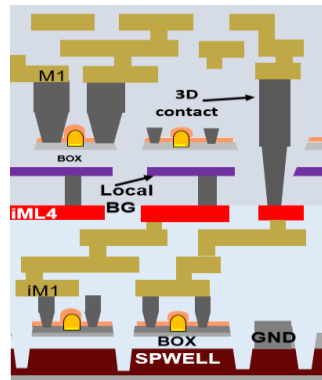


Fig. 71: Schematic stack of CoolCube™ 14nm Design Kit with intermediate vias between the back plane and the upper intermediate metal line.

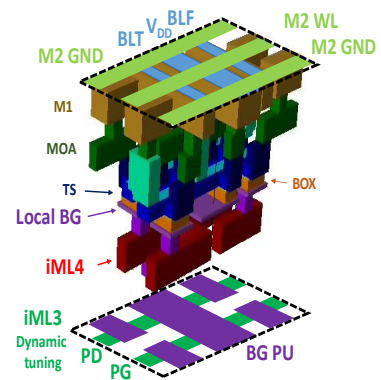


Fig. 72: SRAM 3D layout view with underneath back bias connections (green) routed in the word line direction for dynamic back biasing. PU are modulated in a static manner thanks to a shared well (purple). Taken from [109].

The first step is to verify with the design rules if the back gate of PU, PD and PG can be differentiated. Both PU devices are next to each other, nevertheless one PD is associated to one PU two times in the bitcell as illustrated in Fig. 73-a. However it is possible to define individual back gate below each nMOS transistors, the pMOS transistors having a shared back gate since they do not need to be set apart. Using this configuration, all the PU of a same column are already connected, since the same SRAM cell is repeated (see Fig. 73-b). However, without modification (with a unique back gate contact per column), the PUBG cannot be accessed in a dynamic way since the BG material is more resistive than iML. If a dynamic assist is required, back-gate contacts can be distributed to conduct signal more efficiently in this PU back-gate column.

The second step is to connect, using intermediate metal lines the two (a line or column) PG, respectively PD together. The connections can either be horizontal or vertical, the most important condition is that the distance between the 32nm width metal lines is higher than 32nm. A minimum area of $0.0046\mu\text{m}^2$ is required. In the $0.180\mu\text{m}$ height, the maximum number of lines can be computed from Eq. 4 and is 2.3. Thus, a maximal of 2 independent lines can be included in this bitcell. However, if the lines can be shared with neighbourhood cells, three iML4 lines can be integrated as designed in Fig. 73-c. Nevertheless, in Fig. 73-c, the PD and PG are connected together which is not wanted, so iM3 must be used to dissociate PD from PG.

$$W_{\min} = (W_{\text{spacing}} + W_{\text{lines}}) * N_{\text{lines}} + W_{\text{spacing}}$$

Eq. 4

The additional W_{spacing} takes into account the neighbourhood cell, if the lines are not shared between them.

In the 0.362 width, it is possible to design up to 5 iML4 lines. However, we would like to have them in the PD-PG area and connected to each other as represented in Fig. 73-d in the same bitcell. It is not possible to achieve this with the specified design rules. That is why the (c) configuration is envisioned, where the dynamic assessment of back gate is parallel to the WL, which is impossible in planar FDSOI technology with wells. Using iM4 and iM3 it is possible to dissociate the PG from the PD if required.

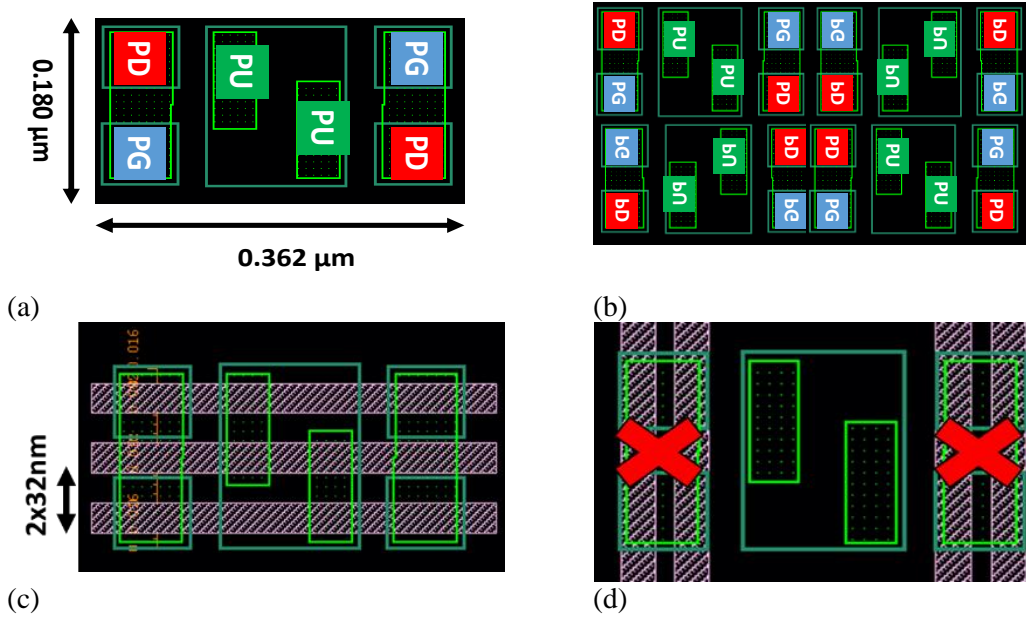


Fig. 73: Different routing configurations for the top-tier SRAM. (a) Description of the bitcell size and transistor position. The PU are close to each other whereas the PG and PD are paired. The back-gate are represented in green. (b) Representation of four adjacent cells to highlight the need of dedicated circuitry for PD and PG back gate respectively. (c) iML4 horizontal (parallel to the wordlines) proposition. (d) iML4 vertical proposition.

We have some design guidelines for the back-bias assist: BGPU are by-default connected in column for a static assist and two groups of back-plane can be dynamically connected. In the next paragraph, SPICE simulations are performed to investigate the interest of a back-bias assist.

ii. Fine grain and versatile back-bias assist

Thanks to the local back plane, the threshold voltage of each SRAM transistor type can be modulated independently in a wide voltage range with no risk of forward biasing any diode between the wells. The voltage back-biasing range here is arbitrarily chosen from $-V_{DD}$ to $+V_{DD}$. Detailed SPICE simulations in typical case for the HD bitcell are performed. Fig. 74 presents the gain in % of each metric (SNM, WNM, Iread and Iwrite) for different back-bias conditions for PU/PG/PD. The first feedback owing to this 3D representation, is that the threshold voltage of the PU must be lowered ($V_{BPU} < 0$) to improve all figures of merit (except leakage, not shown here). This cannot easily be achieved using a gate-first FDSOI process. However, this can be performed in 3D by using a PU-dedicated back plane with a constant bias ($V_{BPU} = -0.8V$) applied in all operation modes. Additionally, PD (or PG) threshold voltage can be modulated dynamically according to the SRAM operations to improve margins and currents. It can be imagined, for example to switch from a low leakage mode (stand-by mode) to a write-assist mode during a write operation and then to a read-stability assist when reading.

To define the assist conditions, the idea is similar to the aforementioned assist mode. In our case, the PD/PG strength ratio is modulated by an independent back-bias and not by changing the voltage

potential. Note that the two types of assist can thus be combined. Two types of assist for read and write operation can be differentiated: stability or drive current (speed) assist. Based on these considerations, three promising assist modes (with different V_{BPG} and V_{BPD}) are selected for their write (A1: $V_{BPG}=0.8V$ and $V_{BPD}=-0.8V$) and read stability (A2: $V_{BPG}=0V$ and $V_{BPD}=0.8V$, A4: $V_{BPG}=-0.8V$ and $V_{BPD}=0.8V$) as well as for their improved read time (A3: $V_{BPG}=0.8V$ and $V_{BPD}=0.8V$) assist performance. The changes in the butterfly curve for A1 and A2 are illustrated in Fig. 75. As far as the write assist is concerned, the strength of the PU ($V_{BPU}=0.8V$) is boosted in comparison of the one of the PD ($V_{BPD}=-0.8V$), increasing the write ability but being detrimental to the read stability. However, during read operation, it is possible to switch to assist mode A2 where the strength of the PD ($V_{BPD}=0.8V$) is increased with respect to PG strength ($V_{BPG}=0V$). In this latter case, the back-gate bias of PG is not negative in order not to degrade the read time. That is why an additional read stability assist A4 is proposed with $V_{BPG}=-0.8V$.

This versatile assist configuration yields +17% WNM, +28% I_{WRITE} for A1, +4% SNM for A2 +17% SNM for A4 and +28% I_{READ} for A3 at $V_{DD}=0.8V$ and $V_{well}=\pm V_{DD}/GND$ vs. the reference configuration with a single back-plane biased at 0V (Fig. 76). It should be noted that the assist mode A4 improves the SNM by 17% with a 10% I_{READ} penalty which can be interesting for a slower but low-power operation mode ($V_{DD}=0.8V$). The assist bias values applied to the different terminals in this 3D-monolithic structure and the associated results are summarized in Fig. 76.

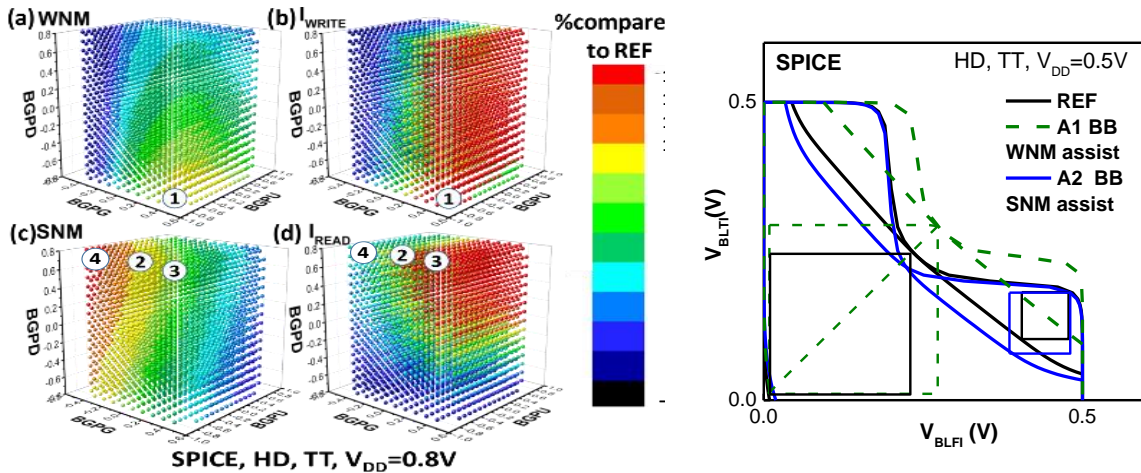


Fig. 74: Sensitivity (%) of (a) WNM, (b) I_{WRITE} , (c) SNM and (d) I_{READ} to independent back-biasing (on PD, PG, PU) (nominal configuration is at $V_{BG}=GND$) (SPICE) taken from [78]. Three assist modes are highlighted: A1, A2, A3 and A4.

Fig. 75: The impact on the butterfly curve for stability assists (A1, A2) taken from [78].

Furthermore, when compared to planar FDSOI (where $V_{BPG}=V_{BPD}$), 3D-monolithic integration offers more freedom owing to its ability to individually back-bias NMOS transistors. For instance at $V_{DD}=0.5V$, the best planar FDSOI configuration achieves (*w.r.t.* REF $V_{WELL}=0V$) +1.4% WNM ($V_{BPMOS}=0.5V$, $V_{BNMOS}=-0.5V$), +25% SNM ($V_{BPMOS}=-0.5V$, $V_{BNMOS}=-0.5V$), +72% I_{WRITE} ($V_{BPMOS}=-0.5V$, $V_{BNMOS}=0.5V$), +72% I_{READ} ($V_{BPMOS}=0.5V$, $V_{BNMOS}=-0.5V$). These gains are to be compared with +23% WNM (A1), +32% SNM (A4), +79% I_{WRITE} (A1), +78% I_{READ} (A3) for 3D configurations. In addition, to reduce the leakage current in the standby mode, the V_T of all transistors should be increased, which can easily be achieved in FDSOI structure with NMOS and PMOS independent well (reverse back biasing). However, this standby configuration is applicable with 3D monolithic structure regardless if the transistors are RVT (conventional well, regular V_T) or LVT (flip well, low V_T). In fact in planar LVT the RBB range (reverse back bias) is limited by the diode formed by the p-well (under PMOS transistor) and N-well (under NMOS transistors) (Fig. 77).

	REF	Stability enhancement			Speed (current) enhancement	
Operation	All	Write (A1)	Read (A2)	Read (A4)	Write (A1)	Read (A3)
PG (V)	0	0.8	0	-0.8	0.8	0.8
PU (V)	0	-0.8				
PD (V)	0	-0.8	0.8	0.8	-0.8	0.8
SNM (mV)	146	130	164	171	130	124
WNM (mV)	336	431	300	295	431	315
I_{READ} (μA)	17.4	20	18.6	15	20	22
I_{WRITE} (μA)	30.6	33	30	24	33	30

Fig. 76: Applied voltage on back-gate for assist A1, A2, A3 and A4.

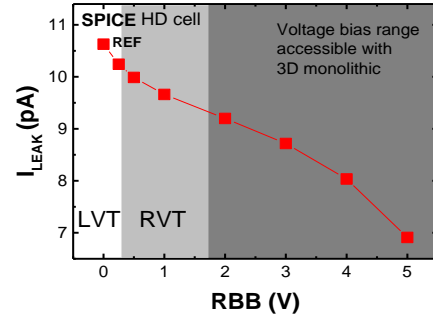


Fig. 77: I_{LEAK} as a function of applied reverse back bias. The voltage bias is informative and taken from [110].

Fig. 78 depicts the analyzed metrics in different biasing configurations with respect to reference case in function of V_{DD} showing an increasing improvement obtains with the V_{DD} decrease. There is an advantage in using back-bias assist for typical conditions, especially at low V_{DD} . We will now consider variability in our simulations to make sure that this gain in typical conditions also translates into a gain for non-standard devices. Monte Carlo simulations (1000 samples) were performed to evaluate the minimum operating voltage ($V_{min} = \min(V_{min-SNM}, V_{min-WNM}, V_{min-FOM})$ being the voltage for $\mu_{FOM} - 6\sigma_{FOM} = 0$) for the different assist configurations. V_{min} represents the lower-limit supply voltage to ensure that 99.99966% of the cases can be read or written. The minimum supply voltage for the HD REF bitcell is 0.56V. The write counterpart can be easily improved with negative bitline (NBL) assist ($-\Delta BL$ increases the strength of the PG) as seen in Fig. 79. A V_{min} reduction of 92 mV is seen with the A4 assist configuration and 60mV for the A2 bias scheme. Back-biasing techniques are thus efficient to boost write or read stability and to lower down the minimum operating voltage.

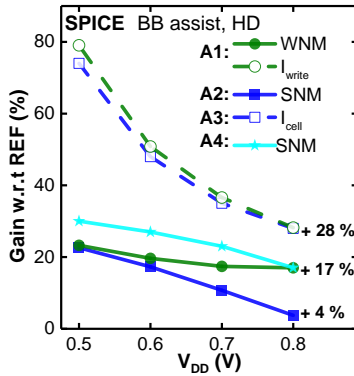


Fig. 78: WNM/SNM/ I_{write} / I_{read} improvement vs. V_{DD} w.r.t. REF (SPICE). Figure from [78].

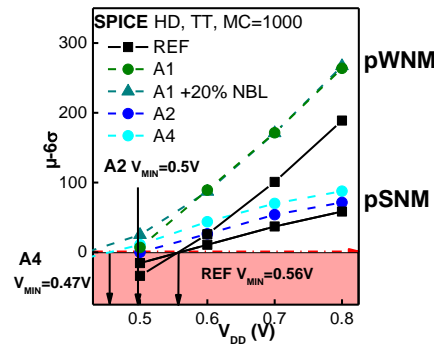


Fig. 79: Partial (to have a Gaussian distribution) SNM/WNM (at $\mu - 6\sigma$) as a function of V_{DD} . Voltage Range (VR) of back-bias are indicated. V_{min} is lowered up to 92mV with back biasing (A4, SPICE). Figure from [78].

The corresponding layout (common for A1-A2-A3-A4) was designed, connecting two groups of local (to-the-bitcell) back planes for PD and PG through internal vias without area penalty (Fig. 80). Back-plane lines parallel to the BLs are used to distribute a static PU bias. Moreover, the two dynamic signals are routed by iML3 in the WL direction within the SRAM height (whereas wells are typically in the BL direction in planar technologies). Thus, back biasing allows boosting a selected row in top-tier without disturbing other rows and without impacting the cell footprint. Since the access for BGPD and BPGP is dynamic and A1-A2-A3-A4 detain the same layout, one can switch between a read stability assist and a read time assist during the read operation or use a write assist when writing.

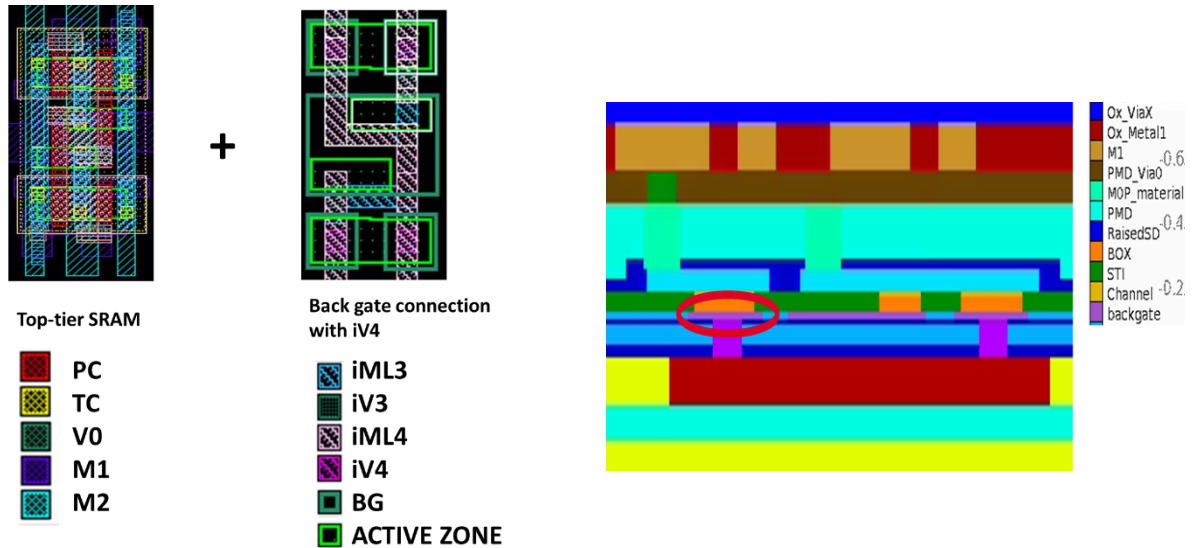


Fig. 80: Corresponding layout for A1-A2-A3-A4 assist mode.

Fig. 81: Clever tool cross-section used to compute the parasitic elements. The additional via is in purple.

iii. Parasitic capacitances reduction

In order to evaluate the capacitance gain provided by having a local back-plane instead of a continuous one, *i.e.* a common back plane running beneath all the devices (or a single well in planar), back-end parasitics have been extracted using clever tool presented in 3-c (Fig. 81) and included in the SPICE netlist. Fig. 82 presents the differences between the chosen configurations, the continuous BP being the reference.

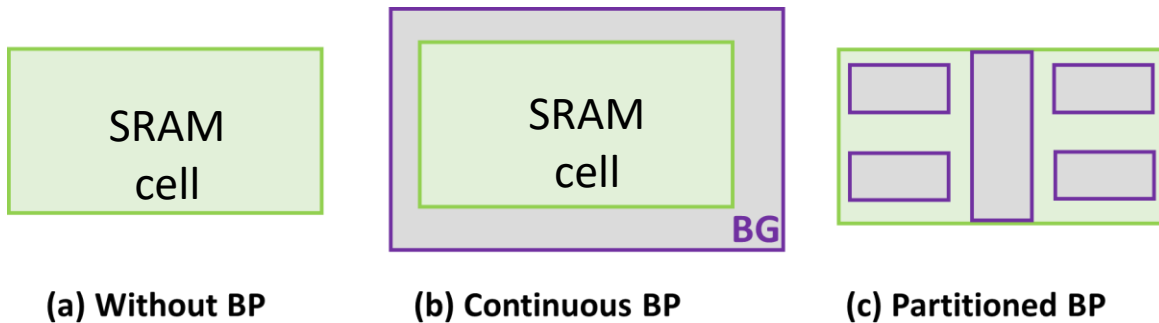


Fig. 82: Schematics of the three back-plane (BP) configurations investigated in this work, (a) without BP, (b) with a continuous BP (REF) and (c) with a partitioned BP, corresponding to the assist layout previously designed.

Fig. 83 gives as an example the BL capacitance values, showing that compared to a continuous back plane, the BL capacitance is reduced by 7%.

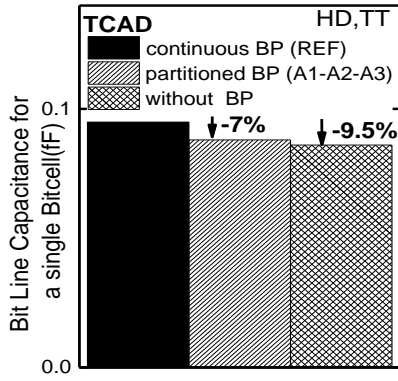


Fig. 83: Bitline capacitance computation for a single bitcell with different back plane configurations (TCAD). Taken from [78].

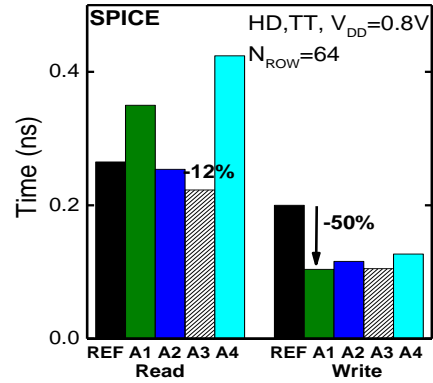


Fig. 84: Read/write time (SPICE). A3 is particularly interesting to boost cell-reading time. Taken from [78].

Using the obtained values allows an estimation of the dynamic bitcell performance for the selected biasing conditions. Read and write times are extracted on 64x64 SRAM matrix. A 12/50% read/write time improvement is achieved for A3 and A1 assist (*w.r.t.* reference cell at $V_{DD}=0.8V$) (Fig. 84). In A4, an increase of the read speed by 59% *w.r.t.* REF is observed, owing to the strong RBB on the PG, however, the point of this configuration is the V_{MIN} minimization, therefore the loss in terms of access time is less critical. A2 emerges as a good compromise for the read operation, achieving a better read stability and being slightly faster (-4%) than the reference. A2 detains also less leakage since, contrary to A3, the PG is not forward biased. In the write operation, A1 increases the write stability and the write speed (-16%) *w.r.t.* the reference. In addition, the demonstrated assist technique can be combined with WL underdrive, negative BL or other standard assist techniques [110] for further performance and stability improvement.

In this part, we showed that SRAM stability margins are highly vulnerable to process induced variability. The use of the well as a back plane in planar structures can mitigate this variability. However, the back gate polarization degree of freedom provided by top-tier SRAMs integrated in 3D is a real asset, enabling a dynamic polarization for a versatile and fine grain assist. In next part we will turn the variability between adjacent devices in the SRAM cell into an asset to generate a unique identification key for chips.

6- Variability as an asset: FDSOI SRAM PUF

In the previous part we saw technics to increase the margins limited by variability. In this part, variability is rather taken as an advantage to create a unique key of identification for circuits. In the first part, the achievement of such a functionality with SRAM is explained. In the second part, the different process lever to increase the variability in a dedicated part on the chip for this kind of operation are explained. Finally, emulation results obtained with the 14nm SPICE model are presented.

a. PUF: SRAM based fingerprint

With the IOT devices spreading, there is an industrial need for a low cost way of chip identification. In fact, wearables and portable devices contains user's data and has a regular access to the cloud. That is why an embedded private key is required to allow the IC recognition and to protect stored data. Process induced variability, when device manufacturing, can be exploited to generate a unique and non-predictable fingerprint. Such digital fingerprints are the output of a Physical Unclonable Function (PUF) to a specific input. The PUF when submitted to a known input will deliver an unpredictable but repeatable output. For instance, threshold voltage measurement is used for RFID [111] since the V_T variability comes mainly from random dopant fluctuation which is not spatially correlated. Su *et al.* [112] propose a dedicated design relying on mismatch and cross-coupled NOR gates to generate IDs. However to use existing devices without area overhead, Holcomb *et al.* [113] use SRAM power-up state to generate an identifying fingerprint.

An SRAM is composed of two inverters in series with two additional transistors to access the data. When the SRAM is not powered (*i.e.* $V_{DD}=0$) each internal nodes labelled BLLi and BLRi are discharged low (*i.e.* to gnd) '00'. When power is applied, the state (BLLi, BLRi) will be either '01' or '10' depending on process variation mismatch and noise. The final state will depend on the balance between the two inverters. Two cases can be distinguished in Fig. 86. The first one (a) presents a '1' skew cell where the cell is biased enough by process variation to be resilient to noise. The cell result when power-up is repeatable and thus, this cell can be used for identification. On the contrary, the (b) case cell is neutral and will indifferently gives a '0' or a '1'. So, a first insight for a technology- SRAM PUF friendly is that the skew must be higher than the noise.

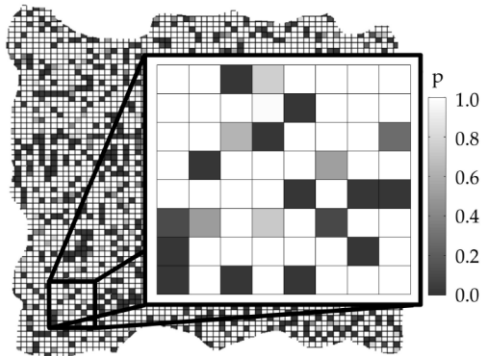


Fig. 85: a 64-bit fingerprint taken from [113]. The shade of grey indicates the probability of powering-up to 1. The desired pattern is 32 bits black and 32 bits white to maximize the security. When the cell is grey, the power-up state is not repeatable. If the randomness is process biased, the pattern will be easier to reproduce.

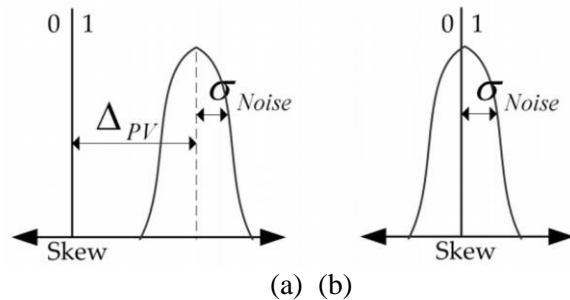


Fig. 86: Two cases are exposed and taken from [113]. The first one (a) presents a '1' skew cell: the noise is not sufficient to shift power-up state. Such a cell can be used for identification. In (b) case, the cell is neutral and will indifferently gives a '0' or a '1'.

To generate the fingerprint, a 64-bit SRAM array is powered-up one hundred time and the result is illustrated in Fig. 85. The probability of obtaining a '1' state is indicated by the shades of grey. The white and the black squares represent an SRAM with a repeatable power-up state. The grey ones are not skewed enough to be used as ID. To maximize the security, the perfect pattern would be 32 bits white and 32 bits black. However such a characteristic is not enough to ensure security. For instance, we can

imagine a pattern with the 1st half white and the 2nd half black which could be process dependant and be the same for each die. That is why, to quantify both reproducibility and unique character, the hamming distance is used. It is defined as the number of different bits between two power-up. Ideally, on the same fingerprint, the hamming distance is zero (no differences between two power-up). However, for different fingerprints, the hamming distance should be half of the array size if there is no asymmetry of the design or process bias.

To conclude, from an input power-up pattern, each SRAM array detains a specific answer which is reproducible if the variability of the SRAM is high enough compared to noise. However, Selimis *et al.* [114] evaluate a 90nm commercial 6T-SRAM for PUF applications, in particular, the sensitivity to temperature, supply voltage, voltage rump-up and ageing. In fact, the power-up state is sensitive to such variations and a fuzzy extractor [115] performing code correction error is implemented to counteract these limitations. In the next part, a proposition to enhance variability in dedicated FDSOI SRAM without additional circuit will be proposed to target such applications.

b. Single dopant transport

Some technological modifications to a baseline process can be done in order to increase the variability. For instance, O'uchi *et al.* demonstrate polycrystalline-Si channel FinFET SRAM based PUF [116]. A systematic comparison between poly- and monocrystalline-Si FinFET PUF cells is done. The poly-Si cell improves the intra-PUF hamming distance to 1/3.4 of that of the monocrystalline-Si cell, exploiting the variability of poly-Si. Also, ageing can be used to increase the variability on the chip [117]. In this study we rather propose to exploit a specificity of ultra-scaled FDSOI devices: single-donor ionization energy. As the transistors scale down, the channel is doped by a few discrete atoms and is prone to random dopant number fluctuation. When a single dopant is present in the channel it can participate to the conduction when ionized [118]. Resonant transport occurs and a peak of conduction is seen at $V_G = E_I$ (ionization energy) for low temperatures (10K) [119]. At 300K thermal broadening smears the peak, impacting SS and effective V_T [120]. So the signature of a single dopant in the channel at ambient temperature is a degraded SS and a smaller V_T : a leakier MOS. This degradation is influenced by the number of peaks, intensity and V_G position. The V_g position is determined by the distance between dopant and Gate and their ionization energy. The distance between dopant and gate is random and associated to ion implantation process. However, the dopant must be coupled to source and drain reservoirs, thus the distance between dopant and either source or drain must be lower than two times Bohr radius (2.2nm for As). It leads to a sizing constraint on gate length: $L_G < 10\text{nm}$. Also, the transistor width must be small to avoid the presence of several dopants in the channel. An advantage in using FDSOI devices is that the ionization energy is constant in bulk Si (~53.7 meV for As [121]) but not in SOI where a value of 108 meV is found in [122]. Another degree of randomness is added thanks to the choice of SOI wafers.

To conclude this part, the variability induced by the presence of a single dopant in a FDSOI ultra-scaled transistor will be considered to create SRAM based PUF. It requires only a light additional implantation step to dope only the SRAM. Next part will present the emulation of such devices thought SPICE simulations.

c. Emulation of leaky devices to assist technological choices

Using the 14nm FDSOI SPICE model introduced in part 3-b, we would like to emulate the presence of a single dopant in the channel in the high-density cell. First, the introduction of dopants will induce a shift of the threshold voltage of the transistors. We have to make sure that in the general case, *i.e.* without resonant transport, the SRAM bitcell is operational. Then, the SPICE parameters to emulate the transistor degradation due to the presence of a single dopant in the channel will be identified and discussed. Then, the simulation environment including Monte-Carlo simulations and the definition of a

metric V_{NM} are presented. At this stage, only skew linked to channel doping is considered. After, based on noise consideration and different power-up scenarios, the need of an additional degradation mechanism is highlighted. To finish with, leaky devices are emulated.

i. Impact of channel doping on SRAM devices

In the ideal case, a single doping atom should be implanted in the $L_G \sim W \sim t_{Si} \sim 10\text{nm}$ transistor channel. It means a doping concentration of the order of $N=10^{18}$ at/cm³. The induced threshold voltage shift for FDSOI transistor correspond to $(q \cdot N \cdot t_{Si}) / (2 \cdot C_{ox})$, 16mV for $N=10^{18}$ at/cm³ as illustrated in Fig. 87. This negative ΔV_T (P doped for PMOS and N doped for NMOS) has small repercussion on the butterfly curve, maintaining a good read margin (Fig. 88). However, for bulk devices, the V_T shift for the same doping is much higher and reduce drastically the read margin. For PUF applications, it is important to be able to read the information set by power-up. Nevertheless, if the SRAM is dedicated to PUF application, a destructive read operation does not matter. However, usually, an overhead circuit is present to select the good bitcells (the skewest cells) in a matrix to form a subset for PUF. In this case, part of the SRAM matrix will be used as a conventional SRAM, so the read margin is important. In addition, the introduction of doping does not degrade the symmetry of the cell (the butterfly crossing point being on the diagonal $V_R=V_L$), which avoids a technological skew towards a preferential state.

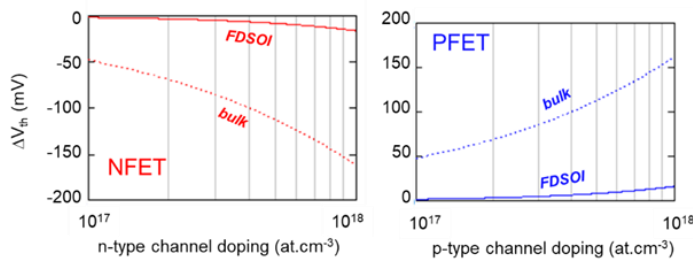


Fig. 87: Threshold voltage shift for FDSOI devices ($EOT=1\text{nm}$, $t_{Si}=7\text{nm}$) and bulk one ($EOT=1\text{nm}$). The NMOS have an n-type doping and the PMOS, a p-type.

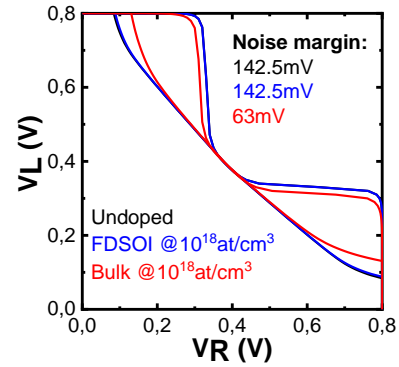


Fig. 88: Butterfly curve of the HD cell at $V_{DD}=0.8\text{V}$ with the associated read margins. The nominal cell is in black, the FDSOI cell with a $N=10^{18}\text{at/cm}^3$ is in blue and the corresponding bulk cell is in red. A detrimental reduction of the read margin is seen for bulk devices.

ii. Simulation environment

Using the 14nm FDSOI SPICE environment, we would like to reproduce the SRAM power-up and then read the information set in the bitcell. For this, we simulate a power-up ramp from 0V to V_{DD} in a time t_{pwup} (abrupt if equals to 0s, see Fig. 89) and a read operation is performed at t_{read} . To assess the reproducibility, this pattern is reproduced several times. Also, to take variability into account, Monte-Carlo simulations are done (if not indicated, $MC=1000$). On Fig. 89 we can see that an occurrence stabilizes at $V_{DD}/2$: it corresponds to $MC=0$, where no variability is introduced and the cell is perfectly stabilized. However, this ideal case is not representative of the reality. In the other cases, the cell will always shift either to 0 or to 1, to take a convention, we consider a 0-skew cell when the right node is 0. Fig. 90 presents the final distribution between GND and VDD for both right and left node and technological skew can be verified.

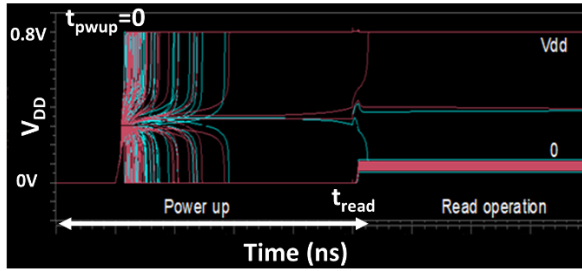


Fig. 89: Waveform of the simulated power-up and read operation. The different parameters are the supply voltage V_{DD} , the power-up ramp t_{pwup} , and when is performed the read operation. Here $MC=100$.

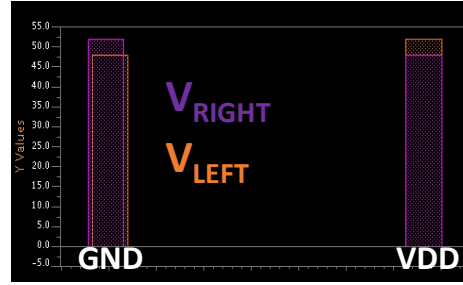


Fig. 90: Final distribution of the right (left) node potential. In the best case the repartition is half-half to maximize the information.

However, for computation reasons, we would like to determine from the butterfly curve if the cell is skewed. In addition, if the DC butterfly characteristic could be linked to power-up state, it could be a powerful metric for fast characterisations. In [123], the strength of the mismatch is determined by the distance between the separatrix and its ideal position. Based on this observation, a metric V_{NM} is defined as the shortest distance (orthogonal projection) between the butterfly curve cross-point and the diagonal (see Fig. 91). In the previous paragraphs the butterfly curve was plotted using reading conditions ($WL=1$). In the power-up case, WL is set to 0 according to our power-up scheme, so we rather use the butterfly curve in power-up conditions like in Fig. 91. To make sure it was the appropriate figure to consider, we performed MC simulations with the described power-up scheme ($V_{DD}=0.8V$, $t_{ramp}=60ns$) to analyse if the V_{NM} is correlated to the cell power-up state. As seen in Fig. 92 the V_{NM} extracted with the read butterfly curve is not representative of the final state contrary to the V_{NM} extracted with power-up conditions. However, for V_{NM} close to 0, there are some cells which do not polarized according to their V_{NM} preferential state (more visible in Fig. 93). For the metric V_{NM} to be reliable, these particular points must be understood properly.

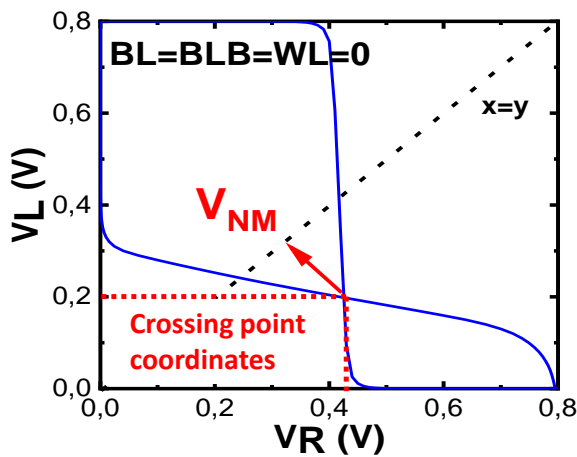


Fig. 91: Definition of the V_{NM} metric to quantify if a cell is skewed or not. From this butterfly power-up curve, a metric called V_{NM} is defined as the crossing point distance (V_R , V_L) to the separatrix.

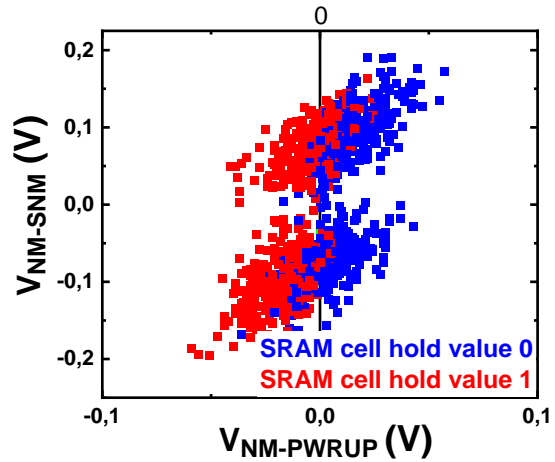


Fig. 92: SRAM cell hold value according to V_{NM} power-up and V_{NM} read conditions when the SRAM is powered-up and read.

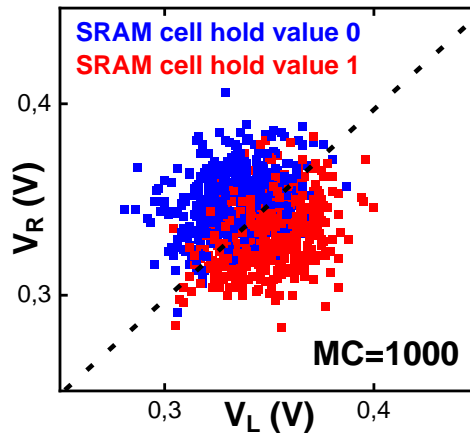


Fig. 93: Crossing point coordinates (V_L , V_R) with the power-up SRAM cell for $MC=1000$.

As far as the power scheme is concerned, several studies show the importance to define precisely the V_{DD} ramp to. Elshafiey *et al.* [124] demonstrate that the start-up value of an SRAM PUF depends on the SRAM power supply rising time and can be optimized to reduce the undetermined cells down to 5%. In fact, depending of the supply power-up ramp time, there is two operation regions. The first one is dominated by capacitance and threshold voltage variations and the second one by threshold voltage variation only. Each cell can be skewed differently for each region, leading to a different power-up state and a different probability to flip. It is showed that a higher rise time will consider only the threshold voltage variations and reduce the undetermined cells. Furthermore, Wang *et al.* [125] analyse different power-up scenario to highlight PUF state sensitivity. An abrupt 0 to V_{DD} (respectively V_{DD} to 0 for V_{SS}) ramp will polarize the cell according to PMOS (respectively NMOS) threshold voltage difference. On the contrary, an extremely long V_{DD} ramp (of the order of the second) will result in a power-up state induced by both NMOS and PMOS variations. For intermediate rise time, the number of undetermined states is unneglectable. That is why, a proper V_{DD} ramp must be chosen to maximize the repeatability. To see if our undetermined points were not linked to the abruptness of the V_{DD} ramp, a ramp of the order of the second have been chosen. Fig. 94 presents the probability of obtaining a logic state '1' as a function of the V_{NM} for different ramps duration. We do observe that the longer the ramp duration, the lower the closer we were to the ideal curve defined as $P(1)_{V_{nm}>0}=1$ and $P(1)_{V_{nm}<0}=0$. However, even for ramps of the order of the second, the error rate is still 9.9%, so the V_{NM} might not be the best metric to consider. Nevertheless, if the V_{NM} is lower than -0.012 (respectively higher than 0.012), the probability to obtain a 0 (respectively a 1) equals to one for $t_{RAMP}=1ms$. So, a $\pm 12mV$ margin can be defined around $V_{NM}=0$ and the further points can be considered reliable and repeatable. The V_{NM} considered here is the V_{NM} at $V_{DD}=0.8V$ but the ramp, especially the longest one, are continuous from GND to VDD. Fig. 95 presents the V_{NM} evolution for 45 MC samples for V_{DD} from 0.3V to 0.8V. We do observe that the V_{NM} is generally higher for lower V_{DD} and saturates from $V_{DD}=0.5V$. For lower V_{DD} values such as 0.1V, the V_{NM} could not be determined since the butterfly curve section closes and they are several crossing points.

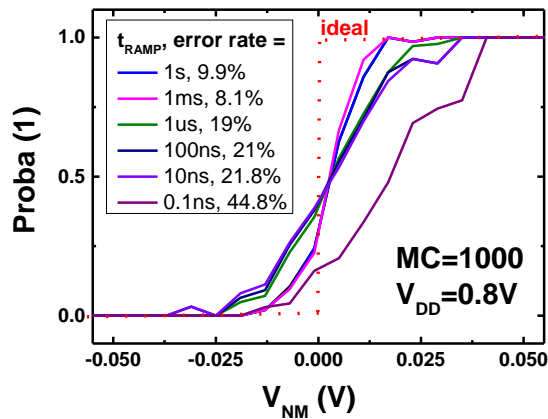


Fig. 94: Probability to obtain a one knowing the V_{NM} value for different V_{DD} ramp durations.

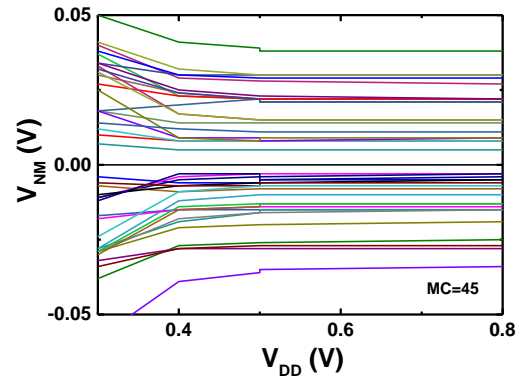


Fig. 95: V_{NM} measured for 45 points at different V_{DD} . The lower the V_{DD} the larger (generally) in absolute value the V_{NM} is.

Since the V_{NM} metric is not sufficient to determine accurately the power-up state for a conventional power-up operation, we decided to test another power-up scheme. In fact, we do observe on Fig. 92 that the V_{NM-SNM} values are in a higher range than the $V_{NM-PWUP}$ and even better, there is a separation between the two distributions (for $V_{NM-SNM}=V_{NM-PWUP}$). So, instead of just turning ON V_{DD} , we propose to turn ON also bitlines and wordlines to be in a READ operation configuration and benefits from the additional variability of the PG. The power-up scheme is described in Fig. 96. With this new power-up scheme, the Fig. 92 is reproduced in Fig. 97 with a long duration ramp, $t_{RAMP}=1s$. The first observation is that some devices (green points) stabilizes at the crossing point of the butterfly curve, *i.e.* they do not switch towards ‘1’ or ‘0’. However, the number of mistakes (Error rate $Er=3.7\%$, accounting for the undetermined points) is lower than the previous power-up scheme which was 8.8% in the best case. From now, we will consider such a power-up scheme but with a $t_{RAMP}=10ns$ (abrupt) which yields the same results.

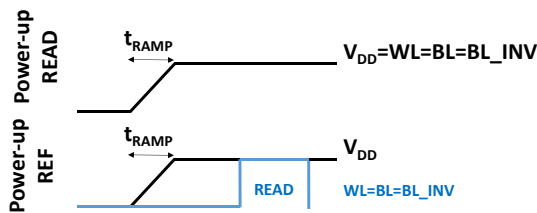


Fig. 96: “READ” power-up scheme compared to the standard one.

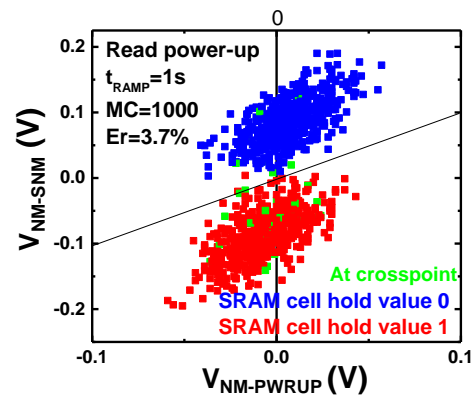


Fig. 97: SRAM cell hold value according to V_{NM} power-up and V_{NM} read conditions when the SRAM is powered-up using the READ power-up scheme.

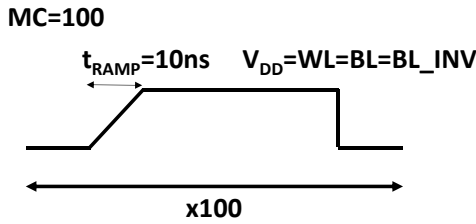


Fig. 98: Waveforms used to investigate the reproducibility of the power-up.

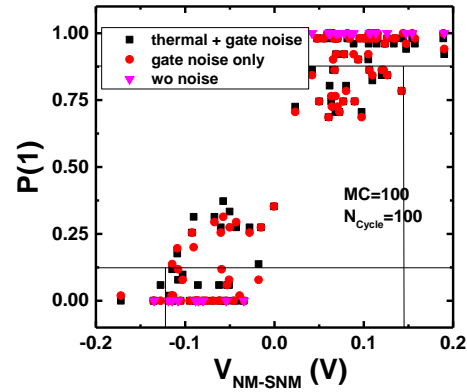


Fig. 99: Probability to obtain a one as a function of the V_{NM-SNM} with the noise model activated.

Now, we will consider the noise to verify the reproducibility of a power-up and its tolerance or not to noise. Cycles of power-up are scheduled 100 times for 100 devices (see Fig. 98). Then the probability of obtain a 1 is sorted out. Concerning the noise, the different flags at our disposal to modulate it in the UTSOI model are:

- SWIGN: Boolean to activate or not the gate noise model
- FNT: thermal noise coefficient $nT = 4kBT/KCFNT$ which can be activated or not.
- FNTEXC: excess noise coefficient
- NFA, NFB, NFC: flicker noise coefficient

Without changing the noise parameters by-default, the gate noise and the thermal noise are activated or not. The result is given in Fig. 99. The main idea is activating the noise will endanger the reproducibility of the power-up. Experimentally we cannot control the level of noise which is technology dependent, so from such a graph, we can just extract a margin V_{noise} such as $P(1)_{V_{nm} > V_{noise}} = 1$. For instance, on the graphic Fig. 99 a $V_{noise} = 150$ mV can be extracted. The cells detaining a V_{NM-SNM} value above in absolute value V_{noise} will switch predictability towards 1 or 0 according to the V_{NM} sign.

To conclude this part, the emulation SPICE environment is set-up and a metric V_{NM} is proposed to predict the power-up state. However, this metric seems to be incomplete and subjected to noise, so that to predict the power-up state, enough margin, noted V_{noise} must be considered. In the next paragraph we will consider the case where some devices detains a single dopant in their channel and analyse if the distortion of the SRAM cell is sufficient to achieve this V_{noise} margin.

iii. Emulation of resonant transport

The presence of a dopants in the channel will degrade significantly the transistor at ambient temperature. To reproduce this behavior, the electrical gate length is reduced. As seen on Fig. 100, from a nominal electrical gate length of 34nm (30nm physical gate length and 2nm per side underlap) to a 14nm electrical gate length, the threshold voltage have been reduced by 250mV and the subthreshold slope degraded by around 20mV/dec for NMOS transistors. This degradation on the subthreshold slope as well as the threshold voltage will account for a resonant transport. Fig. 101 presents the repercussion of this degradation on the butterfly curves. If only one of the PD is leaky, the butterfly curve will be distorted since the PD will impose the ground quicker from one side. If one of the PG is leaky the read margin is degraded since the PG/PD ratio is changed but the butterfly curve is almost not impacted.

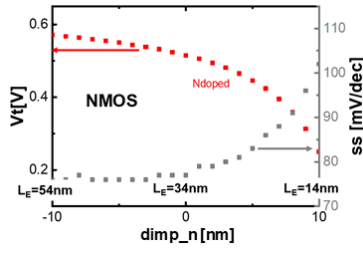


Fig. 100: V_t and SS of an NMOS transistor as a function of electrical gate length. The smaller the gate length the higher the parameters are degraded. A single dopant in the channel is emulated with a $L_E=14nm$ to take into account the ambient temperature degradation induced by resonant transport peaks.

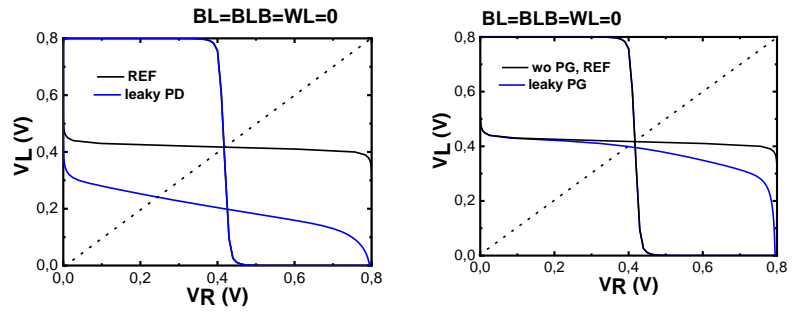


Fig. 101: Power-up butterfly curves with a degraded PD (a) and a degraded PG (b). The deformation is more important for PD since the PD takes part of the inverter pair to set the memory point.

So we would like to estimate the degradation on the V_{NM} : if the degradation is enough, $V_{NM} > V_{noise}$ and the cell will be entirely predictable. Extreme cases with a nominal gate length of 30nm and a degraded one of 14nm for each transistor will be studied. However, to simplify the problem and not study the 64 configurations, we analyse the situation where only one side of the inverter is degraded (8 cases). In fact, if both inverter are degraded the same way, the read margin value will change but the V_{NM} will be equals to 0 as illustrated on Fig. 102 and Fig. 103.

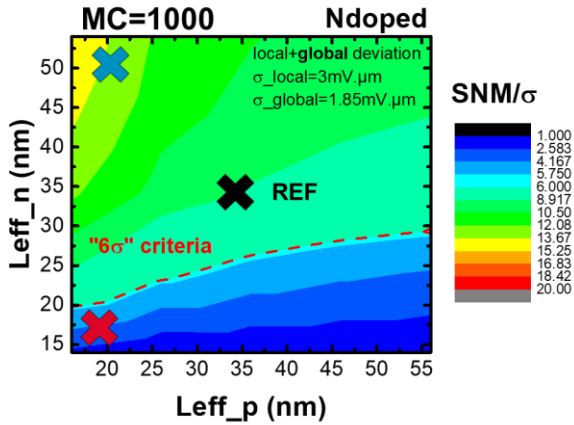


Fig. 102: Verification of the read operation. Local and global deviation is considered and the industrial criteria $SNM/\sigma > 6$ is represented by a red line.

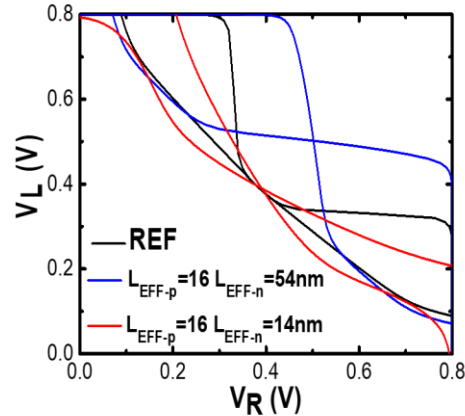


Fig. 103: Butterfly curves for various NMOS and PMOS degradations. The V_{NM} values are all equals to 0mV in the case of a symmetrical degradation.

For the eight configurations, summarized in the table below, we see that most of them satisfy the condition $|V_{NM-SNM}| > V_{noise}$.

Transistors degraded	REF	PU1	PD1	PG1	PU1 & PG1	PG1 & PD1	PU1 & PD1	PU1&PG1 & PD1
V_{NM-SNM} (mV)	0	72	-282	247	247	70	-161	-148

iv. Conclusion:

In this part we proposed to use SRAM power-up state as a digital fingerprint (PUF) of a device. Based on SPICE simulations, a metric (V_{NM}) have been proposed to predict the final transient power-up state from a DC characterisation of the SRAM. Different sensibility studies (power-up scheme, capacitances, noise) have been done to analyse the robustness of V_{NM} . Also, single dopant devices (fabricated with channel implantation) are emulated and are efficient to skew the bitcell and being tolerant to noise. However, device optimization is important to conserve a sufficient read margin.

7- Conclusion of chapter II

We presented the actual 3D VLSI digital planar design flow and the opportunity brought by 3D monolithic integration to pursuit Moore's law. Then based on our 3D environment simulations, we discussed the interest of sharing resources and signals between the two tiers and we studied in details a versatile fine grain back-bias assist for 3D top-tier SRAM. To finish with, we investigated the interest of single dopant transport on planar SRAM based PUF. In fact, 3D monolithic integration leverage the possibilities for design engineers paving the way for high density, low power and high performance specified circuits. However, this is not possible without the fabrication of CMOS transistors over CMOS. The next chapter presents the 3D monolithic process flow where the top-tier is done at low temperature (under 500°C, 2hours) to preserve the bottom tier. In particular the fabrication (and characterization) of low temperature junctionless devices is explained.

Take away of chapter II:

- Minor modifications can be added to planar VLSI process flow to enable 3D monolithic circuit design. Performances or thermal driven algorithms for place and route are proposed in the literature.
- Moving from planar dies to 3D monolithic dies manufacturing one is cost efficient.
- 3D monolithic designs reduce (in general) the area and the overall wire length leading to higher performances.
- Thermal dissipation (hot spot and peak temperatures disparities between tiers) is not an issue.
- Resources such as power rail or clock signal can be shared between tiers without a significant area overhead.
- Dynamic local back-bias are of great interest since they can modulate the threshold voltage of devices independently. A SRAM top-tier assist have been proposed to reduce the minimum operating voltage by 92mV.
- Planar SRAM variability in FDSOI technology can be used to create digital fingerprints. This fingerprint is more robust (less sensitive to ageing and more reproducible) when combined with single dopant transport.

Chapter III: Fabrication of junctionless transistor in the scope of 3D monolithic integration

3D monolithic IC design can improve at the same time performance, power and area compared to planar one. However, to process sequentially the top transistor without degrading the bottom one, a maximum thermal budget of 500°C, 2hours have been identified and still remain challenging. The aim of this chapter is to fabricate transistors compatible with a 3D monolithic integration and to characterize them. In the first part, the state of the art of 3D monolithic integration demonstration is done and junctionless transistors which are good candidates for a low-temperature integration are presented. In the second part, TCAD simulation of junctionless devices are presented to explain its physical behavior. In the third part, junctionless transistors are compared to standard one (inversion-mode) in terms of mobility, capacitances, variability, reliability and noise. After, TCAD simulations allows the sizing of the future device, targeting digital and analog applications. Then, the process flow is developed, presenting the different low-temperature bricks. To finish with, the processed devices are electrically characterized with an emphasis on variability and logic and analog applications.

1-	State of-the art	77
a.	3D sequential integration demonstration: review of literature	77
i-	Deposited top-tier channel material	78
ii-	Reported top-tier channel material	78
b.	Junctionless transistors	80
i-	Short presentation of the JL transistor (JLT) architectures	80
ii-	Polycrystalline materials	81
iii-	Other materials	82
2-	TCAD simulations	83
a.	Chosen device architectures	83
b.	Physical Model used and justification	85
3-	Junctionless MOSFET operation	86
a.	Sub-threshold region: depletion	86
b.	From threshold voltage to flatband voltage: volume conduction	87
c.	Above flatband voltage: accumulation region	88
d.	Analytical models	88
4-	Characteristics of Junctionless devices	89
a.	Effective channel length modulation	89
b.	Mobility	89
c.	Capacitances	91
d-	Variability	92
d.	Reliability	94

e.	Noise	95
f.	Junctionless transistor applications	96
5-	Device sizing	98
a.	Tri-gate junctionless sensitivity to silicon thickness and doping level	98
b.	n over p channel	100
i-	PN junction physics	100
ii-	CMOS Integration	102
iii-	Sizing of the different layers: TCAD simulations	104
c.	Performances of the different structures compared to IM devices	105
6-	Fabrication process flow	108
a.	Gate first integration at high temperature	108
b.	Channel material	109
c.	Active zone patterning	113
d.	Gate stack	114
i-	Gate stack materials	114
ii-	Gate stack etching	116
e.	Spacer	117
f.	Junction engineering SPER	118
g.	Thin silicides	122
7-	Overview of studies related to 3D monolithic integration	123
8-	Electrical results	125
a.	Device fabrication	125
b.	Digital Figure-Of-Merit of Junctionless nMOS	126
i-	Electrical performances	126
ii-	Mobility	128
iii-	Overlap capacitance	129
c.	Analog Figure-Of-Merit of junctionless nMOS	130
i-	Analog gain leveraged by back-bias	130
ii-	Reliability and noise	131
iii-	RF Figure-Of-Merit of junctionless nMOS	133
9-	Conclusion of Chapter 3	135

The previous chapter presented the different opportunities brought by 3D-monolithic integration with a highlight on digital circuits. In fact, stacking instead of shrinking continues the reduction in die size and die power, allows the integration of heterogeneous material and offers new architectures to improve performances [126]. The main technology challenge for this integration is the thermal budget constraint of the top-level process integration. This chapter starts with a short review of 3D sequential integration demonstration and junctionless devices. Then the physical properties of junctionless devices are presented with the help of Technology Computer-Aided Design (TCAD) simulations. Afterwards, TCAD sizing, process flow and electrical results are explained and discussed.

1- State of-the art

The following state of the art is composed of two parts. The first one is directly related to 3D monolithic integration, which focuses on the 3D demonstrations. The second one discusses quickly the junctionless architectures and their applications to highlights their potential for 3D monolithic integration.

a. 3D sequential integration demonstration: review of literature

3D monolithic integration consists in stacking active layers on top of each other in a sequential manner [127]. The top active layer can either be created by direct deposition or by wafer bonding. For both cases, Fenouillet-Beranger *et al.* identifies a maximum thermal budget for top-tier processing in order to avoid bottom CMOS degradation [128]. In fact, the annealing temperature and its duration cannot exceed 500°C, 2h without damaging the stability of bottom devices Ni_{0.85}Pt_{0.15} silicide [129], [46]. Note that, Ni_{0.9}Co_{0.1} silicide is stable up to 800°C [130] and thus could reduce this thermal budget constraint. However, Fig. 105 highlights that higher thermal budget can be applied with shorter durations. For example, thanks to its low-light-depth penetration, a laser (wavelength: 308nm, pulse duration ~200ns) can even melt the top silicon layer without affecting the underneath layer [131]. The most critical steps, when considering thermal budget, in a transistor process flow are the spacer formation (~630°C), the selective epitaxy (SiGe 30% at 650°C or Si at 750°C) and the dopant activation step (>1000°C) [128]. Low temperature device processing (*i.e.* full standard flow with limited thermal budget) is not trivial and will be assessed in part 6-.

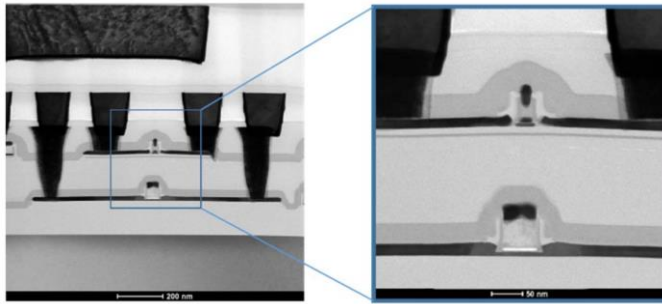


Fig. 104: TEM cross-section of a 3D monolithic integration demonstration in [132]. Two devices are stacked on top of the other featuring high-k metal gate stack.

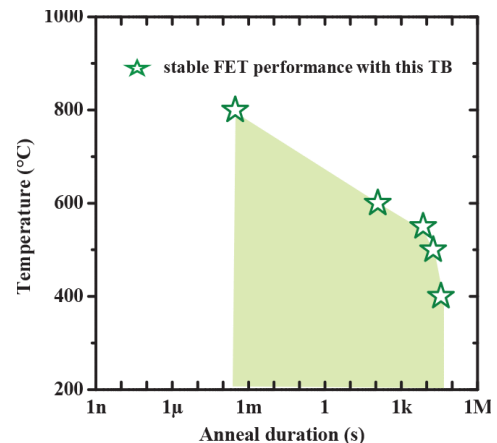


Fig. 105: FET Thermal budget processing window to ensure stability, taken from [128]. The thermal limit has been established at 500°C, 2hours. High temperature are feasible for a very limited amount of time.

The following review of 3D sequential integration demonstration is divided into two parts according to the process used for top-active creation. The first one deals with deposited top-tier and the second one with wafer bonding. A small emphasis is done on heterogeneous integration.

i- Deposited top-tier channel material

To create the future channel material, amorphous silicon (a-si) can be deposited at low temperature (see section 6-b.i-) directly on the intermediate dielectric oxide. In fact, a monolithically integrated thin-film-transistor (TFT) for 3D FPGA is reported in [133]. In this demonstration, amorphous silicon (a-si) is directly deposited on top of Cu interconnects (nine layers) and patterned below 400°C to form a-si TFT. At $V_F=3.3V$ supply voltage, an I_{ON}/I_{OFF} ratio superior to 2000 is achieved. However, the mobility in a-si is much lower than in poly silicon and monocrystalline silicon, yielding to a lower drive current. To boost transistor performances, green nanosecond laser anneal ($\lambda=532nm$) for highly crystallized and large-grained ($>1\mu m$) epi-like Si channel preparation can be used [134]. Poly-si transistors will suffer from poly-si grain size variability and grain boundaries [135]. A three time degradation of threshold voltage variability compared to single crystal one is demonstrated. Nevertheless, the μ -Czochralski process with a grain-filter structure (narrow cavities) is a solution, allowing the formation of location-controlled Si grain up to $6\mu m$ diameter [136]. Fig. 106 presents the different seed window techniques. The control of 2D location and size of Si grain, and thus grain boundaries location, allows building transistors on single grain.

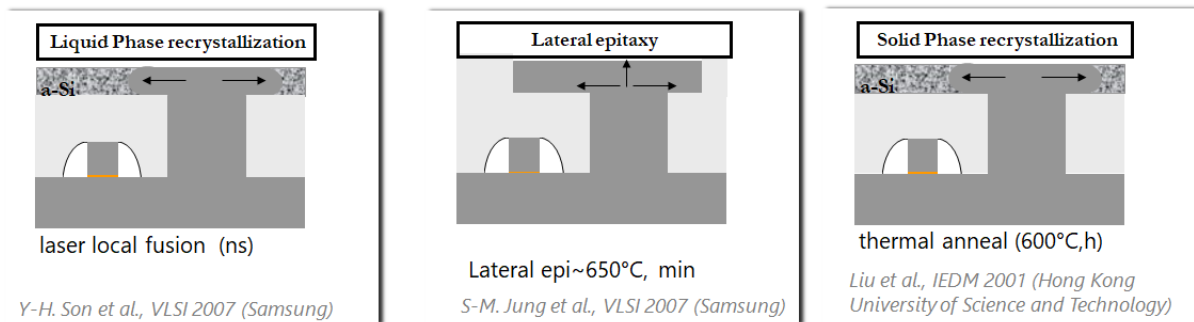


Fig. 106: Seed windows techniques, taken from Batude IEDM 2019 tutorial [137]. The common idea is to benefit from an underneath seed (localized on the bottom layer) to “copy” the crystalline information. Recrystallization of the top amorphous layer can be done either in the liquid phase (with localized melt) or in the solid one (with long thermal anneal).

Also, a 72M bit density 3D SRAM is demonstrated by Jung et al [138]. The creation of the top-tier channel material is done with Laser-induced Epitaxial Growth (LEG). In fact, after intermediate dielectric layer (ILD) planarization, seed holes are patterned and filled when the amorphous silicon top-layer is deposited. Thus, the bottom layer acts as a crystalline seed when top-layer recrystallizes, under the laser annealing, to provide high quality Si channel layer [139].

To conclude, amorphous silicon can be deposited at low temperature and recrystallized without degrading bottom tier transistors. Furthermore, grain position can be controlled at the expense of space loss (seed window and grain junctions). However, poly-Si transistors suffers from degraded variability or/and low density. That is why a high quality and uniform active is required to take fully advantage of 3D monolithic integration for high performances applications. Such a technology is attractive for low cost and variability tolerant applications where density and high performances are not required.

ii- Reported top-tier channel material

Fig. 107 presents the typical wafer bonding process flow to obtain a perfectly monocrystalline active layer [140]. After pre-metal dielectric Chemical Mechanical Polishing CMP, direct top substrate bonding is carried out before obtaining the future channel either by grinding and etching or by Smart Cut™ [141]. Compared to poly-Si deposition, this approach requires the use of a donor (usually SOI) wafer, and thus is more expensive. However, the crystalline quality of the bonded channel is higher.

Furthermore, to take benefit from inter-tier interconnections, the introduction of metal lines (fabricated with Back-End-Of-Line BEOL tools) between the tiers (performed in top-tier Front-End-Of-Line FEOL tools) leads to contamination issues [142]. Now, we will distinguish reported silicon channel from more exotic materials such as GeOI.

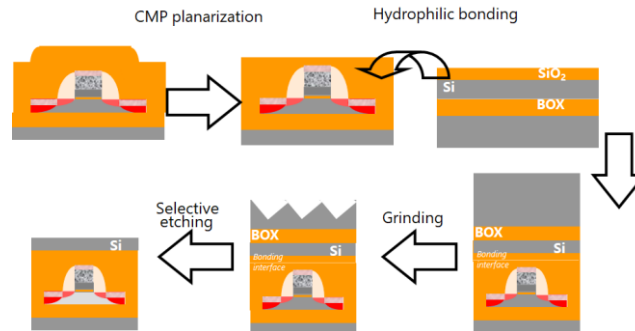


Fig. 107: Wafer bonding process flow taken from [137]. After CMP bottom tier planarization the future top tier material (SOI wafer) is bonded onto the bottom tier. Most of the Si bulk is grinded before selectively etching the BOX to let a thin silicon layer.

A wide range of semiconductors, such as silicon, III-V, carbon nanotubes, can be heterogeneously integrated together thanks to 3D monolithic. In fact, each layer could be independently optimized for specific functionality. For instance, Batude *et al.* [143] demonstrated the integration of p-GeOI MOSFET (top-tier) on n-SOI MOSFET (bottom-tier) for high performance purposes. The 200mm GeOI wafer is bonded onto processed wafers. Similarly, germanium PMOS is transferred onto NMOS [144] or GaN NMOS and Si PMOS are co-integrated thanks to 3D layer transfer [145]. In addition, rather than using conventional Si transistor as top devices, NW Cheng *et al.* reports a monolithic heterogeneous integration of BEOL Power gating transistors of carbon nanotube (CNT) networks [146]. CNT are grown on donor substrate and release in a solution and then can be directly deposited onto processed wafers. Up to five vertically-interleaved layers with three different technologies (silicon, CNTs, III-V) are integrated in [147].

As far as Silicon monocrystalline channel is concerned, Brunet *et al.* demonstrated a full 3D CMOS over CMOS on 300mm wafers [132]. Top devices feature high performance Fully-Depleted Silicon On Insulator FDSOI process requirements like High-k/metal gate and raised source and drain. The maximal thermal budget is 650°C, 2min. Low-temperature silicon epitaxy is feasible [142] and can be doped and activated at low temperature thanks to Solid Phase Epitaxy Regrowth [148], [149]. Similarly, Vandooren *et al.* also demonstrated for the 1st time 3D stacked FinFETs at 45nm fin pitch and 110nm gate pitch technology on 300mm wafers [150]. The top tier is composed of junctionless devices, fabricated under 525°C, without performance degradation. A junctionless transistor is a device featuring a uniformly doped channel and acts as a gate resistor [151]. Typical channel doping values in literature to ensure a correct operation are around 10^{19} at/cm³. With its ease of fabrication (lack of source and drain implantation and annealing), junctionless transistors are promising candidate for 3D monolithic integration [152]. Also, Vandooren *et al.* proposed a buried metal line for junctionless top-tier planar devices which acts as a back gate for dynamic V_{TH} tuning but also as a shield for RF applications [153].

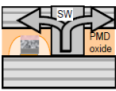
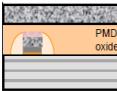

	Seed Window (SW)	Poly-Si + Recrystal.	Wafer Bonding
Description			
Density	limited	same as bottom	same as bottom
Crystal quality	Defects close to SW	Random defects	Perfect ~SOI supply
Thickness control	10 nm range	nm range	A range
Thermal budget	Seems incompatible w/ bottom max. TB	OK w/ ns laser	< 400°C

Fig. 108: Recap table of the advantages of presented techniques [137]. Even if wafer bonding is the more expensive one, its density integration and crystalline quality is an asset for 3D monolithic integration.

To conclude this part, several groups have demonstrated 3D monolithic integration using different techniques. Seed window techniques is efficient to create a good crystal quality at the expense of density (Fig. 108). Amorphous silicon deposition and recrystallization do not affect the top-layer density but suffers from a poor crystalline quality. Finally, wafer bonding offers the best crystal quality with excellent thickness control and low thermal budget. To answer the challenges brought by 3D monolithic, our choice is to focus on junctionless devices and lower the process integration down to 400°C. Before discussing the device fabrication, a non-exhaustive review of junctionless transistors fabricated without thermal budget constraint is exposed. The main idea is to give insights about junctionless performances to select the best architecture and channel material suitable for 3D monolithic integration.

b. Junctionless transistors

The purpose of this part is to discuss the different junctionless (JL) device architectures/materials with associated performances to identify the requirements in terms of device fabrication. We will first assess the different types of architecture before considering polycrystalline materials and more exotic ones.

i- Short presentation of the JL transistor (JLT) architectures

In the literature, many different junctionless architectures have been proposed and are represented in Fig. 109.

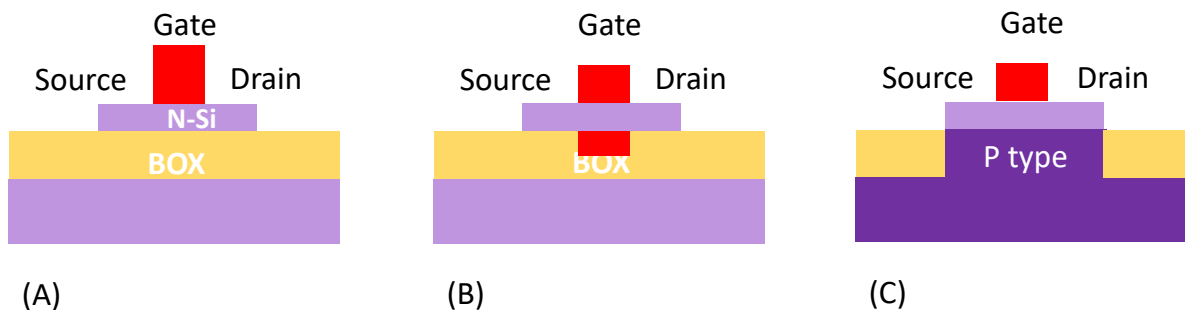


Fig. 109: Presentation of the investigated architectures. A is a planar trigate FDSOI transistors, B is a Gate-all-around (GAA) transistors and C is a bulk junctionless transistor.

(A) configuration in Fig. 109 consist of trigate (A) or planar devices (one gate) on a buried oxide (BOX). To fabricate **planar SOI** junctionless transistor, the silicon must be thin enough to be able to turn the

device OFF (see part 5-a). Barraud *et al.* fabricated JL trigates with gate length down to 13nm [154], an EOT=1.2nm and a channel thickness of 9nm. Sub-threshold Slope of SS<70mV/dec and $I_{ON}/I_{OFF} > 10^6$ are achieved at $L_G=13$ nm.

The **GAA architecture** (B) offers the best electrostatic and hence short channel effect (SCE). Both monocrystalline and polycrystalline transistors have been fabricated. Horizontal GAA NW can be staked to increase the drive current per device footprint [155], [156]. A JL stackable silicon-oxide-nitride-oxide-silicon (SONOS) memory (vertical-Si nanowire) is demonstrated in [157]. The JLTs show SS <70mV/dec and are particularly interesting for 3D stacked memory applications. Germanium is an interesting alternative to silicon to boost transistor performance for PMOS due to higher mobility. Wong *et al.* [158] demonstrated p-channel junctionless GAA germanium nanowire transistors with $I_{ON}=390\mu A/\mu m$ and $I_{ON}/I_{OFF}>10^6$ for 250nm gate length (L_G). However, n-channels germanium FET suffers from poor performance due to the presence of a high density of interface states near the conduction band edge at the germanium-insulator interface [159]. That is why n-channel germanium junctionless transistor are interesting since the conduction occurs in the volume (see part 3). N-channel germanium GAA JLT were first demonstrated by Wong *et al.* [160], yielding $I_{ON}=1235\mu A/\mu m$ ($V_G-V_T=V_{DS}=1V$), $I_{ON}/I_{OFF}=2.10^6$, SS=95mV/dec and $L_G=60$ nm. Planar transistors N and P- JLT have been fabricated on germanium-on-insulator wafer by Ren *et al.* [161] ($I_{ON}/I_{OFF}\sim 10^5$).

Bulk junctionless (C) transistors are feasible but a PN junction underneath the device is required to isolate the source, drain and channel from the substrate [162]. For instance, the 1st demonstration of junctionless accumulation-mode bulk FinFETs is composed of a hybrid channel created by ion implantation [163]. In [163], devices with a fin width of $W=16$ nm, show SS= 68mV/dec, DIBL=9mV/V and $I_{ON}/I_{OFF}> 10^6$. The device is still considered junctionless since there is no junctions in the direction of the current flow. Cheng and al., fabricated pJLT using a hybrid poly-si fin channel, performing SS=64mV/dec, $I_{ON}/I_{OFF}>10^7$, DIBL=3mV/V at $V_G=-4V$ [164], [165], [166]. Going further, Li and al. [167] demonstrated a hybrid P/N/P double nanosheet channels with $I_{ON}/I_{OFF}>10^7$, SS=176mV/dec and DIBL=13mV/V.

ii- Polycrystalline materials

As stated in the previous part, poly-Si or poly-Ge have much lower performance than single-crystal transistors but the cost of fabrication should be lower in the context of 3D monolithic integration since the channel material can be directly deposited on bottom tier. From performances side, a polycrystalline film is composed of several small crystallites separated by grain boundaries. The grain boundaries (dangling bounds at the edge of the crystallites) will trap free carriers such as electrons in n-type doped poly-Si [168], [169]. The formation of a potential barrier due to the trapped carriers reduces the overall carrier mobility. However, in junctionless transistor, the channel is doped around 10^{19} at/cm³ which is enough to saturate the dangling bounds. Su *et al.* [170] shows an apparent mobility for n-channel GAA poly-Si up to five times larger than inversion-mode GAA. The GAA architecture is also demonstrated by Liu and al. [171] (SS=105mV/dec, DIBL= 83mV/V, $I_{ON}/I_{OFF}=7\times 10^8$ at $V_G=4V$ and $V_D=1V$), or by Kuo [172] (SS~75mV/dec, $I_{ON}/I_{OFF}\sim 8\times 10^7$ at $V_G=1.5V$). Similarly, the pi-gate architecture (variation of TG architecture where the gate is extended into the buried oxide to form a π shape) provides good electrostatic control of the channel. With such an architecture, Hsieh *et al.* [173], [174] demonstrated SS= 61mV/dec (poly-Si). Planar devices are feasible if the channel thickness is low enough to ensure depletion at OFF state (see working principle in section 3). Poly-Si thin film transistors have emerged and feature excellent electrostatic control. For instance, with poly-si JLT (silicon channel thickness $t_{Si}=1.5$ nm) a 30mV/dec ss is observed [175]. The sub-60-mV/decade SS is attributed to the impact ionization effect resulting from the high lateral electric field at the drain side at OFF-state. In addition, a 2.4nm ultrathin channel trench poly-Si JLT is demonstrated by Yeh *et al.* [176] and features SS=100mV/dec, DIBL~0mV/V and $I_{ON}/I_{OFF}=10^6$ at $L_G=0.5\mu m$. Furthermore, Lin *et al.* [177] recorded

a 8nm n-type poly-Si JLT thin film with the following characteristics: $SS=240\text{mV/V}$ $I_{ON}/I_{OFF}>10^7$. Note that the JL devices shows 23 higher drive current than inversion-mode at $V_G=4\text{V}$.

Polycrystalline germanium has the potential for higher current drive and low temperature process flow. For instance the poly-Ge nanowire are formed at 550°C and processed below 300°C in [178]. Usuda *et al.* [179], obtained high mobility (200 and $140\text{ cm}^2\text{V}^{-1}\text{s}^{-1}$ for electrons and holes and $I_{ON}/I_{OFF}>10^4$) with flash annealing of polycrystalline germanium allows. Laser annealing is used to recrystallize large-grain poly-Ge in [180], showing $SS=237\text{mV/dec}$ ($V_D=1\text{V}$), $\text{DIBL}=101\text{mV/V}$ and $I_{ON}/I_{OFF}=6\times 10^4$ for $L_G=50\text{nm}$.

iii- Other materials

Since there is no requirement to form junction or to use costly doping gradients, a variety of materials can be used. For instance, III-V semiconductors can take benefits from the junctionless integration, avoiding the need of source and drain implantation and thermal activation. As an example, $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ detains a high mobility ($4000\text{cm}^2\text{V}^{-1}\text{s}^{-1}$ at $N_D=10^{19}\text{at/cm}^3$) and $I_{ON}/I_{OFF}\sim 10^6$ and $SS=88\text{mV/dec}$ is observed [181] [182]. In addition, 2D transition metal di-chalcogenides such as molybdenum disulfide (MoS_2) JLT offers better I_{ON}/I_{OFF} than the inversion-mode counterpart does [183]. To finish with, various applicative domains have demonstrated the utility of junctionless. For instance JL CNT have been fabricated and tested as sensors for cholesterol [184]. Or JLT in indium-tin oxide (ITO) and zinc oxide (ZnO) can be done on polymer and paper substrate for cheap flexible transistor manufacturing [185]. Biodegradable JLT have been made in indium-zinc-oxide (IZO) [186] featuring $SS=130\text{mV/dec}$, $I_{ON}/I_{OFF}>10^6$ at $V_G=1.5\text{V}$.

To finish, the junctionless transistor architecture is widely studied because of its ease of fabrication, allowing the integration of new materials at low temperature. The performance of poly-Si JLT are better than standard poly-Si thin film transistors, which makes poly-si JLT suitable for low-cost applications and wearable electronics. However, for 3D monolithic integration scope, high performance and density are pursued. Thus, from this point of view, monocrystalline JLT on a buried oxide (BOX) are more adapted and are already identified as a candidate for 3D monolithic. This is the reason why we adopt in this manuscript, an SOI architecture, rather nanowire than planar to create low-temperature JL transistors. However, the impact of channel doping on conventional figures of merit (I_{ON} , I_{OFF} , $SS\dots$) must be analyzed in-depth to ensure a proper transistor operation. The guideline of the next part is the Technology Computer-Aided Design (TCAD) environment presentation to simulate the behavior of junctionless devices. After, the JLT specificities compared to inversion-mode will be point out and then the TCAD sizing analysis of process parameters in order to guaranty a proper operation will be presented.

2- TCAD simulations

The motivations of the following simulations are to ensure the correct sizing prior fabrication and to point out in an educative way the specificities of JLT. “Technology Computer-Aided Design (TCAD) refers to the use of computer simulations to develop and optimize semiconductor processing technologies and devices” [187]. Such a tool is essential to silicon engineers to explore new devices concept, characterize electrical behavior of semiconductor for fast prototyping and study sensitivity to process variation [188]. Sentaurus device developed by Synopsys is used in this work to simulate electrical characteristics of both junctionless and inversion-mode devices.

a. Chosen device architectures

We highlighted in previous part the interest of a monocrystalline FDSOI architecture. For sake of simplicity, the TCAD study is carried out only on n-MOS transistors with the structure described in Fig. 110. The uniformly doped channel is a silicon rectangle of doping N_D , height t_{si} , width W and length L_G+2*L_{SD} , on top of a buried oxide (BOX, of thickness t_{box}). The gate stack (from bottom to top) is composed of an high-k oxide $\epsilon_0(HfO_2)=3.9$ defined by an equivalent oxide thickness (EOT) and a metal gate with ϕ_m workfunction. The length of the gate is noted L_G . By default, the values are given in nanometer. The “contacts” of length L_{cont} , here the electrodes where voltages will be imposed are assumed to be perfect (no additional resistance or Schottky contact). Four terminals are defined: the gate (on top of metal gate), the bulk (silicon under the BOX), source and drain (with L_{spacer} space with the gate contact). A fine meshing is used in the channel region to obtain accurate results.

From this structure, a nanowire configuration ($L_G=30nm$, $W=20nm$) and planar-like ($L_G=30nm$, $W=230nm$) are identified as references to study the impact of the various process parameters. These dimensions are in the range of the fabrication capability. To compare with standard devices (inversion-mode), an additional structure with undoped channel and doped Source/Drain SD ($10^{20} at/cm^3$) is considered. More than the direct comparison between JL and IM devices, the main goal of this TCAD study is to size the process parameter for a proper junctionless operation. Since the geometry of future devices is fixed by the mask and detains a whole panel of W and L (from $10\mu m$ down to $20nm$), we won't search an optimized L and W to fulfil an industrial target, but rather look for a channel thickness t_{si} and a channel doping N_D which turns OFF the device at $V_G=0V$. Two extreme configurations (nanowire and planar) are taken to ensure the junctionless operation for all the dimensions.

TCAD geometric parameters:

- t_{si} : the SOI base wafer consists in $16nm$ silicon on top of $145nm$ or $25nm$ BOX. Thus, the maximum silicon thickness is $16nm$. If no indication, $t_{si}=11nm$ is chosen as REF.
- N_D : typical values for junctionless transistors are around $10^{19}at/cm^3$ [162]. We will explore doping values ranging from $10^{18}at/cm^3$ to $10^{20}at/cm^3$. Without contrary indication, the reference doping is $5.10^{18} at/cm^3$.
- W and L : to account for the worst degradation, a short gate length $L_G=30nm$ is chosen as a reference. In the same manner, a $W=20nm$ and $W=230nm$ are chosen to simulate a nanowire (NW-REF) and a planar device (PL-REF) respectively. These parameters can be changed to see width or length effects. However, the more you increase W or L , the more the channel volume will increase, leading to a higher time of simulation. That is why planar-like devices are represented by $W=230nm$ and not for instance $W=10\mu m$.

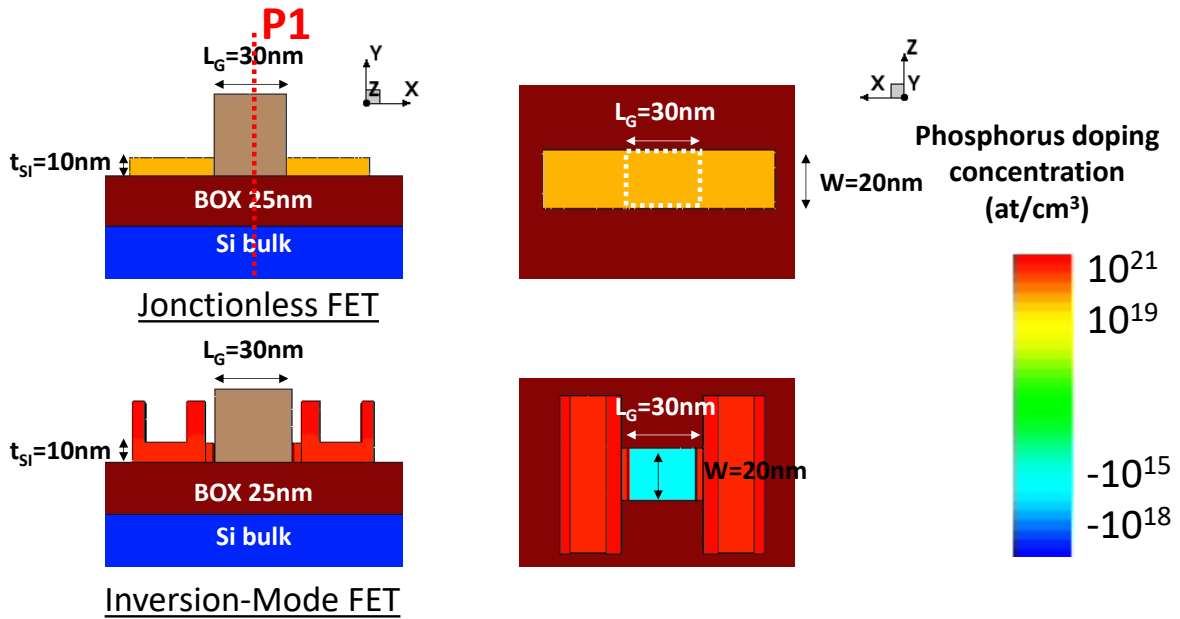


Fig. 110: Presentation of the different simulated architectures. Junctionless devices features a variable silicon thickness (here 10nm), a BOX thickness fixed at 25nm , a variable channel doping at N_D (here 10^{19} at/cm^3), a variable gate length L_G and a width W . Similarly Inversion-mode devices features a variable silicon thickness (here 10nm), a BOX thickness fixed at 25nm , a source drain doping of 10^{21} at/cm^3 , a variable gate length L_G and a width W .

Our process choices:

- Gate workfunction ϕ_m : for n and p co-integration, a midgap material is needed. A $\phi_m = 4.61\text{eV}$ for TiN (integrated at high-temperature) is chosen.
- Equivalent Oxide Thickness EOT: a 1nm EOT is chosen, in coherence with the lots integrated during this PhD.
- Source/Drain to L_G distance L_{spacer} : 12nm .

Our assumptions:

- The form of the active zone is rectangular. In fact, (see Fig. 111), the active zone shape of fabricated devices is similar to a butterfly. It can influence the electrical characteristics [189] but is not taken into account here. In fact, the goal of the simulation is not to fit perfectly the experimental results but to give some insights of junctionless operation and design guidelines.
- Contacts are considered ideal and do not take into account process choices. For instance, salicide process (thin transition metal layer over patterned transistors deposition and anneal) forms a low-resistance transition metal silicide [190].
- Uniform channel doping (such as, uniform width, length...) is considered which is not the case for small dimension devices where average values are no more representative. To tackle the variability, sensitivity studies to different process parameters will be done in part 5-.

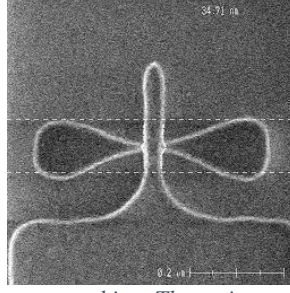


Fig. 111: SEM top view of a transistor after gate etching. The active zone is not rectangular as in TCAD simulation but rather have the shape of a butterfly. It can influence the electrical characteristics [63] but is not taken into account in the TCAD simulations.

b. Physical Model used and justification

Different physical model restricted to silicon are chosen to describe the device in an accurate way. For instance, to take into account the generation-recombination of carriers, the Shockley-Read-Hall SRH model is used. The Scharfetter relation inside this model captures the doping dependence of SRH lifetimes [191]. The standard bandgap model is used with Slotboom model for bandgap narrowing, based on measurements of p-n-p transistors with different doping concentrations [192]. Except for the mobility model, no additional specifications for the physical description have been made.

As far as the mobility model is concerned, all the contributions (temperature, impurities, surfaces...) can be decoupled and combined thanks to the Mathiessen's rule (Eq. 5). For IM devices, the mobility model used is the by-default one depending only on temperature (phonon scattering) expressed in Eq. 6.

$$\frac{1}{\mu} = \frac{1}{\mu_{b1}} + \frac{1}{\mu_{b2}} + \dots + \frac{1}{\mu_{bn}} \text{ with } \mu_{bi} \text{ a contribution} \quad \text{Eq. 5}$$

$$\mu_{const} = \mu_L \left(\frac{T}{300K} \right)^{-\zeta} \text{ with } \mu_L = 1417 \text{ cm}^2/\text{Vs} (470.5) \text{ and } \zeta = 2.5 (2.2) \text{ for electrons (holes)} \quad \text{Eq. 6}$$

Since a JL transistor detains a doped channel, the by-default mobility model is no longer accurate. In fact, the so-called constant mobility model accounts only for phonon scattering and depends just on the lattice temperature. Therefore, it is not adapted for doped semiconductors where carrier scattering by charged impurity ions degrades the mobility. That is why to describe junctionless devices, the Philips unified mobility model proposed by Klaassen is used [193]. This model takes into account the temperature dependence of the mobility, the electron-holes scattering, the screening of ionized impurities by charge carriers and clustering of impurities. In addition, the mobility degradation at interfaces due to the high transverse electric field is considered with the Enhanced Lombardi model. These surface contributions are combined with the bulk mobility according to Mathiessen's rule. Furthermore, the carrier drift velocity is not proportional to the electric field for large electric fields and saturates to a finite speed v_{sat} . High-field saturation (velocity saturation), thin layer and transverse field dependence are specified. The last model used is the thin-layer mobility model describing the degradation due to finite silicon film thickness. It accounts for phonon scattering dependency on quantization and empirical degradation terms.

To conclude, the main differences between the two devices simulated, IM and JL are the doped channel (and thus, the mobility model is adapted) and the addition of doped source and drain for IM. Two sizing flavors have been figured out (planar-like and nanowire) to compare in depth the electrical characteristics. Different parameters, t_{si} , N_D , W and L will vary in order for us to optimize the structure.

3- Junctionless MOSFET operation

This part will enter deeply into the junctionless semiconductor physics, highlighting with the help of TCAD simulations, the divergences between JL and IM in the nanowire configuration. The working principle of IM transistors have been presented in the introduction chapter. Fig. 112 presents the I_D - V_G of the NW-REF structure. From the operation point of view, the JL structure acts as a switch between an OFF state (at $V_G=0V$ gate voltage) and an ON state (arbitrarily chosen at $V_G=0.8V$) for $V_D=0.8V$ drain voltage. One can define a subthreshold slope and a threshold voltage extracted at a given current such as inversion-mode devices. However, contrary to standard devices, the derivative of the drain current with respect to the gate voltage plot (so-called g_m) in Fig. 113 shows two peaks. The dissociation between the two peaks being more pronounced for experimental data, measurements are shown instead of TCAD simulations. Each peak is associated to a threshold voltage. We can then distinguish three regions separated by two threshold voltages, V_T and V_{FB} . The operation of the different regimes will be explained in details in the next sub-sections.

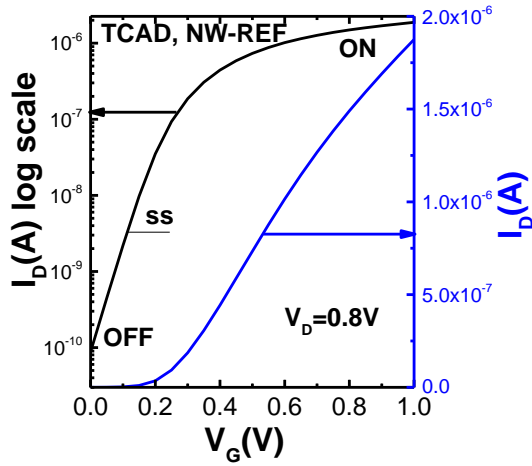


Fig. 112: I_D - V_G of a junctionless transistor at $V_D=0.8V$. ON and OFF state are indicated.

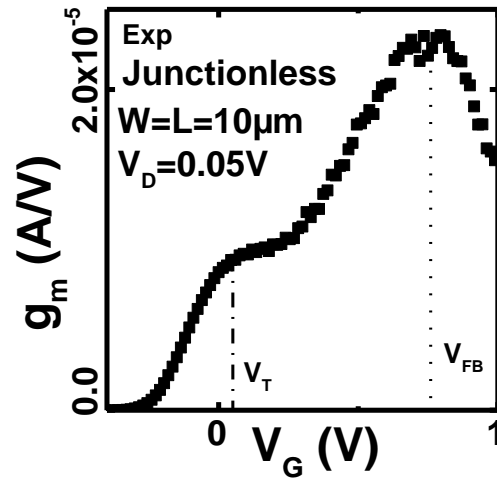


Fig. 113: Experimental data of a junctionless device transconductance g_m ($W=L=10\mu m$) as a function of gate voltage.

a. Sub-threshold region: depletion

First, the OFF state (ie. at $V_G=0V$ and $V_D=0.8V$) is studied. Since the nanowire is heavily doped, the doping density is equal to the electron density. Thus, an electron density cross-section perpendicular to the channel is sufficient to analyze the carrier's density. We plot the electron density cross-section in Fig. 114-(a) of NW-REF and electron density cut line in the middle of the channel C1 (i.e. $t_{si}/2$) in Fig. 114-(b). A smaller electron density for JL is observed on the oxide-silicon interface. From Fig. 114-(b), the electron density reaches $8.10^{15} \text{ cm}^{-3}$ at exactly the middle of the channel, value below N_D . In fact, the difference between metal gate and doped silicon work function is sufficient, if properly designed, to deplete entirely the channel at the OFF state [194]. Note that the depletion occurs from three sides in a trigate configuration contrary to the planar configuration where the full channel depletion is imposed from only one side.

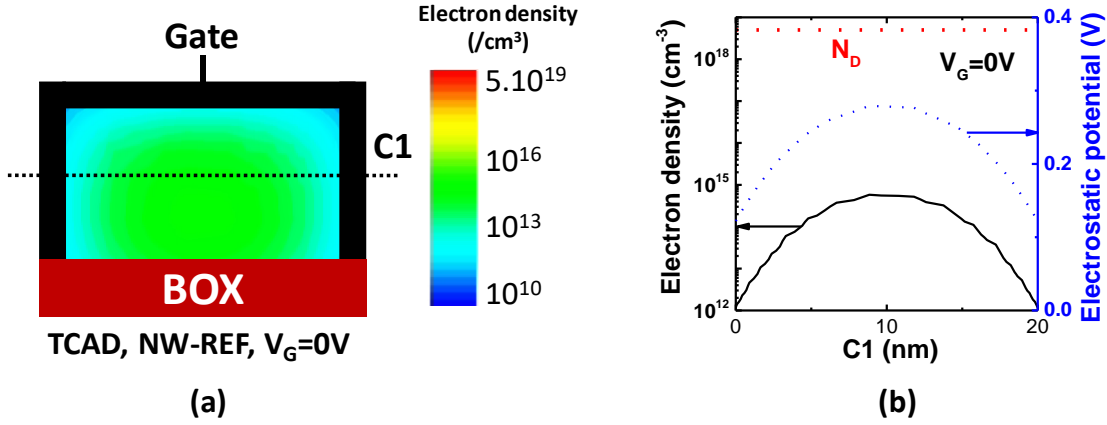


Fig. 114: TCAD, NW-REF ($L_G=30\text{nm}$, $W=20\text{nm}$, $t_{si}=11\text{nm}$, $N_D=7.10^{18}\text{at/cm}^3$) at $V_G=0\text{V}$ (a) electron density (cm^{-3}) planar cut along P1 (b) electron density (cm^{-3}) and electrostatic potential (V) along the cutline C1.

b. From threshold voltage to flatband voltage: volume conduction

Fig. 115 presents the electron density cut plane at $V_G=0.3\text{V}$. Like previously, the electron density is higher in the volume of the channel and is around 2.10^{18} (to be compared with 8.10^{15} at $V_G=0\text{V}$). In fact, as the gate voltage is increased, the depletion imposed by the gate electrode becomes weaker. Therefore, the electron density in the volume of the channel increases up to N_D . A first threshold voltage V_{TH} can be defined, corresponding to the peak electron concentration equals to N_D . Further increase of gate voltage will expand the diameter of the region at $n=N_D$. At some point, the channel will become entirely neutral (*i.e.* no longer depleted) with $n=N_D$ is the whole cross-section. This state corresponds to the flatband voltage V_{FB} . From OFF state to V_{FB} , the conduction occurs in the volume. In the case of an undoped inversion-mode FET, above V_T (in this case, $V_{FB}<V_T$) a surface inversion layer is also formed at this stage ([194] and Fig. 115 (b)). In this NW-REF TCAD simulation, a $V_T=0.4$ and $V_{FB}=0.5\text{V}$ have been extracted.

A V_T formula can be derived for planar devices, from the condition that the film is fully depleted at $V_G=V_T$ (Eq. 7). This equation is also valid for double gate JL-FET when considering $T_{si}/2$ instead of T_{si} since only half of the channel must be depleted. The demonstration is done in Annex II. We may observe the threshold voltage dependency on channel doping N_D and silicon thickness.

$$V_T = V_{FB} - \frac{q \cdot N_D \cdot t_{si}^2}{2 \cdot \epsilon_{si}} - \frac{q \cdot N_D \cdot t_{ox} \cdot t_{si}}{\epsilon_{ox}} \quad \text{Eq. 7}$$

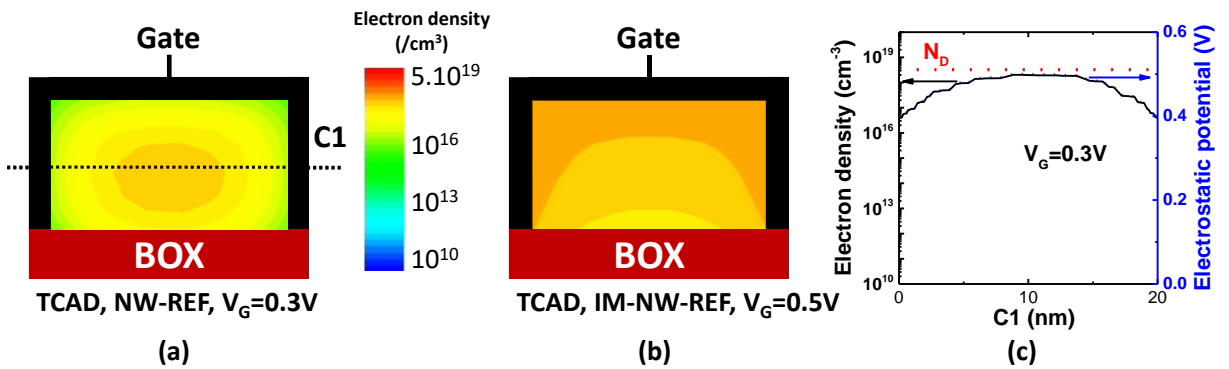


Fig. 115: TCAD, JL-NW-REF ($L_G=30\text{nm}$, $W=20\text{nm}$, $t_{si}=11\text{nm}$, $N_D=7.10^{18}\text{at/cm}^3$) at $V_G=0.3\text{V}$ (a) electron density (cm^{-3}) planar cut along P1 (b) IM at $V_G=0.5\text{V}$ electron density (cm^{-3}) planar cut along P1 (c) electron density (cm^{-3}) and electrostatic potential (V) along the cutline C1.

c. Above flatband voltage: accumulation region

For $V_G=1V$, the electron density cut plane (Fig. 116) shows that the electrons are concentrated in the edges of the tri-gate, the volume being at $n=N_D$. In fact, the increase of gate voltage ($V_G > V_{FB}$) creates accumulation channels. The drive current is higher but the benefits of a volume conduction are lost.

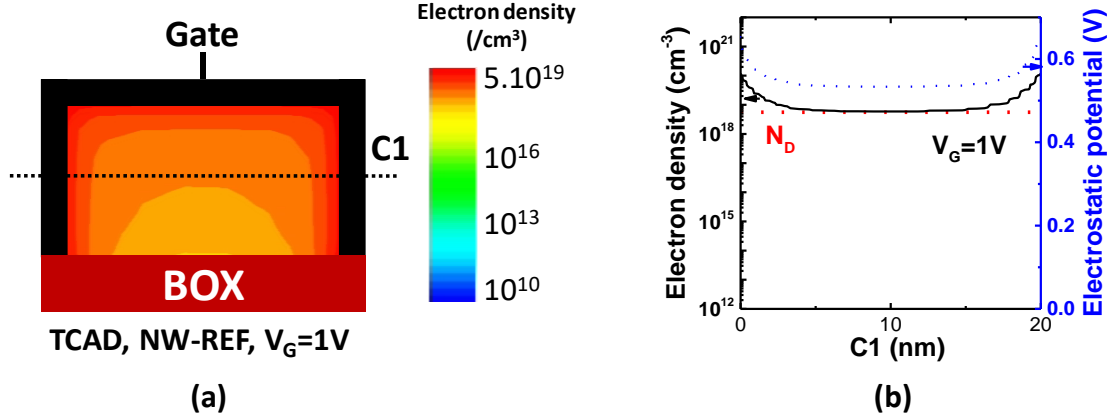


Fig. 116: TCAD, NW-REF ($L_G=30nm$, $W=20nm$, $t_{si}=11nm$, $N_D=7.10^{18}at/cm^3$) at $V_G=1V$ (a) electron density (cm^{-3}) planar cut along P1 (b) electron density (cm^{-3}) and electrostatic potential (V) along the cutline C1.

d. Analytical models

For information, to describe the operation of junctionless transistors, many analytical model (set of equations and associated parameters) have been widely developed in literature. For example, Trevisoli *et al.* proposes analytical models to capture dynamic behavior [195] and tri-gate nanowires drain current, accounting for series resistances [196]. Sallese *et al.* [197] developed a common core model for junctionless nanowires and symmetric double gate FET. Drain current in sub-threshold region is modelled in [198], [199], [200]. But also, trap modelling [201], [202] or channel thermal noise and gate-induced noise [203] are proposed in the literature.

To summarize, device operation relies on fully depleting the channel to turn OFF the device, thanks to the work function difference with the gate material ($V_G < V_T$). When the gate voltage increases, the channel depletion disappears and the current is carried out in the bulk of the channel ($V_{T-acc} > V_G > V_{T-bulk}$), a neutral channel being formed, connecting source to drain. For higher gate voltages ($V_G > V_{T-FB} > V_{T-bulk}$), the current is increased due to the formation of an accumulation layer at the interface of the gate. Note that, Jeon *et al.* [204] proposes an experimental method based on V_{FB} to separate bulk channel current and surface accumulation current.

4- Characteristics of Junctionless devices

The next paragraphs point out the specificities of JLT and a slight comparison with IMT will be done.

a. Effective channel length modulation

If we consider source-drain total current density cuts (Fig. 117), one can notice that for low gate voltages, the depletion imposed by the gate extends towards the source and drain, contrary to inversion-mode devices. In fact, the fringing field lines from the gate edge deplete part of the source and drain in OFF state, increasing the channel length, which become higher than the physical gate length. This modulation is higher in the drain region for n-channel devices since $V_D > V_S$ [205]. Furthermore, Fig. 118 presents the effective channel length (L_{EFF}) as a function of gate voltages extracted by TCAD simulations. We observe that the additional source and drain depletion decreases with gate voltage and disappears when flatband voltage is reached [206]. Trevisoli *et al.* [207] shows that in a 30nm long device the effective length is increased in the subthreshold regime by up to 60nm. This increases of L_{EFF} in OFF state results in an OFF state current reduction and a better I_{ON}/I_{OFF} ratio. As a result of, junctionless devices detains a better short channel effect control, SS, DIBL than their IM counterparts for small dimensions [208]. To boost further this modulation, high-k spacer can be used or dual-k spacer (low-k spacer and high-k spacer) [209]. Saini *et al.* [210] improves ON current by 72.5%, DIBL by 37.8%, SS by 6.5% at $V_{DD}=0.4V$ with dual-k spacer engineering (TCAD simulations).

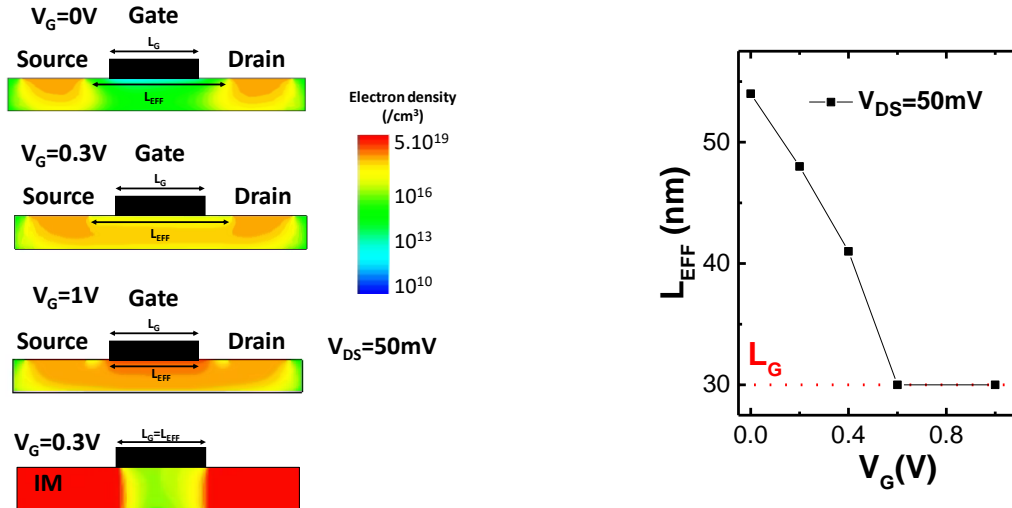


Fig. 117: source-drain density cut at $V_G=0V$ IM and JL for NW-REF ($L_G=30nm$, $W=20nm$, $t_{si}=11nm$, $N_D=7.10^{18}at/cm^3$).
 Fig. 118: L_{EFF} as a function of V_G for NW-REF ($L_G=30nm$, $W=20nm$, $t_{si}=11nm$, $N_D=7.10^{18}at/cm^3$).

In addition, we can think of JLT as a way to mitigate SCE for ultra-scaled devices. For instance, the band-to-band tunneling (BTBT) current is an exponential function of the width of the potential barrier between source and drain and so of L_{EFF} . As an example, Hur and al. [211] analyzed the gate-induced drain leakage (GIDL) between JL and IM vertically stacked nanowires. It was observed that the current is higher for IM than JL and is attributed to the different doping concentration in the extension regions. In fact, the tunneling width is larger for JL devices than IM and the electric field is lower, thanks to channel length modulation.

b. Mobility

The electron mobility ($cm^2/V.s$) captures how quickly an electron moves through the channel when pulled by an electric field (Eq. 8). If the electrons were in a perfect environment, the electric field

(ballistic transport) will increase the electron velocity. However, the same electron in a semi-conductor (crystal lattice) scatters with crystal defects, impurities, phonons... As a result, the electron can lose some energy and change its direction. It impacts the net electron motion. The mobility physical model used for TCAD simulation takes into account temperature, impurities scattering and surface interactions. In fact, Takagi *et al.* [212] proposes an universal mobility model for inversion layer with three major components: coulomb scattering, phonon scattering and surface roughness scattering. Fig. 119 presents the mobility limitations as the electric field increases.

$$\mathbf{v}_d = \mu \cdot \mathbf{E} \quad \text{with } v_d \text{ the drift velocity, } E \text{ the electric field and } \mu \text{ the mobility} \quad \text{Eq. 8}$$

For low values of electric fields, carriers mainly experience scattering due to the presence of ionized doping impurities. In fact, the Coulomb forces will deflect carriers when approaching the impurity. This phenomena is called Coulomb scattering and is proportional to $T^{3/2}/N_D$. JL devices experience more Coulomb Scattering than IM devices because of channel doping. On Fig. 120, one can observed that the mobility (both for electrons and holes) is degraded for $N_D > 10^{16}$ which is the case for JLT. In fact, JL devices mobility is mainly limited by impurities scattering [213]. However, at large gate overdrive (*i.e.* in accumulation regime), the ionized impurity charges are screened by majority carriers, leading to an higher mobility, even higher than bulk mobility [214], [215]. At high field, Doria *et al.* [216], shows that for small gate width, the effective mobility exceed the bulk mobility of 9-10%. It is attributed to Coulomb scattering reduction thanks to screening.

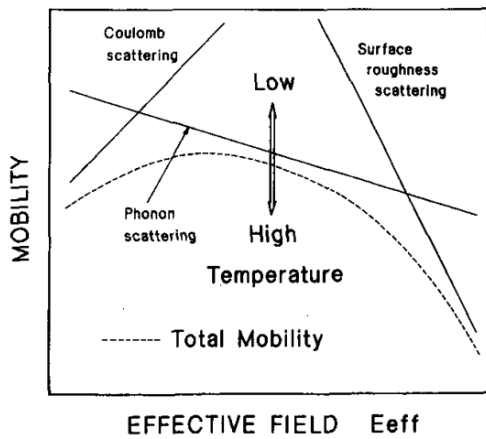


Fig. 119: schematic diagram of E_{EFF} dependence in mobility taken from [212].

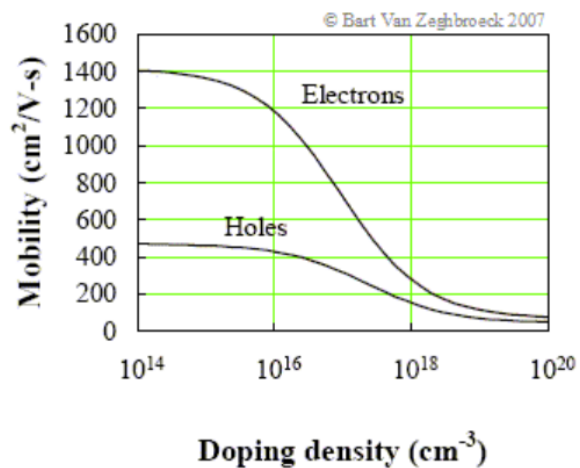


Fig. 120: electrons and holes mobility versus doping density for silicon taken from [217].

Fig. 119 shows the second limitation depending on temperature. In fact, the higher the temperature is, the higher the atom vibration (or pressure waves considered as phonon particles) in the crystal lattice is. Furthermore, a phonon can interact/collide an electron/a hole. That is why at higher temperature, more phonons are generated, reducing the mobility. This phonon-scattering mobility is proportional to $T^{-3/2}$.

As the gate voltage is increased, the carriers are more pushed closer to the silicon-oxide interface. Moreover, the interface quality compared to channel one is lower due to additional defects such as dangling bonds, interfacial roughness. This mobility limitation is called Surface Roughness Scattering (SRS). However, in a junctionless transistor, the conduction occurs in the volume below flatband voltage and its electric field perpendicular to the current flow is equal to zero [218]. That is why, JL transistors experience less SRS and show less g_m mobility degradation from the reduced transverse electric-field compared to IM [219].

To conclude this part, mobility in JL devices at $V_G < V_{FB}$ is degraded due to the channel impurities. However, at $V_G > V_{FB}$, JL mobility is less degraded by SRS and can take benefits from impurities screening. The mobility value can even be higher than bulk one. Experimentally, a technic to separate

bulk and accumulation conduction have been proposed in [220] and in [221]. It relies on the use of a front and a back gate to dissociate the two types of conduction.

c. Capacitances

In this part, we will discuss quickly the differences between IM and JL devices in terms of capacitances. We will put the emphasis on gate capacitance, miller capacitance and parasitic capacitances.

As far as gate capacitance is concerned, compared to IM devices, at low gate voltage, the conductive channel is located in the center of the physical one and the gate oxide capacitance is in series with the depletion one, decreasing the overall gate capacitance C_{gg} [222]. However, this is no longer true for large gate voltage where the conductive channel is close to the oxide interface and when there is no depletion anymore.

The Miller capacitance is the gate/drain capacitance noted C_{gd} and is important for RF applications since it impacts the cut-off frequency and the maximum operating frequency (see part 8-c.iii-). For junctionless devices, the depletion region in the channel extends inside the source and drain for low gate voltages (same as channel length modulation). For NMOS devices, this extension is higher in the drain side than the source one since $V_D > V_S$, so $C_{gd} < C_{gs}$. So like in IM underlapped devices [223], JL transistors shows lower Miller capacitances, making them suitable for RF applications.

Fig. 121 illustrates the different parasitic capacitances for a planar FDSOI transistor (Both Inversion-mode and Junctionless). Two categories of parasitic capacitances are identified. The 1st one corresponds to two parallel electrodes, like the capacity between the gate and the contact C_{pp} . The 2nd category consists in two electrodes perpendicular like C_{OF} between the gate and the source. The different parasitic components are:

- C_{ov} : overlap capacitance between the gate and source-drain extension. There is no overlap capacitance in junctionless devices for $V_G < V_T$.
- C_{OF} : outer-fringe capacitance between gate edges and source or drain though the spacer.
- C_{IF} : inner-fringe capacitance between gate edges and source or drain though the oxide and channel.
- C_{pp} : Gate-contact capacitance.
- C_{corner} (not shown here): corner capacity between the transistor and the gate extension on the BOX or STI (Shallow Trench Isolation).

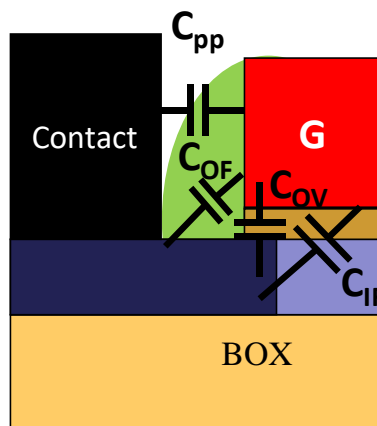


Fig. 121: Capacitance contributions

d-Variability

For industrial purposes, the performance of the manufactured product must be predictable. In fact, the customers desire a specific product with limited variability compared to the typical one. For instance, for a transistor, the threshold voltage must be identical (or almost identical) in all the circuit to work properly. As an example, chapter II presented SRAM operation where if the two inverters are not perfectly matched, a read failure can happen. To avoid discrepancies between devices, the process must be carefully monitored to ensure a functionality and a decent working window.

As an example, let us consider a gate oxide deposition tool (HfO_2 , 2nm). In a perfect world, all the versions on different production site must deposit the same oxide on all processed wafer. However, even if the tools are regularly calibrated, some differences between tools (spatial variations) and in the same tool but for two successive periods (temporal variations) are seen. It can results wafer-to-wafer variations. Furthermore, if the deposition is not perfectly uniform (gradient temperature, gas flux...), discrepancies appears in the same wafer (either inter or intra-die). This variability due to the manufacturing is called systematic variability and is expressed at the wafer level. For instance, die level variability can come from lithography steps because pattern exposure is done die by die. Also, layout variability detains a spatial correlation but is more dependent on density and patterns.

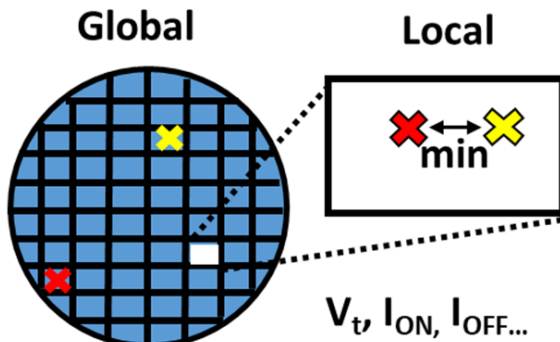


Fig. 122: schematic introducing global and local variabilities.

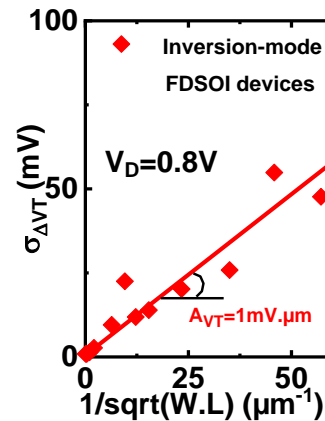


Fig. 123: Pelgrom plot example for inversion-mode devices. The standard deviation of the V_T difference is computed for matched pairs and is proportional to $1/\sqrt{W.L}$.

In addition, when we consider devices separated by the minimum distance allowed by the design rules (Fig. 122), we can notice a local variability of the electrical transistor parameters. Conversely to systematic variability, this variability is stochastic, random (no spatial correlation) and cannot be neglected for technological nodes below 65nm [224]. Usually this variability is quantified with the variation of threshold voltages ΔV_T between paired devices within a wafer. We can notice in Fig. 20 a linear dependence between the standard deviation of threshold voltage difference and $1/\sqrt{W.L}$ with L and W being the gate length and active width of the transistors. In fact for small dimensions (*i.e.* high values of $1/\sqrt{W.L}$), the devices are really sensitive to local fluctuations, for example induced by the local variability of the number of dopants in the channel or the gate roughness (Line Edge Roughness) and feature high variability. Inversely, larges devices (*i.e.* small values of $1/\sqrt{W.L}$) can average all the small-range fluctuations leading to a low mismatch between paired devices. For infinite transistor surface (*i.e.* $1/\sqrt{W.L}=0$), no stochastic variability is seen and thus $\sigma_{\Delta V_T}$ tends to 0. From this plot, called a Pelgrom plot, an A_{V_T} parameter (expressed in $\text{mV} \cdot \mu\text{m}$) is extracted from the slope (Eq. 10 and [225]). This technology dependent parameter ($A_{V_T}=0.95\text{mV} \cdot \mu\text{m}$ for FDSOI has been reported in [226]) takes into account several sources of variations which will be detailed in next paragraph.

The same figure of merit can be extracted for global variability to take into account the die-to-die variations. In this case, instead of considering two paired devices, the standard deviation is done on the whole wafer. Similarly, an A_{VT} can be extracted from the Pelgrom plot. Note that the curve does not necessarily cross the origin. Also, in order to compare the local and global variability, we can consider that $\sigma_{\Delta VT} = \sigma_{VT} \cdot \sqrt{2}$, when there is no correlation between the 2 matching transistor variabilities.

$$\sigma_{\Delta VT} \propto EOT \cdot \frac{\sqrt[4]{N_D}}{\sqrt{W \cdot L}} \quad \text{Eq. 9}$$

$$A_{\Delta VT} = \sigma_{\Delta VT} \cdot \sqrt{W \cdot L} \quad \text{Eq. 10}$$

$$V_{T-JL}(N_D) = V_{FB}(N_D) - q \cdot N_D \cdot t_{si} \cdot 1/C_{ox} \quad \text{Eq. 11}$$

$$V_{T-IM} = V_{FB} + 2\phi_f + \sqrt{2q \cdot n_i \cdot \epsilon_{si} \cdot \phi_f} / C_{ox} \quad \text{Eq. 12}$$

With V_{FB} being flatband voltage, N_D junctionless doping level, C_{ox} oxide capacity, q the elementary charge, ϵ_{si} the permittivity of silicon, $2 \cdot \phi_f$ the surface potential.

To compare JL and IM devices, let's bear in mind the V_T formulas (Eq. 11 and Eq. 12) for junctionless and inversion-mode transistors. Unlike IM transistors, JL devices V_T are dependent on the work function and the depletion charge (*i.e.* $q \cdot N_D \cdot t_{si}$). The threshold variability comes from different sources (see Fig. 124) such as:

- Random dopant fluctuations (RDF): the discrete atoms placement in a channel follows a Poisson distribution law. As the channel scales down, the number of dopants is lower, increasing the relative variation and having a severe impact on V_T (Eq. 9). For instance, a 10^{19} at/cm^3 doping in a $30\text{nm} \cdot 20\text{nm} \cdot 11\text{nm}$ volume (NW-REF) results in 66 dopants. To reduce this variability, N_D could be lowered down to intrinsic silicon values (ideal case). In fact in inversion-mode device FDSOI, the channel is left “undoped”, which means at $N_D = \text{few } 10^{15} \text{ at/cm}^3$, explaining such a low variability value measured on these devices. In junctionless devices, the channel doping is higher. However, in accumulation regime, the screening of the doping impurities (in junctionless devices) reduces RDF variability [227].
- Line-Edge Roughness (LER) and Line-Width Roughness (LWR): for small dimensions, the gate edges cannot be considered straight (*i.e.* equals to a nominal value L) but are rather rough (ΔL deviation from nominal value L). It means that along the width the gate length varies between $L - \Delta L$ and $L + \Delta L$. This gate length variation impacts SCE, SS, $V_T \dots$ Fig. 125 presents the V_T sensitivity for different process parameters for JL and IM devices. One can note that the L sensitivity is slightly lower for JL than IM. In fact, JL threshold voltage relies on channel depletion which is, at first order independent of L . Furthermore, JL are less prone to channel length variation since the electrical channel length is higher than the physical one (channel length modulation around the OFF state).
- Width variation: JL devices threshold voltage is highly sensitive to ΔW (Fig. 125) whereas IM are not. In fact, the channel width contributes actively to depletion in tri-gate configuration and detains a high impact on V_T . That is why monitoring the width uniformity is critical in junctionless devices especially for nano-scaled devices.
- Work function variations (WFV): the work function depends on crystal orientation of the metal gate [228] and induces a V_T variation for nano-scaled devices [229]. Several studies show the importance of WFV for JLT devices [226]. In fact in bulk technologies, with RDF, WFV are the main limitations for variability [230].
- Silicon uniformity t_{si} (and BOX thickness uniformity): FDSOI technology consists in a thin film channel on top of a buried oxide. The SOI wafers manufactured by SOITEC with SMART CUT™ process have a wafer uniformity of $\pm 5 \text{ \AA}$ [231]. This silicon thickness control is crucial, especially for JL devices where the whole channel (*i.e.* t_{si}) must be depleted for OFF state. For instance, TCAD simulations (Fig. 125) shows that a 1nm ($\pm 5 \text{ \AA}$) variation of t_{si} (nominal value

11nm) implies a 20mV variation on V_T . For IM devices, the induced V_T variation is lower and equals to 5mV. The silicon thickness variability can be local or global, impacting the local or global V_T variability. BOX variation can also impact the V_T variability with back-biasing.

- Gate oxide variability: the gate dielectric thickness variation affects V_T by changing locally the EOT [232], [233]. Localized charges (or dipoles for high-k) also play a role.

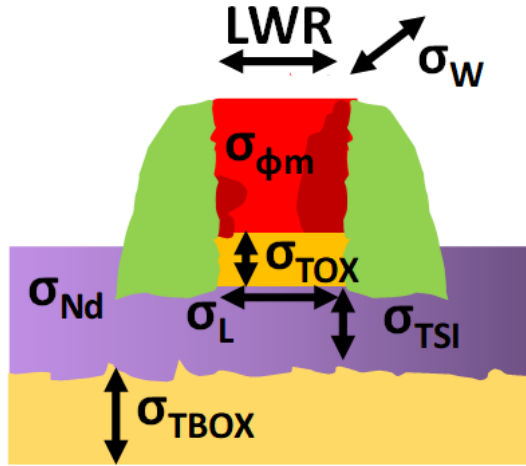


Fig. 124: schematics of variability sources in junctionless transistors.

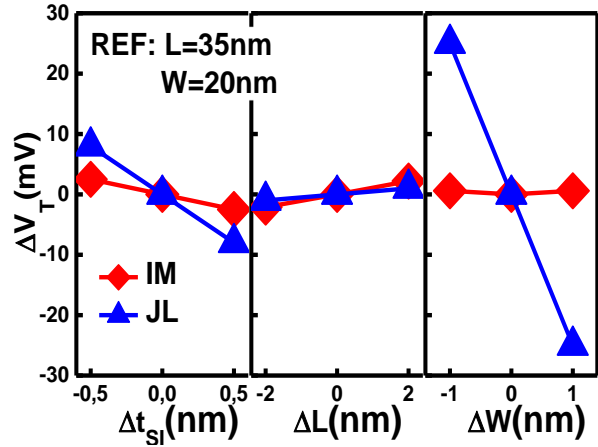


Fig. 125: Comparison (TCAD) between the V_T sensibility of junctionless (JL in blue) and inversion-mode (IM in red) devices ($L=35\text{nm}$, $W=20\text{nm}$) to t_{si} , L and W . A higher sensitivity on W is seen for junctionless devices compared to inversion-mode one.

To conclude this part, variability is a major issue for manufacturing and process are carefully monitored to reduce it. Variation-aware design or different architecture (from bulk to FDSOI) can be used to mitigate variability. As far as junctionless devices are concerned, another variability component is added, compared with IM, namely the random dopant fluctuation, due to a heavily doped channel. In fact, the mismatch between two adjacent devices is enhanced by the fluctuations of the channel dopants. That is why, the extracted A_{VT} parameter is higher for JL devices than IM and correlated to channel doping. For instance, Vandooren *et al.* [234] reported A_{VT} values of $3.1\text{mV}\cdot\mu\text{m}$ for $N_D=9\cdot 10^{18}\text{at}/\text{cm}^3$ contrary to IM-FDSOI typical A_{VT} around $1\text{mV}\cdot\text{V}$ [226]. Variability of junctionless devices will be assessed in detail in chapter IV.

d. Reliability

Reliability and in particular Negative Bias Temperature Instability (NBTI) and Positive Bias Temperature Instability (PBTI) have been assessed in chapter II, part 5-a and annex I. The takeaway is that the threshold voltage shifts in time due to transistor operation; this is caused by the injection of carriers from the channel into the gate oxide. Since for JL devices, the electric field peak occurs in the drain (and not in the channel region as for IM) and is lower, less degradation (HCI) is seen [235], [236]. Furthermore, as far as NBTI is concerned, the conducting channel in JL is far from the interface (at least for $V_G < V_{FB}$), limiting the interactions between the carriers and the interface traps. Toledano-Luque *et al.* [237] demonstrated that JL-pFETs have superior NBTI reliability than IM and pass the 10-year lifetime test up to $V_G=1.2\text{V}$.

e. Noise

Measuring a device noise gives several information about its quality and will be discussed later. Drain current noise measurements consists of a time fluctuation around the mean value (Fig. 126). Prior the analysis, a Fast Fourier Transform (FFT) is done to transform time-domains into frequencies one. The idea is that every time domain signal can be represented by the sum of harmonic oscillations, with associated coefficient in frequency domain. The resulting spectrum in frequency domain is called power spectral density. Fig. 127 shows the drain current power spectral density S_{Id} as a function of frequency. Next paragraph will detail shortly what information can be extracted from such a spectrum. The differences between IM and JL will be highlighted.

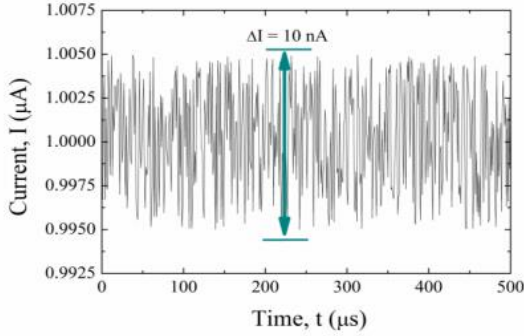


Fig. 126: Example of current fluctuation taken from [238].

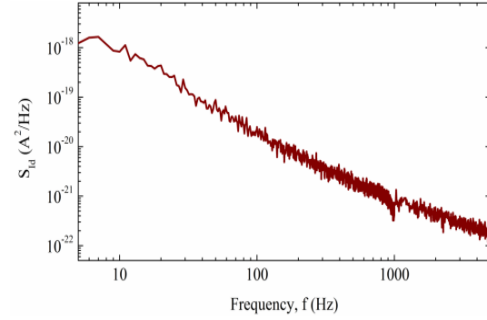


Fig. 127: Drain current power spectral density S_{Id} plotted versus frequency. taken from [238].

- Thermal noise: this noise is caused by the temperature dependent electron motion, resulting in continuous and random fluctuations even without current [239]. The associated spectral density is flat across frequency. The noise induced by these uncorrelated fluctuations is called a “white” noise. However, semi-conductors low frequency noise is dominated by others source of noise [240].
- Generation/recombination noise: this noise is caused by the trapping/detrapping of carriers for a specific trap level. Fig. 128 shows the characteristics of the obtained Lorentzian spectrum, which can be expressed as Eq. 13, [241]. To go further, the generation/recombination induced spectral density caused by N carrier number fluctuations due to their interaction with N_T traps (fills and empty) results in Eq. 14. This spectrum can give insights on trap location and energy level.

$$S_{Id} = \frac{A}{(1 + \frac{f}{f_c})^2} (A^2/Hz) \quad \text{Eq. 13}$$

$$S_N = N_T \frac{\tau}{1 + (2\pi f \tau)^2} \text{ with } N_T = 4\overline{\Delta N^2} \text{ and } \tau = \frac{1}{\frac{1}{\tau_c} + \frac{1}{\tau_e}} \quad \text{Eq. 14}$$

with S_x being x power spectral density, f_c the cutoff frequency and τ_c, τ_e are the capture and emission time of the trap.

- Random Telegraph Signal Noise (RTS): it the particular case where only one trap can be occupied. By analysis of the time domain (see Fig. 129), the emission and capture time and trap position can be figured out [242]. However, this particular noise can occur only in small devices ($<1\mu\text{m}^2$ surface) since a single trap is concerned.

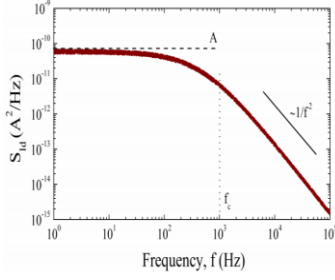


Fig. 128: Drain current power spectral density S_{id} plotted versus frequency. Lorentzian-like spectrum. Taken from [238].

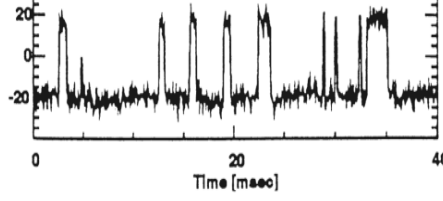


Fig. 129: RTS noise or pop-corn noise in time domain. Reproduction from [238].

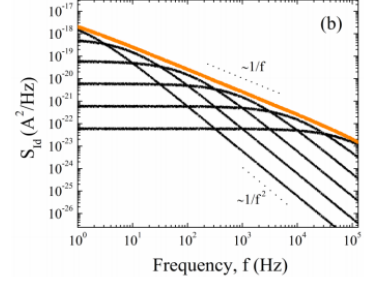


Fig. 130: Six Lorentzian are represented with $1/f^2$ slopes. The resulting spectrum is in orange and have a slope of $1/f$. From [238].

- Flicker Noise: each trap or group of traps with the same time constant results in a Lorentzian. Fig. 130 presents the spectrum resulting of the addition of 6 Lorentzian spectra. It must be noticed that the spectrum is around $1/f^\gamma$ with γ around 1. If $\gamma=1$, the density of traps is uniform in oxide depth and energy. If $\gamma>1$ ($\gamma<1$) the density is increasing (decreasing) deeper in the oxide [243].

As far as JL devices are concerned, LFN is slightly lower than for IM devices [244]–[247]. In fact, by modulating the conductive channel position thanks to back-biasing, Doria and al. showed that the LFN increases when the conductive channel is moved to the semiconductor/oxide interface [248]. For similar reasons, the drain current spectral density S_{id} increases when the accumulation layer is formed (for $V_G > V_{FB}$).

In conclusion, Noise measurements can give some information about trap density, position and energy. JL devices, thanks to their volume conduction, detains a lower low-frequency noise than IM for $V_G < V_{FB}$.

f. Junctionless transistor applications

The next part will consist in sizing the junctionless transistor thanks to TCAD simulation. However, before going further, we have to define some criteria to be able to choose between two different sizings. First of all, we have to define the targeted applications. As seen previously, a strength of junctionless transistor is about lower Miller capacitance which is an asset for RF applications. Junctionless transistors for RF applications are usually associated to ultra-low power analog applications ([249], [250]). For such an application, the main component is about maximum frequency, cut-off frequency, analog gain, noise and variability. Furthermore, low power applications minimizes the power consumption and one metric could be I_{OFF} to lower static power consumption. The ON current are not necessary to be maximized in this case, since analog transistor have usually relaxed width to drive a large amount of current. It is not the case for digital applications where density is an issue and transistor width is limited. In fact for purely digital applications, the typical figures of merits are ON-OFF current, SS, DIBL, V_T to have insights about electrostatic control. In our case, we would like to target digital/analog mixed applications to propose a versatile device featuring a good maximum operating frequency for low power applications while being effective for digital ones. For this, we have at our disposal the following figure of merits:

- OFF current (I_{OFF}): I_D for $V_G=0V$ and $V_D=V_{DD}$ (saturated region or 50mV for linear one).
- ON current (I_{ON}): I_D for $V_G=V_{DD}$ and $V_D=V_{DD}$ (saturated region or 50mV for linear one).
- V_T : V_G for $I_D=W_{tot}/L \cdot 10^{-7}A$.
- Subthreshold slope SS: I_D - V_G slope extracted between $I_D=W_{tot}/L \cdot 10^{-7}A$ and $I_D=W_{tot}/L \cdot 10^{-8}A$ for both saturated condition ($V_D=V_{DD}$) SS_{SAT} and linear region SS_{LIN} ($V_D=50mV$).
- $DIBL=(V_{T-LIN}-V_{T-SAT})/(V_{D-LIN}-V_{D-SAT})$

As far as ON current is concerned, it is not a critical point for low power analog applications, so if it might be used to compare JL and IM devices, but an analysis of OFF current is preferred. In fact, OFF current combined with SS indicates the electrostatic control and the ability to close or not the channel. The idea is to screen quickly the appropriate values of silicon thickness and doping level to achieve a low OFF current and a good electrostatic control. The targeted I_{OFF} value is chosen as $\log(I_{OFF}/W)=-8$. After this first screening, an in-depth analysis of JL with the pre-selected condition will be done.

To conclude this presentation of JL characteristics parts, the advantages of JL transistors are summarized in the table below. The targeted application is mixed digital/analog applications. For this, simulations will be performed to size the future devices and especially the OFF current. The takeaways of this part (for JL devices) are:

- Channel length modulation ($L_{EFF} > L_{G-physical}$ at $V_G < V_{FB}$) can improve the short channel effects.
- Mobility is degraded by Coulomb scattering, when compared to IM devices but increases under high electric field thanks to impurities screening.
- Variability is a major issue, mainly due to random dopants fluctuations.
- Gate capacitance is lower because of volume conduction in the appropriate operation regime.
- The Miller capacitance is lower for JL than for IM or Accumulation Mode devices AM due to the depletion in source and drain regions for $V_G < V_{FB}$.
- Junctionless devices are less prone to low-frequency noise and reliability issues since the conduction occurs in the volume far from the interface for $V_G < V_{FB}$.

	Junctionless vs. Inversion mode
Heterogeneous/3D monolithic	+
Mobility	-
Channel length modulation	+
Matching	-
Source and drain resistance	-
Drive current	-
I_{ON}/I_{OFF}
Miller capacitance	+

5- Device sizing

The previous part presented the particular operation of JL transistors and its strengths and weaknesses. This section tackles the proper sizing to ensure a correct operation, *i.e.* a full depletion at OFF state and a maximum drive current I_{ON} at ON state. As said in part 2-a, the process degrees of freedom concerns mainly the thickness (t_{si}) and the doping concentration (N_D) of the silicon layer. For this, we will discuss first the impact of channel doping in NW-REF and PL-REF, then introduce a stacked architecture before analyzing the performance with respect to inversion-mode devices.

a. Tri-gate junctionless sensitivity to silicon thickness and doping level

TCAD simulations have been carried out with various t_{si} or N_D for $W=20\text{nm}$, $L=30\text{nm}$, $EOT=1\text{nm}$ and $V_D=50\text{mV}$ (REF). The resulted I_D - V_G are presented in Fig. 131 and Fig. 132. One can observe that the thinner the channel is, the better is the I_{OFF} (defined as $I_D(V_G=0V)$) and the lower is the I_{ON} (defined as $I_D(V_G=0.8V)$). Conversely, the thicker the channel is, the higher is the I_{OFF} and the better is the I_{ON} . However, for t_{si} values larger than 15nm , the I_{ON}/I_{OFF} ratio is below 10^4 . This ratio, representative of the dissociation between an ON state and an OFF one has to be maximized. In a similar way (Fig. 132) indicates that the more the channel is doped, the more it will deliver current at ON state but the more it will let current flow at OFF state. In fact, to ensure a good operation, $V_G=0\text{V}$ (OFF state) must be enough to deplete entirely the channel, thanks to the work function difference (see section 3). So, the thinner the channel, easier will be the depletion. In the same state of mind, less dopant will be easier to deplete. From this two graphs, the trade-off between leakage current (OFF state) and drive current (ON state) must be kept in mind. The idea now is to size the transistor layer and doping level to target digital applications.

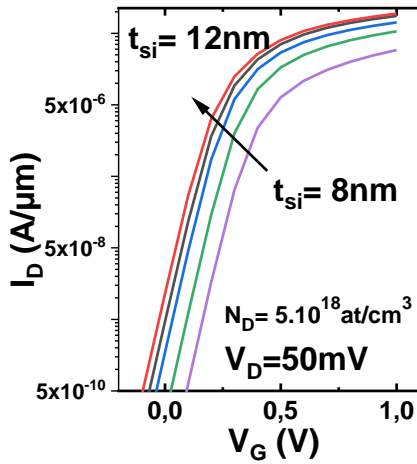


Fig. 131: I_D - V_G for various t_{si} with N_D fixed at 5.10^{18}at/cm^3 $L_G=30\text{nm}$ and width $W=20\text{nm}$.

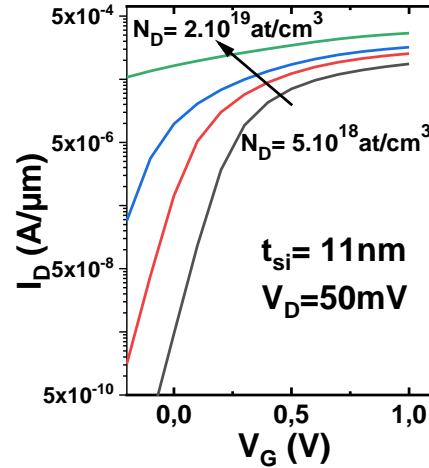


Fig. 132: I_D - V_G for various N_D (5.10^{18} , $7.5.10^{18}$, 1.10^{19} , 2.10^{19}at/cm^3) with t_{si} fixed at 11nm $L_G=30\text{nm}$ and width $W=20\text{nm}$.

For this, we will consider an I_{OFF} mapping with t_{si} ranking from 4nm to 12nm and N_D from 5.10^{18}at/cm^3 to 1.10^{20}at/cm^3 for the planar-like PL-REF. I_{OFF} is considered rather than I_{ON} or I_{ON}/I_{OFF} ratio, to ensure that the drive current is low enough at OFF state to minimize leakage. The upper limit, for analog or digital applications, is fixed at $\log(I_{OFF}(\text{A}/\mu\text{m}))=-8$, materialized by a red line. Also, the lower limit for channel doping is fixed at 5.10^{18}at/cm^3 to make sure that the contact is ohmic and not Schottky. First (Fig. 133), notice that for IM devices, this 8nm t_{si} variation results in less than one decade variation on I_{OFF} . Secondly, for JL devices (Fig. 134), such a variation induces a high range of OFF current. For instance at $N_D > 10^{19}\text{at/cm}^3$, $\log(I_{OFF}/W)$ equals to -11.3 for $t_{si}=4\text{nm}$ and -6 for $t_{si}=12\text{nm}$. This last value

means that the channel is not OFF according to our criteria. Even more, at high channel doping $N_D=10^{20}$ at/cm³, the lowest $t_{si}=4$ nm is not enough to turn OFF the transistor ($\log(I_{OFF})=-5$) and deplete entirely the channel. That is why, using the criteria $\log(I_{OFF}/W) < -8$ we determined a range of couple (t_{si} , N_D) values acceptable. Nevertheless, from process point of view, to avoid raised source and drain formation, which thermally is costly, a consequent silicon thickness is needed for silicide before contacting the S/D. A study presented in part 6-g demonstrated that during the silicide step, the NiPt will react at least on a 5.7nm depth. A full silicide contact is unwanted because of Kirkendall voids [251], so the silicon thickness must be at least 8nm and should be maximized. That is why, the condition of $N_D=7.10^{18}$ at/cm³ and $t_{si}=11$ nm is considered. Nevertheless, the studied structure was for $W=20$ nm and $L=30$ nm which is aggressive and after fabrication will correspond to a small amount of devices. The larger structures $L=W=10\mu$ m might have a different electrostatic, since they configurations is no longer tri-gate but rather planar. To make sure that the pre-selected thicknesses and doping level conditions for $W=20$ nm and $L=30$ nm are applicable for wide devices, similar simulations are done at larger width $W=240$ nm. Enlarging the device will degrade the electrostatic control of the gate from lateral sides. That is why, only the couple (N_D , t_{si}) satisfying $\log(I_{OFF}/W) < -8$ for $W=20$ nm and $L=30$ nm are simulated for $W=240$ nm and $L=30$ nm. The result is depicted in Fig. 135 and we can observe than the margins are dramatically reduced from $W=20$ nm to $W=240$ nm. To satisfy our I_{OFF} condition, the silicon thickness must be kept below 8.5nm.

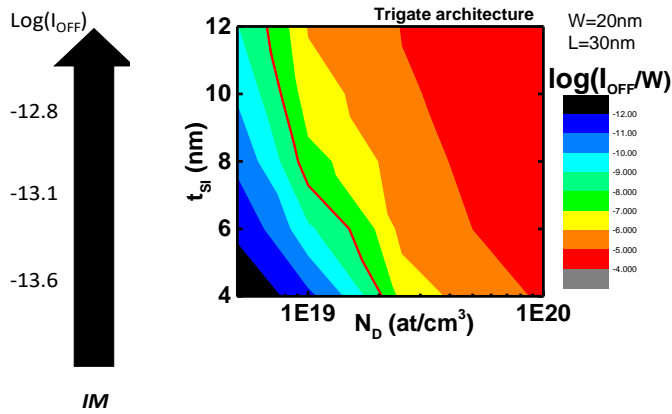


Fig. 133: $\log(I_{OFF})$ modulation for IM devices for t_{si} ranking from 12 to 4nm.

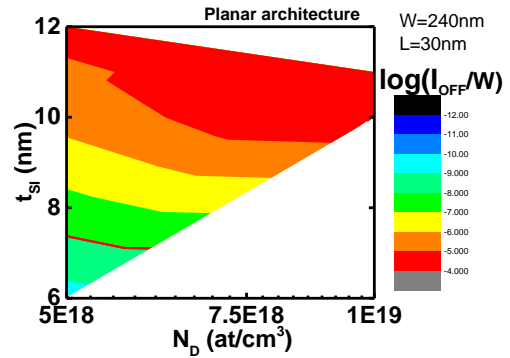


Fig. 135: $\log(I_{OFF}/W)$ as a function of selected N_D and t_{si} for $W=240$ nm, $L=30$ nm, $V_D=50$ mV in a tri-gate configuration. The condition $\log(I_{OFF}/W)=-8$ is materialized by a red line.

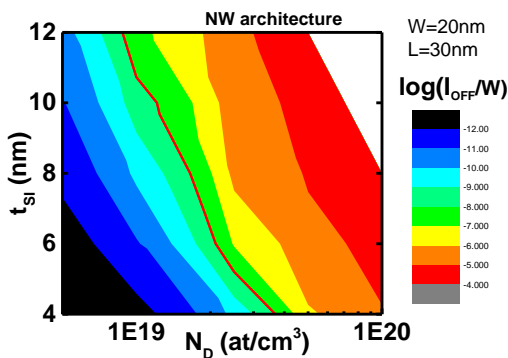


Fig. 134: $\log(I_{OFF}/W)$ as a function of N_D and t_{si} for $W=20$ nm, $L=30$ nm, $V_D=50$ mV in a tri-gate configuration. The condition $\log(I_{OFF}/W)=-8$ is materialized by a red line.

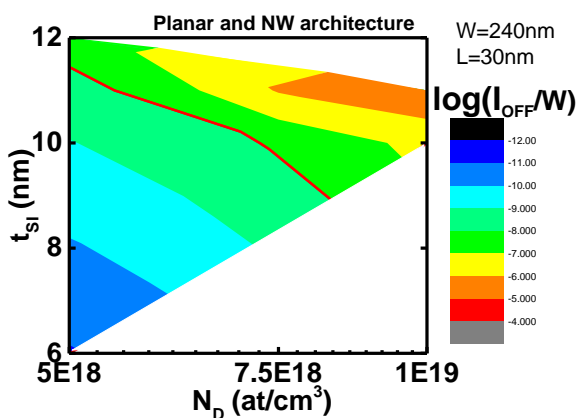


Fig. 136: $\log(I_{OFF}/W)$ as a function of N_D and t_{si} for $W=20nm$, $L=30nm$, $V_D=50mV$ in a nanowire configuration. The condition $\log(I_{OFF}/W)=-8$ is materialized by a red line.

Fig. 137: $\log(I_{OFF}/W)$ as a function of selected N_D and t_{si} for $W=240nm$, $L=30nm$, $V_D=50mV$ in a nanowire configuration. The condition $\log(I_{OFF}/W)=-8$ is materialized by a red line.

However, in this planar-like configuration, the depletion comes only from one side of the device, Fig. 138 summarizes the control gain from one gate configuration to four gate configuration. And this gain is observed in terms of I_{OFF} (Fig. 136), where a lower value is obtained for gate-all-around than trigate than planar for the same gate length $L_G=30nm$ and $W=20nm$. Similar gain is seen for $W=240nm$ (Fig. 137). Changing architecture can diminish the constraints on t_{si} and N_D . So the best architecture for junctionless devices as far as electrostatic control is concerned is GAA. However, GAA integration scheme to form nanowires is more complex than planar architecture since the future gate material must be placed underneath the channel and wrap the silicon channel. That is why, to be as close as the existing planar FDSOI baseline, we rather propose to integrate a p-doped SiGe layer below the n-channel to take benefits from a back-depletion. In this case, depicted in Fig. 139, there is a PN junction perpendicular to current flow. The additional p-layer will deplete part of the n-channel. Thus, the effective channel thickness of the device is lowered (easier to deplete) and the electrostatic control is increased. For PMOS, the opposite structure p-layer over n-layer is proposed for the same purposes. This additional layer can be done either by epitaxy or by implantation. The interest of this architecture will be presented in next paragraph after a reminder of PN junction physics.

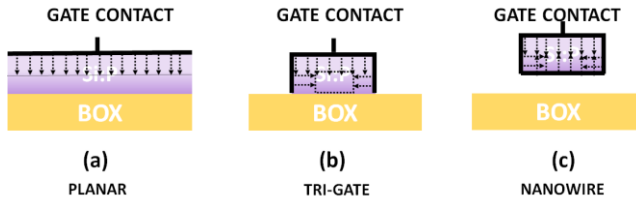


Fig. 138: schematics explaining depletion.

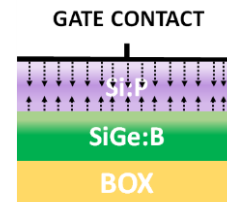


Fig. 139: stack n over p or vice versa.

b. n over p channel

The main idea is to relax the constraint on (t_{si}, N_D) by inserting below the channel, a layer of the opposite doping type to create an additional depletion. First, a brief reminder of PN junction physics is done. Secondly a sizing study of the device is done to target low power applications. Lastly, a CMOS integration is proposed, highlighting the challenges of this structure.

i- PN junction physics

The schematic of a PN junction is presented in Fig. 140. It consist of a material of a N_D donor negative doping concentration in contact with a material of a N_A acceptor positive-doping concentration. Under thermal equilibrium, *i.e.* without external bias applied, the free electrons in the n-type material (majority carriers) are attracted to the positive holes in p-type. The free electrons will diffuse in the p-type material and combine with the holes, forming a negative charge region. In a similar way, the diffusion of holes from the p-type (majority carriers) into the n-type material forms a positive charge region. The charge due to the ionized donors and acceptors causes an electric field, which in turn causes a drift of carriers in the opposite direction [252]. The diffusion of carriers continues until the drift current balances the diffusion current, reaching thermal equilibrium as indicated by a constant Fermi energy (see Fig. 141). As a result, majority charge carriers are depleted in the region around the junction interface, so this region is called the depletion region or space charge region. A potential barrier qV_{bi} forms across the space charge region. V_{bi} is called the built-in potential and is the consequence for holes or electron of

the balance between drift and diffusion (Eq. 15 and for holes, Eq. 16). In fact, the electric field ε is the opposite of the derivative of the potential V with respect to x (Eq. 17). And by integrating between two points (such one and two in Fig. 140) far from the interface, the built in voltage (potential difference between n and p region) is obtained (Eq. 19). In a similar way, by integrating the Poisson equation, the depletion width can be computed and separated into x_n accounting for depletion width in n-type material and x_p (Eq. 20).

$$|I_{diff}| = |I_{drift}| \quad \text{Eq. 15}$$

$$qAD_p \frac{dp}{dx} = qA\mu_p p \varepsilon \quad \text{Eq. 16}$$

$$\varepsilon = -\frac{dV}{dx} = \frac{D_p}{\mu_p} \cdot \frac{1}{p} \cdot \frac{dp}{dx} \quad \text{Eq. 17}$$

$$V_{bi} = V_2 - V_1 = \frac{D_p}{\mu_p} \cdot \ln \frac{p_1}{p_2} \quad \text{Eq. 18}$$

$$V_{bi} = V_T \cdot \ln \frac{N_D \cdot N_A}{n_i^2} \quad \text{Eq. 19}$$

$$x_n = \sqrt{\frac{2\varepsilon_s}{q} \cdot \frac{N_A}{N_D} \cdot \frac{1}{N_A + N_D} V_{bi}} \quad \text{and} \quad x_p = \sqrt{\frac{2\varepsilon_s}{q} \cdot \frac{N_D}{N_A} \cdot \frac{1}{N_A + N_D} V_{bi}} \quad \text{Eq. 20}$$

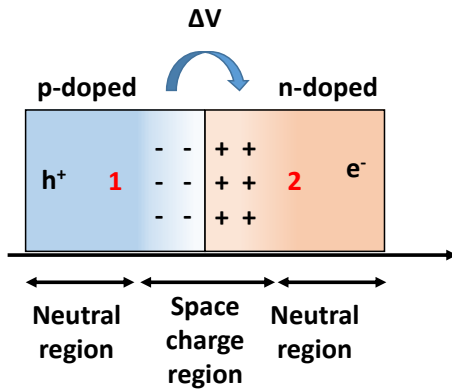


Fig. 140: PN junction schematics.

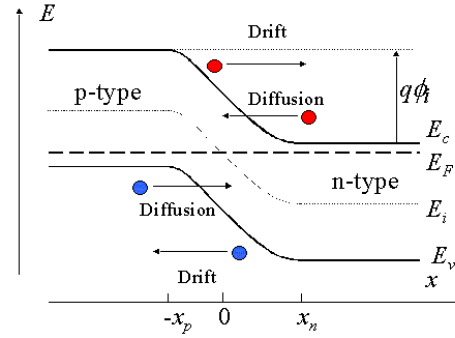


Fig. 141: band diagram PN junction. Reproduction from [217].

Fig. 142 shows the x_n depletion depth into the n-type material for various N_A and N_D doping. We can observe that the depletion region extends more in the less doped side of the junction. To enhance this depletion for our application, the underneath layer must be highly doped. For instance, a $N_D = 2 \cdot 10^{19}$ at/cm³ and a $N_A = 5 \cdot 10^{18}$ at/cm³ electron density cut is presented in Fig. 143 for $W=20$ nm and $L=30$ nm. So for $V_G=0$ V a depletion comes from the p-layer and complete the depletion imposed by the gate relaxing the constraint on t_{Si} and N_D . That is why, we propose to insert a layer beneath each device of the opposite polarity. The nMOS devices with a p layer (n over p stacking) and pMOS devices (p over n stacking) have to be co-integrated in the same wafer. The next part will present an integration scheme with their constraints to set boundaries for device sizing.

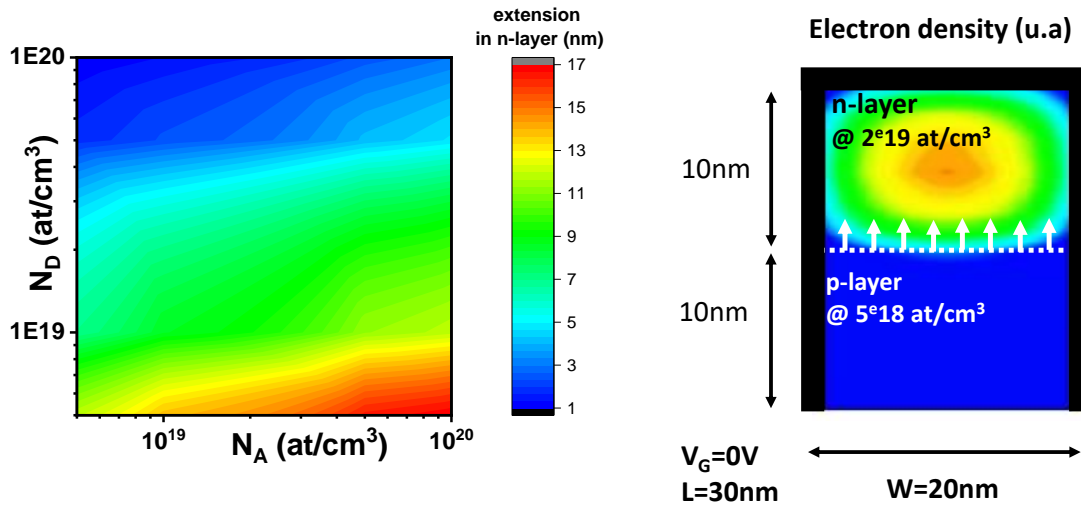


Fig. 142: extension x_n in nm into the n-type material for various N_D and N_A doping. Computed from Eq. 19 and Eq. 20.

Fig. 143: TCAD SD cut along P1 showing the electron density at $V_G=0V$. The depletion coming from the p-type layer is highlighted.

ii- CMOS Integration

In this part, we will highlight the stakes of replacing a Si channel by a bi or tri-layer one and some process solutions.

First, to create the channel, one can think of two techniques: ion implantation and epitaxy. Ion implantation consists in accelerating a certain amount of ions (dose in at/cm^2) and energy E (usually in keV) and collide it at a certain angle with an existing substrate. An annealing is required to move (activation step) the impurities into substitutional sites to allow conduction. The penetration depth of ions will depend on the energy and the tilt (orientation of the crystalline lattice). The concentration will depend at first order of the dose. We can thus define with Kinetic Monte-Carlo simulations (Fig. 144), some implantation conditions corresponding to the bi-layer implantation. To differentiate NMOS from PMOS, the implantation can be masked to create either N channel or P channel. However, ion implantation creates defects and the junction are not abrupt. To fabricate two or three stacked crystalline layers, epitaxy is preferred. Contrary to ion implantation, the doped channels will grow layer by layer on a seed substrate with the same crystalline orientation and limited defects. It is feasible to grow on top of 4nm Silicon, 8nm of phosphorous doped silicon (Si:P) and then 12nm of boron doped $\text{SiGe}_{30\%}$ and 12nm of Si:P. Latter in the process, the top $\text{SiGe}_{30\%}$ layer can be removed selectively to create the future NMOS as illustrated in Fig. 145. In this case, p over n over p devices must be fabricated to take benefits of the depletion is all configurations. This approach is more expensive due to the epitaxy process but the p and n layers are well defined contrary to ion-implantation. As far as 3D monolithic integration is concerned, the layers could be done prior to bonding to fit the thermal budget restrictions (Fig. 145). That is why in this manuscript, the channel material is fabricated without thermal constraints.

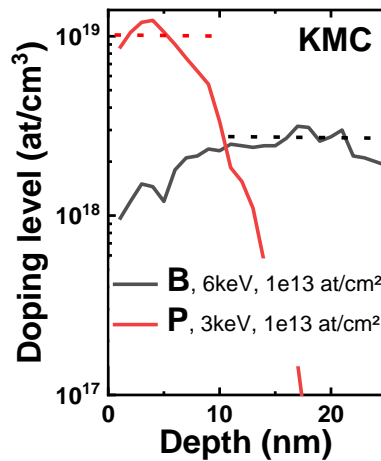


Fig. 144: Doping profile obtained after implantation and spike annealing (KMC plot).

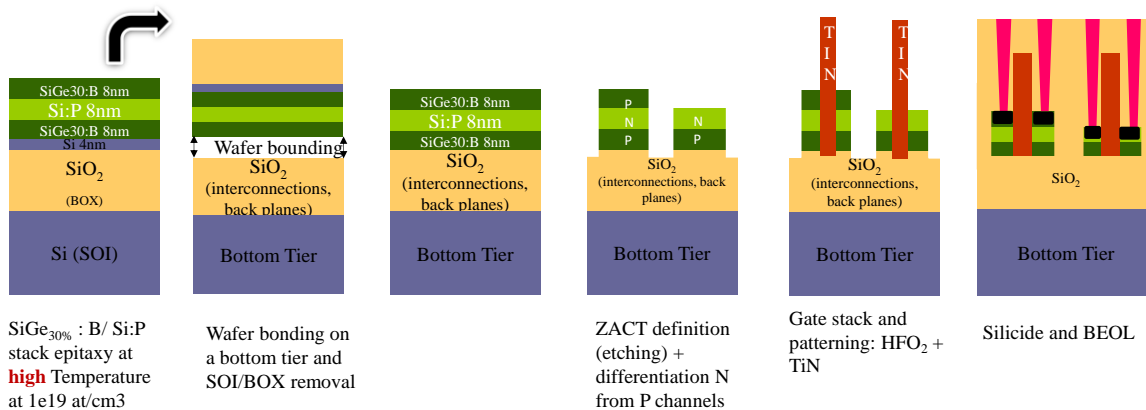


Fig. 145: Simplified proposed process flow to define n over p structures and p over n over p one to allow CMOS integration.

Following the proposed process flow, the active area (width definition) is etched and P and N transistors are differentiated by removing selectively the top SiGe layer of future N devices. One drawback of this tri-layer stack concern the additional topography (3 times higher) which cannot be neglected during etching and deposition steps. Compared to the standard gate first process flow, prior studies are required to etch properly the tri-layers.

Then after the gate stack formation (gate length definition), silicide process is performed to reduce the access resistance. In TCAD simulations, the contacts are assumed to be perfect, *i.e.* just connecting the surface of the source or drain (Fig. 146, (a)). Nevertheless, according to the contact process, the morphology can either be (b) or (c) where the underneath layer is also connected (by top (b) or lateral (c) short-cuts). We did TCAD simulations with (a) and (c) configurations for $W=20\text{nm}$, $L=30\text{nm}$, $t_{\text{Si-n}}=t_{\text{Si-p}}=10\text{nm}$ and $N_D=1.10^{19}\text{ at/cm}^3$. In Fig. 147, we can observe that from the perfect contact (REF in black) to the all-around contact (green curve), the transistor is no longer capable of closing OFF the channel. Total density cut plane (Fig. 148) extracted on these two configurations at $V_G=-0.5\text{V}$ indicates a bipolar conduction in the case of all-around contact. In fact a volume conduction confined in the p-type layer appears for negative V_G , characteristic of a junctionless operation. And this parasitic current increases with the p-type layer doping N_A . For positive V_G the transistor acts as usual.

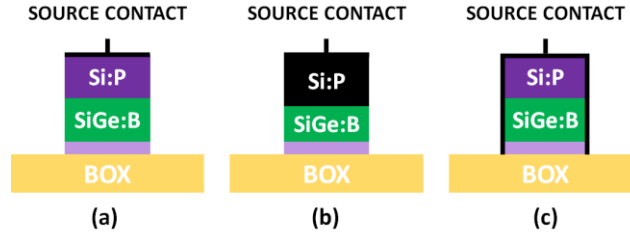


Fig. 146: contact schematic (a) perfect TCAD case (b) recessed contact and (c) all-around contact

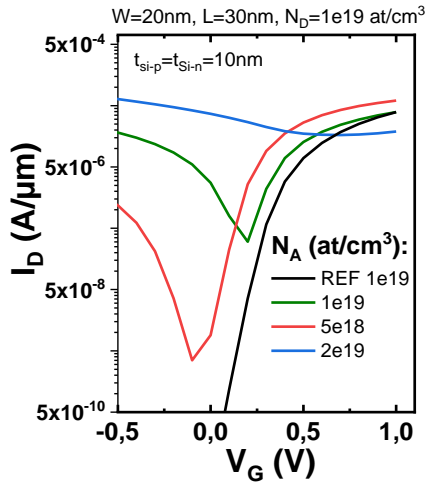


Fig. 147: I_D - V_G for different p -layer N_A doping with either all-around contact or surface contact (REF in black). Bipolar conduction is observed with all-around contact and is more important with p -layer doping.

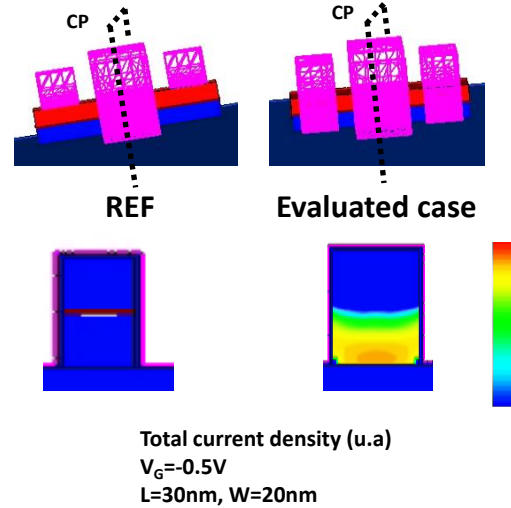


Fig. 148: TCAD simulated cases: (a) perfect and previous TCAD case (b) evaluated case with an all-around contact and their associated TCAD cut plane (cp) at $V_G = -0.5V$ to highlight the Bipolar conduction of all-around contacts.

To avoid the configurations (b) and (c), the source and drain side must be protected by a spacer when silicide is done and silicide must be thinner than the top layer thickness. These two conditions for device processing will be tackled in the process-focused part (6-g). From the simulation point of view, this limitation sets a lower bound for t_{Si} and t_{Si-top} since the silicide process penetrates.

Now we will address the performances of such devices in the situation where the channel is made by epitaxy (n over p or p over n over p devices).

iii- Sizing of the different layers: TCAD simulations

There is a need to study both NMOS and PMOS since in the proposed integration layer, the channel of the n device is the underlying channel of the p one. But first, we will focus on n over p stacking. The studied structures have the following characteristics: $W=20nm$, $L=30nm$, a channel composed (from BOX to gate) of a p layer of thickness t_{p-1} and doping level N_{A-1} , of a n layer (N_D , t_n) and of an optional p-layer t_{p-2} and N_{A-2} . Due to the number of unknowns (N_{A-1} , N_D , N_{A-2} , t_n , t_{p-1} , t_{p-2}) we will fix the parameters from bottom to top to ensure a good OFF current. To have insights about the influence of the underneath p layer (for NMOS), Fig. 149 presents $\log(I_{OFF}/W)$ for various N_{A-1} and t_{p-1} for $N_D = 10^{19} \text{ at/cm}^3$ and $t_n = 1 \text{ nm}$. A higher doping of the p layer will lead to a higher depletion in the n channel (as stated in Fig. 142) and thus decreases the I_{OFF} . Nevertheless, the variation of N_{A-1} and t_{p-1} do not impact much the OFF current in the n channel, so a N_{A-1} of $1e19 \text{ at/cm}^3$ and a t_p of 10 nm , which are representative of the future process flow are taken to express the advantage of this underneath layer. In fact, with this p layer, the OFF current is lowered as seen in Fig. 150. The constraint on doping can be relaxed and a higher doping

can be chosen to lower the access resistance. In fact, even with a channel doping of $N_D=2.10^{19}$ at/cm³ the OFF constraint is respected. As previously said, the silicon thickness must be enough to withstand the silicide process and avoid a bipolar conduction, so we choose $t_n=12$ nm and a doping of $N_D=1.10^{19}$ at/cm³. With these fixed values, the same analysis is performed, N_{A-2} and t_{p-2} varying between 5.10^{18} and 2.10^{19} at/cm³ and 8 and 12nm respectively. The result is presented in Fig. 151 and based on I_{OFF} , and to minimize the access resistance $N_{A-2}= 1.10^{19}$ at/cm³ and $t_{p-2}= 12$ nm. The final structure dimension is presented in Fig. 152.

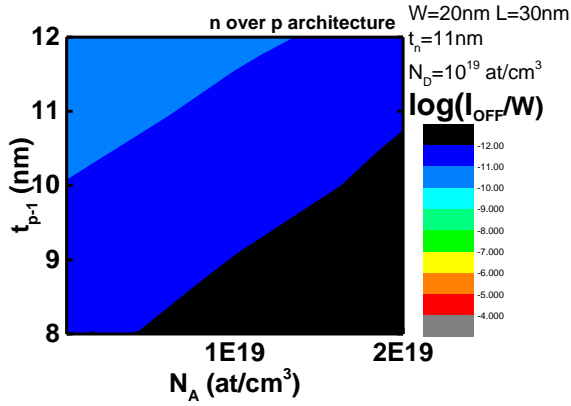


Fig. 149: $\log(I_{OFF}/W)$ as a function of N_A and t_{p-1} for $W=20$ nm, $L=30$ nm, $t_n=11$ nm, $N_D=10^{19}$ at/cm³ and $V_D=50$ mV in a n over p configuration.

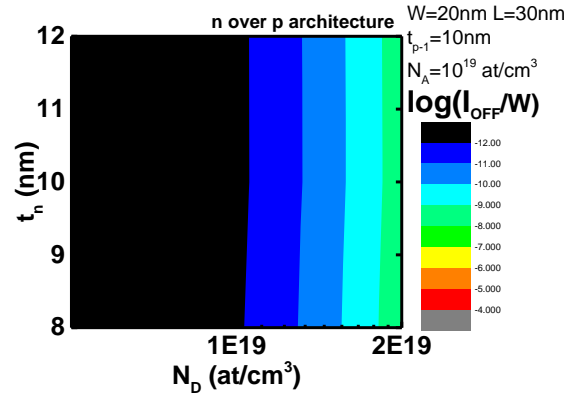


Fig. 150: $\log(I_{OFF}/W)$ as a function of N_D and t_n for $W=20$ nm, $L=30$ nm, $t_{p-1}=10$ nm, $N_A=10^{19}$ at/cm³ and $V_D=50$ mV in a n over p configuration.

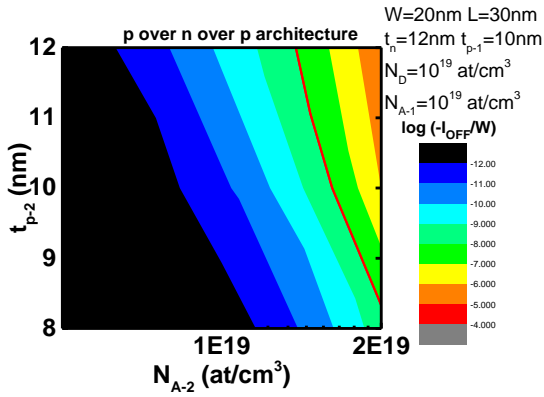


Fig. 151: $\log(I_{OFF}/W)$ as a function of N_{A-2} and t_{p-2} for $W=20$ nm, $L=30$ nm, $t_{p-1}=10$ nm, $t_n=12$ nm, $N_{A-1}=10^{19}$ at/cm³ and $N_D= 1. 10^{19}$ at/cm³ $V_D=50$ mV in a stacked p -n -p configuration.

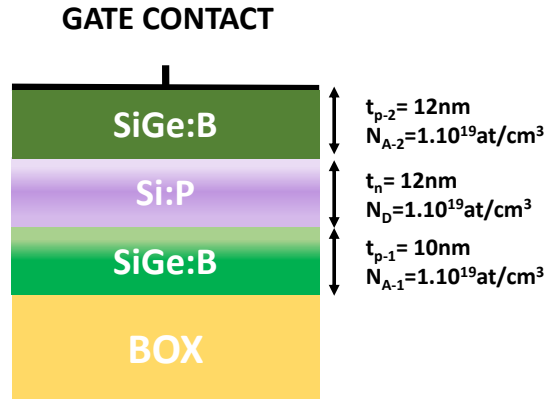


Fig. 152: Presentation of the final sizing which will be analyzed more in depth in next part.

In this part we proposed devices with stacked n-p layer channel to improve the electrostatic control of devices with an associated process flow to create CMOS devices highlighting the process development needs. Good performances targeting low power applications are evidenced even with penalties on capacitances. Now, based on the identified trade-off, the performances between JL devices, n-over-p JL devices and inversion-mode devices are compared.

c. Performances of the different structures compared to IM devices

JL tri-gate devices (TG-JL: $N_D=7.10^{18}$ at/cm³, $t_{si}=11$ nm) and stacked n/p devices (n/p-JL $N_{A-1}=10^{19}$ at/cm³, $t_{p-1}=10$ nm, $N_D=2.10^{19}$ at/cm³, $t_n=12$ nm) have been selected for their I_{OFF} without considering ON current and electrostatic control (SS, DIBL...). That is why in this part, we will analyze

further the performances of the selected dimensions for NMOS and compared them to NMOS IM devices (TG-IM: $t_{si}=11\text{nm}$ and undoped channel). We will focus only on $W=20\text{nm}$ width, considering threshold voltages, SS in linear ($V_D=50\text{mV}$) and saturated ($V_D=0.8\text{V}$) region and DIBL.

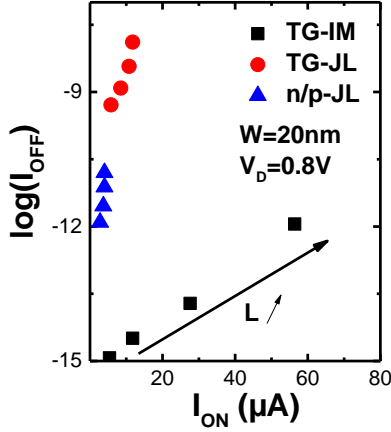


Fig. 153: I_{ON} - I_{OFF} for $W=20\text{nm}$ and various gate length for the three analyzed configurations. Inserting the p layer beneath the n channel (n/p-JL case) lowers the OFF current (compared to TG-JL) for JL devices.

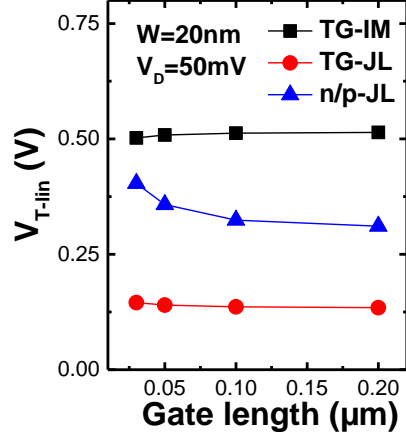


Fig. 154: Extraction of threshold voltage at constant current (at $I_D=10^{-7}W/L$) for the three analyzed devices ($W=20\text{nm}$ and various gate length). JL device retains a lower V_T compared to IM devices but this negative shift can be compensated by the insertion of the p layer beneath the channel.

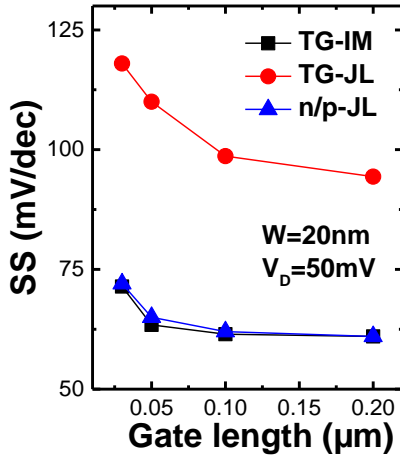


Fig. 155: extraction of the subthreshold slope in linear regime as a function of the gate length. A degradation is seen for small gate length. TG-IM and n/p-JL features similar subthreshold slope close to the ideal value of 60mV/dec , indicating the good electrostatic control.

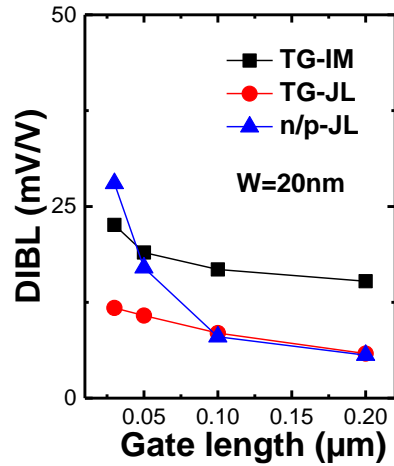


Fig. 156: Computation of the DIBL as a function of the gate length. A degradation is seen for small gate length as predicted by theory. However, JL transistors achieves lower DIBL values attributed to channel length modulation for uniformly doped devices.

The traditional I_{ON} - I_{OFF} figure of merit is presented in Fig. 153. We can notice that inserting the p layer beneath the n channel (n/p-JL case) lowers the OFF current (compared to TG-JL) for JL devices at the expense of the ON current. There is still a lack of drive current for JL devices but due to the scaled dimensions, this is attributed to the not additional doped source and drain, increasing the access resistances. If we have a look now at the linear threshold voltage extracted at constant current (Fig. 154), we observe the V_T shift for JL devices due to the channel doping but which could be compensated with the insertion of a p layer beneath the channel. In this precise case, the V_T between TG-IM and n/p-JL are not aligned but since the SS is identical (presented in Fig. 155), we could size the n/p device to detain the same I_{OFF} and thus similar V_T . To finish with, Fig. 156 presents the DIBL from 30nm gate length to 200nm. We do observe that the DIBL values is smaller for JL device than IM one. This is attributed to channel length modulation in JL devices.

From this presentation of performances, we validated the sizing of the different layers for JL transistors. Devices have been fabricated to electrically analyze the differences between JL and IM devices. The next part will present the general gate first fabrication process and some adaptation and process development to lower the process thermal budget to make it compatible with 3D monolithic integration.

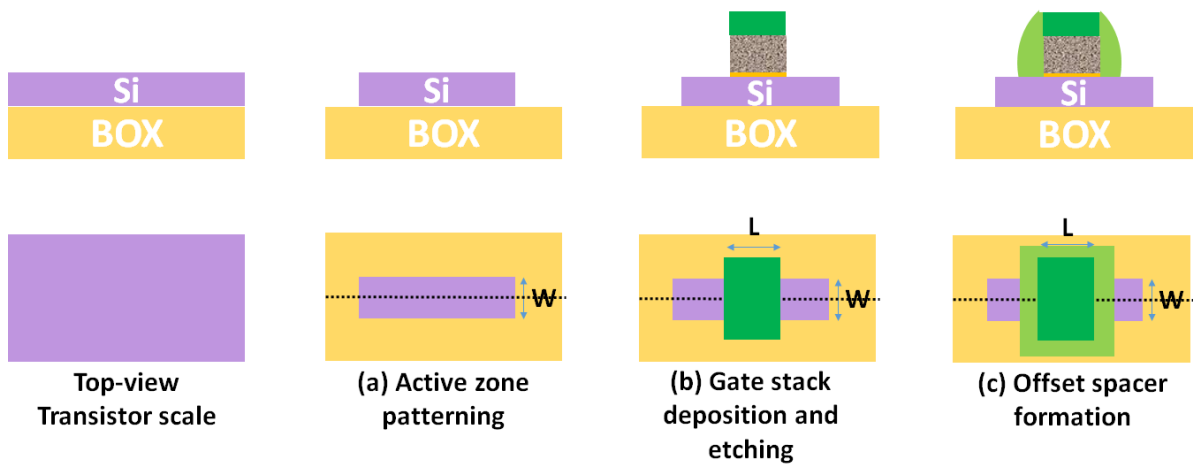
6- Fabrication process flow

In this part, we will first present the process steps needed for the transistor fabrication (process flow). Secondly, each brick will be developed in details and indications and studies to lower its thermal budget in the scope of 3D monolithic integration will be exposed.

a. Gate first integration at high temperature

The main steps of the high temperature process of reference are presented in Fig. 157. In this chapter, a gate-first FDSOI architecture is considered, but in chapter IV, a gate last architecture is realized and will be explained later. Starting from a silicon on insulator blanket wafer, the first step is to define the future active zone (a). By doing so, the neighborhood devices are electrically separated (no silicon connection between them, only oxide). This isolation scheme is referred as mesa isolation (our case). It is also possible to dug trenches and fill them with an isolation material. This step defines the future width of the transistor (see top-view). The second step consists in the gate stack deposition to form the future Metal-Oxide of the MOS transistor. Usually, it consists of a high-k material deposition tuned to achieve a certain EOT (here 2nm HfO_2), a metallic material deposition (here, TiN), poly-Si and a hard mask material. The gate stack is then etched to form the gate with a gate length L (b). The third step (c) is the formation of a spacer to create an offset between the gate and source and drain. Then the fourth step is to grown silicon (doped in-situ or not) selectively by epitaxy on top of source and drain, to raise them before implantation (d). The fifth one (e) is about source and drain implantation to form the transistor junctions and also lower the access resistance. The highest value of dopants is wanted in source and drain to minimize access resistance but are undesirable near the channel and the drain/source to avoid HCI. However, this step is divided into two sub-steps to avoid dopant diffusion into the channel. That is why a first Lightly Doped Source and Drain (LDD) implantation is carried out before an additional spacer creation (f) and a Highly Doped Source and Drain (HDD) implantation. Spike annealing is required in both cases to activate the dopants. The resulting junction detains a high doping level far from the conductive channel (where the future contact will be) and a lower one near the junction. The last step is the silicidation of the contact area in order to decrease further the access resistance, Pre-Metal Dielectric (PMD) deposition, contact etching and filling (g).

A study called Hot Temperature Reference (HT-REF) is fabricated with this process flow to answer the question what is the impact of a heavily doped channel introduction. Further details are given in parts 7-.



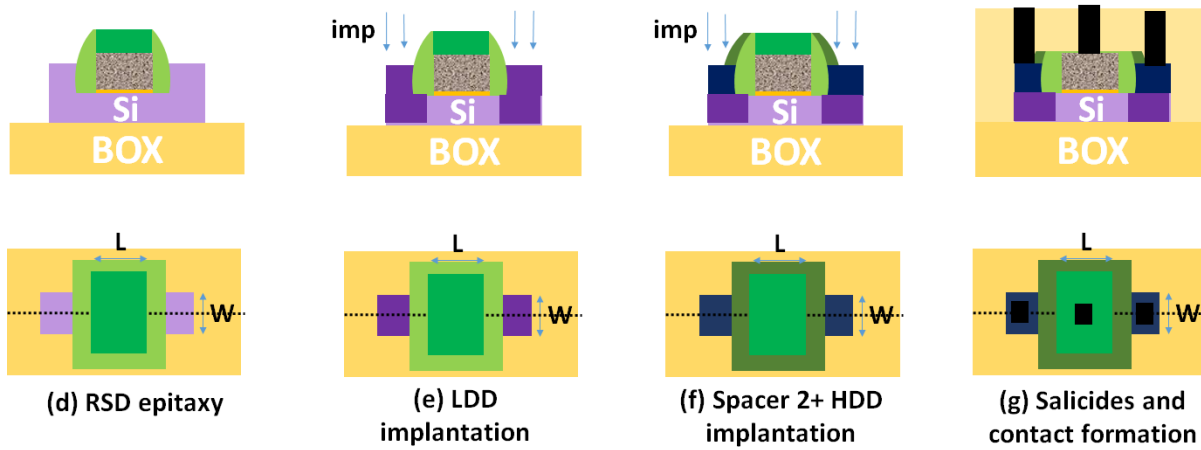


Fig. 157: Standard process flow for 28nm integration (gate first), top-view and cross-section.

The following parts will present the thermal budget of all these steps and how to lower it to make this process flow suitable for 3D monolithic junctionless transistors. Let's keep in mind that the thermal budget must be lower than 500°C, 2hours. The comparison of our choices with state-of the art references will be done.

b. Channel material

Junctionless transistors must detain a doped channel. As seen in part 1-a, this doped channel can be directly deposited on bottom tier at low-temperature or can be done prior bonding either by epitaxy or implantation and high temperature annealing. In the former case, the material deposition and processing cannot exceed the limited thermal budget. In the latter case, no particular restrictions is seen since the channel material is prepared before bonding.

i- Poly-si deposition

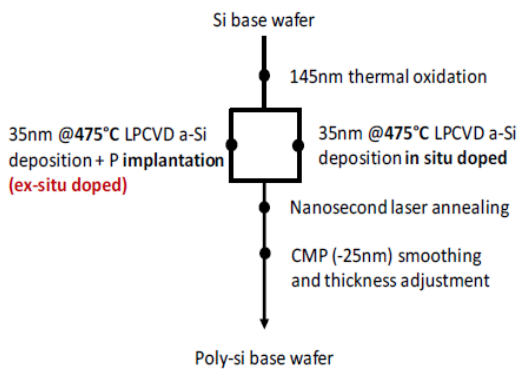


Fig. 158: low-temperature (<475°C) integration scheme. In-situ and ex-situ doping are investigated in this work.

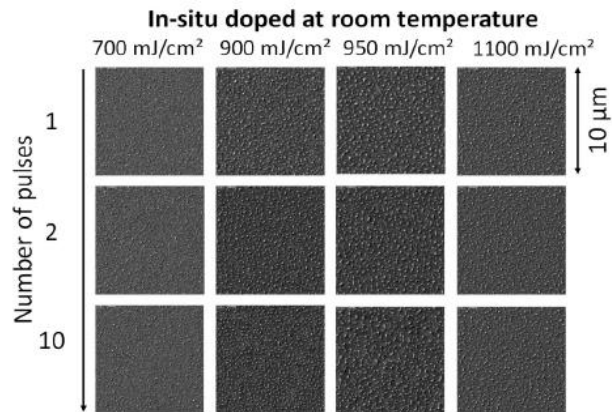


Fig. 159: Scanning Electron Microscopy (SEM) images showing the impact of cumulated pulses for different laser energy density values. More pulses will lead to larger grain size. Figure from [148].

As seen in 1-b, poly-Si junctionless transistors can achieve good performances for applications where variability is not an issue. In literature, several groups deposit the channel material directly on the bottom tier without damages [134], [253]. However, this integration scheme suffers from a poor poly-silicon roughness. For instance, typical RMS values are 0.7nm [134]-1.2nm [253] (Green Nanosecond Laser

Crystallization (GNS-LC) + Chemical Mechanical Polishing (CMP)) or 0.6nm [254] (HPA trimming). Channel roughness is a critical issue for junctionless transistors since its threshold voltage is highly dependent on silicon thicknesses. To overcome the variability, we propose the following process (Fig. 158): low-temperature amorphous silicon (a-Si) deposition followed by nanosecond laser annealing and CMP. The specifications are the following: a maximum thermal budget of 500°C, 2hours, and a 12-15nm thick a-Si layer doped at $1e19$ at/cm³ for future JL device formation with a ≤ 0.5 nm RMS variation to lower device variability. This work has been presented in SOI-3D-Subthreshold Microelectronics Technology Unified Conference in 2019 [148].

A 35 nm thick a-Si layer is deposited at 475°C on an oxidized blanket bulk wafer. Ex-situ doping (using ion implantation) and in-situ doping have been compared. For in-situ doping deposition, a 90sccm phosphorus flow is used in order to achieve a 10^{19} at/cm³ doping concentration [255]. For ex-situ doping, the implantation conditions have been simulated by Kinetic Monte-Carlo TCAD (Silvaco).

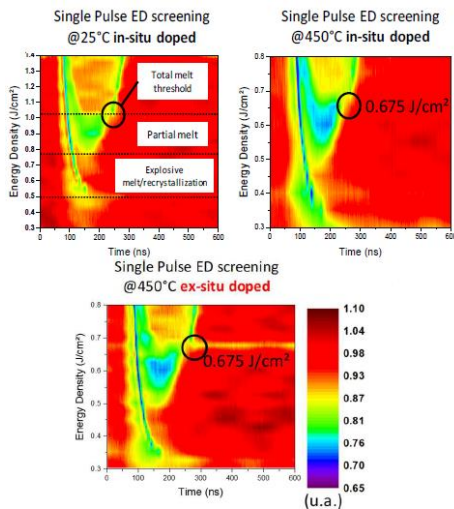


Fig. 160: Time Resolved reflectometry: energy density screening. Three regimes are observed: explosive melt/spontaneous recrystallization, partial melt and total melt. No significant difference is observed between in- and ex-situ doping. Figure from [148].

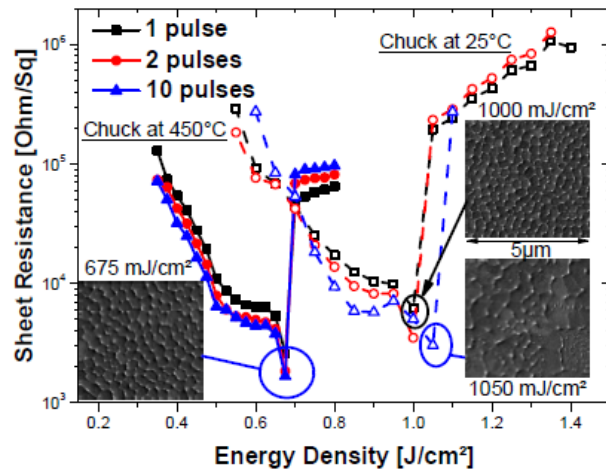


Fig. 161: Impact of chuck temperature and number of pulses on sheet resistance for in-situ doped Si. The higher the number of pulse, the lower the resistance is. Also, a 450°C chuck will shift the curve as far as the laser energy density is concerned [148].

Thanks to low depth penetration, UV-NLA is suitable for crystallizing a top a-Si layer while preserving the integrity of the bottom tier [256]. For these reasons, an excimer laser (308nm wavelength and optimized pulse duration 160ns [257]) is used to activate the dopants and recrystallize the a-Si layer. Based on (Figs 3-6). We can play on different parameters to tune the grain size/ film resistivity:

- **Laser energy:** on the Time Resolved Reflectometry (TRR) analysis (energy density screening from 0.3 to 1.4 J/cm²) three regimes can be identified. In fact, in-situ reflectometry monitoring allows us to detect film melting (Reflectivity decreases). For low energy density (< 0.775 J/cm²), no melting is detected though TRR. An explosive melt followed by a spontaneous recrystallization can occur for such energy densities (see Fig. 160). For intermediate energy densities (preferred processing window), a part of the a-Si film melts and recrystallizes, forming grains and lowering the film resistivity (see Fig. 161). For higher energy density (> 1.025 J/cm²), the layer is entirely melted before the end of the laser pulse resulting in an amorphous film. In this work, we want to maximize the grain size to create single-grain channel. That is why, the working point of energy density is chosen at the minimum of the resistivity curve. However, for safety reasons, a lower energy is also considered.
- **Number of pulses:** a second laser annealing (two pulses) will preferentially melt the smaller grains which will coalesce with bigger grains leading to an overall grain size increase [258]. For

instance Fig. 159 presents the impact of one, two and ten pulses. In fact, each pulse increases the grain size, but also double the processing time. That is why four cumulative pulses have been chosen to optimize the grain size.

- Stage temperature:** the idea consists in increasing the stage temperature at 450°C to reduce the thermal gradient undergone by the film. It will slow down the cooling, yielding larger grain and thus lower resistivity ([259] and Fig. 162). For instance, at the melt threshold, melting time is 156ns (chuck at 450°C) compared to 130ns at 25°C. We chose to have a stage temperature at 450°C in all cases.

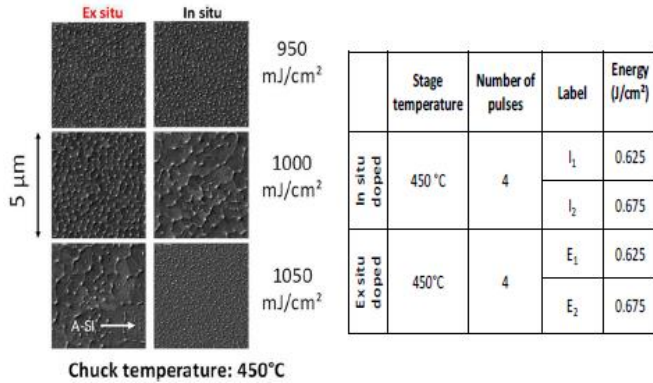


Fig. 162: SEM comparison between in- and ex-situ doping. No significant differences is seen.

Fig. 163: Selected conditions to maximize grain size and obtain a low resistivity and lower processing time.

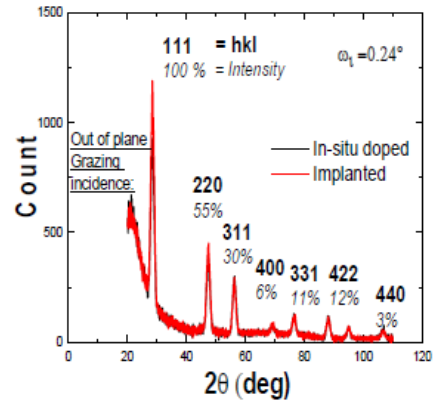


Fig. 164: XRD out-of plane grazing incidence patterns. Miller index and intensity are taken from [260]. Figure from [148].

Fig. 162 and Fig. 164 compare in- and ex-situ doping in terms of energy response and grain size. No significant difference is seen. To probe further, two energy density conditions are studied (0.625J/cm² for I₁, E₁, and 0.675J/cm² for I₂, E₂, as defined in Fig. 162) with a stage temperature of 450°C using four cumulative pulses. Patterns from grazing incidence X-ray diffraction in-plane and out-of-plane geometry (Fig. 164) correspond to poly-Si with no texture, indicating no preferential direction regrowth during annealing.

Extracted from 5µm*5µm (AFM)		In-situ doped		Ex-situ doped	
		I ₁	I ₂	E ₁	E ₂
Before CMP	R _{MAX} (nm)	50	63	49	71
	R _Q (nm)	9	9.6	7	10
After CMP	R _{MAX} (nm)	4	2.3	1.6	1.7
	R _Q (nm)	0.25	0.29	0.2	0.2

Fig. 165: Roughness measured by Atomic Force Microscopy (AFM) on 5x5µm² scan before and after CMP. No difference is seen between in and ex situ doped wafers. The best case obtained roughness is 0.2nm.

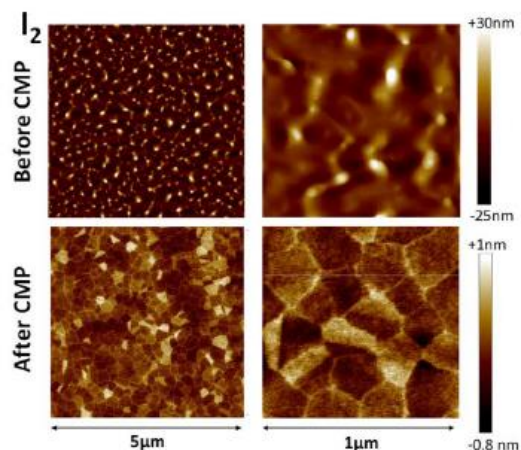


Fig. 166: Atomic Force Microscopy (AFM) before and after CMP. The CMP was efficient to reduce the R_{max}. The same grain size (around 200nm) is seen before and after CMP, meaning that the grain morphology is invariant with depth. Figure from [148].

To adjust the thickness and to lower the roughness, 25nm silicon is removed by CMP. For instance, for the E1 condition, the peak-to-valley thickness variation (R_{MAX}) is reduced from 49 nm (before CMP) to 1.6 nm (after CMP) and the Root Mean Square ($RMS=R_Q$) from 7 nm to 0.2 nm (Fig. 165). Furthermore, Atomic Force Microscopy measurements (Fig. 166) evidence that grain size is unchanged after CMP, highlighting that the film morphology is homogeneous within the depth. In addition, thicknesses measurements indicates that a 13nm poly-si thickness is achieved.

To conclude, it is possible to create a 475°C 13nm doped poly-si layer with optimized grain size. No specific difference is seen between in-situ and ex-situ doped. One major advantage of this approach is the cost since it doesn't evolved a SOI donor wafer. However, future devices will suffer from an additional variability due to the presence of grain boundaries.

ii- Channel creation wo thermal budget constraints

In this case, where the substrate doping is done prior bonding, there is no constraint on the thermal budget for wafer preparation. As seen in Fig. 167, the wafer can be pre-process at high temperature before report. In the literature, several groups proposed technics to bound the wafer on-top of the other and are accessible in CEA-LETI [132]. However, 3D monolithic wafer processing (bottom tier processing + wafer bounding + top tier processing) requires much more processing time than planar one. That is why, in this thesis work, no 3D monolithic wafers have been realized but rather unipolar (no CMOS integration) planar 300mm wafer.

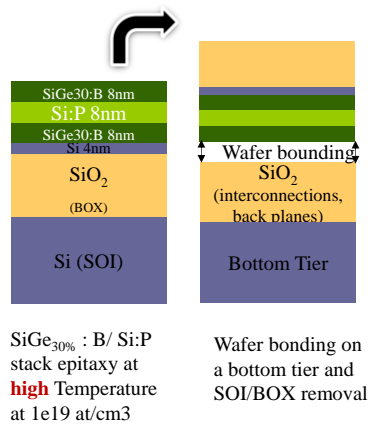


Fig. 167: Wafer bounding process illustration.

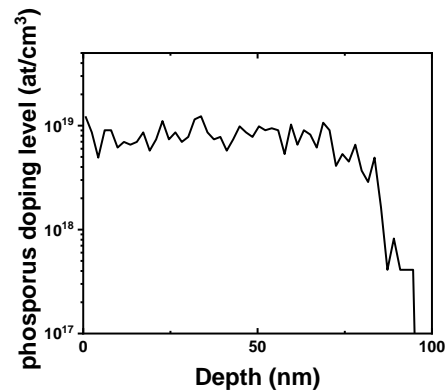


Fig. 168: Dopant profile for Si:P determined by Secondary Ion Mass Spectrometry (SIMS). The $10^{19} \text{at.cm}^{-3}$ doping level is achieved.

We propose two hot temperature processes to create the future channel material:

- Ion implantation and spike annealing (1050°C, 30s). The annealing will redistribute the dopants, forming a uniformly doped channel at N_D . Various N_D are investigated but the reference doping is $7 \cdot 10^{18} \text{at.cm}^{-3}$. Phosphorus (respectively Boron) is chosen for N devices (respectively P).
- In-situ doped epitaxy with a 4nm silicon seed. On the SIMS profile (Fig. 168), we can observe that silicon is doped with phosphorous at $10^{19} \text{at.cm}^{-3}$ and $\text{SiGe}_{30\%}$ is doped with boron at $10^{19} \text{at.cm}^{-3}$. A good silicon thickness uniformity is seen on the 300mm wafer and illustrated in Fig. 169. These layers can be stacked to create PN junctions perpendicular to current flow and modulate the threshold voltage (see Fig. 139).

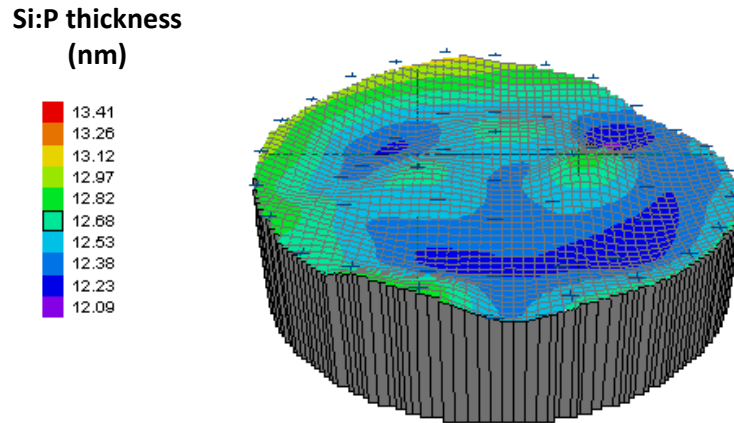


Fig. 169: Thickness measurements of silicon phosphorus doped deposited by epitaxy on a 300mm wafer. The range of value is 1nm.

Different studies are associated to these choices in order to make the comparison between the two technics. The first (and third) with a low temperature channel, called Junctionless Low Temperature 1 (JL-LT 1, respectively JL-LT 3), is done by ion implantation. The second low temperature study channel, JL-LT 2, is done by epitaxy. Further details are given in section 7-.

c. Active zone patterning

This process step consists in defining different islands (or mesa) being the future device silicon channel and source/drain (active zone). To define them, a process called photolithography is used. The idea is to transfer a geometric pattern from a photomask (optical mask) into a substrate thanks to light exposure (Fig. 170). For this, a photosensitive chemical photoresist is spread by spin-coating on the wafer before being exposed to light. They are two types of resins: positive and negative one. Here a positive resin is given as an example, *i.e.* the light-exposed part will be soluble and the non-exposed part will remain. The smallest feasible dimension is called the critical dimension CD and depends on the ability of the light system to project a clear image of a small feature. Optics states that CD is proportional to the light wave length λ and inversional proportional to the numerical aperture. That is why, current lithography tools uses Deep Ultra-Violet (DUV, $\lambda \sim 193\text{nm}$) to reach nanometer dimensions. Once the resist is exposed, it is also possible to trim the resist to reduce the dimensions (4) at the expense of density. Then, the underneath material is etched, following the pattern. Several layers can be etched, transferring this pattern, layer by layer. The last step consists in removing the photoresist thanks to a resist stripper liquid.

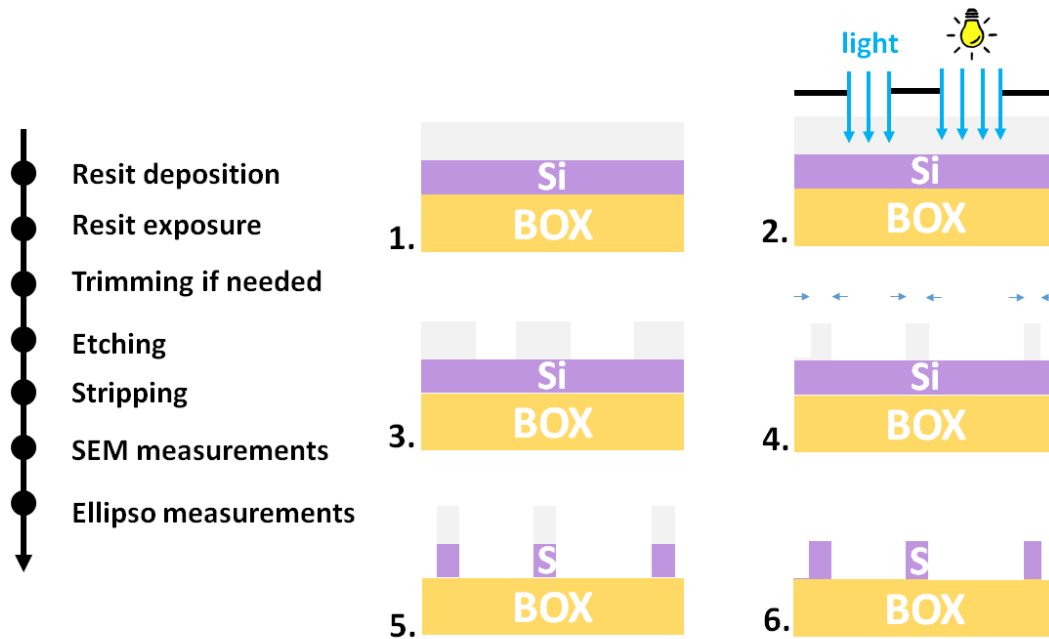


Fig. 170: Active zone creation process flow.

In our case, the obtained CD for channel width is 80nm for the photoresist and 20nm after trimming. CD-SEM measurements are done on specific dies and devices of the wafer to extract a mean and a standard deviation values. The exactly same dies will be measured in electric test once the process is finished. In a similar way, the silicon thickness consumption is monitored among the 300mm wafer to have an idea of silicon variation.

The active zone is etched below 500°C, so there is no need to lower down the process temperature.

d. Gate stack

After active zone creation and thus the channel definition with a specific width, gate stack is deposited and etched. It will determine the gate length noted L_G . In the first part we will discuss the material usually chosen for gate stack deposition and motivate our choices. In the second part, gate stack etching will be tackled, highlighting the engineering work required.

i- Gate stack materials

The gate stack will govern the future device electrostatics. A good choice of materials is determinant. For a standard gate first flow, to create a Metal Oxide Semiconductor transistor, a dielectric layer (historically SiO_2) and a metal gate electrode are needed.

As far as the dielectrics is concerned, historically SiO_2 was deposited. However, to work at a lower voltage the thickness of SiO_2 have been drastically reduced, increasing the gate leakage current [261]. That is why high-permittivity (high-k) materials have replaced the conventional SiO_2 layer, enabling an Equivalent Oxide Thickness (EOT) reduction with correct thickness [262]. Since our applications are mainly digital and low power, an EOT of 1nm have been chosen, corresponding to 2nm HfO_2 deposition on a SiO_x interfacial layer. To obtain such a small thickness with a good control, the HfO_2 is deposited atomic layer by atomic layer (Atomic Layer Deposition). Then a nitriding at 250°C can be performed or not followed by an annealing at 600°C, 2min.

Metal gate electrodes were at first done in doped poly-Si. However, it cannot be used for advanced nodes, where the poly-Si/high-k contact creates Fermi level pinning due to the formation of dipoles at

the interface [263]. In addition, Poly-Si degrades the electron mobility with high-k (and thus limit the circuit speed), contrary to metal gates [264]. Furthermore, metal gates feature a lower gate resistance than poly-Si. The first criteria to choose the metal gate is the metal work function Φ_m . Φ_m will dictate the future threshold voltage. For fully depleted undoped-channel devices, the usually chosen gate material is close to mid-gap to achieve low V_T for both PMOS and NMOS [265]. In our case, the channel is doped, which will shift the V_T . However to co-integrate NMOS and PMOS (CMOS), the same metal gate material must be preferentially used. TCAD simulations (part 5-a) consider a mid-gap material, such as TiN (integrated at high-temperature) and show well-operating devices. Furthermore, the TiN detains a thermal stability above 700°C [266], even if in the present case this aspect is let apart since the process flow temperature is under 500°C. In superposition of this TiN layer (by reference, 5nm), a 50nm poly-si layer is deposited and the so-called hard mask which ease the etching process.

Starting from this gate stack deposition baseline, some modifications have been done to lower the thermal budget under 500°C, 2hours. As far as temperatures are concerned, the HfO₂ is deposited at 400°C, the annealing after nitration is typically done at 600°C, the TiN at 400°C and the poly-Si at 630°C. The poly-Si can be deposited amorphous at 500°C and recrystallized (see 6-b). In this latter case, the poly-Si will suffer from roughness. That is why, we made the choice to integrate a thicker TiN layer (around 30nm) instead of 5nm TiN + poly-Si. The sizing of this TiN layer comes from two considerations. First, the thickness must be enough to withstand contact bricks. Secondly, for the spacer formation, it is preferable to choose a similar thickness as the baseline to lower the engineering work on this future brick. The modifications to the baseline are depicted in Fig. 171 and Fig. 172. Next part present the etching of this gate stack. Slight modification of top SiN/oxide between the morphological batch JL-LT 1 and electrical one JL-LT 2 and 3 to adjust the top oxide consumption during processing. In fact, for contact brick, no oxide must remain.

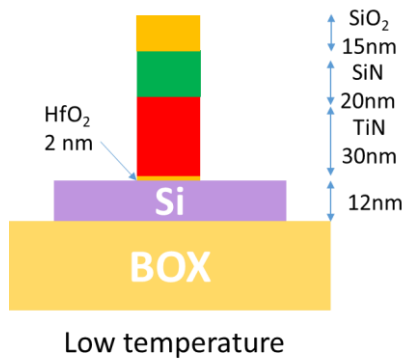


Fig. 171: Schematic of the deposited gate stack for low temperature junctionless transistors and high temperature one.

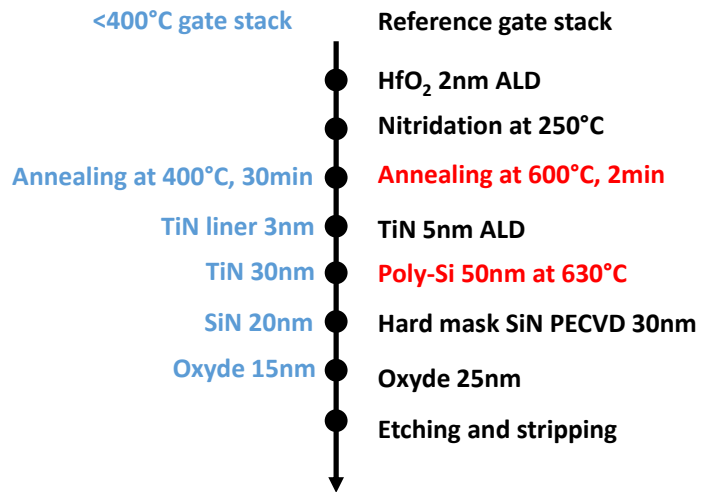


Fig. 172: Gate stack process flow. Variants between low-temperature process flow and baseline are highlighted.

ii- Gate stack etching

Two strategies have been developed to etch the 30nm TiN/HfO₂ gate stack. The standard anisotropic etching process is not selective to HfO₂, it means that the HfO₂ is not an etch stop layer and without additional monitoring, the active zone is also etched and just the gate stack remains (see Fig. 173).

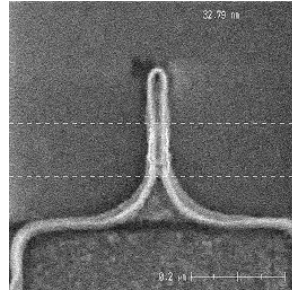


Fig. 173: SEM picture after gate etching with standard process: the silicon channel/HfO₂ does not act as a stop layer, so no active area is remaining.

To avoid this, the etching process must be stopped before degrading the gate oxide. However, as seen in Fig. 176, at this step, there is still TiN remaining on the edge of the silicon active zone which electrically connect source and drain (short). Thus an over etch is performed to get rid of the TiN spacer. So the idea is to etch almost all of TiN gate with a non-selective anisotropic process and to over-etch the remaining TiN (few nanometers) with an isotropic process, selectively to HfO₂ (see Fig. 174). Due to polymerization and the presence of a hard mask, the top of the gate stack is less attacked by the isotropic process leading to a T shape gate (see Fig. 175). Such a shape is interesting for RF applications because of a short electric length with a low gate resistance. After optimization, a recess of 10nm by side is seen using this technique.

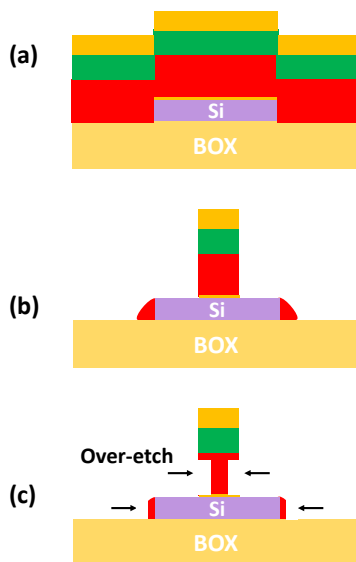


Fig. 174: gate etching process explanation. (a) Presents the deposited stack. (b) The first step consist in an anisotropic etch. However, this etching step is not selective to HfO₂. If we stop just before attacking the HfO₂, a TiN spacer is seen on the edge of the silicon active zone. (c) a TiN over-etch step selective to HfO₂ but isotropic is done, creating a T-shape gate.

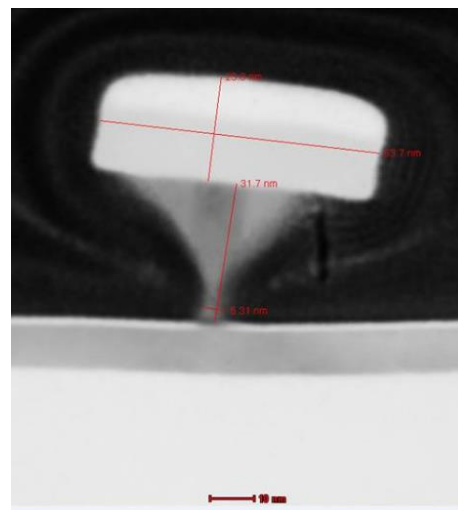


Fig. 175: T-shape gate TEM cross-section. An electrical gate of 5nm is obtained while the top of the gate measures 60nm. This condition too extreme have been worked out to obtain only a 10nm recession by edge (see Fig. 177)

On the other hand, we performed a study using different etching chemicals (BCl₃/Cl₂ instead of HBr/Cl₂) with or without over-etch and with and without substrate bias. Fig. 176-a highlights that without

modifications to the standard process a TiN spacer is remaining on the edge of the active zone as stated previously. Fig. 176-b and -c shows that adding an over-etch step get rid of the TiN spacer but consume the silicon channel. This consumption is mitigated by biasing the substrate but is still detrimental for the structures (-9nm). The best solution (Fig. 176-c) is to use BCl_3/Cl_2 chemicals to obtain a selective anisotropic process. The gate profile is rather straight as illustrated in the TEM cross-section in next part (Fig. 178).

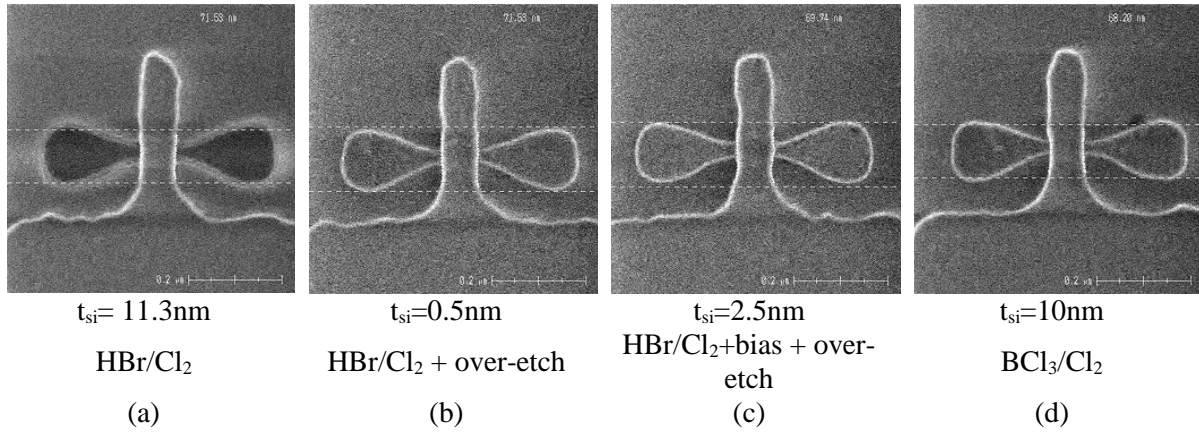


Fig. 176: Chemicals used for gate etch with the associated SEM top-view and silicon thickness measurements. (a) No over-etch is performed, a TiN spacer on the edge of active zone can be seen. (b) Over-etch is performed, no TiN spacer is seen but the active zone have been damaged. (c) To increase the selectivity of HBr/Cl_2 with respect to silicon, an additional bias is used. The silicon is less damaged but it is not thick enough for next steps (like raised source and drain). (d) Another etching chemical BCl_3/Cl_2 is used. No TiN spacer is seen and the silicon consumption is moderate. This is the chosen condition.

In this part, we proposed a gate stack processed below 400°C and two different etching strategies to obtain a T-shape gate or a straight gate. Next part, we will analyze the impact of these profiles on spacer shape.

e. Spacer

The role of this sidewall spacer for standard devices is to prevent any short-cut between the gate and the source/drain and also to prevent the region near the channel to be highly doped by the implantation of the source and drain. It also allows raised source and drain formation by epitaxy before the fabrication of an additional spacer (so-called “spacer 2”) for the HDD implantation. In the junctionless case, the source and drain implantation is not mandatory since the source and drain are already doped (but still advised see next sub-section). Additionally, the channel thickness is between 12-15nm and thick enough for a thin silicide process. That is why the epitaxy and spacer 2 are not done in our proposed flow.

However, the gate stack is fully metallic in our case. It raises another constraint: when the silicide is formed on the source/drain region, there is a selective removal of the non-reactive metal. If there is a path though the TiN gate, all the TiN will be removed. That is why the spacer also encapsulate the gate to protect it from chemicals used for the silicide module. Thus the sizing of the spacer must be done carefully and depends of the etching rate of the chosen material. To increase the density of the deposited SiN at low-temperature, cycles of 2nm deposition and plasma treatment to densify the SiN is done. The fullsheet etching rate of this material is 3.8nm/min (HF 0.5%). However in our case the deposition is also lateral and the T-shape gate can screen the densification on lateral sides and thus increasing the etching rate. That is why, we prefer to oversize the spacer to make sure the TiN is fully encapsulated especially that in junctionless devices there is no need to implant close to the gate. That is why a 40nm SiN layer is deposited at 400°C and etched resulting in a 30nm spacer as seen on TEM cross-section in Fig. 177. Fig. 178 presents the result for straight gate spacer etching when a 20nm SiN layer is deposited. The chosen material is rather conform, *i.e.* it follows the topography of the wafer and is deposited on

the sidewall of the gate. In the case of the T-shaped gate, a void is formed when SiN is deposited due to its peculiar shape.

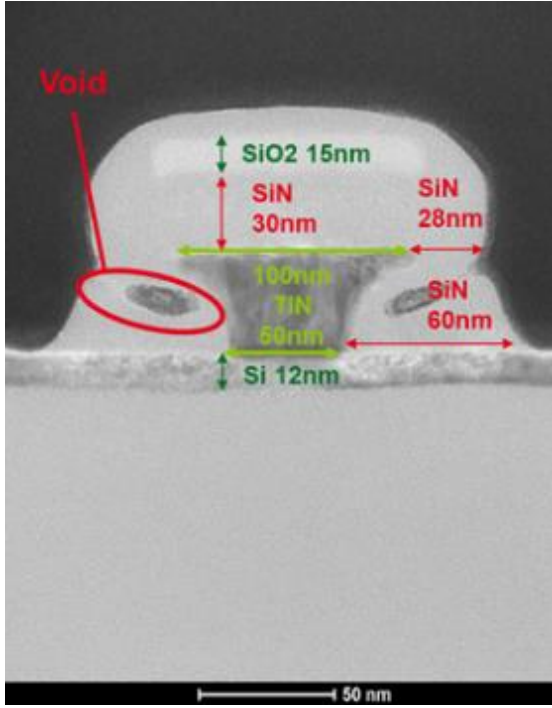


Fig. 177: TEM cross-section of the T-shape gate after sidewall spacer etching (40nm deposited).

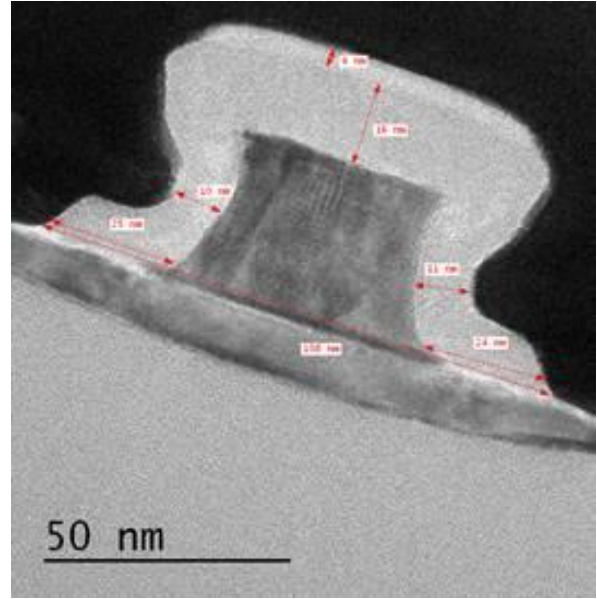


Fig. 178: TEM cross-section of the straight gate after sidewall spacer etching (20nm deposited).

f. Junction engineering SPER

After defining the spacer, source and drain implantation can be performed to reduce the access resistance. In fact, nowadays, the current degradation is mainly due to high-access resistance [267]. The resistance can be expressed by Eq. 21, one part accounting for the channel resistance (which is modulated by the gate) and another part representing the resistance to access the channel (Fig. 179). In the access resistance, we can dissociate the contact resistance R_{co} (dependent on contact size and materials) from the interface resistance between silicide and source/drain and from the resistive silicon piece between the contact and the channel R_{spa} (below the spacer). For small channel dimensions R_{access} becomes comparable to $R_{channel}$ and cannot be neglected [268] especially for thin films. That is why the resistivity under the spacer must be lowered, *i.e.* this region must be heavily doped. As far as junctionless transistors are concerned, this region is by default uniformly doped at N_D . However, N_D is generally around 10^{19} at/cm³ which ensures an ohmic contact but could be increased further to lower R_{spa} . To decouple the impact of channel doping and access resistance, we propose to fabricate a purely junctionless devices and devices with additional S/D implantations. Note that the latter case do not have a uniform doping anymore and thus has a lower channel length modulation.

$$R_{ON} = V_D/I_D = R_{channel} + R_{access} + R_{silicide} \text{ with } R_{access} = R_{co} + R_{spa} \quad \text{Eq. 21}$$

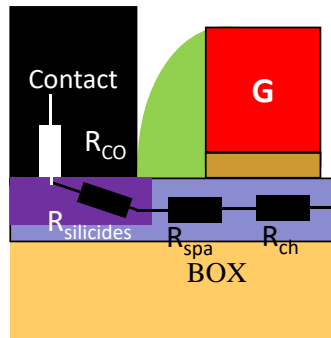


Fig. 179: Representation of the ON resistance contributions for junctionless devices.

In a conventional hot process flow, a doped in-situ epitaxy to raise S/D is done before LDD and HDD implantation and annealing (dopant activation). The limiting thermal budgets are the epitaxy and the spike annealing (1050°C, few seconds). In our case, no epitaxy is done since the silicon thickness is enough for silicide process. As far as dopant activation is concerned, it remains possible to move dopants from interstitial sites to substitutional one by Solid Phase Epitaxy Regrowth (SPER) at 500°C [148]. The SPER technique is described in Fig. 180. First an implantation is realized with a double functionality: to introduce the dopant into the crystalline substrate and to amorphize part of the layer. If the doping specie is not heavy enough, a neutral one such as germanium can be used to amorphize partly the layer. A crystalline seed (at least 3nm [269]) must be maintained to recrystallize the amorphous layer. This recrystallization starts from the amorphous/crystalline *a/c* interface and activates efficiently the dopants located in the amorphous layer as evidenced by L. Pasini *et al.* [270]. It can create end-of-range defects (extended defects, below the previous *a/c* interface) [271]. The maximum dopant concentration is defined as the clustering limit, which is the maximum doping level before forming clusters deactivating dopants. The limit has been established at 6×10^{20} at/cm³ in [270] for phosphorous and at 3×10^{20} at/cm³ in [272] for Boron at 600 °C. The recrystallization rate will depend on the implanted specie [273], the recrystallization temperature, crystalline orientation and stress. In fact the SPER rate increases with the temperature, following an Arrhenius-like law [274]. According to the crystalline plane regrowth direction, one atom (100), two (110) or three (111) are needed to form undistorted bounds [271]. That is why the crystallization velocity is anisotropic and is faster for the <100> than <110> than <111> (with speed ratio of 20:10:1, respectively). Our fabricated devices have a channel orientation of <110>. To conclude, in our case, the SPER rate of $N_{\text{phosphorous}} = 2.10^{20}$ at/cm³ (respectively $N_{\text{boron}} = 2.10^{20}$ at/cm³) at 500°C is 2nm/min (respectively 6nm/min). That is why, we chose to oversize the SPER annealing: 30 minutes at 500°C.

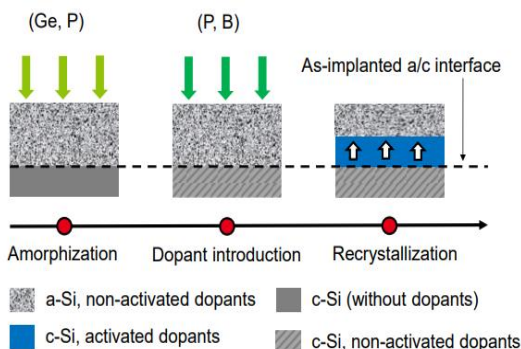


Fig. 180: Illustration of the SPER process taken from [271].

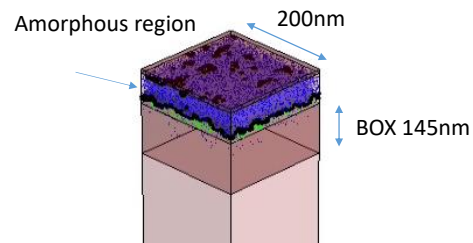


Fig. 181: KMC simulation result: the amorphous depth can be computed with the associated dopant profile.

In the present case, the top layer is around 12nm. Thus, we should experimentally amorphize around 8-9nm and let 3-4nm seed. In the case of stacked layers, the silicon thickness is either 24 or 36nm, so that the amorphization can be deeper. Furthermore, to take all the benefits from silicidation, the silicide

diffusion must be limited to a highly doped region. That is why in stacked layer case, we choose to amorphize the whole top layer. To determine the amorphization thickness t_{amo} we can either use dedicated software to simulate the process or determine it after processing by a TEM observation. Here, we chose Kinetic Monte Carlo (KMC) simulations to compute t_{amo} and the doping profile. An example of KMC simulation is shown in Fig. 181. The different parameters to take into account are:

- Energy (keV): the higher the energy, the deeper dopants will be injected. The energy is tuned in order to meet our t_{amo} and dopant profile targets.
- Dose (at/cm²/s): the dose rate is proportional to the implantation current (tool-dependent) and inversely proportional to the tool-scanning area. For instance at CEA-LETI, the corresponding implantation current for low energy implantation (<20keV) is 500 μ A for Ge and 5mA for P and B. The dose is tuned to obtain the desired dopant profile.
- Tilt ($^{\circ}$): depending of the tilt values, the dopants will encounter more atoms when travelling into the crystalline structure. In our case this value is kept by default at 7 $^{\circ}$.
- Implanted species: for NMOS, the implanted specie is Phosphorous. For PMOS, Germanium is associated to Boron to amorphize the layer.
- Geometry: for our purposes, a bare silicon square (200nm*200nm) of 12nm height (or for stacked structures, 36nm with the associated stack) have been considered. We did not define a transistor structure to study precisely the junction profile since the source/drain – channel interface has the same doping specie. In our case, the junction is rather an n⁺-n than an n⁺-I where the junction is more critical.

The following study have been performed to define the implantation conditions. The chosen energy and dose conditions lead to the Fig. 182 dopant profile and are summarized in Fig. 183. A batch composed of resistance and kelvin cross structures have been fabricated to test these different conditions.

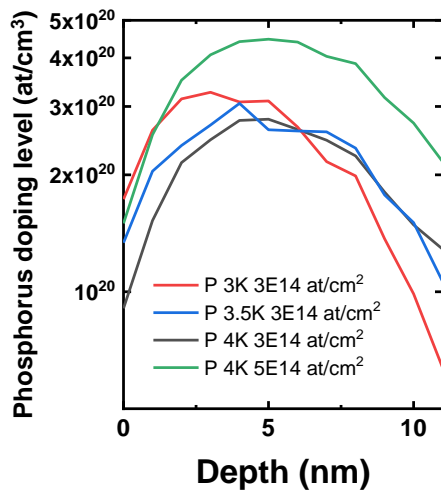


Fig. 182: KMC profile for various phosphorus implantation conditions. The profile shifts in depth for higher energy and for higher dose, the doping level increases.

N channel:

No IMP	x
ac 7.6 nm	P 3K 3E14 at/cm ²
ac 8.54 nm	P 3.5K 3E14 at/cm ²
ac 9nm	P 4 K 3E14 at/cm ²
ac 11nm	P 4K 5E14 at/cm ²
ac 15.6 nm	P 5K 5E14 at/cm ²

Fig. 183: KMC chosen implantation conditions.

We have at our disposal silicon rectangle structures with two contacts of length L_C and spaced from a variable length L (Fig. 184). In this work, the silicon rectangle is composed of either Si:P or SiGe:B 8nm epitaxial layer on top of 4nm intrasec silicon. Then implantation either phosphorus or germanium and boron are performed according of the defined conditions. A SPER annealing at 520 $^{\circ}$ C for 20min is done with different waiting time between the same implantation process and the annealing. The idea was to determine if a time constraint was needed between the two process steps. Also, silicide process is done to get closer to future transistor processing. Fig. 185 shows the measured resistance R as a function of the contact spacing L . From this graph, we can observe a linear dependence between L and

R. The slope of the curve indicates the resistance of the silicon barrel and the resistance for $L=0\mu\text{m}$ is two times the contact resistance R_{co} . In fact R_{co} depends of the contact length L_c . From this specific example, we can observe that the R_{co} depends on the size of the contact $L_c \cdot W_c$ ($W_c=0.09\mu\text{m}$). However, once normalized by the contact area, the remaining value is the same. That is why we will only focus in the study on one contact dimension: $L_c=0.09\mu\text{m}$. Similarly, the extracted silicon resistance R_{si} is the same for the three structures, indicating that the contact shape do not impact the resistance of the barrel.

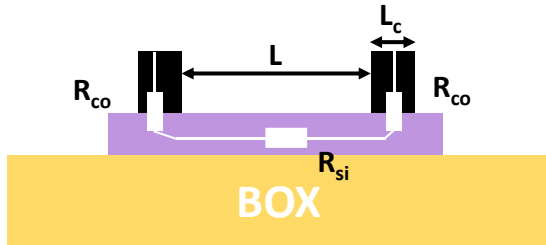


Fig. 184: Electrically characterized structures to evaluate the access resistance.

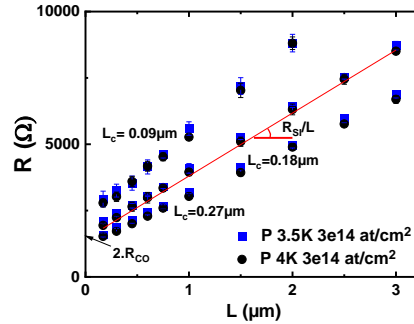


Fig. 185: Resistance as a function of distance between contacts (L) for different sizes of contact length. From this plot, R_{co} and R_{si}/L can be extracted.

Let us analyse now the differences between the implanted conditions for $L_c=0.9\mu\text{m}$. One aspect of this study was to determine if a time constraint between implantation step and annealing was required. The extracted R_{si} presented in Fig. 186 does not depend on the sequencing time between implantation and SPER process, indicating that the recrystallization is the same between these four conditions. From now the mean value between the four samples will be taken to analyze in depth the impact of the different implantation conditions. We will focus on conditions for future nMOS with thin 12nm channel of Si:P (the first four conditions of Fig. 183). The extraction of R_{si} and R_{co} (Fig. 187) for the condition I1 where the contacts are directly done on the substrate, which doping is estimated at 7.10^{18} at/cm³ after epitaxy, indicates high values for both contact and silicon resistance. The goal of the study is to define source and drain implantation conditions to lower these resistances which are in series in a conventional device. This is achieved by using the conditions I2, I3 and I4 which lower both resistances. Thus, with I3 and I4, we were able to provide an additional doping by amorphizing and recrystallizing the channel. It means that the seed thickness $12-7.6= 4.4\text{nm}$ $12-8.54= 3.5\text{nm}$ (KMC) for I3 and $12-9= 3\text{nm}$ were enough. Please note that in a transistor integration the crystalline seed can provide from the transistor channel and is not necessary from source and drain area. However, we can note that the lower the crystalline seed, the lower the resistance is. In fact, even if the implanted dose is the same, a higher energy will amorphize more film and a larger thickness will be recrystallized.

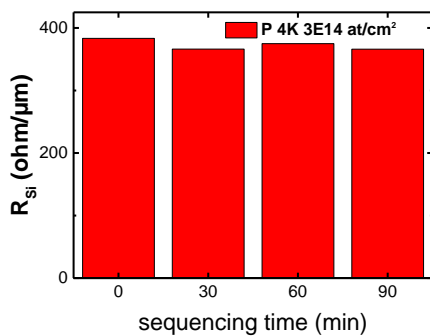


Fig. 186: Extracted R_{si} as a function of sequencing time between implantation and SPER processes.

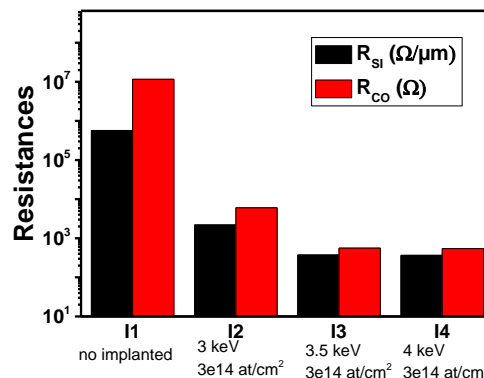


Fig. 187: Extracted R_{si} and R_{co} for the different implantation conditions.

We performed sheet resistance measurements on full sheet wafers and we compared them with the depth of amorphization t_{amo} computed by KMC simulations in Fig. 188. We can observe that a linear dependence is seen between the depth and the sheet resistance for depth range around 7-11nm. For the largest value $t_{\text{amo}} = 15.6\text{nm}$, possible because the 12nm Si:P layer is done by epitaxy on top of a 8nm Si:Ge layer, we do not observe a reduction of the resistance compared to $t_{\text{amo}} = 11\text{nm}$. In parallel, the consumption of the silicon thickness during implantation process have been monitored and for $t_{\text{amo}} = 15.6\text{nm} - 1.1\text{nm}$ is measured whereas for $t_{\text{amo}} = 11\text{nm}$ only -0.3nm is seen. Thus, $t_{\text{amo}} = 15.6\text{nm}$ condition lead to $12\text{nm} - 1.1\text{nm} = 10.9\text{nm}$ Si:P remaining recrystallized film which is similar to the $12\text{nm} - 0.3\text{nm} - 1\text{nm} = 10.7\text{nm}$ recrystallized for $t_{\text{amo}} = 11\text{nm}$. This could explain the saturated value of sheet resistance. To see the impact of the dose, an additional implantation condition (I5) have been measured. I5 features a dose of 5.10^{14} at/cm³ and an energy of 4 keV. The sheet resistance before silicide is in average 360 ohm/sq to be compared to 570 ohm/sq for a dose of 3.10^{14} at/cm³. However this difference can be explained by the difference of amorphization thicknesses. This means that increasing the dose will not increase the doping level but just change t_{amo} . In fact, when considering Phosphorus solid solubility (*i.e.* the electrically active doping level) in silicon matrix extrapolated for $T = 500^\circ\text{C}$ (Fig. 189), we observe that around $8-9.10^{19}$ at/cm³ are activated in best case. As seen in Fig. 182, I4 phosphorus profile is already above this limit. That is why, increasing the dose from I4 to I5 does not increase the doping level but just change t_{amo} , changing the silicon resistance. Based on these measurements, we identified optimized conditions (I4) implantation to lower source and drain access resistance.

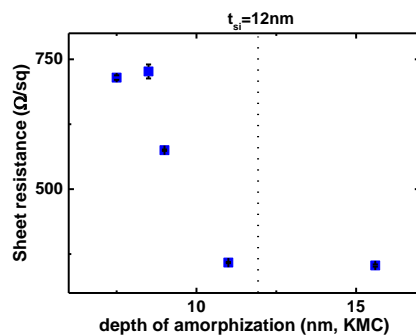


Fig. 188: Sheet resistance (on full sheet wafer) as a function of depth of amorphization computed by KMC simulation.

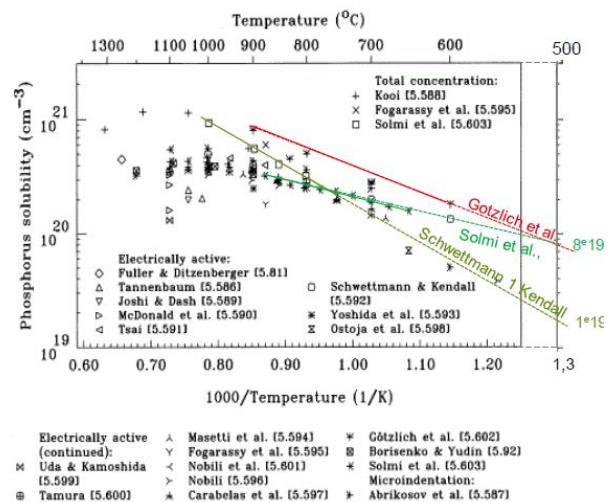


Fig. 189: Solubility of Phosphorus in silicon taken from [275]. The values for 500°C have been extrapolated.

In our fabricated junctionless transistors the SPER module increases the thermal budget from 400°C to 500°C . However, it is not a necessary step for junctionless devices so some devices are done without implantation source and drain and are totally processed below 400°C .

g. Thin silicides

Salicide (Self Aligned siliCIDE) is done prior forming the contact to lower the contact resistivity between metal and silicon [276]. In fact, the NiPt uniformly deposited on the wafer will react with exposed silicon part to form a less resistive phase. The process is described in Fig. 191 and consist in chemical cleaning and NiPt deposition, then a first rapid thermal annealing at 230°C for 20s to form the silicide. The silicide is formed by reactive diffusion. Then the non-reactive NiPt is removed and an additional annealing (390°C , 30s) is done to stabilize the silicide in its less resistive phase. In the junctionless case, since the gate is purely metallic, only the source and drain must be exposed. That is why contrary to IM devices, there is no need to remove the hard mask. The silicide process thermal

budget is below 500°C, 2h. However, as explained in 5-b, the formed silicide must be thin enough ($t_{\text{silicide}} < 10\text{nm}$) not to contact underlying layers. In order to assess t_{silicide} , we deposited on blank silicon wafer different NiPt thicknesses t_{NiPt} and performed different annealing duration and temperature. Then the sheet resistances have been measured and compared to XRR measurements. For instance, the C1 condition leads to silicide depth of 5.7nm, determined by reflectometry. Fig. 191 presents the different results. We observed that small $t_{\text{NiPt}}=2\text{nm}$ results in the same sheet resistance no matter the thermal budget. It means that all the NiPt diffuses. The retained condition is $t_{\text{NiPt}}=4\text{nm}$ and $\text{RTA1}=200^\circ\text{C}$, 20s to ensure film continuity among the wafer.

●	NiPt deposition		C1	C2	C3	C4	C5	C6
●	RTA 1 20s at 230°C		x	x				
●	NiPt removal				x	x		
●	RTA 2 30s at 390°C						x	x
●		RTA1	x		x		x	
●		(°C, 20s)		x		x		x
●		R_s	51.8	51.4	35.5	51	30.3	53

Fig. 190: NiPt silicide process flow.

Fig. 191: Summary table presenting the anneal temperature and duration to achieve a thin silicide film. The retained condition is $t_{\text{NiPt}}=4\text{nm}$ and $\text{RTA1}=200^\circ\text{C}$.

In conclusion of this part, the standard FDSOI gate first process flow have been modified to create low temperature junctionless transistor (down to 400°C without source and drain access optimization). Engineer process development, before and for processing have been done mainly on channel material, gate-stack etch and silicide. Each brick development represents upstream batches to choose the best condition and thus months of fabrication. For the next part of the process, no specific developments are required for the back-end-of line since the temperature is already limited to maintain the silicide stability. The next paragraph will present all the technology variants done for this thesis work to justify the potential interest in junctionless transistors for 3D monolithic integration.

7- Overview of studies related to 3D monolithic integration

Without going into processing details, the main goals of each study is described below. Starting from a baseline (Hot temperature reference), the low-temperature process flow are fully custom and evolve from junctionless low-temperature 1 to 3. The different studies sequentially answer the following questions. What differences can we expect with a heavily doped channel and no additional source and drain implantation (HT-REF)? How can we lower down the temperature of junctionless transistors (Access resistance and JL-LT 1)? What is the impact of channel doping level and source and drain doping level (JL-LT 2 and 3)? Is a change of gate metal work function relevant (JL-LT 3)? Without indications, only NMOS are fabricated.

- **Hot temperature reference (HT-REF):** the main goal is to compare inversion-mode IM and junctionless transistors JL (uniformly doped channel). An additional technical flavor called JAM where the channel is doped and the source and drain are highly doped, is done to decorelate the impact of heavily doped channel and poor access resistance on typical figures of merit. A low temperature brick (SPER for source and drain implantation) is included in JL split to assess future low temperature performance. Otherwise, the process flow used is the gate first FDSOI baseline used in the laboratory with minor modifications.
- **Access resistance:** this study is done for two purposes. The first one is to compare different implantation conditions for source and drain doping at low temperature. It gives also insights

on this importance of doping and annealing succession. The second one is to validate the thin film silicide block. Its maximal thermal budget is 525°C, 30min.

- Junctionless low-temperature 1 (JL-LT 1): this morphological batch is fully realized under 400°C. The aim is to develop the low-temperature process flow without necessarily obtaining electrical values. The main work has been the development of a T-shaped gate without poly-si. Two different size of SiN spacer are tested. Thanks to this prior study, the process flow have been modified to achieve better results.
- Junctionless low-temperature 2 (JL-LT 2): it implements a slightly modified process flow as JL-LT in particular for gate etching. N-MOS (Si:P channel made by epitaxy) and p-MOS (SiGe:B channel made by epitaxy) are fabricated. Different conditions of source and drain implantations are chosen. The maximal thermal budget is 500°C.
- Junctionless low-temperature 3 (JL-LT 3): different technological variants have been implemented to fully study junctionless transistors. They consist in: different channel doping done by implantation to see the impact of N_D , W or TiN liner to study work function impact and source and drain implantation to provide either a 400°C transistor with poor access or a 500°C one with optimized source and drain resistance. It uses the same process flow as JL-LT 2.

The next part will present the electrical results of the different studies.

8- Electrical results

This part presents the experimental electrical results of the fabricated transistors. A comparison between junctionless, accumulation-mode and inversion-mode devices, processed without any thermal budget constraints is done. Unfortunately, no electrical data concerning the low-temperature transistor is presented.

a. Device fabrication

To compare electrically the behavior between junctionless (JL) devices and more conventional inversion modes (IM) ones, we fabricated IM and JL nanowire nMOS down to $W=20\text{nm}$ channel width and $L=15\text{nm}$ gate length (Fig. 192). Fig. 193 presents the process flow, which corresponds to the standard gate first integration flow presented in 6-a. The junctionless devices are made by epitaxially growing an 8nm thick in-situ phosphorous (P) doped Si film on 4nm undoped SOI layer. Excellent crystalline quality is obtained (Fig. 192). After a full transistor process integration the final channel doping level is uniform and equal to 7.10^{18} at/cm³, with our device sizing. For IM devices, the silicon channel is thin down to 12nm. All the devices feature the same gate stack with HfO₂ dielectrics, TiN + poly-Si capping and the same 12nm thick spacer.

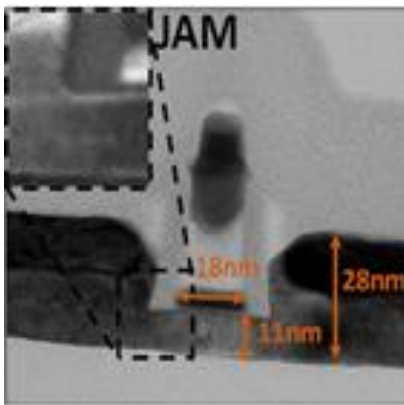


Fig. 192: TEM cross section of JAM device. Transistor have been fabricated down to a $(N+-i-N+)$, JAM $(N+-N-N+)$ and JL (N) gate length of 18nm and width of 20nm.

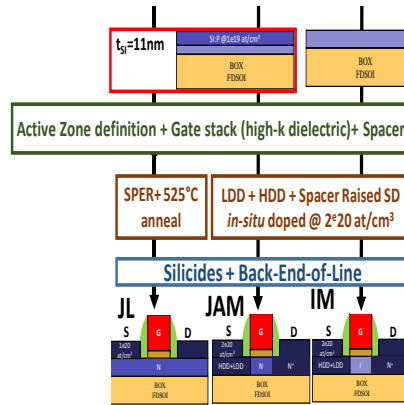


Fig. 193: Detailed Process flow for IM $(N+-i-N+)$, JAM $(N+-N-N+)$ and JL (N) devices.

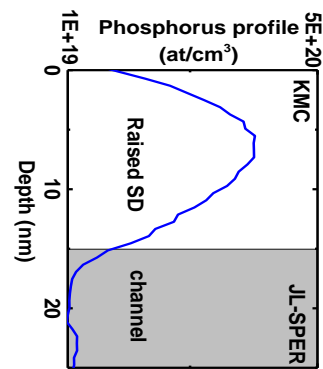


Fig. 194: KMC profile of the phosphorus implantation performed on junctionless (JL) devices. The idea was to dope only the raised source and drain region to maintain a purely junctionless channel.

In order to assess the impact of the channel doping and the source/drain resistance in the junctionless transistor performance, we fabricated also the so-called Junctionless Accumulation Mode (JAM) transistors by adding the same 15nm thin Raised Source/Drain (RSD) and HDD + LDD doping processes as the inversion-mode (IM) references. Both JAM and IM saw a final dopant activation anneal at 1050°C. Purely Junctionless transistors (JL) have no extra doping under the spacer. However, a 5keV P implantation in the 15nm thick RSD followed by a Solid Phase Epitaxy Regrowth (SPER) annealing at 525°C 30 min was carried out, in order to only dope the RSD (see Kinetic Monte-Carlo profile in Fig. 194) and avoid any lateral doping diffusion. This SPER brick gives insights for future JL integrated at low-temperature.

In this work, the color convention is the following: red for IM, black for JAM and dark blue for JL. All the technological variants (IM, JL and JAM) are at least done on two 300nm wafers to ensure repeatability. We will first study the device performance for various dimensions and then tackle digital applications (*i.e.* ultra-scaled devices) and analog (*i.e.* larger) ones.

b. Digital Figure-Of-Merit of Junctionless nMOS

We have at our disposal a large panel of gate length (from $L=10\mu\text{m}$ to $L=18\text{nm}$) combined with different gate width (from $W=10\mu\text{m}$ to $W=20\text{nm}$). In the digital case, most of the performances will be addressed for scaled devices, *i.e.* $W=20\text{nm}$ or $W=240\text{nm}$. In this part, we will first explain the electrical performances and then extract mobility and capacitances, comparing inversion-mode devices and junctionless transistors.

i- Electrical performances

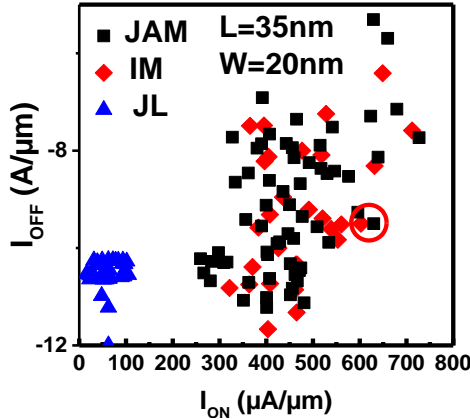


Fig. 195: I_{ON} - I_{OFF} for $L=35\text{nm}$ and $W=20\text{nm}$. JL devices suffer from source and drain access resistance at $V_{DD}=0.8\text{V}$.

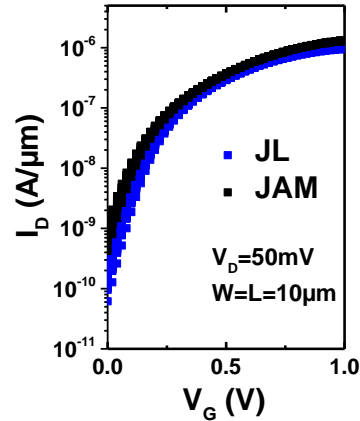


Fig. 196: I_D - V_G for $L=W=10\mu\text{m}$. JL and JAM characteristic matches.

In this digital part, the reference device sizing is $W=20\text{nm}$ and $L=35\text{nm}$, as for the TCAD structure NW-JL. Fig. 195 presents the I_{ON} drain current for $V_G=V_D=0.8\text{V}$ as a function of I_{OFF} ($I_D(V_G=0\text{V})$). For such dimensions, we can see a discrepancy between JL devices and IM-JAM transistors. In fact JL transistors drives much less current. Since JAM performances are equivalent to IM, this lack of current is not attributed to channel doping but rather to access series resistances. In fact as illustrated in Fig. 196 there is no differences between JL and JAM devices for large dimensions ($W=L=10\mu\text{m}$) where the access resistance are neglectable. We extracted the access resistance for $W=20\text{nm}$ and $W=240\text{nm}$ for various gate lengths ($L=18\text{nm}$ to $L=100\text{nm}$). Fig. 197 indicates that IM and JAM transistors retains similar access resistance for all the dimensions, suggesting similar SD implantation. However for JL devices, the extracted R_{SD} is orders of magnitude higher than IM/JAM ones. It can be linked to SD implantation process which is not optimized as IM/JAM one to lower access resistance but rather to detain a uniformly doped channel. This high access resistance can limit the measurements especially for small dimensions where the channel resistance modulated by the gate become small with respect to R_{SD} . We also ensured the quality of the contact (ohmic or Schottky) by plotting the ON current as a function of leakage current (Fig. 95). No correlation is seen between the two quantities indicating that the contact is not Schottky.

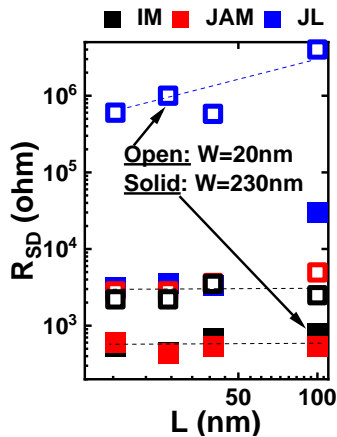


Fig. 197: R_{SD} values for $W=20\text{nm}$ and $W=240\text{nm}$ and L from 18nm to 100nm . JAM and IM transistors shows similar access resistance unlike JL transistors which are orders of magnitude higher.

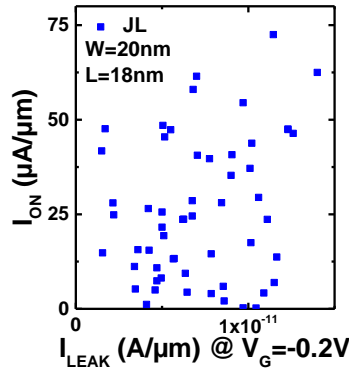


Fig. 198: I_{ON} - I_{LEAK} for JL devices at $W=20\text{nm}$ and $L=18\text{nm}$. No correlation is seen between ON current and leakage current, indicating no Schottky contact.

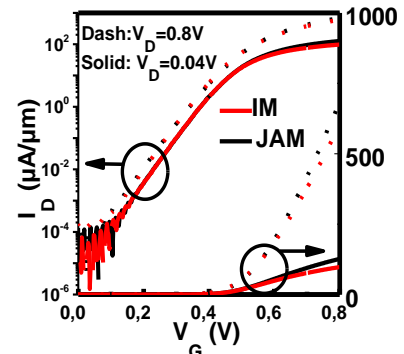


Fig. 199: I_D - V_G for the selected devices on : I_{ON} - I_{OFF} plot . JAM achieves $I_{ON}=560\mu\text{A}/\mu\text{m}$ at $V_{GS}=0.8\text{V}$ and $I_{OFF}=1.3\times 10^{-10}\text{A}/\mu\text{m}$ at $V_{GS}=0\text{V}$.

The I_D - V_G of two selected devices (circled in Fig. 195) IM and JAM is presented in Fig. 199. There is no significant I_{ON} - I_{OFF} difference between IM and JAM devices, reaching up to $I_{ON}=560\mu\text{A}/\mu\text{m}$ at $I_{OFF}=1.3\times 10^{-10}\text{A}/\mu\text{m}$ at $V_{DD}=0.8\text{V}$ supply voltage for the JAM device. Fig. 200 and Fig. 201 show the distribution of JAM leakage current whose mean value is equal to $121\text{pA}/\mu\text{m}$.

At $W=20\text{nm}$, the DIBL and the sub-threshold slope have been extracted for various gate length (Fig. 202). A similar SS and DIBL is seen for JAM, IM and JL transistors (not extracted below 80nm due to high source and drain access resistance) indicating that the channel doping does not degrade the electrostatic control for $W=20\text{nm}$. This is consistent with $R_{ON}(L)$ presented in Fig. 203 showing similar channel and external resistance between IM and JAM FETs for $W=20\text{nm}$. Additionally we extracted the EOT for larger dimension from capacitances measurements (Fig. 204). A similar EOT of 1nm is extracted for all devices, suggesting a similar gate stack. In conclusion, for $W=20\text{nm}$, there is no electrostatic differences between doped channel devices and undoped one, only the source and drain optimization matters.

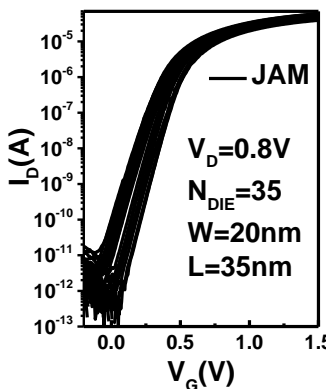


Fig. 200: I_D - V_G for 35 JAM devices.

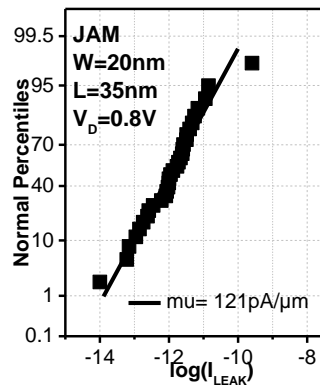


Fig. 201: I_{LEAK} distribution.

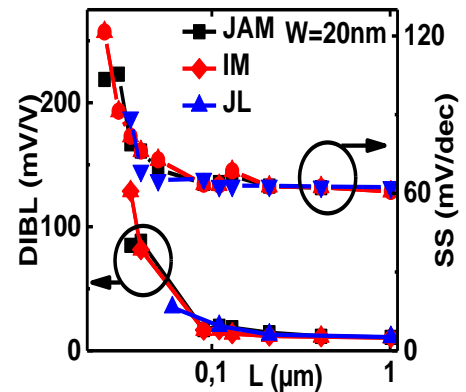


Fig. 202: Measurement of DIBL and SS (in the linear regime) as a function of L_g .

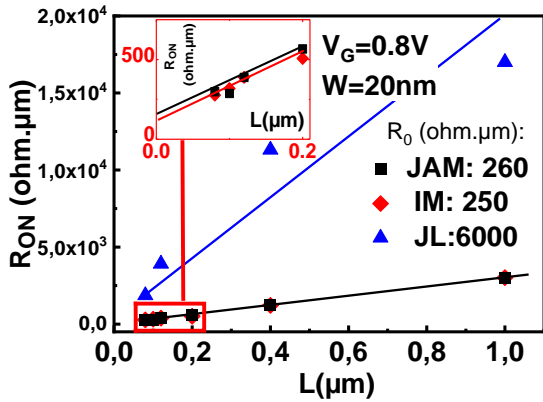


Fig. 203: Measurement of R_{ON} as a function of gate length L for $W=20nm$ and $V_G=0.8V$.

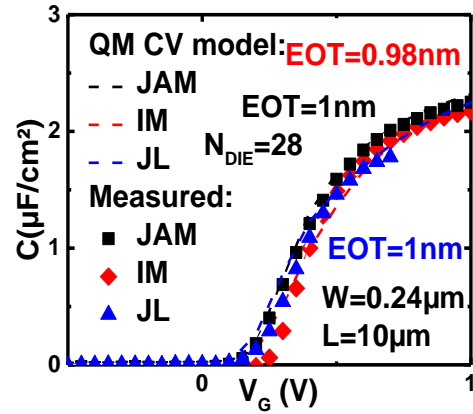


Fig. 204: Gate capacitance vs. V_G . Similar $EOT=1nm$ is measured for all devices.

ii- Mobility

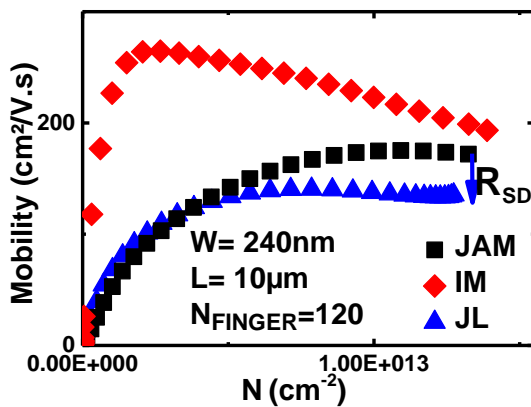


Fig. 205: Mobility vs. carrier charge density. The mobility of heavily doped channel device is lower as expected and a further degradation due to R_{SD} is seen for purely junctionless devices.

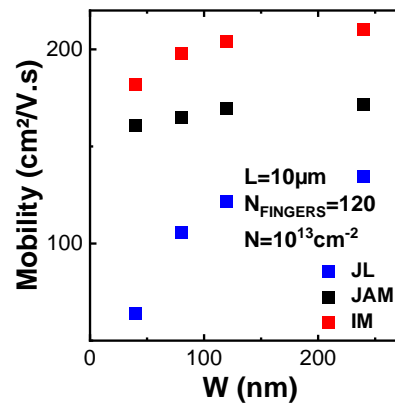


Fig. 206: Mobility vs. device width at fixed carrier charge density ($10^{13}cm^{-2}$). A degradation due to R_{SD} is seen for purely junctionless devices and is more pronounced as the width is small.

Nevertheless, a well-known drawback of junctionless transistors is the lower electron mobility with respect to undoped channel. As compared to IM devices, Fig. 205 shows that the JAM and JL mobility is mainly impacted at low carrier charge density, evidencing Coulomb scattering due to high channel doping as explained in section 4-b. However, for large carrier charge density the discrepancy between doped channel devices and undoped one reduces drastically. For instance at $N=1.3 \times 10^{13} cm^{-2}$ carrier density, only a 9% mobility degradation is measured for JAM, compared to IM. In standard devices, the reduction of mobility for large carrier charge density is attributed to surface roughness scattering. For junctionless devices the conduction occurs in the volume for moderate electric field and could explain a lower degradation at high carrier density. However, for higher electric field, a surface accumulation layer is formed pushing the conductive channel towards the surface. But in junctionless devices unlike in inversion-mode one, the high concentration of majorities carriers (e^- for nMOS) forms a neutralizing screen around positively charges ionized donor atoms, reducing their scattering cross section and thus Coulomb scattering. Thus, this lower degradation at high carrier density could be explained by the bulk conduction, impurity screening and lower surface roughness scattering. Please note that for JL devices, a slight mobility degradation is seen and attributed to the access resistances. Furthermore this degradation is more important for smaller width as highlighted in Fig. 206. For JAM and IM transistors,

the mobility difference tends to reduce with width even if there is still a gap between the measured values. However, this large channel mobility degradation is not translated into $I_{ON}-I_{OFF}$ for ultra-scaled dimensions ($W=20\text{nm}$ and $L=18\text{nm}$).

iii- Overlap capacitance

We measured capacitances (Fig. 207) at $V_G=-0.4\text{V}$ and we observed a linear dependence between the capacitance and the transistor width. From this curve, the C_{GDS} (expressed here in $\text{fF}/\mu\text{m}$) is extracted for $L=35\text{nm}$ (Fig. 208). JL transistors has a $0.06\text{fF}/\mu\text{m}$ lower C_{GDS} , which cannot be explained only by the fringe components (Fig. 209) but rather by a depletion region extended below the spacer for JL transistors.

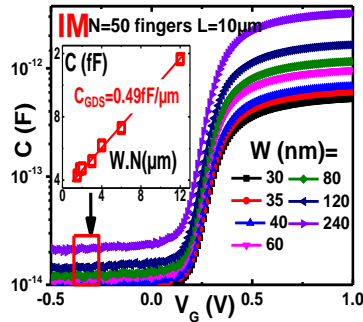


Fig. 207: Gate to channel capacitances for IM devices at various W (and $L=10\mu\text{m}$). C_{GDS} is extracted at $V_G=-0.4\text{V}$.

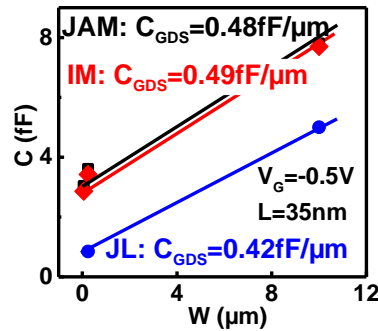


Fig. 208: $C(W)$ and C_{GDS} extraction at $L=35\text{nm}$. JL transistors show a lower capacitance C_{GDS} than IM and JAM. It is attributed to the absence of junction.

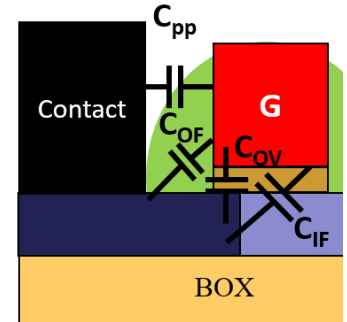


Fig. 209: Schematics with parasitic capacitance contributions. (reminder from part 4-c).

In this part, we studied in depth the differences between junctionless devices and inversion-mode ones for digital applications. A lower mobility is seen for doped-channel devices but it is not translated into the $I_{ON}-I_{OFF}$ FOM for $W=20\text{nm}$ and $L=18\text{nm}$. Excellent performances and electrostatic control are seen for JAM and IM devices. However, for JL devices, the performances are limited by the high access resistances making them not suitable for advanced digital applications. Nevertheless with the appropriate source drain optimization, doped channel devices are good candidates for digital applications.

c. Analog Figure-Of-Merit of junctionless nMOS

In this part, analog performances are analyzed. For such applications, the transistor dimension is less critical since current drive is more important than density. If no contrary indication the transistor width is 240nm and L from 80nm and 10 μ m. First, analog gain comparison between devices is presented with the use of back-bias to enhance it. Secondly, the reliability and the noise are extracted before ending by RF measurements.

i- Analog gain leveraged by back-bias

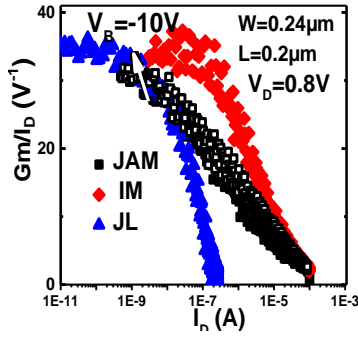


Fig. 210: G_m over I_D as a function of I_D for $W=0.24\mu\text{m}$ and $L=0.2\mu\text{m}$. JL and JAM plateau is slightly lower than IM one.

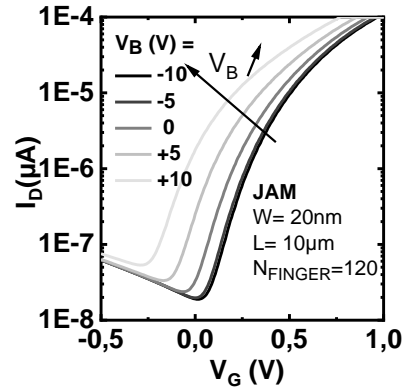


Fig. 211: I_D - V_G at various back-bias for JAM transistor at $W=20\text{nm}$ and $L=10\mu\text{m}$. A negative back-bias improves the SS slope.

For analog applications, we consider a nominal analog transistor of $W=0.24\mu\text{m}$ width, (planar SOI configuration instead of a trigate nanowire structure chosen for digital part). A well-known analog figure of merit is the A_{v0} gain (in dB) defined as $A_{v0}=20 \log(g_m/g_d)$. g_m is the transconductance and equals to $\frac{\partial I_d}{\partial V_g}$ and g_d is the output transconductance defined by $\frac{\partial I_d}{\partial V_d}$. Fig. 210 presents g_m/I_D as a function of I_D where IM plateau for low values of I_D is slightly higher than JAM and JL one and close to the ideal one. This can be explained by IM subthreshold slope of 61mV/dec vs. $SS=64\text{mV/dec}$ for JL/JAM (SS)

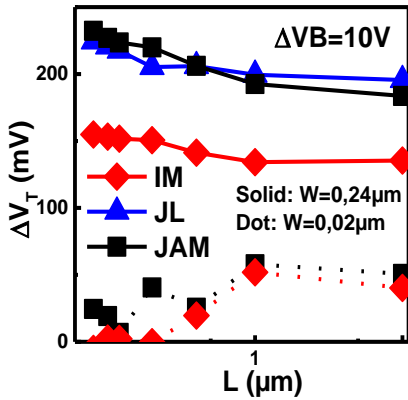


Fig. 212: Back bias efficiency for $W=0.02\mu\text{m}$ and $W=0.24\mu\text{m}$ as a function of L . Wider devices are more sensitive to back-bias.

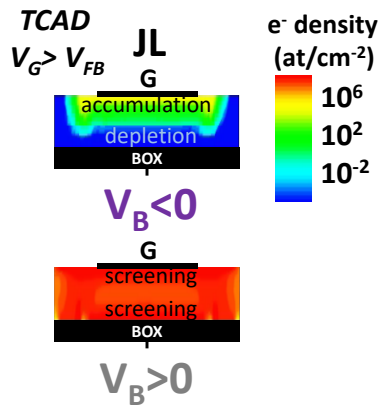


Fig. 213: TCAD simulation to see the impact of back-biasing on junctionless devices. Electron density cut are provided for $V_G > V_{FB}$.

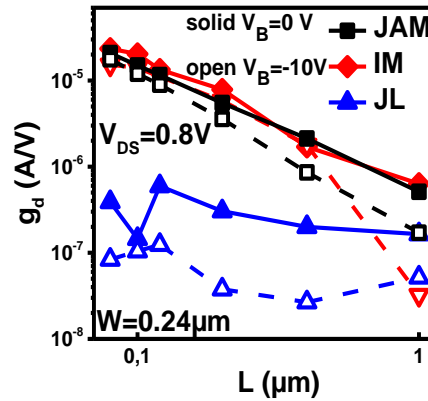


Fig. 214: g_d as a function of L for $V_B=0$ and $V_B=-10\text{V}$. JL devices have a lower g_d attributed to the extension of the depletion region in source and drain region.

devices for such dimensions. To gain on SS , we propose to use back-bias to adjust the threshold voltage and tune performance [277]. We can observe on Fig. 211 that a negative back-bias improves the SS for JAM transistors. Fig. 212 presents the V_T variation for 10V back-bias variation (BOX thickness is 145nm). We observe that back-bias is more effective for wider ($W=240\text{nm}$) than for narrower devices

($W=20\text{nm}$) and it is more effective on JL/JAM than on IM transistors. Markedly, a negative back-bias applied on JAM moves the bulk conduction channel upwards towards the gate (as simulated by TCAD in Fig. 213), which results in an improvement of the electrostatic control. Not only the subthreshold slope is improved but also the output conductance g_d as seen in Fig. 214. Also JL devices have a lower g_d and thus a higher early voltage E_a due to the extension of the depletion in source and drain region. Fig. 215 recaps the gain with and without back-bias for specific geometries. As a result, JAM FETs reach analog performances that are slightly better than IM devices, up to an $A_{v0}=20 \log(g_m/g_d)=68.8\text{dB}$ gain.

Gain A_{v0} (dB)	V_B (V)	L=100 nm	L=200 nm	L=400 nm
IM	0	1	60	50
	-10	18	65	53
JAM	0	1	12	51
	-10	64.5	68.8	55
JL	0	59	47	38
	-10	61.5	65	41

Fig. 215: Gain A_{v0} for different gate lengths and $W=0.24\mu\text{m}$. $V_B<0\text{V}$ improves the analog gain.

ii- Reliability and noise

We have performed Positive Bias Temperature Instability (PBTI) and Hot Carrier Injection (HCI) measurements. Fig. 216-a presents the degradation of threshold voltage for a stress time t_{stress} equals to 300s and a stress voltage applied on the gate V_G from 1.2V to 2V at $T=125^\circ\text{C}$. We can see that JAM and IM have similar degradation. For these devices, we extrapolated a similar PBTI (88 years lifetime at $V_{DD}=0.8\text{V}$) for IM and JAM devices, demonstrating a negligible impact of the channel doping. The power-law extrapolation fits well the data points (Fig. 216-b). However, the JL threshold voltage shift is not sufficient for Time-To-Failure extrapolation. We speculate it may be due to the thermal budget difference, mainly due to the 1050°C spike annealing absence. Fig. 217 presents the HCI test performed at 125°C and for drain voltages ranking from 1.2 and 2V. Better HCI is measured for JL as compared to IM and JAM. It can be explain by a lower and shifted to the drain (not underneath the gate dielectric as for IM/JAM) peak electric field. The five working years industrial criteria is met for both PBTI and HCI.

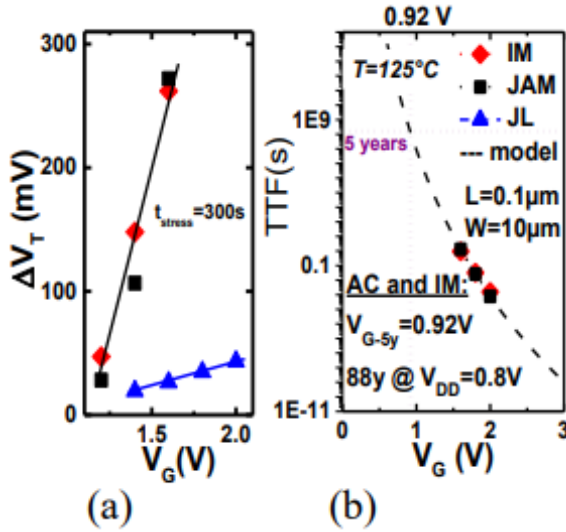


Fig. 216: (a) ΔV_T as a function of V_G for a 300s stress time for $L=0.1\mu\text{m}$ and $W=10\mu\text{m}$ nMOS. (b) Time-To-Failure for PBTI. The 5-year criterion is met and up to 88 years reliability is seen at $V_D=0.8\text{V}$.

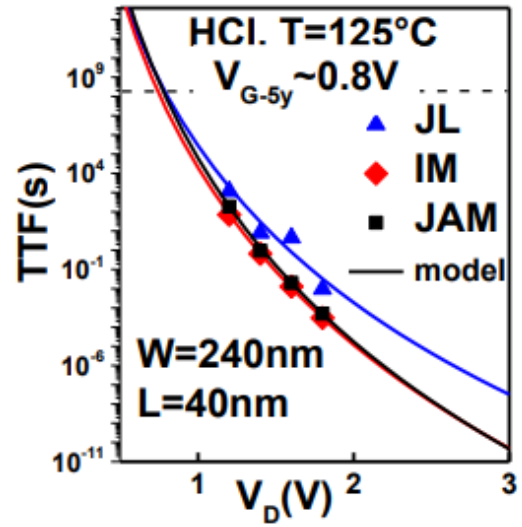


Fig. 217: Time-To-Failure for HCl for $W=240\text{nm}$ and $L=40\text{nm}$ nMOS. JL devices are less degraded than IM and JAM. It can be explained by the deported electric field peak. The 5-year criterion is met for all devices at 0.8V .

For analog applications, the ratio signal/noise quantifies how the signal can be differentiated from the background noise. In an ideal case, the noise must be kept as low as possible. To ensure that the presence of a doped channel do not degrade this figure of merit, we have measured low-frequency drain current noise (Fig. 218). They show a 31-die average $1/f$ signature and a slightly lower input-referred gate voltage noise level (SV_g) for JAM.

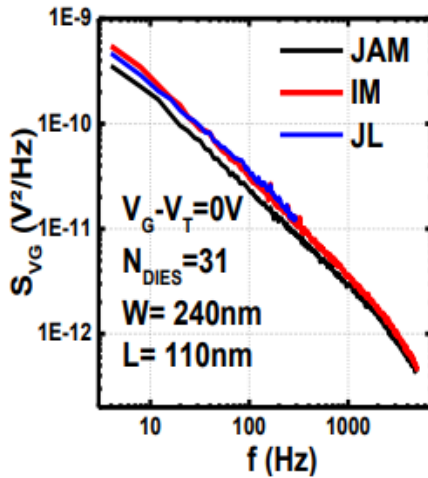


Fig. 218: Input-referred gate voltage power spectral density versus frequency at $W=240\text{nm}$ and $L=110\text{nm}$. JAM shows slightly less low-frequency noise than IM and JL.

Carrier number model + Carrier Mobility Model + R_{SD}

$$\frac{S_{I_d}}{I_d^2} = \left(\frac{gm}{I_d}\right)^2 \cdot S_{vfb} (1 + \alpha) \left(\frac{I_d}{gm}\right)^2 + S_{Rsd} \left(\frac{I_d}{V_d}\right)^2$$

$$S_{vfb} = \frac{q^2 kT \lambda N_t}{W L C_{ox} F^V}$$

$$\alpha = \alpha_{sc} \mu_{eff} C_{ox}$$

$$S_{Rsd}$$

With N_t = volumetric oxide trap density
 γ characteristic exponent ~ 1
 λ tunnel attenuation distance $\sim 0.1\text{nm}$

Fig. 219: Description of the model used to fit the drain low-frequency noise measurement. The source drain excess noise is considered. Three parameters can be extracted: N_t the oxide trap density, α_{sc} the remote Coulomb scattering coefficient and S_{Rsd} the contribution of SD excess noise.

We used the Carrier number fluctuations with Correlated Mobility Fluctuations model explained in [278], and Fig. 219 taking into account the series resistance noise (SR_{sd}). We fitted the normalized drain current noise at $f=10\text{Hz}$ (Fig. 220) to extract the volumetric oxide effective trap density N_T , and the remote Coulomb scattering coefficient α_{sc} for all wafers. We extracted a value of $N_T \approx 7.5 \cdot 10^{17} \text{eV/cm}^3$ for all cases, reflecting a similar interface quality, independently of the conduction mode. This value is also very close to state-of-the-art N_T values of high-k-metal-gate CMOS technologies [279]. Concerning α_{sc} , a very similar value ($\approx 4 \cdot 10^3 \text{Vs/C}$) is extracted for all wafers, showing that the remote Coulomb scattering is not affected by the different conduction modes. Finally, SR_{sd} has a significant impact only

for JL, which can be linked to non-optimized source/drain doping [280] and confirms previous discussion about the impact of access resistance.

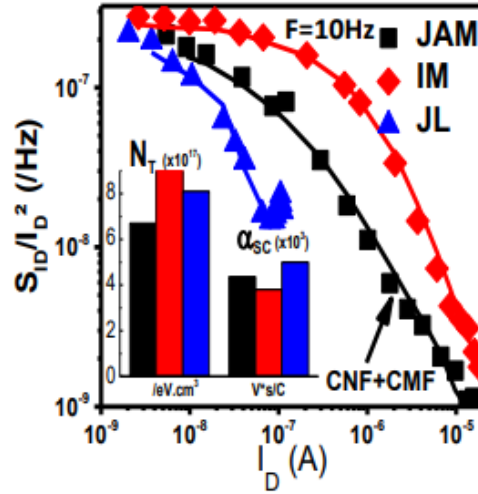


Fig. 220: Normalized drain current power spectral density versus I_D at $W=240\text{nm}$ and $L=110\text{nm}$. Inset: Extracted values of N_t and α_{sc} .

iii- RF Figure-Of-Merit of junctionless nMOS

Two metrics representing RF performances are the maximum operating frequency f_{\max} and the cut-off frequency f_T . f_T is defined as the frequency for which the current gain equals unity. For instance f_T is the maximum useful frequency for amplifiers. Its expression is given by Eq. 22 and is proportional to g_m . However, the maximum operation frequency f_{\max} is inversely proportional to g_{ds} and C_{gd} , The equivalent circuit used to extract f_T and f_{\max} is presented in Fig. 221.

$$f_T = \frac{g_m}{2 \cdot \pi \cdot C_{gs}} \quad \text{Eq. 22}$$

$$f_{\max} = \frac{f_T}{2 \cdot \sqrt{g_{ds}(R_g + R_s) + 2 \cdot \pi \cdot f_T \cdot R_g \cdot C_{gd}}} \quad \text{Eq. 23}$$

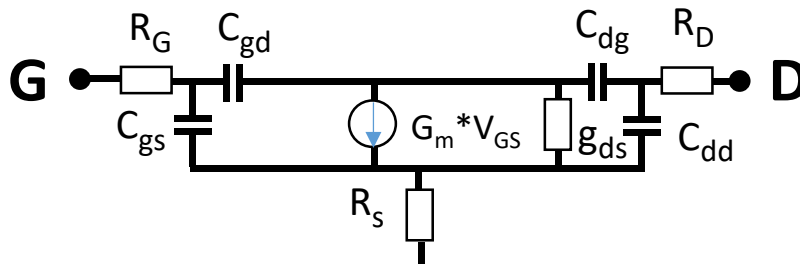


Fig. 221: equivalent circuit of the measurement setup.

RF measurements have been performed on IM and JAM transistors (Fig. 222 and Fig. 223) for $W=240\text{nm}$ and $L=30, 40$ and 60nm . We measured the cut-off frequency at $f_T=130$ GHz for JAM vs. 136 GHz for IM for $W=240\text{nm}$ and $L=30\text{nm}$. Based on Eq. 22 this can be explained since junctionless transistors detains a lower g_m (lower mobility) than inversion mode one. But JAM exceeds IM devices in terms of f_{\max} . It is attributed to a lower parasitic capacitances C_{GD} . In fact, only a few GHz are

compromised for f_t when using JAM allowing to obtain better f_{max} . We here measure a record $f_{max}=182\text{GHz}$ for junctionless nMOS.

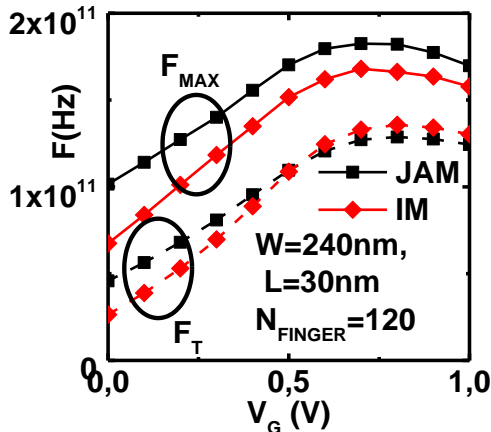


Fig. 222: F_{MAX} and F_T as a function of V_G . f_{max} is higher for JAM than for IM contrary to f_t for $W=240\text{nm}$ and $L=30\text{nm}$.

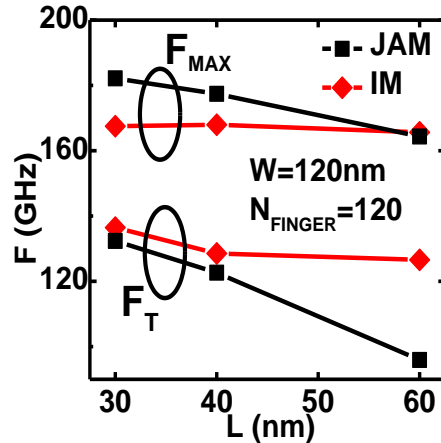


Fig. 223: F_{MAX} and F_T comparison for different L_G . The JAM gain on f_{max} , is more pronounced at small dimensions.

In this part, the interest of junctionless devices for analog applications is raised. In fact, they detains a superior reliability and RF capability than IM devices for similar noise and analog gain.

To conclude the comparison between junctionless, junctionless-accumulation-mode and inversion-mode devices, the main characteristics to bear in mind are:

- JL transistors suffer from a high access resistance, impacting mobility, ON-current and high field variability. However, for analog application its output transconductance is one order of magnitude lower than the others devices, leading to a good gain which can be modulated by back-bias. The overlap capacitances is two times lower than IM. Superior HCI reliability is seen for JL than IM and JAM and is attributed to the electric field peak shift to the drain.
- JAM transistors have optimized source and drain resistance and despite a lower mobility especially at low field, they feature similar performance as IM devices for scaled dimensions. A slight gain on gate-to-drain capacitances is translated on a maximum operating frequency gain (the maximum frequency reaching 182GHz). The electrostatic control is depreciated for large width, compared to narrow devices, as predicted by TCAD. However, the analog gain of JAM devices ($Av_0=68.8$ dB with back-bias) outperformed the IM one.

9- Conclusion of Chapter 3

This chapter presented the assets of junctionless transistor for 3D monolithic integration, their fabrication and electrical performances.

In a first time, **bibliography research and TCAD simulation** showed the interest of such a device for low temperature integration, featuring good performances. Different specificities with respect to inversion-mode devices have been highlighted. I would like to put the emphasis on the **mobility and variability degradation** seen by junctionless transistor due to their heavily doped channel. However, due to an absence of source to channel junction, a lower Miller **capacitance** is predicted as well as a **resilience to HCI**, making such a device attractive for **analog/RF applications**.

In a second time, the gate first FDSOI fabrication process is exposed. The modifications done to lower the **thermal budget down to 400°C** are explained. In particular, challenges concerning the **gate etching and silicides** are presented. Several process optimizations, taking into account of JL particular operation, are explained. Our choices are motivated either by simulations or by preliminary batches. Different process technology variants are implemented to study the impact of channel doping, source and drain resistance and metal gate work function.

In a third time, the **electrical results** associated to the fabricated batches are presented. The comparison between inversion-mode, junctionless accumulation-mode and purely junctionless devices without temperature constraints agreed with previous simulations. The **degradation** of mobility and transconductance is shown experimentally. Furthermore, the **gain** on typical analog figures of merit such as **f_{\max} or A_{v0}** is demonstrated.

I would like to indicate precisely what work is mine. I performed all the TCAD simulations with the help of the simulation laboratory, especially to define the proper structure and conditions. I modified the gate first FDSOI process flow to lower down the process temperatures. Such modifications required upstream work of integration experts who gave me all the insights for technological choices. As far as batches processing is concerned, this work belongs to cleanroom technician, expert and my reactivity in case of problems. Most of the process characterization (such as thicknesses measurements or SEM pictures) are done by dedicated people. Validation of technological steps or non-standards measurements are mine. Regarding the electrical characterization, if not indicated, I realized the measurements and analyzed the data with the help of CEA characterization laboratory and integration laboratory. Concerning the RF measurements, J. Lugo and R. Youcef did all the characterizations and post treatment. For the noise measurements, I would like to thank IMEP-LAHC platform and team, in particular the PhD student A. Tataridou and C. Theodorou. Without all these people help, my work would have been limited.

As stated in the introduction, the next chapter will go further to propose an In-Memory Computing (IMC) solution based on the co-integration of Junctionless transistors and Resistive Random Access Memory (RRAM). Such a high density architecture can overcome the so-called memory wall by gathering computation and memory units.

Chapter IV: Assessment of an ultra-dense Non-Volatile Memory cube for In-Memory Computing applications

In-Memory Computing (IMC) is foreseen as an alternative to the traditional transistor scaling to break the so-called “Memory Wall”. In fact, gathering the memory and computation part enables improving delay and energy by reducing the data transfer. The aim of this chapter is to propose an ultra dense 3D structure, called MY-CUBE, which gathers a memory and computation part. The first section of this chapter consists in a literature review of exiting IMC solutions and our choices. The second section, based on TCAD and SPICE simulations, demonstrates the IMC feasibility in such a structure. The third one presents the process flow for MY-CUBE as well as a planar variant integration. The last section tackles the topic of variability in all operation regime, to verify if junctionless transistors are compatible with such an application.

1-	State of the art of In-Memory-Computing existing solutions	138
a.	Existing In-Memory Computing implementations	138
i.	Memristors for IMC	138
ii.	Boolean logic	139
b.	IMC existing solutions: examples	141
c.	IMC materials/ selectors	142
i.	Memristors materials	142
ii.	Focus on OxRAM technology	143
iii.	Selectors	145
d.	My-Cube project: choices	146
i.	Stacked nanowires	147
ii.	Memory element	147
iii.	Boolean logic: Scouting logic	147
2-	Sizing simulations	149
a.	Simulated pillar structure	149
b.	Definition of SPICE simulation inputs	149
i.	JL performances at W=50nm	150
ii.	Drive current for stacked nanowires at W=75nm	151
iii.	OxRAM distribution extraction	153
c.	Scouting logic in the pillar	154
d.	MY-CUBE: read and write schemes	154
3-	Processing of stacked structures	156
a.	Gate-All-Around stacked nanowires detailed process flow	156
b.	Modification to standard process flow to integrate memory elements	157

4-	Variability	159
a.	Standard evaluation of the mismatch: Pelgrom plots	159
b.	Gate input referred normalized matching parameter	161
c.	Drain current local and global variability in all-regimes	163
d.	Variability of JAM devices for IMC	165
5-	Conclusion of chapter IV	166

1- State of the art of In-Memory-Computing existing solutions

In widely used Von-Neumann architectures, the data is stored in memory and is transferred to computational blocks, resulting in around 50-80% energy waste for memory access [281]. This challenge, called “Memory wall”, have been already addressed in computing systems. For instance, multi-core processors increases the parallelism and thus reduces data latency. However, even with multi-core processing, part of the chip cannot be used due to power restriction, the so-called “dark silicon”, which is predicted to represent around 21% of the chip at 22nm node [282]. To handle the future “data deluge” coming from IOT and 5G, alternative computing paradigms emerge. One promising solution consists in gathering memory and computational parts in a circuit, breaking the conventional Von-Neumann architecture. This new computing paradigm is called In-Memory Computing (IMC) and promises substantial gains on data energy and latency. In this part we will dress the state of the art of the existing solutions. We will first explain the relevance of memristors for IMC and the different computation/logic available. Then we will address the different materials (especially in the memory side), which are available to perform IMC. Afterwards, some state-of-the-art solutions will be discussed. To finish with, MY-CUBE device structure is detailed with the different aspects to tackle in order to enable IMC.

a. Existing In-Memory Computing implementations

In this subsection, computing solutions, including Boolean logic, enabling IMC are presented in a first time. In a second time, the memory materials will be discussed with an emphasis on Oxide-based RAM technologies. The last part will present the proposed structure, which will be analysed in this chapter.

i. Memristors for IMC

Memristors detain many advantages such as CMOS process compatibility, zero standby power, great scalability and high density of integration, enabling new computing paradigms [283]. The memristor consists in a two terminal elements (bottom electrode and top electrode) which detains a hysteresis loop as illustrated in Fig. 224. All of the 2-terminal non-volatile memory devices fit into this category. Based on this structure, several designs have been proposed and can be classified according to the type of operation performed. Thus “Boolean logic” can be dissociated from “Implication logic” and from “threshold/majority” one. The two later ones use voltages to represent data, making Boolean logic more appropriate for IMC applications.

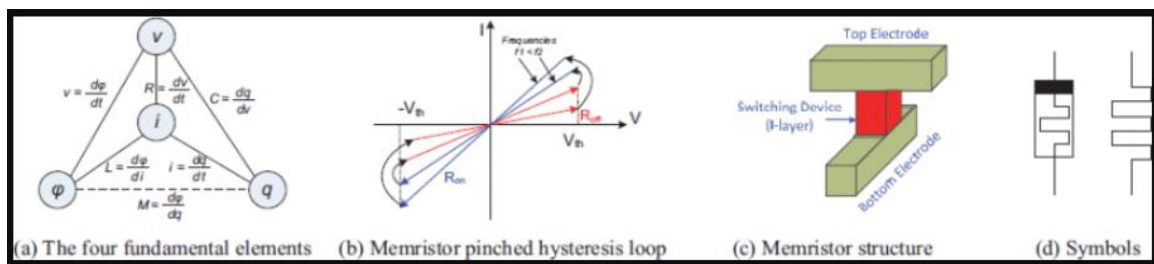


Fig. 224: Memristor introduction: relations between V , I , q and φ , hysteresis loop, structure and symbols (Figure from [283]).

- **Boolean logic:** the idea is to perform the conventional mathematical algebra primitives, like conjunction (AND), disjunction (OR) and negation (NOT). Different implementations are proposed in the literature. Vourkas *et al.* [284] proposes a memristor crossbar circuit where the logic gates are implemented by replacing the standard CMOS pull-up and pull-down network with memristors. In [285], Memristor Rationed Logic proposes OR and AND gates based on memristive logic and mixed with CMOS inverter to avoid the additional circuitry required for

the memristive network/CMOS compatibility. However there is also memristor only logic family. For instance, memristor-aided logic (MAGIC) is presented in [286] where memristors are inputs with previously stored data and an additional memristor serves as an output. To finish with, by using appropriate signals on a crossbar array, it is possible to implement Boolean functions. Xie *et al.* [287] demonstrate that only 7 steps are needed to implement any Boolean function.

- **Implication logic:** $p \text{IMP} q$ “p implies q” is equivalent to “if p then q”. In fact, IMP and FALSE operation form a computationally complete basis. Two main families of memristors can be distinguished. The first one is called stateful logic [288], [289] where the logic state is represented by the resistance of the memristor. The second one is called “complementary Resistive Switch” [290] and relies on the combination of antiseriial resistive switches in a passive crossbar array to avoid the sneak path currents through neighbouring cells.
- **Threshold/Majority logic:** threshold logic relies on the assembly of threshold gates where the output changes if the arithmetic sum of weights inputs exceed a threshold. Majority logic is the particular case where all the inputs are binary and the weights are equals. Among this logic category, we can distinguish programmable CMOS/Memristor logic [291] and Hybrid current mirror logic [292]. For current mirror approach, the weights are represented with memristance so that Ohm’s law converts voltage signal inputs into current which can be summed and compared to a threshold current $I_{\text{threshold}}$. Current mirror are used to perform the full operation (weights, sum and comparison). As far as programmable CMOS/Memristor logic is concerned, memristive devices implement ratioed diode-resistor logic and CMOS logic is used for signal amplification and NOT gate.

From my point of view, the threshold/majority logic requires several threshold elements (CMOS or current mirror) which will be difficult to integrate and expand in the third dimension. In the case of Boolean logic the resistance states (low or high) are used to represent the conventional ‘0’ and ‘1’ logic state. Several crossbar implementation have already been demonstrated, indicating the ease of fabrication for 3D structures. That is why, in the next part we will focus on Boolean logic, giving the example of the so-called MAGIC and Pinatubo/ Scouting logic approaches.

ii. Boolean logic

In this part, we will expose two representative examples of Boolean logic which are MAGIC and Pinatubo/Scouting logic.

In the MAGIC approach, the inputs and outputs of logic gates are the logical states of the memristors (high ‘0’ or low ‘1’). Different memristors for inputs and outputs are needed and the logic gate output is the final logical state of the output. Fig. 225-a presents the example of a NOR gate composed of two inputs in_1 and in_2 and one output. The initialisation step consists in writing a low resistance value ‘1’ into the output and if necessary write the correct inputs values. Then the evaluation is performed by applying a voltage pulse V_0 at the Gateway. If $in_1=in_2='0'$ (high resistance state), no current flows though the memristor (or is lower than memristor output threshold) and the output is left at ‘1’. If either in_1 or in_2 is in a low resistance state ‘1’, the current will flow though and switch the output state to ‘0’. However, we do not want to change the input values. For this, the memristor threshold ($V_{T,OFF}$ and $V_{T,ON}$) and the chosen voltage pulse V_0 must verify $V_0 < \min [R_{OFF}/R_{ON} \cdot V_{T,OFF}, V_{T,ON}]$. Similarly, to switch the output, $V_0 > 2V_{T,OFF}$. The realized function is an NOR gate and can be extended to a number N of inputs. Additional MAGIC gates, NOR, NAND, OR, AND and NOT are presented in [286]. For instance, the topology of the NOR gate can be used to create an OR gate with an initialization of the output to ‘0’. This approach is efficient for IMC but relies on writing operations which can be detrimental for the memristor component.

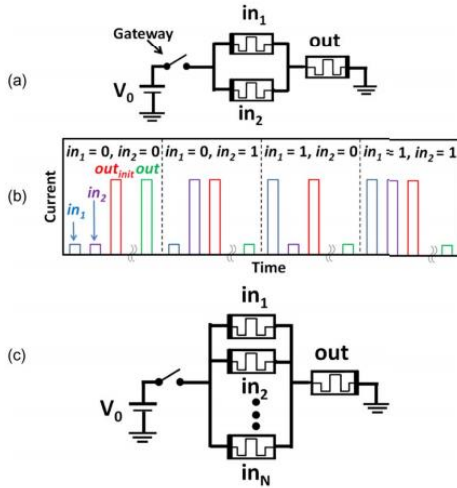


Fig. 225: (a) Schematics of a two inputs NOR gate composed of two inputs memristor in_1 and in_2 and an output memristor out . (b) Simulations of a two input NOR gate for all the combinations. (c) Extension to N inputs. Reproduction from [286].

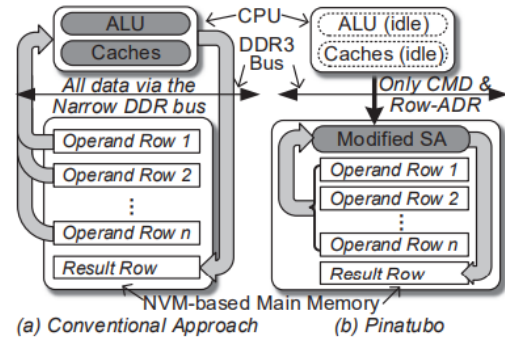


Fig. 226: Modification of the Von-Neumann architecture. By modifying the Sense Amplifiers, bitwise operations are performed. Figure from [293].

As far as the Pinatubo / Scouting logic approaches are concerned, Shuangchen Li *et al.* [293] proposed Pinatubo, a Processing In Nonvolatile memory ArchiTEcture for bUlK Bitwise Operations, including OR, AND, XOR, and INV operations. Instead of integrating logic into the memory, Pinatubo redesigns the read circuitry to compute the bitwise logic of two or more memory rows (Fig. 225). In fact, the sense amplifier is designed to distinguish the resistances $HRS||HRS$, $LRS||HRS$ and $LRS||LRS$ for two rows. If the memory window is high enough, it can support multi-row OR operations. Since the working principle is similar to Scouting logic, it will be detailed in the next paragraph. Concerning the performance, $\sim 500x$ speedup and ~ 28000 energy savings are seen on bitwise operations and in overall 1.12 speedup and 1.11 energy savings for the processor [293].

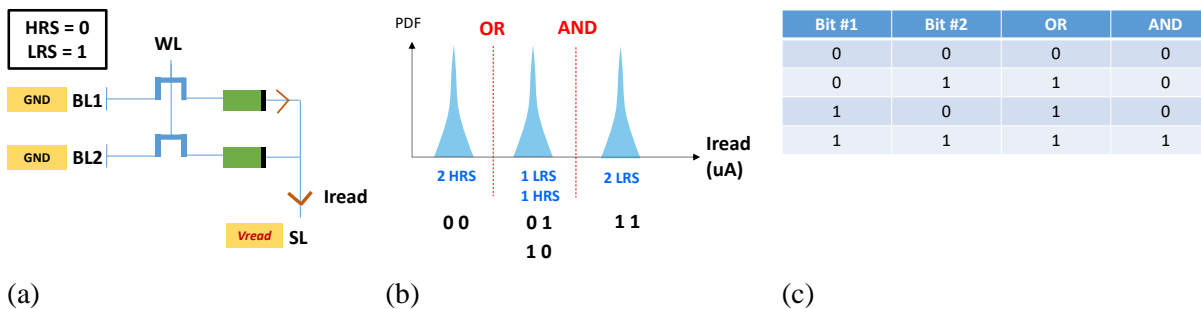


Fig. 227: (a) Two 1T1R cell electrical schematic (b) cumulative distribution of read current (c) Truth table of the Boolean operation.

The principle of Scouting logic (SCL) is depicted in Fig. 227. Similarly to Pinatubo, the main idea is to perform Boolean operation on the resistive states (low, ‘1’ or high, ‘0’) of the two bits which can be 00, 01, 10, 11 by reading them [294]. In a classical 1T1R column, two rows are simultaneously activated ($WL = '1'$), so that the corresponding memristors are subjected to the same read voltage V_{READ} (Fig. 227-a). Depending on the combinations of the two accessed memristors (2 HRS, 1HRS + 1 LRS, or 2 LRS), the total current flowing through the Source Line (SL) will take different mean values. Thus, a current reference is chosen between each current distribution to represent a Boolean operation (Fig. 227-b). AND, OR, XOR are then simply achieved by sensing the SL current and comparing it to the appropriate reference(s). For instance, to perform an ‘OR’ operation, we compare the SL current I_{read} to the leftmost reference I_{refOR} : if I_{read} is smaller than I_{refOR} , it means that the two accessed memristors are in the

HRS state (i.e. both represents logic state ‘0’), and the “OR” output is therefore ‘0’; if I_{read} is bigger than I_{refOR} , at least one of the memristors is in logic state ‘1’, so the “OR” output is ‘1’. However, this approach is feasible only if the distributions are disjointed but can be extended to n inputs if the read current distributions are tighten enough. To finish with, AND and OR operations can be combined to form more complex one. For instance, XOR operations can be expressed as $A \text{ XOR } B = (A \text{ AND } \text{NOT}(B)) \text{ OR } (\text{NOT}(A) \text{ AND } B)$. The reading circuitry consists in sense amplifiers which detain a lower delay and smaller area than the modified Pinatubo sense amplifier.

One of the main advantages of Scouting logic/Pinatubo compared to MAGIC logic is about device endurance which is not impacted by computing since it does not require writing sequence.

b. IMC existing solutions: examples

This section will give some implementations with memristors enabling IMC for various applications. The structures will be discussed to provide insights about performance gain. Concerning 1T1R structures, Xue *et al.* [295] demonstrate IMC and in particular multiply and accumulate operations, in a 55-nm 1-Mb RRAM macro. The 1T1R structure is given as an example in Fig. 228, note that each 1T1R cell can be selected in the matrix thanks to bitlines, wordlines and sourcelines. The time to perform this operation is 14.6ns and the peak energy efficiency is 53.17 TOPS/W (Tera Operation Per Second/Watt). A similar result is presented in [296] using a 65nm CMOS technology in where the read delay is 14.8ns for Convolutional Neural Network (CNN) operation. Also a dual mode to perform either Boolean operations (AND, OR...) or more complex operations (adder/multiplier) is proposed in [297]. This approach combines self-write-termination circuits, multiple logics current-mode sense amplifiers and dual mode wordline for SET and RESET operations. Thanks to this structure, both memory operation and IMC one can be done. The 16Mb RRAM is composed of 1T1R HfO RRAM and 0.15 μ m CMOS technology and achieves IMC operations in less than 14ns. To finish with, Mochida *et al.*[298] integrate an analog RRAM-based 4M synapses achieving 66.5 TOPS/W in a 40nm technology. 1T1R structures provide low sneak path current of unselected cells but one of their major drawbacks concerns the silicon footprint required to integrate both transistors and memory elements. To overcome this density issue, Luo *et al.* [299] propose a 8 layer 3D vertical RRAM with a self rectifying behavior. However, the sub μ A operation current targets low power applications rather than high performance one.

To conclude this section, through these examples, we do observe that IMC is energy efficient, providing a large number of operations per second and per watt. The next part will present the materials for memristors.

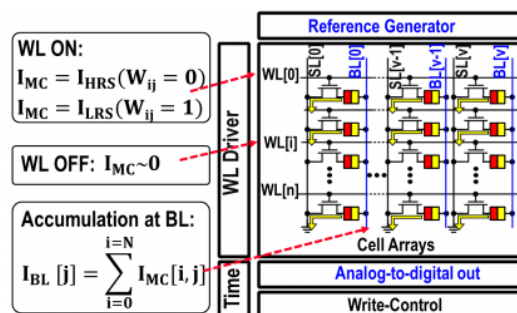


Fig. 228: 1T1R structure presentation. Figure from [295].

c. IMC materials/ selectors

We saw the different solutions for IMC based on memristors. In this part, we will dress a short review of all the memristor types before focusing on the OxRAM technology. However, the creation of large matrix leads to more leakage current and is feasible only if sneak paths are managed. For this, a selector device can be inserted in series with the memory element. This solution will be detailed in the last part.

i. Memristors materials

By definition, the memristor is a two terminal element presenting a hysteresis characteristics as presented in Fig. 224. The first fabricated memristor [300] consists in two layers of TiO_2 (doped and undoped) sandwiched between two platinum electrodes. With suitable voltages, one could switch between two states: a high and a low resistance one. For IMC applications which require a high density of memory elements, the memristor should be scalable and fabricated in compact structures, for instance in a crossbar array. More importantly, they should be BEOL compatible (processed at low temperature) to enable 3D stacking. There is also a need of non-volatile memory element, in order to store the resistive state without wasting power. Among the emerging non volatile memories, the rest of the section will provide a quick overview of Phase-Change Memory (PCM), Spin-Torque-Transfer Magnetic Memory (STT-MRAM) and Resistive Memory (RRAM). The main criteria of comparison between these non-volatile memories are the ability to have a dissociable high resistive state (HRS) and low resistive state (LRS), represented by the ratio $R_{\text{HRS}}/R_{\text{LRS}}$ and called “memory window”, the number of times the device can switch between these two states, called the “endurance” and the energy required to switch between the states.

- **PCM:** it is composed of two electrodes sandwiching a chalcogenide glass that can switch between a crystalline phase and an amorphous one. The crystalline state features a low electrical resistance state whereas the amorphous detains a high resistance state. The ratio between these two states is higher than RRAM one, but they suffer from a high resistance state value drift over time [301]. Also, the PCM is programmed by Joule heating and needs high programming current even for ultra-scaled dimensions [302] which is not compatible with a large low-power IMC bloc.
- **STT-MRAM:** the device structure is two ferromagnetic layers separated by a thin insulator layer. The data ‘0’ or ‘1’ is stored in the magnetisation of ferromagnetic materials. More precisely, the magnetisation of the free layer is switched while the other one is left unchanged. If the magnetisation is the same, the electrons have a high probability to pass through the device (low resistivity state). If not, there is no current conduction and the device is in a high resistivity state. It provides a low-energy programming, high speed and excellent programming endurance [303]. Nevertheless, the resistance ratio between high and low resistivity state is lower than the other technologies.
- **RRAM:** it consists on a Metal Insulator Metal structure where a thin metal oxide layer is sandwiched between two metal electrodes. In the metal oxide layer, a conductive filament can be formed or dissolved modulating the resistance of the layer. In Oxide-based RAM (OxRAM) technology, the conductive filament is composed of oxygen vacancies in the oxide layer whereas for Conductive-Bridge RAM (CBRAM), it relies on the migration of metallic cations. Generally, OxRAM presents better endurance ($>10^8$ cycling operation) than CBRAM ($<10^4$) but worst dissociation between conductive and not conductive state [304].

To conclude this part, PCM are interesting in terms of memory windows but suffers from a large programming power which is not compatible with IMC where large memory arrays are required. On the contrary, STT-MRAM detains a small memory window but is incompatible with large array

computation. That is why, OxRAMs are promising elements with respect to the trade-off between memory window and endurance. Thus we decided to favor the endurance characteristic (OxRAM). The next part will focus on OxRAM technology.

ii. Focus on OxRAM technology

Working operation:

The OxRAM device physics is explained in this subsection. The device detains two terminals, called a top electrode and a bottom electrode made of metal and an oxide material is in between. The interest of this non-volatile memory is to switch from a high resistivity state (HRS) to a low one (LRS), representing the logic state '1' (LRS) and '0' (HRS). This is performed by the formation or not of a conductive filament of oxygen vacancies. Fig. 229 presents the switching process with the oxygen ion migration and diffusion. On a pristine cell, featuring a high resistance value, the filament must be formed by applying a forming voltage V_F on the top electrode. Note that for HfO_2 based memory, the forming voltage is linearly dependant on the thickness of the film. A forming operation have been demonstrated in a 3nm thick HfO_2 film [305]. A soft dielectric breakdown occurs and oxygen ions drift to the anode interface due to the high electric field. Afterwards, the oxide/top electrode interface behaves as an oxygen reservoir. With the conductive filament presence, the cell lets the current flow and is in a low-resistance state LRS. To break the filament and reverse the process, a negative voltage V_{RESET} is applied between top and bottom electrodes. The cell is now in a high resistance state HRS. Now, it is possible to SET again the LRS into the cell by applying V_{SET} (instead of V_F to form the filament) which is lower than V_F . Switching back and forth between LRS and HRS is performing a switching cycle. The maximum number of switching cycles is called programming endurance (or just endurance) and depends on the technology.

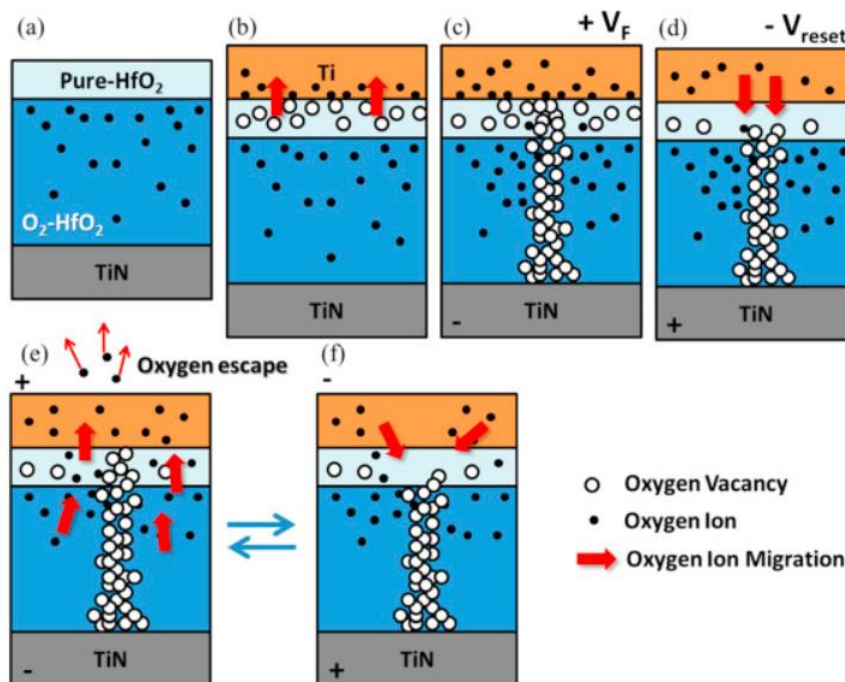


Fig. 229: Schematic description of the RS mechanism of the device. (a) Before and (b) after the Ti top electrode deposition. (c) CF grows from TiN to Ti electrodes under a positive forming voltage on it. (d) A negative voltage is applied on it for rupture of the CF. (e) CF formation and some oxygen ions release during set process. (f) CF ruptures during reset process. Figure and legend from [306].

As far as the materials are concerned, several stacks are proposed in the literature. For instance, Lee *et al.* [305] propose an HfO_2 based memory with TiN electrodes. The TiN/TiO_x/HfO₂/TiN structure yields high ON/OFF resistance ration ($>10^3$), fast switching speed (5ns), endurance ($>10^6$ cycles) and reliable

data retention (10 years at 200°C). In fact, even if several metal oxides materials exhibit resistive switching behavior, HfO₂ have been widely studied and detains attractive advantages for RRAM. Such devices are simple to integrate and detains a low operating power, high speed and high non-linearity [307], [308]. Concerning the top and bottom electrodes, noble metal can be used such as Pt, Au or Ti, TiN and TaN.

Programming conditions and resistance distribution:

In RRAM technologies, the HRS and LRS resistances values depend on the programming conditions, *i.e.* the applied voltage (V_{SET} or V_{RESET}), the programming time and the programming current (called compliance current I_{CC}). The compliance current is necessary to prevent an abrupt increase of current causing the failure of the cell. This current is imposed by the serial integration of a selector, such as a diode or a transistor. Some considerations about the influence of the programming condition on the resistance distribution are:

- Programming time: it depends exponentially on programming voltage [309], so that usually this time is fixed and the programming voltage is changed.
- I_{CC} : it will determine the LRS resistance values during SET operation. The relationship between I_{CC} and LRS resistance mean value is power law. In fact, increasing I_{CC} results in lower LRS resistance values (R_{LRS}) [310].
- V_{RESET} : using higher voltages during RESET operation results in higher HRS resistance value (R_{HRS}) [311].

As said previously, the HRS and LRS accounts for ‘0’ and ‘1’ logic states and must be distinguished one from the other. A typical figure of merit to maximize, is the ratio R_{HRS}/R_{LRS} , called memory window. However, it has been demonstrated that there is a trade-off between the Memory Window and the endurance. In fact, for a higher memory window, higher I_{CC} and V_{RESET} are required which endangers the endurance of the cell. Fig. 230 presents the typical endurance for two different stacks, illustrating the trade-off between endurance and memory window. A low memory window is critical for large memory array due to sneak path issues [312]. However, it is possible to integrate a transistor for each memory element, creating a so-called 1T1R structure to mitigate this leakage issue at the expense of density [313].

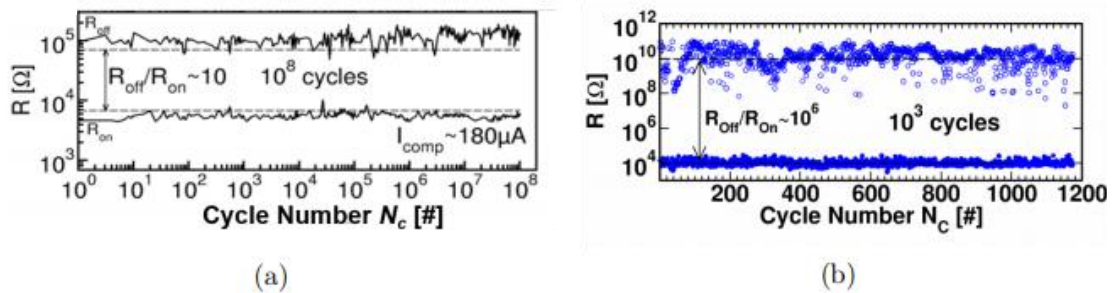


Fig. 230: Typical endurance characterisations performed on (a) a GeS_2/Ag (b) a $HfO_2/GeS_2/Ag$ Resistive Memory (RRAM) stack. While it is possible to sustain a low resistance ratio R_{off}/R_{on} of 10 during 10^8 switching cycles, only 10^3 switching cycles can be performed with a large resistance ratio of 10^6 . Reproduction from [314].

Variability:

As illustrated on Fig. 230 a given RRAM presents different LRS and HRS values for each cycle. This variability is referred as cycle-to-cycle variability and is attributed to the stochastic nature of the conductive filament during formation and dissolution. Additionally, resistance variability takes place across devices among the memory array. This device-to-device variability ensues from manufacturing variability. Fig. 231 presents the resistance distribution for HRS and LRS for a 4kbit TiN/HfO₂/Ti/TiN

RRAM array. Between the median HRS and LRS values, a resistance ratio of 2500 is measured and is drastically reduced to 600 when considering 3σ device-to-device variation. To mitigate this, Grossi *et al.* demonstrate that the higher the compliance current I_{CC} , the lower the forming resistance values and the tighter the resistance distribution [315].

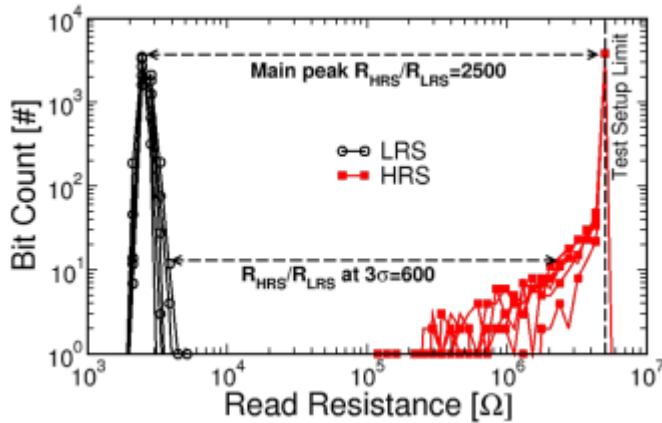


Fig. 231: resistance distribution for HRS and LRS for a 4kbit TiN/HfO₂/Ti/TiN RRAM array after one RESET/SET cycle. Reproduction from [315].

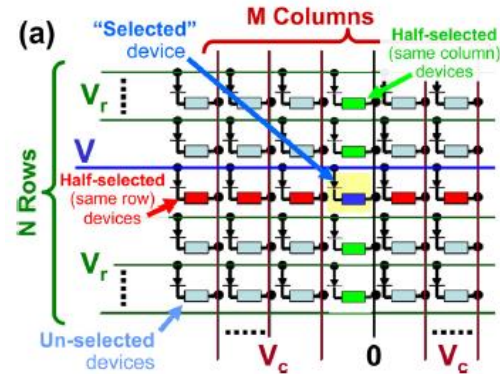


Fig. 232: Presentation of the 1T1R array with voltages to select a particular device. Figure from [316].

iii. Selectors

To suppress efficiently the sneak path and select a particular device into the memory array, a selector must be integrated next to each memory element (1S1R). It should be at least a two terminal device that can reliably, repeatedly and readily switch between two resistance states with a large resistance contrast. It must be scalable to avoid a large area overhead. However, the selector is not useful only for the leakage current but also to deliver the right current amount to the resistive element. Fig. 232 presents the problem in an array: when a device is selected by its row and column coordinate, the cells in the same row or column are half selected and if the selector element is not properly sized unwanted writing of cells can occur. To choose a selector, the main criteria [316] are:

- High ON state current density
- Low OFF state leakage node
- BEOL compatibility
- Switching voltage: its threshold must be below the RRAM one to form correctly the memory point
- Switching speed, endurance, yield and variability... (the properties of the access element must be better or equivalent to the memory element not to degrade the matrix)

Different types of selectors are available in the literature and their characteristics are described in Fig. 233. The physics of the devices won't be presented and we will rather focus on the performance required for selectors. Representative examples taken from the literature are given:

- PN poly-Si diode: [317] 8MA/cm² ON current (+2V) and OFF current 100A/cm² (-2V), 4F² cell size.
- CuO/InZnO diode: [318] 10⁴A/cm² for 3V, room temperature, 10³ selectivity.
- Metal/Semi-conductor/Metal junction: for instance, based on Schottky barrier tunneling [319], $I_{ON}/I_{OFF} \sim 10^7$, $I_D = 0.2 \mu A/\mu m$ [320], Back-end of line compatible (250°C), 10⁵ A/cm² for 1V.
- Ovonic threshold switch (OTS) : [321] RESET speed of 9ns, endurance of 10⁶ cycles.

- Mixed ionic electronic conductor (MIEC): [322] <400°C process integration, scalable, high current density $J > 50 \text{ MA/cm}^2$, endurance $10^{10}/10^5$ for low/high current.
- Field Assisted Superlinear Threshold (FAST): [323] [324] $SS < 5 \text{ mV/dec.}$, ON/OFF ratio= 10^7 , sub-50ns operations, > 100M endurance and integration temperature less than 300°C.
- Insulator Metal Transition (IMT): [325] 10^6 switching cycles, fast switching speed (22ns).
- Threshold Vacuum Switch (TVS): [326] $> 10^8 \text{ A/cm}^2$, selectivity of $> 10^5$, $> 10^8$ cycles.
- Transistors: conventional transistors detains excellent ON/OFF current ratio, the ability to tune threshold voltage (with doping for instance), large ON current but relatively large cell size. To reduce the dimensions, one solution demonstrated by Wang *et al.* [327] is to integrate the transistor vertically. However this proposition is not compatible with BEOL multi-level stacking due to the thermal budget for processing. However, 3D monolithic integration could leverage multi-level stacking of 1T1R arrays.

To conclude about the selector element choice, OTS, MIEC and FAST are very promising since they ensure a large ON current density along with a good endurance for a small cell dimensions. However, transistors cannot be set apart, since they provide an excellent ON/OFF current ratio with a tunable V_T and can be processed at low temperature. However, there are still challenges to overcome to create stackable 1T1R arrays with a low silicon footprint.

	Si diode	Oxyde diode	MSM	OTS	MIEC	FAST	IMT	TVS	Transistor
J_{ON}	Red	Red	Yellow	Green	Green	Green	Green	Green	Green
ON/OFF current	Green	Red	White	Yellow	Red	Green	Red	Yellow	Green
V_T flexibility	Red	Red	Red	Yellow	Green	Green	Green	Yellow	Green
Endurance	No data	White	White	Green	Green	Green	Yellow	Green	Green

Fig. 233: Summary table of the advantages (in green), neutral (in yellow) or disadvantages (red) of the selector devices.

To enable IMC in an array, one needs a resistive element in series with a selector for leakage and selection issues. In this part we presented the different resistive elements with an emphasis on OxRAMs and a non-exhaustive overview of selectors was done. It emerges that a good candidate for memory part is the OxRAM due to its ease of fabrication, low power, endurance and scalability. From the selector side, OTS, MIEC and FAST are promising devices but transistors are still competitive due to their excellent electrostatic control, apart from silicon footprint. The best situation will be to provide a 1T1R 3D array which compensates this lack of density by going into the third dimension.

d. My-Cube project: choices

In this context, the project My-Cube financed by a European Research Council grant, aims to co-integrate memory element and transistors into a 3D cube, towards a functionality-enhanced system with a tight entangling of logic and memory for IMC. It relies on three key enabling technologies presented in the previous paragraphs (or introduction): non-volatile resistive memory, energy-efficient stacked nanowires transistors and 3D monolithic integration. Combined together, it is possible to create an ultra-dense 3D structure as depicted in Fig. 234 where each bitcell (1T1R) can be addressed by a bitline, a wordline and a sourceline. Unlike 3D sequential integration, all the transistors can be fabricated at the same time without additional wafer bonding and lithography. However, depending of the configuration (which will be discussed later), some rows or columns of transistors share the same gate, source and drain. It is thus possible to select a particular cell by applying the appropriate voltages on wordlines, sourcelines and bitlines. The memory element is laterally integrated at the source of the transistor.

Before describing accurately the structure, we will present the different technology choices for the transistor and the resistive element.

i. Stacked nanowires

As discussed in the selector part, the main characteristics required for this transistor are high ON state current density, low OFF state leakage, high switching speed, high endurance, high yield and low variability... Gate-all-around structures offer an excellent electrostatic control and ON/OFF current can be tuned with transistor sizing (nanowire thickness, gate length and width). Concerning the endurance, GO1 devices (with a thin gate oxide) are compatible with OxRAM requirements. In fact, up to 10^7 switching cycles have been demonstrated on 1T1R structures in [313] showing no sign of premature degradation for GO1 devices. This thin gate oxide will enable us to drive a large ON current for large gate overdrive. Due to their ease of fabrication, junctionless GO1 transistors are proposed to avoid the transistor source doping and the bitline doping. However, doping being a major drawback for doped channel devices, an in-depth study will be performed to study the impact of using junctionless devices on variability. This variability analysis will be performed in section 4- after the structure analysis and sizing (in part 2-b).

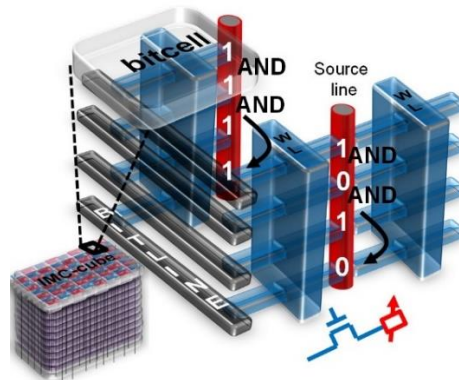


Fig. 234: My-Cube topology artist view.

ii. Memory element

Due to their ease of fabrication (CVD deposition) OxRAM technology, and especially HfO_2 based, is chosen among RRAM for this 3D structure. One major challenge is the lateral integration of the memory element in the drain of the device. The bottom electrode will be formed by the silicide process and the deposition of TiN. Then HfO_2 and Ti and TiN are then deposited to complete the stack. Due to its specific structure, the conductive oxygen vacancy filament will be confined to the transistor drain. Thus its position will be controlled and the variability should be reduced. Different technological variants concerning the size of the layers will be investigated.

iii. Boolean logic: Scouting logic

We have focused on the Pinatubo/Scouting logic approach to minimize the writing operation to preserve the OxRAM endurance. It will result in read operations in the matrix to perform AND and OR operations. However, we have to make sure that the dissociation between each state (*i.e.* for 2 memristors: 00, 01 or 10 and 11) is effective. For this, we need to consider the OxRAM HRS and LRS distributions and see if, when reading, there is no overlap. This issue will be tackled in the next section.

In this section we presented the technological choices for My-CUBE project. One of these choices consists in the integration of junctionless devices with OxRAMs. The next section will analyse the pertinence of these choices to enable IMC.

2- Sizing simulations

Before analysing the whole MY-CUBE structure, a single pillar is considered. The aim is to prove the pillar functionality before considering the cube depth. For this, after presenting the bit-cell topologies, junctionless transistor measurements are exploited by TCAD simulations to define the inputs (especially currents and resistance distributions) for SPICE simulations. The SPICE simulations will use scouting logic (see part 1-a.ii) to perform Boolean operations in the pillar.

a. Simulated pillar structure

The pillar topology and equivalent electrical scheme is represented in Fig. 235. Conversely to GAA transistors, each source and drain of the stacked nanowires is independent and address an OxRAM, whose materials (oxide and top electrodes) are deposited in a vertical pillar. The bottom electrode of the memory being localized at the drain side of each transistor. So, the final pillar structure includes two times n stacked nanowires with a common gate (referred as WL1, WL2), separate drains (referred as BitLines BL1a to BLna, and BL1b to BLnb) and a common pillar called source line (SL) gathering the sources.

To reset a particular bit-cell, V_{reset} is applied on the associated BL while the others are left at GND like the SL. The corresponding transistors are turned ON with $WL_i = V_{\text{DD}}$. Programming and read operations on the pillar are performed classically, like in standard 1T1R memories. SET, RESET or READ voltages are applied to BLs or SLs. The bit-cells of the same pillar which are unused are inhibited with $V_{\text{BL}} = V_{\text{SL}}$, while access transistors of unused pillars are OFF.

Like in MY-CUBE integration, in this structure, the stacked nanowires are preferentially junctionless which relaxes the constraint in term of S/D doping for multiple stacked nanowires. The gate oxide is thin since GO1 devices were already proven compatible with OxRAM endurance requirements [313].

For this study, a structure with four layers is studied, this number being chosen arbitrarily, up to seven stacked nanowires have been demonstrated in [18]. The peripheral circuit is not considered yet, to prove the IMC concept in this pillar.

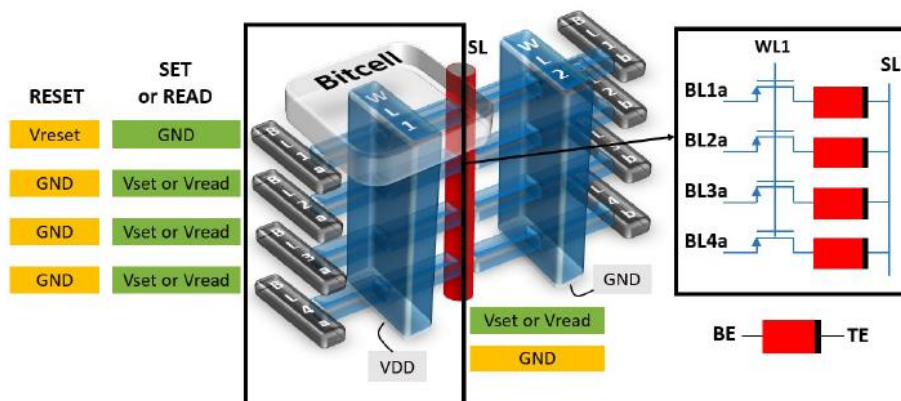


Fig. 235: 3D pillar structure scheme with 4 layers. Bitlines (BL), wordline (WL) and source line (SL) voltages to program (SET/RESET) and to read are indicated. The equivalent electrical schematic is given.

b. Definition of SPICE simulation inputs

To define the SPICE simulation inputs, we consider silicon-based measurements of Chapter III junctionless accumulation-mode transistors (JAM). The process flow is outlined in Chapter III. However, these devices were in a tri-gate configuration and not gate-all-around and the nominal width was $W=50\text{nm}$ which is smaller than the nominal width targeted in MY-CUBE ($W=75\text{nm}$). That is why

we performed TCAD simulations to have insights on the drive current for a GAA configuration at $W=75\text{nm}$.

i. JL performances at $W=50\text{nm}$

We have at our disposal experimental nMOS with different gate lengths ranking from 60nm to 200nm for $W=50\text{nm}$. Fig. 236 presents the $I_{\text{ON}}-I_{\text{OFF}}$ for the two smallest dimensions. To drive more current, we decided to work at a high gate overdrive: $V_G=1.5\text{V}$ and $V_D=1.3\text{V}$. For $W=50\text{nm}$ and $L=80\text{nm}$ (respectively $L=60\text{nm}$), the transistors drive in average (on 28 measurements), $87\mu\text{A}$ ($120\mu\text{A}$) ON current for an OFF current of 10^{-9}A (10^{-6}A). Both trade-offs are interesting either for a low-power consumption pillar or to deliver high drive current. This drive current should be sufficient to form the OxRAM. However, if the leakage current is too important, one might not be able to read or perform operations in the cube due to sneak paths. We will study more in details the sizing $W=50\text{nm}$ and $L=80\text{nm}$ which corresponds to a low power consumption configuration. I_D-V_G and I_D-V_D of the aforementioned dimensions are given in Fig. 237 and Fig. 238. The stakes of this study is to find a correct drive current to set a logic state into the memory element, while ensuring reliability to endure the IMC scheme and correct variability to reduce the resistance distribution.

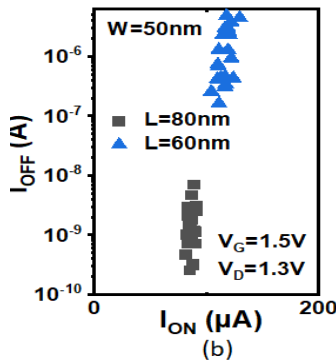


Fig. 236: $I_{\text{ON}}-I_{\text{OFF}}$ for $W=50\text{nm}$ and $L=60\text{nm}$ and 80nm . Figure from [328].

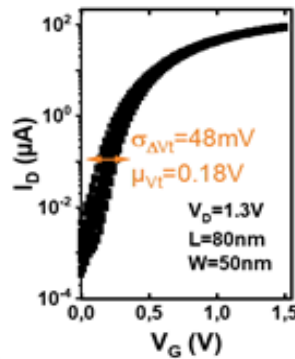


Fig. 237: I_D-V_G for $L=80\text{nm}$ and $W=50\text{nm}$ and $V_D=1.3\text{V}$. Figure from [328].

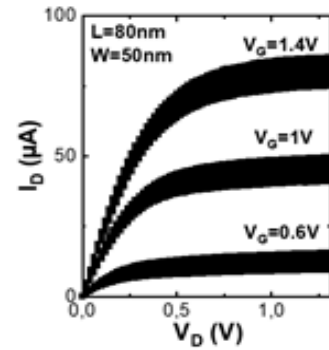


Fig. 238: I_D-V_D for $L=80\text{nm}$ and $W=50\text{nm}$ and various V_G . Figure from [328].

Cycling tests have been performed on junctionless devices only to verify this GO1 endurance compatibility. A pulse duration of 100ns 10^7 times of value V_{DD} is applied on the gate as presented in Fig. 239. To see the degradation of the selected device, I_D-V_G curves are realized before and after stress and for each cycling decade. The voltage applied on the gate V_{DD} is set at 1.5V (like previous measurement) but $V_{\text{DD}}=1.8\text{V}$ and $V_{\text{DD}}=2\text{V}$ are also investigated. The resulting stress on the device ($W=50\text{nm}$ and $L=110\text{nm}$) is seen in Fig. 240 where the arrow indicates the directions of the measurements (first $V_{\text{DD}}=1.5\text{V}$ 10^7 times, then $V_{\text{DD}}=1.8\text{V}$ and finally $V_{\text{DD}}=2\text{V}$). For $V_{\text{DD}}=1.5\text{V}$, no degradation is seen on the device, the initial curve and the final one (after 10^7 cycles) being perfectly superposed. For $V_{\text{DD}}=1.8\text{V}$ (in blue), a slight degradation is seen resulting in a V_T shift of 5mV between initial curve and final curve (after 10^7 cycles). For $V_{\text{DD}}=2\text{V}$, this degradation is worst and results in a 10mV shift. However these ranges of V_T shifts are acceptable for 10^7 cycles. To conclude, given that we would like to work at $V_G=1.5\text{V}$, the junctionless transistor will not be the limiting element of 1T1R endurance.

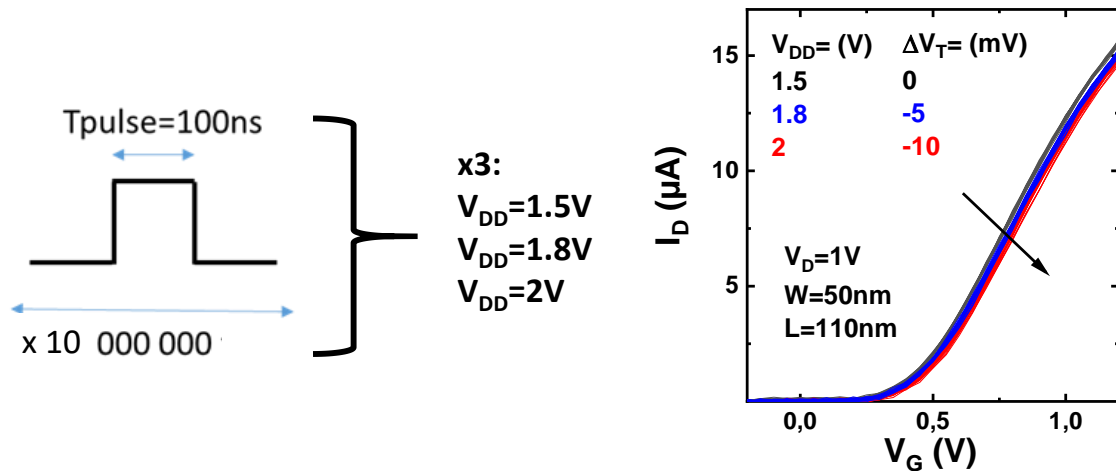


Fig. 239: Stressing scheme applied on the gate of the transistor to evaluate its endurance.

Fig. 240: Resulting I_D - V_G during stress sequence. Three different stress voltages are applied.

ii. Drive current for stacked nanowires at $W=75\text{nm}$

To increase the drive current, we propose to study the case of a larger width, $W=75\text{nm}$. Note that with the stacked nanowire GAA configurations, large width are not feasible due to the mechanical constraints during the nanowires release step. In fact in literature for advanced nodes which requires thin films, the width of fabricated transistors are limited: 50nm in [19], 75nm in [329], 100nm in [330]. Concerning the drive current, Tri-Gate-JL (TG-JL) transistor at $W=50\text{nm}$ and $L=80\text{nm}$ (REF device) delivers $75\mu\text{A}$ where the standard compliance current to form the memory point is at least $100\mu\text{A}$. However, we do not have the silicon devices corresponding to $W=75\text{nm}$, so that we would like to extrapolate the characteristics for this enlarged dimension. The reference structure are presented in Fig. 241 and features a silicon channel of 11nm , doped at $N_D=7.10^{18}\text{at/cm}^3$ and a width W of 50nm and a gate length of 80nm . The TCAD simulation environment for tri-gate devices is identical as the one introduced in chapter III. However, for the gate-all-around configuration, some minors' modifications are done and explained in the next part.

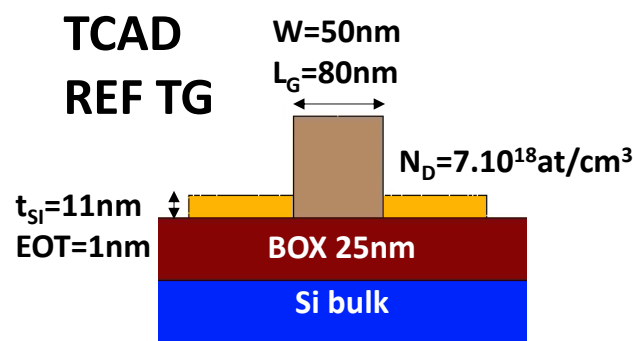


Fig. 241: Presentation of the REF TCAD structure featuring a silicon channel of 11nm , doped at $N_D=7.10^{18}\text{at/cm}^3$ and a width W of 50nm and gate length of 80nm .

TCAD simulations parameters: differences from chapter III:

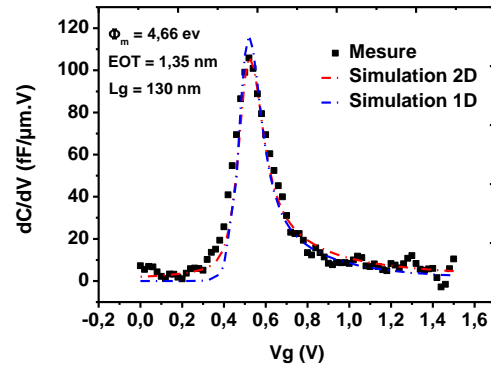
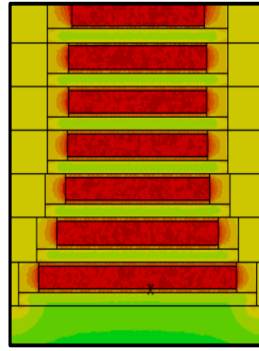
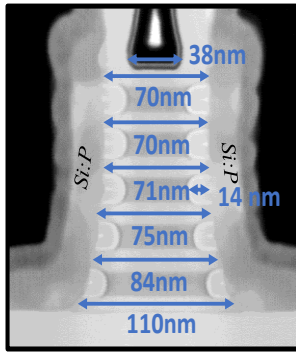


Fig. 242: TEM cross-section of stacked nanowires which is reproduced for the 2D simulation.

Fig. 243: Extraction of Φ_m from dC/dV curve as a function of V_G .

In the pillar the JL transistors will rather be in a NW-Gate-All-Around (NW-GAA) configuration and their width can be tuned to deliver more current and also to reduce the variability. To take into account the particularities of Gate-last Gate-all-around architectures in simulations, the Φ_m have been extracted from a transistor with 240 channels and a top width of $W=38\text{nm}$ (see Fig. 242). A 1D simulation considering 6 GAA and 1 FDSOI channel is done by CEA characterisation laboratory. However to consider channel edge effects, spacer and fringe field, a 2D simulation based on TEM dimensions (see Fig. 242) and using FlexPDE software is done. The results are depicted in Fig. 243 and the extracted Φ_m is equal to 4.66eV . This value will be used in NW-GAA TCAD simulations. For trigate (TG) configurations, the previous value of Φ_m (4.61eV) used in chapter III will be kept.

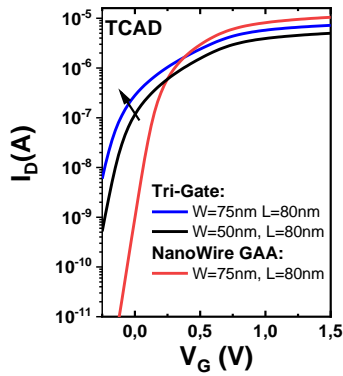


Fig. 244: I_D - V_G for the different structures. An electrostatic control gain is seen from the Tri-Gate configuration to the NanoWire Gate-all-around one, $V_D=50\text{mV}$.

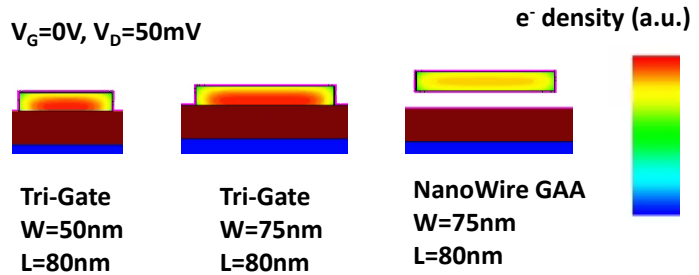


Fig. 245: Electron density cutplane for $V_G=0\text{V}$ and $V_D=50\text{mV}$.

Estimation of the transistor drive current:

I_D - V_G in linear regime ($V_D=50\text{mV}$) are given in Fig. 244 and the electron density cut are given in Fig. 245 for $V_G=0\text{V}$. The OFF and ON currents ($V_G=0\text{V}$) are presented in Fig. 246 for the different configurations. Compared to the TG-JL REF, JL-TG @ $W=75\text{nm}$ drives 50% more current but at the expense of a higher OFF current. Going to a JL-GAA-NW configuration increases both electrostatic control and drive current (-3 decades on I_{OFF} and +70% on I_D). Even more, since the electrostatic control is better, the channel doping N_D can be increased to obtain +150% drive current for the starting I_{OFF} . Applying these gains between TCAD structures to our measured JL-TG, four SET conditions (μA) for

pillar transistor drives have been defined: weak (70 μ A), Light typical (100 μ A), Strong Typical (150 μ A) and Strong (200 μ A).

Configuration	TG EXP	TG	TG	GAA	GAA
W (nm)	50	50	75	75	75
N _D (at/cm ³)	7.10 ¹⁸	7.10 ¹⁸	7.10 ¹⁸	7.10 ¹⁸	10 ¹⁹
$\log(I_{OFF})$ (A)	-9	-7	-6.5	-10	-7
I_D @ $V_D=1.3V$, $V_G=1.5V$ (μA)	87	50	87	86	126

Fig. 246: Summary table of the TCAD simulations.

iii. OxRAM distribution extraction

Grossi *et al.* demonstrate that the higher the compliance current I_{CC} , the lower the forming resistance values are and tighter the resistance distribution is [315]. In fact in Fig. 247, we observe that a large I_{CC} forming will result in low read resistances values and compact distributions. That is why for the SPICE simulations, a different resistance distribution for each previously defined SET conditions must be considered. The OxRAM are fabricated [315] by 10nm HfO₂/Ti 10nm/TiN stack deposition on top of a TiN bottom electrode and arranged into 4kbits 1T1R array. Resistance distributions (mean μ and standard deviation σ) for previously defined SET conditions are extracted for a 100ns pulse width and a 2V source line voltage. The RESET conditions for $V_{bl,reset}=2.5V$ and $T_{pulse}=100ns$ corresponds to a lognormal HRS distribution with parameters $\mu=120k\Omega$ and $\sigma=0.63$. The resistance distribution (mean and standard deviations) for each SET conditions defined in part 2-b.ii are given in Fig. 247.

To conclude this part, we selected SET current conditions and their associated resistance distribution based on junctionless measurements for $W=50nm$, $L=80nm$, $V_D=1.3V$ and we performed the extrapolation to a larger width and gate-all-around configuration. In the next part, SPICE simulations will be done to verify if the different conditions enable SCL operations.

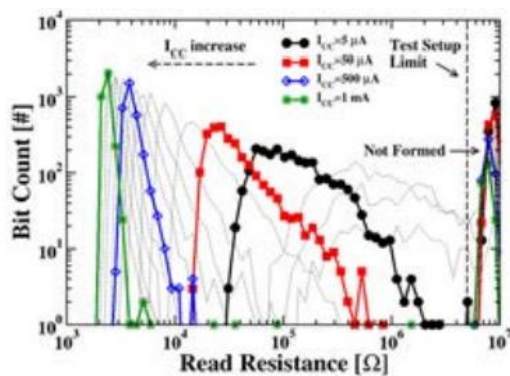


Fig. 247: Forming with increasing I_{CC} and $V_{BL}=4V$: read resistance distributions evolution with $T_{pulse}=100ns$. Figure from [315].

SET Condition	Compliance current	Resistance parameters
	(μA)	$\mu(k\Omega) / \sigma(k\Omega)$
Strong	200	5.2/0.58
Strong Typical	150	5.7/0.73
Light Typical	100	8/1.3
Weak	70	10/2

Fig. 248: Summary table of the SET conditions.

c. Scouting logic in the pillar

The next section will present briefly our work driven by M. Ezzadeen *et al.* to prove the feasibility of SCL in the pillar. The SPICE simulations use an OxRAM model based on experimental distributions and a 300nm thick oxide transistor model from a commercial design kit. Variations were considered up to 3 sigma, and Monte Carlo simulations were performed with 1000 runs.

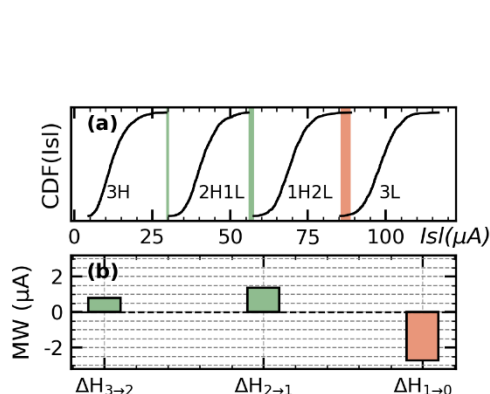


Fig. 249: Scouting logic results with Light Typical SET on three levels represented by current distributions (a) or by Memory Windows values (current margin between two consecutive operations) between current distributions (b). Figure from [328].

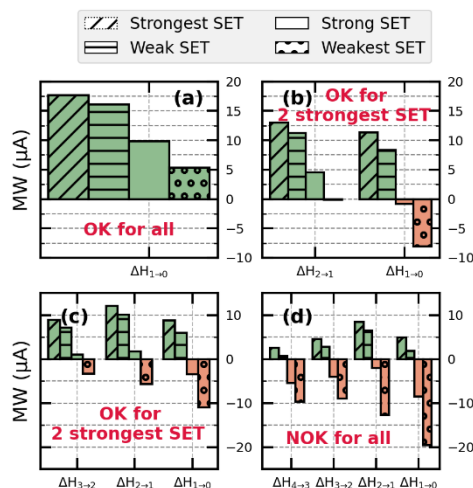


Fig. 250: Memory Window as a function of SET conditions, from one (a) to four (d) activated levels, with simple coding and a bitline voltage, $V_{scl}=0.5V$. Figure from [328].

As said previously, to implement SCL successfully on a given number of levels, current distributions corresponding to the different combinations of HRS and LRS states must not overlap when the pillar is read. Fig. 249 shows the simulated current distributions when performing SCL on three levels, with Light Typical SET. Four different combinations of states can be obtained: all in HRS, all in LRS, one HRS & two LRS and two HRS & one LRS. In this particular case, we observe that the third distribution and the fourth one overlaps, which means that the state 3LRS and 2LRS1HRS cannot be distinguished properly. However, the marge between two consecutive states, called here the Memory Windows (MW), are preserved between the first, second and third distributions.

Fig. 250 presents the same simulation for 4 layers and our four preselected SET conditions with a sourceline voltage of 0.5V. We notice that classical read operations can be achieved by all SET conditions. The two strongest SET conditions enables scouting logic with up to three parallel levels. Note that, as expected, MW are higher for stronger SET conditions. Of course, these results have to be completed by taking into account the variability of the read circuitry.

d. MY-CUBE: read and write schemes.

We demonstrated the functionality of a single pillar, where the bitline, wordline and sourceline were driving only respectively one, four and eight devices. However, to extend it into the third dimension (called here the depth), we need to think about the connections to be able to write (SET and RESET) a single device. Simultaneous cell reading should be possible to perform Scouting logic. For the sake of simplicity, a cube of two layers only is considered.

To be able to write a particular cell, the WL and SL directions must be perpendicular and BL parallel to SL, addressing a single column (Fig. 251). By applying VDSET on the corresponding SL, GND to the selecting BL and VDSET to the others which are adjacent to the selected SL and VGSET to the correct WL, one can write a unique bitcell as described in Fig. 252. RESET operation is performed similarly by

polarizing the WL to VGRESET and the bitline to VDRESET as presented in Fig. 253. Concerning reading operation, it is possible to use the same configuration as the SET operation by replacing the WL voltage to VGREAD, the SL by VDREAD and the unselected BL of the selected pillar by VDREAD. Fig. 254 presents how to read the whole pillar. Also, this cube allows a high operation parallelism. In fact, while performing a read operation in a $x_0y_0z_0$ bitcell, it is possible to read other bitcell which are not in the x_0 plan simultaneously. A parallel programming is feasible in the selected z_0 plan without parasitic SET or RESET operation.

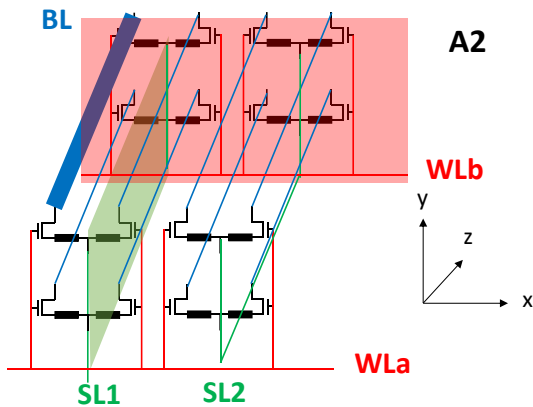


Fig. 251: Possible addressing scheme A2. WL, BL and SL plane are highlighted.

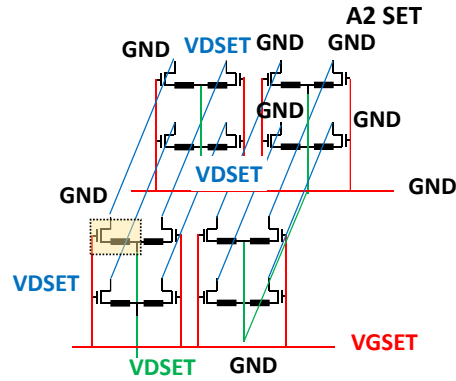


Fig. 252: SET operation in A2 addressing scheme.

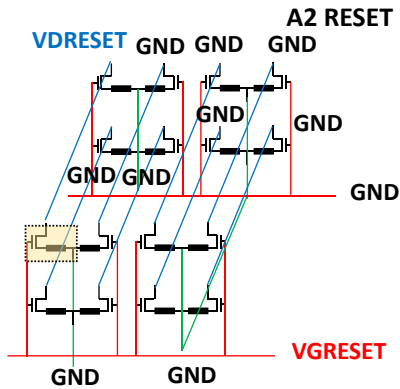


Fig. 253: RESET operation in A2 addressing scheme

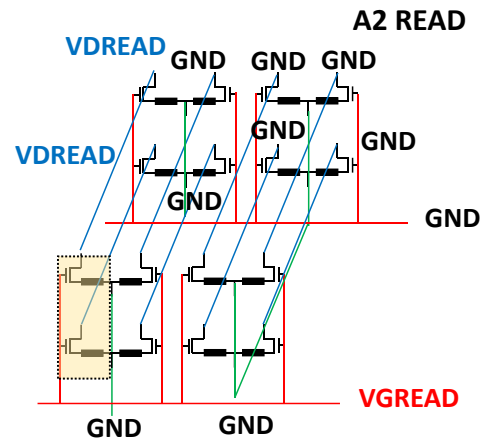


Fig. 254: READ operation in A2 addressing scheme.

In this section we demonstrated that MY-CUBE pillar was compatible with IMC and that according to silicon-based measurements and extensive simulations up to 3 layers can be computed at the same time. An optimum topology have been proposed to read and write into MY-CUBE structure. So, the choice of junctionless transistor and OxRAM technologies is relevant. In fact, the junctionless transistors could deliver experimentally enough drive current with the appropriate biases and even more current are predicted for a GAA configuration. The next section will present how to process stacked structure to manufacture the array.

3- Processing of stacked structures

The goal of this part is to propose layouts and associated process flow for MY-CUBE array. Starting from the Gate-All-Around stacked nanowires process flow, we will propose some modifications to create in a first time, 1T1R transistors. This simplified devices will allow us to screen the different materials and sizing to experimentally choose the best trade-off for the full MY-CUBE structure manufacturing.

a. Gate-All-Around stacked nanowires detailed process flow

In this part, the process flow to create Gate-All around stacked nanowires is described. Unlike the gate first process flow presented in chapter III for low-temperature transistors, this process is called “gate last”. In fact, a sacrificial gate is used during the process but is filled with gate material just before the formation of the source and drain contacts. The process flow is described in Fig. 255 and will be briefly commented.

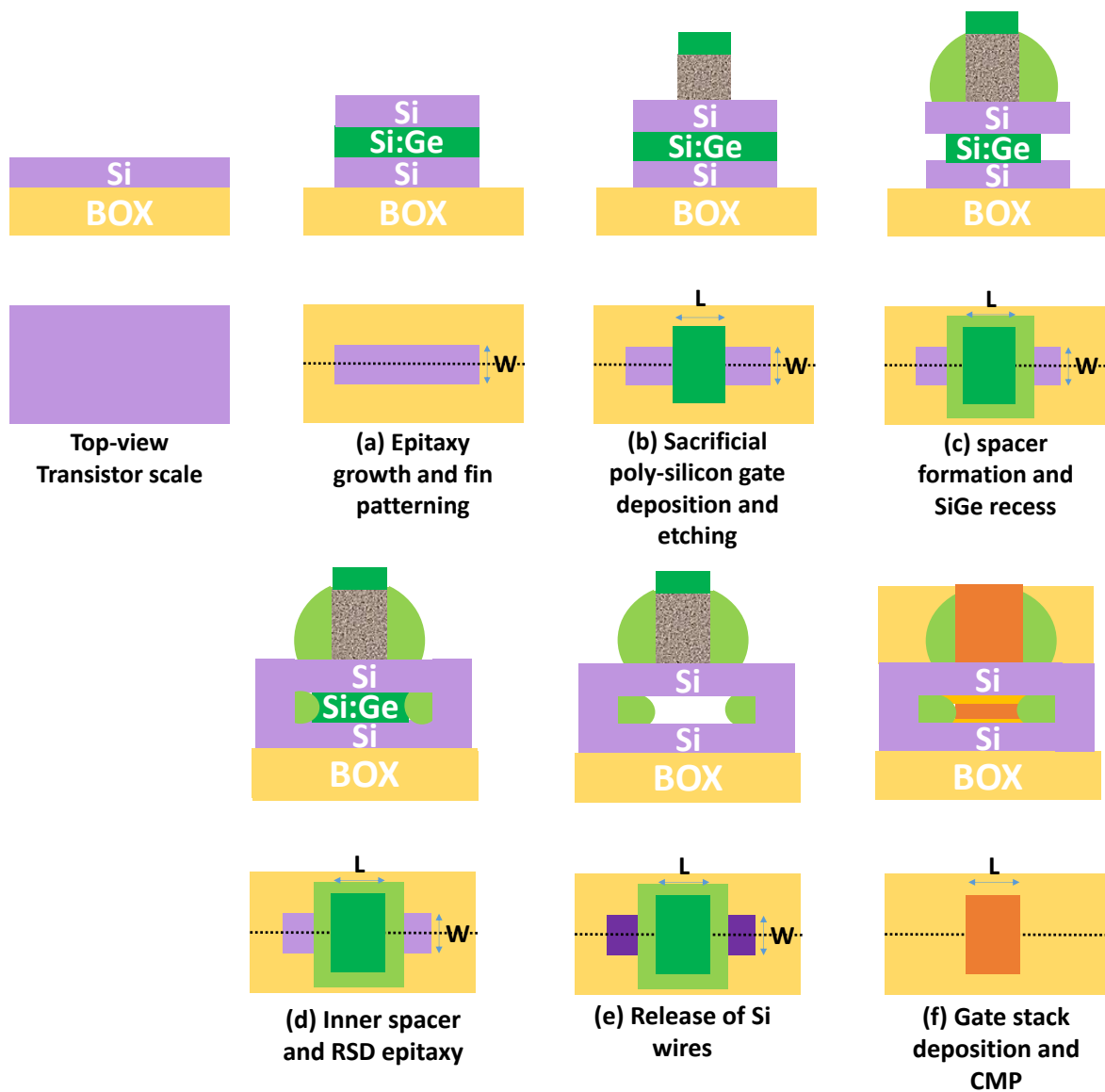


Fig. 255: Presentation of stacked nanowire process flow.

The first step (Fig. 255-a) consists in an epitaxial growth of $(\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si})$ multilayers. For the sake of simplicity, Fig. 255 presents the case where only two nanowires are stacked. However, up to seven

stacked nanowires have been demonstrated in the literature in [18]. The silicon in this superlattice will be the future channel material and the sacrificial SiGe layer will be removed latter. Then (Si_{0.7}Ge_{0.3}/Si) multilayers are patterned to define the transistor width. The second step (Fig. 255-b) is about the creation of a SiO₂/poly-si dummy gate formation. The material are deposited and planarized though a CMP process to counteract the large topography induced by the multilayers and finally patterned. Then (Fig. 255-c) a spacer is defined before etching partially the SiGe/Si layers in the source/drain region, recessing the SiGe layers laterally and forming an inner spacer in these cavities.

The fourth step (Fig. 255-d) is about the epitaxial growth of raised source and drain, connecting all the wires together. Fig. 256 presents a TEM cross-section after RSD definition. To finish with, the dummy gate is taken away and the Si nanowires are released by etching selectively the SiGe layers during the replacement metal gate module. This is followed by gate stack deposition: HfO₂, TiN and W, wrapping the Si wires and planarization. To finish with, Back End Of Line contacts and metal lines are fabricated. A TEM cross-section of the final structure for two stacked nanowires is presented in Fig. 257. The elements characterisation highlight the conformity of the gate stack which wraps the wires. Note that the bottom channel is a tri-gate configuration and not a GAA one.

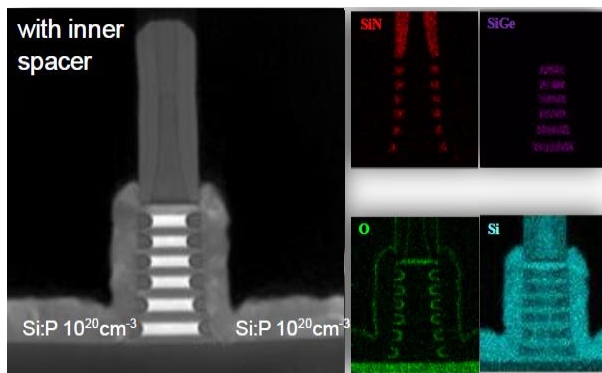


Fig. 256: TEM Cross-section of a 7 stacked nanowire transistors before SiGe removal. Figure from [18].

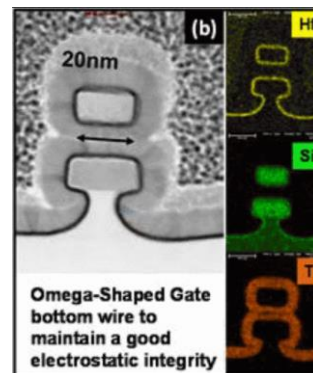


Fig. 257: TEM Cross-section of a two stacked nanowires transistors after whole processing. Figure from [329].

b. Modification to standard process flow to integrate memory elements

Starting from this process flow, some modifications are done to integrate the memory element laterally to each nanowire drain. For the sake of simplicity, we will consider the case where the number of wires is equals to one (trigate configuration). There is no significant process integration difference between stacked GAAs and My-CUBE until the formation of the source and drain contacts. First the source contact must be dissociated from the drain contact and is realised conventionally as illustrated in Fig. 258-1. Then the source contact is etched down to the BOX instead of stopping onto the raised source and drain. By doing so, the memory element can be integrated laterally directly next to the transistor drain end. In a nanowire configuration, this lateral integration will dissociate each nanowire from the drain side. If not, all of the stacked nanowire drains would be electrically connected to a “big” memory element. In the last step (Fig. 258-4), the memory element is formed by silicide at the drain extremity, conformal HfO₂ deposition, Ti/TiN and W filling. The thickness of each layer can be tuned to choose the appropriate forming voltage according to the transistor drive current. I managed the process integration of such a batch. Different process variants were done such as the HfO₂ thickness which varies from 5nm to 10nm. Concerning the transistor itself, the doping level of the channel varied from 5.10¹⁸ to 5.10¹⁹ at/cm³ and the gate oxide is either GO1 or GO2.

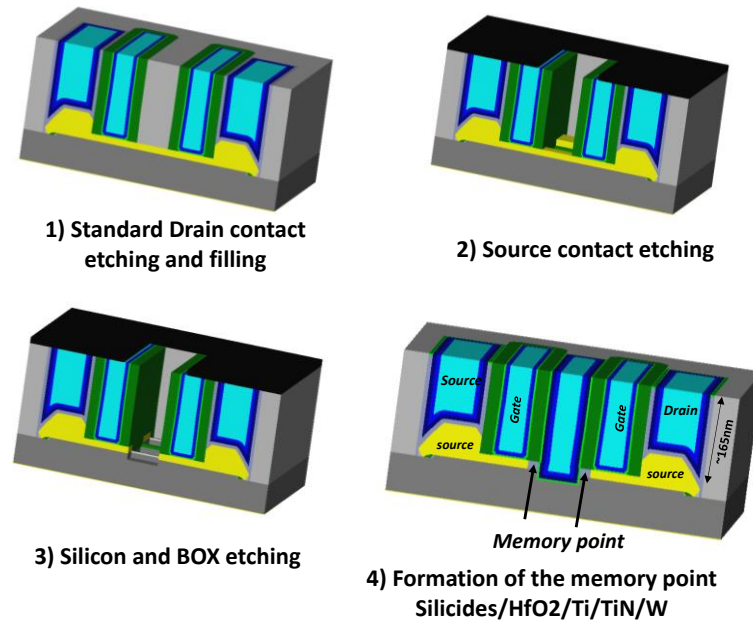


Fig. 258: Presentation of the modified Stacked nanowire process flow in the particular case of a single nanowire.

Morphological studies have been performed in the cleanroom to validate this process flow, especially the dissociation of source and drain part. A tilted SEM image is given in Fig. 259 at the dummy gate removal step. These independent 1T1R structures detain a simplify process flow compared to the matrix and are useful to screen transistor sizing (t_{si} and N_D for instance) and memory element (oxide thickness...). The batches were not completed at the end of my thesis and there is no associated electrical results. However, as a perspective of this work, this 1T1R structure will provide insights about the materials and sizing to select for the junctionless transistor as well as for the memory element in the scope of a matrix integration.

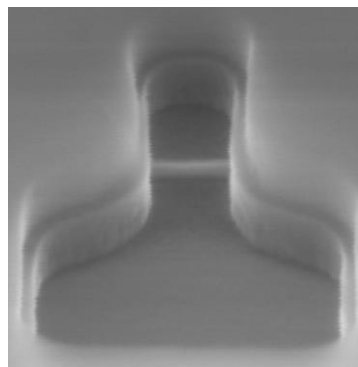


Fig. 259: dummy gate removal step.

To go further, a My-CUBE layout, corresponding to the previously introduced structure in 2-d which enables efficiently Scouting logic, as well as a possible process flow is proposed in Annex III.

From the previous sections, we demonstrated the feasibility of scouting logic up to three layers in this structure and preliminary batches are done to select the best sizing for junctionless transistor and OxRAM. However, a major known drawback of junctionless devices is the threshold voltage mismatch degradation due to channel doping level. This additional variability could be translated into ON current variability broadening the OxRAM resistive states distributions which can be detrimental for SCL. So far, the measurements of ON current variability were correct for our targeted application but were done on a limited number of individual structures for specific voltage conditions. That is why, there is a need of an in-depth characterisation in all operation regime of drain current of junctionless transistors with dedicated mismatch structures which will be presented in the next paragraph.

4- Variability

In this part, we will tackle both local and global variability for Inversion-Mode devices (IM), Junctionless Accumulation Mode (JAM) and purely Junctionless devices (JL). For more details about the device process flow, please refer to chapter III section 6-. The idea is to verify if the choice of a junctionless structure for the IMC is relevant. In fact, to tighten the OxRAM resistance distribution, we need a low variability on ON current. In this section, we will first introduce the standard way of evaluating the threshold voltage mismatch. Then we will study the mismatch in all operation regime by using the gate input referred normalized matching parameter. After the variability of drain current is investigated and modelled. To finish with, ON current variability is studied with the experimental conditions defined in subsection 2-b.i.

Specific transistor structures are designed for mismatch measurements. In fact, to consider the variation between two adjacent devices, the so-called local variation, the “pairs” of transistors (denoted MOS1 and MOS2) are designed with the following characteristics:

- Pairs of transistors, spaced with the minimum distance allowed by design rules.
- The environment is identical for both devices
- The devices are electrically independent with symmetric connections.

Due to the higher density in matching devices than in isolated devices measured in the previous chapter, some sizing differences might appear. For instance the width might be larger in dense area than for isolated devices, leading to slight performance differences. Furthermore, since a standard deviation is computed and should be representative of a technology, a large number of dices are measured in order to ensure a significant population statistics. In this work 112 paired devices N_{dies} are systematically measured on the whole 300mm wafer. For each technological variant (IM, JL and JAM), two wafers are considered.

a. Standard evaluation of the mismatch: Pelgrom plots.

First, let's have a look at I_D - V_G curves for various dimensions to intuit the variability. Note that when we consider all the transistor I_D - V_G and not the I_D - V_G difference between paired devices, we talk about global variability and not (local) mismatch. Fig.260-a ($W=230\text{nm}$ and $L=47\text{nm}$, D1) and Fig.260-b ($W=230\text{nm}$ and $L=18\text{nm}$, D2) show that JL transistors are less prone to short-channel effect since the subthreshold slope even for $L=18\text{nm}$ is not degraded. It can be attributed to channel length modulation. However, JAM transistor detains more variability in the sub-threshold regime than IM, which seems even worse for smaller gate length. In addition, JL sensitivity on access resistance (for $V_G > 0.5\text{V}$) is already seen on D1 and D2 with large gate width of $0.23\mu\text{m}$. This variability is reinforced for lower dimensions ($W=20\text{nm}$), as seen on D3 and D4. However, IM and JAM have the same behavior for ultra-scaled devices (D3 and D4) and are less sensitive to SCE. Their variability seems higher in the subthreshold regime than in the ON-state regime.

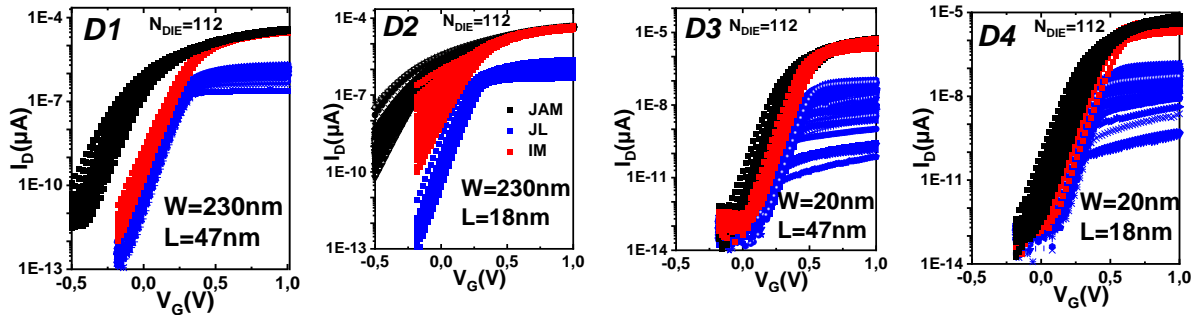


Fig.260: I_D - V_G for various dimensions, highlighting JL immunity to SCE and higher RSD than IM and JAM, $V_{DS}=50mV$. Figure from [331].

To go further, Fig.261 presents the Pelgrom plot of global + local variability (for definition, see subsection chapter III 4-d). The threshold voltage is extracted with the constant current method at $I_{th}=10^{-7} \cdot W/L$. As seen on the I_D - V_G curves, smaller device surface (larger $\frac{1}{\sqrt{W \cdot L}}$ values) leads to higher V_T deviation. During the device fabrication, the different width and length variations have been measured by SEM on other structures. We estimate that the gate length and active width within-wafer uniformity are about $\Delta L=4nm$ and a $\Delta W=2nm$. As said in chapter III 4-d, the ΔL or ΔW induced variability becomes predominant for smaller dimensions. However, for infinite devices ($\frac{1}{\sqrt{W \cdot L}} = 0$), a global variability offset is seen on the wafer. In reality, the biggest transistor dimension is $W=L=10\mu m$. Nevertheless, at such a sizing, the ΔL or ΔW are not significant and can't explain this offset. In fact the variability sources for large devices are silicon thicknesses, gate stack oxide thicknesses. In our case, the t_{si} monitored during the process indicates a 1nm variation on the 300nm wafer. We have carried out TCAD simulations to study the sensitivity of V_T on such a variation. They (Fig.262) show that for large TG-REF devices a $\Delta t_{si}=1nm$ implies a ΔV_T of 15mV in the junctionless case and a ΔV_T of 4.8mV for IM. JL devices threshold voltage is highly sensitive on t_{si} and this sensitivity is included into the V_T formula (Eq. 24 and Eq. 25). These simulated values correspond to the measured ones. Note that JL and JAM detains the same offset, meaning that this additional variability is not caused by source and drain region but rather by channel doping or gate stack.

Let's tackle now the local variability only. The associated Pelgrom plot is presented in Fig.263. The extracted A_{vt} values are $1mV \cdot \mu m$ for IM, $1.4mV \cdot \mu m$ for JAM and $1.7mV \cdot \mu m$ for JL. IM devices detains less variability than junctionless ones due to their undoped channel (no Random Dopant Fluctuation). The difference between JAM and JL could be explained by additional variability linked to source and drain resistance since the impact is seen for $V_G > 0.4V$ on Fig.260-d. This explanation will be confirmed by in-depth analyses (presented in the following sections).

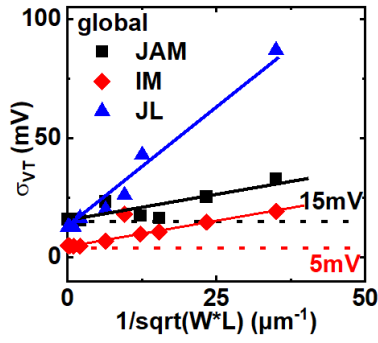


Fig.261: Pelgrom plot. The local and global variability is taken into account. The same variability offset is seen for JL and JAM transistor, $V_{DS}=0.8V$. Figure from [331].

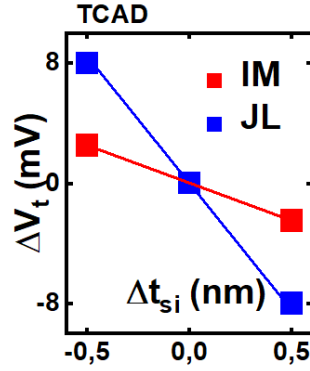


Fig.262: TCAD simulation to analyze the sensitivity on V_t of silicon variations measured on the wafer. The same offset as the Pelgrom plot is seen.

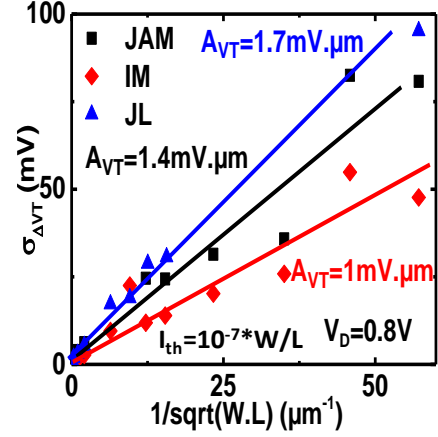


Fig.263: Pelgrom plot (local variability determined from V_t variation of matched pairs). The threshold voltage is extracted with the constant current method at $I_{th}=10^{-7}.W/L$.

In addition to, the constant current threshold voltage extraction gives an indication of variation for a certain amount of current but do not rely on a physical extraction of the threshold voltage. For instance, junctionless devices feature two threshold voltages, which might be associated to a different variability. When we recall the V_T and V_{FB} equations (Eq. 24 and Eq. 25), not the same dependency is seen with respect to the donor doping level N_D . Starting from this, $\frac{\partial V_T}{\partial N_D}$ and $\frac{\partial V_{FB}}{\partial N_D}$ are different and cannot be approximated to V_T extracted by constant current method and derived with respect to N_D . To go further, the variability in all regimes is considered in the following paragraph.

$$V_T(N_D) = V_{FB}(N_D) - q \cdot N_D t_{si} \cdot \frac{1}{C_{ox}} \quad \text{Eq. 24}$$

$$V_{FB}(N_D) = k \cdot T \cdot \ln\left(\frac{N_D}{N_i}\right) + \varphi_m \quad \text{Eq. 25}$$

b. Gate input referred normalized matching parameter

The gate-input referred normalized matching parameter $iA_{\Delta V_g}$ ($mV \cdot \mu m$) is defined by Eq. 26. Please note that $iA_{\Delta V_g}$ ($V_G=V_T$) corresponds to the A_{vt} parameter. Fig.264 presents the $iA_{\Delta V_g}$ for large devices ($W=L=10\mu m$) and all the technological variants. As far as IM is concerned, $iA_{\Delta V_g}$ is constant in the subthreshold region and increases for $V_G > V_{T-IM}$. Contrary to this behavior, junctionless devices reach a local maximum for their first threshold voltage before increasing again for larger gate voltages. In order to explain this behaviour, as well as the long and large channel matching performance, we have performed TCAD simulations. The experimental JL $iA_{\Delta V_g}$ feature is well reproduced with TCAD simulations where for JL devices, only a doping variation have been considered. For IM TCAD simulations, a t_{si} variation is assumed.

$$iA_{\Delta V_g} = \frac{\sigma\left(\frac{\Delta I_D}{I_D}\right)}{\frac{g_m}{I_D}} \cdot \sqrt{W \cdot L} \quad \text{Eq. 26}$$

To dissociate the impact of the bulk conduction and the accumulation conduction on the mismatch, a back-bias have been applied. In fact, a negative back-bias will move the conductive channel closer to the interface and suppress the volume conduction. Fig.265 shows that the hump on the g_m figure that is characteristic of this volume conduction can be suppress with $V_B=-10V$ ($BOX=145nm$). Conversely, a $+10V$ V_B will increase this volume conduction. This modulation of the conduction type is seen on $iA_{\Delta V_g}$

variability (Fig.266): for $V_B=-10V$, the gate input normalized matching parameter detains the same behavior as IM ones. Furthermore, the variability is accentuated for $V_B=+10V$. It shows the necessity of taking into account this specific feature linked to junctionless operation.

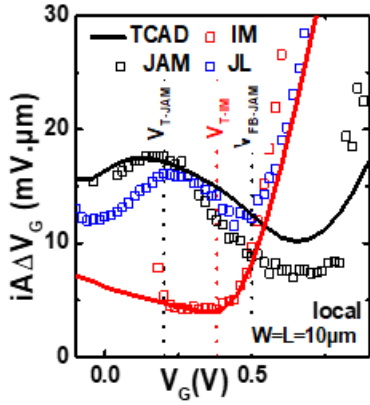


Fig.264: $i\Delta V_G$ as a function of gate voltage for large dimension $L=W=10\mu m$. JL and JAM features a local maximum around V_T and a minimum for $V_G > V_{FB}$. TCAD sensitivity simulation reproduce well this behavior. Figure from [331].

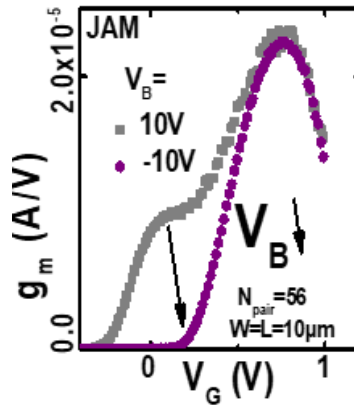


Fig.265: $g_m(V_G)$ for $V_B=10V$ and $V_B=-10V$, modulating the position of the conduction channel. Negative back-bias suppresses the JL characteristic hump by moving the conduction. Figure from [331].

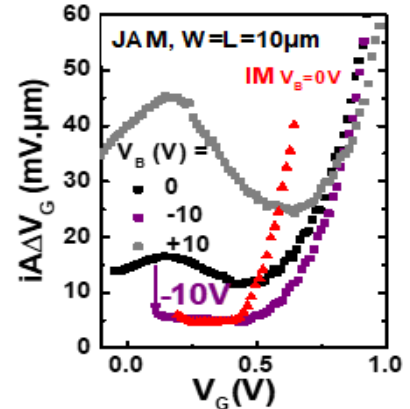


Fig.266: $i\Delta V_G$ as a function of gate voltage for various back-bias in planar devices. The negative back-bias suppresses the variability associated to V_{T-JL} . The obtained $i\Delta V_G$ has the same behavior as IM one. Figure from [331].

Fig.267 presents $i\Delta V_g$ for all the chosen dimensions. Please note that the minimum of $i\Delta V_g$ (for $V_G=V_T$) corresponds to iA_{VT} and this value is consistent with Pelgrom plots. In fact, if we take the mean of $\min(i\Delta V_g)$ for all dimensions, the A_{VT} is found back. For JL transistors, the double hump is seen for $W=240nm$ but no more for $W=20nm$. This is attributed to the prevalence of R_{SD} at such dimensions. On the contrary for JAM devices, the hump is seen at $W=20nm$ and not $W=240nm$. The parabolic form for $W=240nm$ is similar as the one seen for IM devices where the electrostatic control is poor. In fact, the ideality factor n variations are more important. Also, IM devices curves at $W=20nm$ detains a ‘‘plateau’’.

Furthermore, the same measurements are done at $V_D=0.8V$ (not shown here). We can observe that JL transistors achieves lower variability value. This is explained by the channel length modulation enhanced at high V_D .

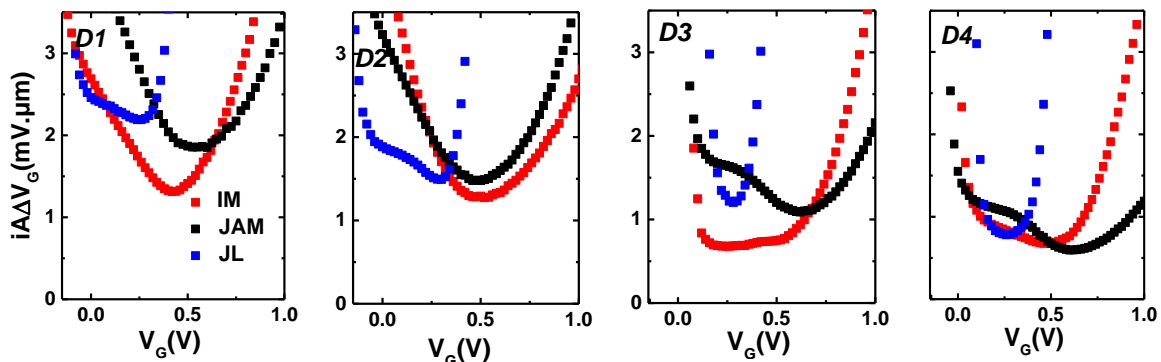


Fig.267: $i\Delta V_G$ as a function of gate voltage for various dimensions in linear regime $V_{DS}=50mV$. Figure from [331].

However, to dissociate the contribution to each MOSFET parameter and confirm what we figure out from the $i\Delta V_g$ curves, a modelling of the drain current is done and explained here-below.

c. Drain current local and global variability in all-regimes

The general drain current mismatch model is taken from [332] and was deriving for inversion-mode transistors. It considers the MOSFET drain current sensitivity to parameters such as V_T , β and R_{SD} [333]. Eq. 27 is the Taylor approximation describing the drain current variation. After calculating the derivative, the drain current mismatch in linear region is expressed as Eq. 28. To model all the operating regions, some terms are added in [334] (Eq. 29).

$$\left(\frac{dI_D}{I_D}\right) = \left(\frac{1}{I_D} \cdot \frac{\partial I_D}{\partial V_T}\right) \cdot dV_T + \left(\frac{1}{I_D} \cdot \frac{\partial I_D}{\partial \beta}\right) \cdot d\beta + \left(\frac{1}{I_D} \cdot \frac{\partial I_D}{\partial R_{SD}}\right) \cdot dR_{SD} \quad \text{Eq. 27}$$

$$\sigma^2\left(\frac{\Delta I_D}{I_D}\right) = \left(\frac{g_m}{I_D}\right) \cdot \sigma^2(\Delta V_T) + [1 - g_d \cdot R_{SD}]^2 \cdot \sigma^2\left(\frac{\Delta \beta}{\beta}\right) + g_d^2 \cdot \sigma^2(\Delta R_{SD}) \quad \text{Eq. 28}$$

$$\begin{aligned} \sigma^2\left(\frac{\Delta I_D}{I_D}\right) &= \left(\frac{g_m}{I_D}\right) \cdot \sigma^2(\Delta V_T) + \left[1 - \frac{I_D}{V_D} \cdot R_{SD}\right]^2 \cdot \sigma^2\left(\frac{\Delta \beta}{\beta}\right) \\ &+ \left[\ln\left(\frac{I_D}{I_{D,th}}\right)\right]^2 \cdot \left[\exp\left(\frac{-I_{D,th}}{I_D}\right) - 1\right]^2 \cdot \sigma^2\left(\frac{\Delta n}{n}\right) \\ &+ \left(\frac{I_D}{V_D}\right)^2 \cdot \sigma^2(\Delta R_{SD}) \text{ where } n = \frac{q}{k \cdot T \cdot SS} \text{ and } \beta = \mu_0 \cdot C_{ox} \cdot V_D \cdot \frac{W}{L} \end{aligned} \quad \text{Eq. 29}$$

This model depends on four fitting parameters:

- V_T variability: when normalized by $\sqrt{W \cdot L}$, corresponds to A_{VT} .
- β variability: reflects the variability of mobility.
- Ideal factor n variability: reflects the variability of the SS. Its domain of application is in the sub-threshold regime and is monitored by a threshold current $I_{D,th}$.
- R_{SD} variability: becomes predominant when $\frac{I_D}{V_D}$ overcomes the other contributions, especially at large V_G .

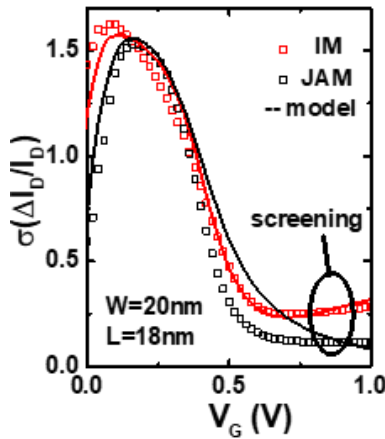


Fig.268: Drain current variability as a function of gate voltage. The model fits well IM and JAM transistors. Figure from [331].

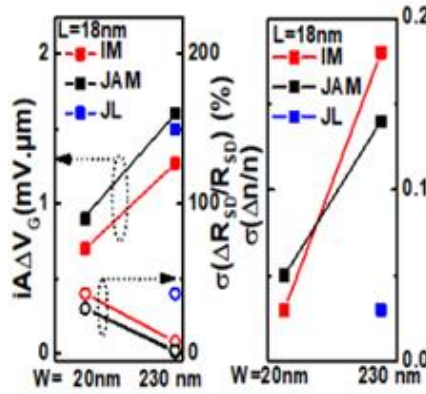


Fig.269: Universal model parameter extraction. A higher resistance variability is extracted for JL devices. Figure from [331].

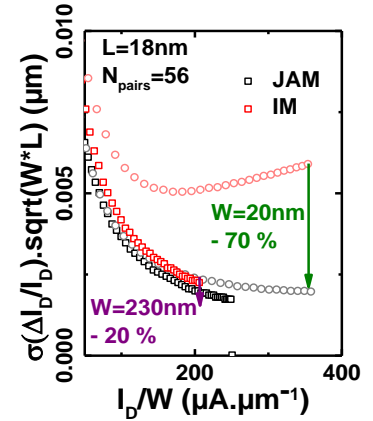


Fig.270: Total current variability as a function of drive current. Up to -70% gain on variability is seen for JAM w.r.t. IM at the same drive current at $W=20\text{nm}$ and $L=18\text{nm}$. Figure from [331].

In this work, the current difference between two paired-device can be analysed in a range of up to several decades. That is why, the drain current mismatch is evaluated with the log difference: $\frac{\Delta I_D}{I_D} = \ln\left(\frac{I_{D2}}{I_{D1}}\right)$ as in [335].

Fig.268 shows the fitting of D4 for JAM and IM transistors (JL being too much impacted by access resistances). Even if the model is not exact for junctionless transistors and does not take into account the two different variabilities associated to V_T and V_{FB} , the data points fit with the model. Fig.269 recaps the matching parameters extracted with this fitting for the previously chosen dimensions. It confirms the intuition developed in previous part. In fact, the variability associated to R_{SD} is much higher in JL than JAM and IM (which are equivalent). However, the ideal factor variability is higher for IM and JAM than JL. As far as A_{VT} is concerned (in $mV \cdot \mu m$), the values are coherent with the ones extracted in the Pelgrom plot.

Let's put the emphasis on Fig.270 where for high value of V_G , JAM devices shows significant lower value of drain current variability. As seen on mobility, this effect is attributed to impurity screening once the accumulation layer is formed, reducing the RDF-induced variability, in agreement with theoretical predictions [336]. In fact, when the drain current variability is plotted versus the drive current (Fig.270), one can observe that for the same drive current, up to -70% variability gain is seen on JAM devices vs. IM. This effect seems more pronounced for small dimensions. For larger one, the gain is above 20/30%.

Furthermore, similarly as for local variability, the global variability is up to 30% better for JL/JAM than IM at high current for $W=L=10\mu m$ mainly because of the dopant screening (Fig.272). In addition, we found back the same offset values as in the Pelgrom plot when considering ΔV_G ($\Delta V_G = \frac{\sigma(I_D)}{g_m} \cdot \sqrt{W \cdot L}$) as a function of V_G for $W=L=10\mu m$. (Fig.271).

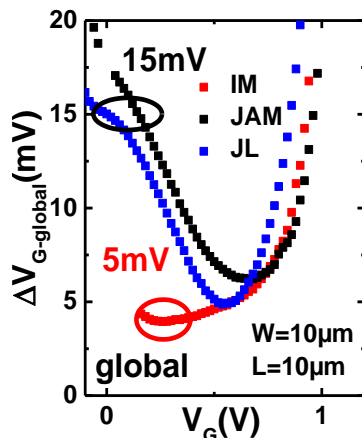


Fig.271: ΔV_G global. The same offset value as on the Pelgrom plot is seen. At low- V_G , JL devices detain three times more variability than IM, attributed to t_{si} variation. Figure from [331].

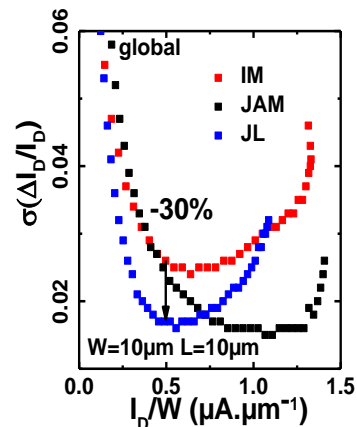


Fig.272: total + global current variability as a function of drive current. Even for large devices, a -30% gain on global variability is seen for JL and JAM compared to IM. Figure from [331].

The next subsection will present the variability results for the selected conditions in subsection 2-b.i which corresponds to the SET operation.

d. Variability of JAM devices for IMC

To be able to SET the OxRAM and to form the oxygen filament, enough current must be driven at the ON state of junctionless transistors. Compare to the previous mismatch results, the gate voltage is increased up to 1.5V as well as the drain voltage, up to 1.3V. This large gate overdrive condition delivers sufficient current ($\sim 75\mu\text{A}$). However this current increase could come along with a degraded variability. To provide insights about variability at large gate overdrive, we drawn the Pelgrom plot (Fig. 273) for $V_D=1.3\text{V}$, considering both local and global variability. The V_T variability for $L=80\text{nm}$ and $W=50\text{nm}$ is 48mV as indicated on the I_D-V_G plot in Fig. 237. To have more information about the ON current variability at $V_D=1.3\text{V}$ and $V_G=1.5\text{V}$, the ON current standard deviation is represented as a function of the ON current in Fig. 274 for $W=50\text{nm}$ and $W=240\text{nm}$ and various gate lengths. We do observe that for each gate length, the variation of current is proportional to the delivered current. However, for a same level of ON current variability, $W=240\text{nm}$ delivers more current. Based on the Pelgrom plot, a 42mV V_T variability is extracted for a larger width $W=75\text{nm}$ and less than $8\mu\text{A}$ ON variability are predicted. If we consider $+4\mu\text{A}$ around the nominal value $75\mu\text{A}$ on the graph Resistance as a function of I_{CC} presented in [315], around 150 ohm additional variability (on sigma) is roughly estimated for the LRS state. As a reminder, the weak condition ($I=70\mu\text{A}$) detains a 2000 variability for the LRS. The variability on HRS seems to be neglectable.

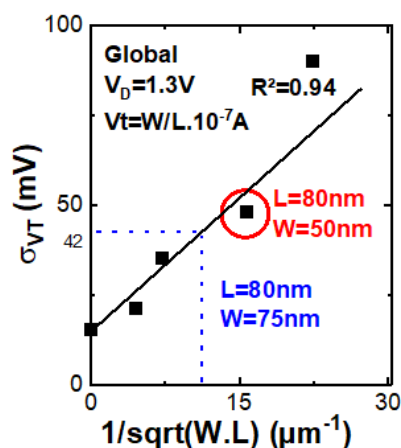


Fig. 273: Pelgrom plot of JAM devices considering local and global variability and a 1.3V drain voltage.

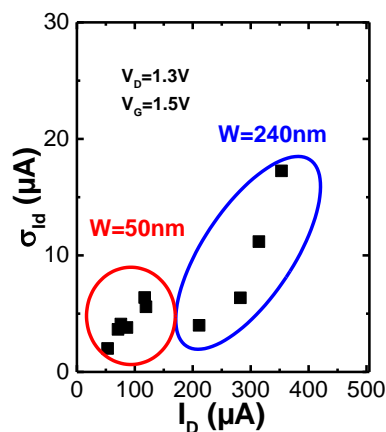


Fig. 274: ON current ($V_G=1.5\text{V}$ and $V_D=1.3\text{V}$) variability as a function of ON current for $W=50\text{nm}$ and $W=240\text{nm}$ and various gate lengths.

In this section we demonstrated that the variability of junctionless devices is higher than conventional one for sub-threshold operation but not at high gate voltage. In the scope of IMC, the junctionless transistor will be used ON to SET or RESET the OxRAM cell. In fact, the ON current variability of the transistor can impact the resistance distributions, which is in the junctionless case even better than IM devices: junctionless transistors are a great candidate for My-Cube structure. In fact, we verified that the ON current variability of junctionless transistor at large gate overdrive will not impact a lot the future OxRAM distribution.

5- Conclusion of chapter IV

In this chapter, IMC is envisioned as the solution to break the so-called “memory wall”. Gathering memory and computational blocs could reduce drastically the amount of energy (and increase the bandwidth) wasted during data transfer. A state of the art presenting the new computing paradigms is provided, highlighting the interest of Scouting logic to perform Boolean logic operation into a memristor array. Then a review of memristor devices proposes to use OxRAM technology for the memory element thanks to their scalability, endurance and compatibility with CMOS technologies. Combining, junctionless stacked nanowires transistor with lateral OxRAM, an ultra-dense low power cube is proposed for IMC applications. Then, by means of simulations (SPICE and TCAD), a subset of MY-CUBE structure have been proven compatible with Scouting logic. After, planar 1T1R structure are fabricated to screen the different sizing feasible for junctionless transistor and memory element. To finish with, to verify the junctionless compatibility with the low ON current variability requirement, a study of drain current variability in all operation regime is carried out. The specificities of JL transistors are highlighted: a higher sensitivity to silicon thickness at low gate voltage but a lower variability for large gate voltage attributed to Coulomb scattering screening in the accumulation layer.

The key points are:

- Nowadays, most of the energy is wasted for data transfer between memory and computational parts.
- In-Memory Computing, as opposed to Von-Neumann architecture, proposes to get closer memory and computational parts.
- In the scope of MY-CUBE project, an ultra dense and low-power structure is proposed to leverage IMC. This structure combines state of the art devices: junctionless stacked nanowires and OxRAM technology.
- Scouting logic have been demonstrated by mean of simulations (TCAD and SPICE) in a MY-CUBE pillar: up to 3 stacked layers can be used.
- A process flow as well as preliminary studies to fabricate the structure is exposed.
- The compatibility of junctionless devices with this type of applications, especially concerning the variability, have been verified.

The perspectives of this work consist in:

- Electrical characterisation of the 1T1R structures which presents several technological variants such as junctionless channel doping, gate oxide thickness, OxRAM HfO₂ thicknesses.
- Consideration of a peripheral circuit for scouting logic simulations.
- Benchmarking of MY-CUBE structure in terms of area and energy efficiency.
- Proposition of a process flow including the formation of metallic bitlines to avoid a huge access resistance.

REFERENCES:

- [1] H. Mujtaba, « AMD EPYC Rome CPUs Feature 39.54 Billion Transistors, IOD Detailed », *Wccftech*, oct. 22, 2019. <https://wccftech.com/amd-2nd-gen-epyc-rome-iod-ccd-chipshots-39-billion-transistors/> (consulté le sept. 04, 2020).
- [2] T. Ghani *et al.*, « A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors », in *IEEE International Electron Devices Meeting 2003*, déc. 2003, p. 11.6.1-11.6.3, doi: 10.1109/IEDM.2003.1269442.
- [3] M. A. Quevedo-Lopez *et al.*, « High performance gate first HfSiON dielectric satisfying 45nm node requirements », in *IEEE International Electron Devices Meeting, 2005. IEDM Technical Digest.*, déc. 2005, p. 4 pp. - 428, doi: 10.1109/IEDM.2005.1609369.
- [4] T. Huynh-Bao *et al.*, « Statistical Timing Analysis Considering Device and Interconnect Variability for BEOL Requirements in the 5-nm Node and Beyond », *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 25, n° 5, p. 1669-1680, mai 2017, doi: 10.1109/TVLSI.2017.2647853.
- [5] S. Manipatruni, D. E. Nikonov, et I. A. Young, « Beyond CMOS computing with spin and polarization », *Nat. Phys.*, vol. 14, n° 4, p. 338-343, avr. 2018, doi: 10.1038/s41567-018-0101-4.
- [6] « ENIAC », *Wikipédia*. mai 08, 2020, Consulté le: juill. 23, 2020. [En ligne]. Disponible sur: <https://fr.wikipedia.org/w/index.php?title=ENIAC&oldid=170603125>.
- [7] « Logic Technology - Taiwan Semiconductor Manufacturing Company Limited ». https://www.tsmc.com/english/dedicatedFoundry/technology/logic.htm#l_7nm_technology (consulté le avr. 24, 2020).
- [8] R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassous, et A. R. LeBlanc, « Design of ion-implanted MOSFET's with very small physical dimensions », *IEEE J. Solid-State Circuits*, vol. 9, n° 5, p. 256-268, oct. 1974, doi: 10.1109/JSSC.1974.1050511.
- [9] G. E. Moore, « Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff. », *IEEE Solid-State Circuits Soc. Newsl.*, vol. 11, n° 3, p. 33-35, sept. 2006, doi: 10.1109/N-SSC.2006.4785860.
- [10] Jing Wang et M. Lundstrom, « Does source-to-drain tunneling limit the ultimate scaling of MOSFETs? », in *Digest. International Electron Devices Meeting*, déc. 2002, p. 707-710, doi: 10.1109/IEDM.2002.1175936.
- [11] A. Hikavy, I. Zyulkov, H. Mertens, L. Witters, R. Loo, et N. Horiguchi, « Use of high order precursors for manufacturing gate all around devices », *Mater. Sci. Semicond. Process.*, vol. 70, p. 24-29, nov. 2017, doi: 10.1016/j.mssp.2016.10.044.
- [12] O. Weber, « FDSOI vs FinFET: differentiating device features for ultra low power IoT applications », in *2017 IEEE International Conference on IC Design and Technology (ICICDT)*, mai 2017, p. 1-3, doi: 10.1109/ICICDT.2017.7993513.
- [13] C. Auth *et al.*, « A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors », in *2012 Symposium on VLSI Technology (VLSIT)*, juin 2012, p. 131-132, doi: 10.1109/VLSIT.2012.6242496.
- [14] R. Carter *et al.*, « 22nm FDSOI technology for emerging mobile, Internet-of-Things, and RF applications », in *2016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 2.2.1-2.2.4, doi: 10.1109/IEDM.2016.7838029.
- [15] N. Planes *et al.*, « 28nm FDSOI technology platform for high-speed low-voltage digital applications », in *2012 Symposium on VLSI Technology (VLSIT)*, juin 2012, p. 133-134, doi: 10.1109/VLSIT.2012.6242497.
- [16] O. Weber *et al.*, « 14nm FDSOI technology for high speed and energy efficient applications », in *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, juin 2014, p. 1-2, doi: 10.1109/VLSIT.2014.6894343.
- [17] F. Andrieu *et al.*, « Fully depleted Silicon-On-Insulator with back bias and strain for low power and high performance applications », in *2010 IEEE International Conference on Integrated Circuit Design and Technology*, juin 2010, p. 59-62, doi: 10.1109/ICICDT.2010.5510295.
- [18] S. Barraud, « 7-levels-stacked nanosheet GAA transistors for high performance computing, » », *Symposium on VLSI Technology (VLSI Technology)*, 2020.

-
- [19] N. Loubet *et al.*, « Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET », in *2017 Symposium on VLSI Technology*, juin 2017, p. T230-T231, doi: 10.23919/VLSIT.2017.7998183.
- [20] C. texte fournit des informations générales S. ne peut garantir que les informations soient complètes ou exactes E. raison de cycles de mise à jour variables et L. S. P. A. D. D. P. R. Q. C. R. D. L. Texte, « Thème: L'utilisation des smartphones en France », *Statista*. <https://fr.statista.com/themes/2758/l-utilisation-des-smartphones-en-france/> (consulté le juill. 23, 2020).
- [21] J. Desjardins, « How Much Data is Generated Each Day? », *Visual Capitalist*, avr. 15, 2019. <https://www.visualcapitalist.com/how-much-data-is-generated-each-day/> (consulté le juill. 23, 2020).
- [22] « About - IEEE International Roadmap for Devices and Systems™ ». <https://irds.ieee.org/> (consulté le juill. 23, 2020).
- [23] P. Batude *et al.*, « 3D monolithic integration », in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, mai 2011, p. 2233-2236, doi: 10.1109/ISCAS.2011.5938045.
- [24] M. Vinet *et al.*, « Opportunities brought by sequential 3D CoolCube™ integration », in *2016 46th European Solid-State Device Research Conference (ESSDERC)*, sept. 2016, p. 226-229, doi: 10.1109/ESSDERC.2016.7599627.
- [25] J. Shi *et al.*, « A 14nm FinFET transistor-level 3D partitioning design to enable high-performance and low-cost monolithic 3D IC », in *2016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 2.5.1-2.5.4, doi: 10.1109/IEDM.2016.7838032.
- [26] P. Vivet *et al.*, « Monolithic 3D: an alternative to advanced CMOS scaling, technology perspectives and associated design methodology challenges », in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, déc. 2018, p. 157-160, doi: 10.1109/ICECS.2018.8617955.
- [27] G. Yeap, « Smart mobile SoCs driving the semiconductor industry: Technology trend, challenges and opportunities », in *2013 IEEE International Electron Devices Meeting*, déc. 2013, p. 1.3.1-1.3.8, doi: 10.1109/IEDM.2013.6724540.
- [28] « Moore's law scaling dead by 2021, to be replaced by 3D integration - ExtremeTech ». <https://www.extremetech.com/extreme/232342-moores-law-scaling-dead-by-2021-to-be-replaced-by-3d-integration> (consulté le juill. 24, 2020).
- [29] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, et E. Eleftheriou, « Memory devices and applications for in-memory computing », *Nat. Nanotechnol.*, mars 2020, doi: 10.1038/s41565-020-0655-z.
- [30] « Simulating a Ring Oscillator ». http://web.engr.uky.edu/~elias/Tutorial_F05/introsim.htm (consulté le mai 14, 2020).
- [31] V. F. Pavlidis, I. Savidis, et E. G. Friedman, « Chapter 19 - Case Study: 3-D Power Distribution Topologies and Models », in *Three-Dimensional Integrated Circuit Design (Second Edition)*, V. F. Pavlidis, I. Savidis, et E. G. Friedman, Éd. Boston: Morgan Kaufmann, 2017, p. 565-603.
- [32] V. Gutnik et A. P. Chandrakasan, « Embedded power supply for low-power DSP », *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 5, n° 4, p. 425-435, déc. 1997, doi: 10.1109/92.645069.
- [33] P. S. Nair, S. Koppa, et E. B. John, « A comparative analysis of coarse-grain and fine-grain power gating for FPGA lookup tables », in *2009 52nd IEEE International Midwest Symposium on Circuits and Systems*, août 2009, p. 507-510, doi: 10.1109/MWSCAS.2009.5236045.
- [34] R. Kulkarni et S. Y. Kulkarni, « Power analysis and comparison of clock gated techniques implemented on a 16-bit ALU », in *International Conference on Circuits, Communication, Control and Computing*, nov. 2014, p. 416-420, doi: 10.1109/CIMCA.2014.7057835.
- [35] J. Lee, B.-G. Nam, S.-J. Song, N. Cho, et H.-J. Yoo, « A Power Management Unit with Continuous Co-Locking of Clock Frequency and Supply Voltage for Dynamic Voltage and Frequency Scaling », in *2007 IEEE International Symposium on Circuits and Systems*, mai 2007, p. 2112-2115, doi: 10.1109/ISCAS.2007.378516.
- [36] S. Kim, J. Kim, et S.-Y. Hwang, « New path balancing algorithm for glitch power reduction », *IEE Proc. - Circuits Devices Syst.*, vol. 148, n° 3, p. 151-156, juin 2001, doi: 10.1049/ip-cds:20010343.
-

-
- [37] S. N. Pradhan et P. Choudhury, « Low power and high testable Finite State Machine synthesis », in *2015 International Conference and Workshop on Computing and Communication (IEMCON)*, oct. 2015, p. 1-5, doi: 10.1109/IEMCON.2015.7344528.
- [38] S. Panth, K. Samadi, Y. Du, et S. K. Lim, « Tier-partitioning for power delivery vs cooling tradeoff in 3D VLSI for mobile applications », in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, juin 2015, p. 1-6, doi: 10.1145/2744769.2744917.
- [39] H. Sarhan, S. Thuries, O. Billoint, et F. Clermidy, « An Unbalanced Area Ratio Study for High Performance Monolithic 3D Integrated Circuits », in *2015 IEEE Computer Society Annual Symposium on VLSI*, juill. 2015, p. 350-355, doi: 10.1109/ISVLSI.2015.102.
- [40] S. Sawicki, G. Wilke, M. Johann, et R. Reis, « 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)*, déc. 2009, p. 852-855, doi: 10.1109/ICECS.2009.5410761.
- [41] G. Berhault, M. Brocard, S. Thuries, F. Galea, et L. Zaourar, « 3DIP: An iterative partitioning tool for monolithic 3D IC », in *2016 IEEE International 3D Systems Integration Conference (3DIC)*, nov. 2016, p. 1-5, doi: 10.1109/3DIC.2016.7970013.
- [42] N. K. Macha et M. Rahman, « Cost projections and benefits for transistor-level 3-D integration with stacked nanowires », in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, oct. 2017, p. 1-3, doi: 10.1109/S3S.2017.8309235.
- [43] X. Dong, J. Zhao, et Y. Xie, « Fabrication Cost Analysis and Cost-Aware Design Space Exploration for 3-D ICs », *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 29, n° 12, p. 1959-1972, déc. 2010, doi: 10.1109/TCAD.2010.2062811.
- [44] F. Andrieu *et al.*, « A review on opportunities brought by 3D-monolithic integration for CMOS device and digital circuit », in *2018 International Conference on IC Design Technology (ICICDT)*, juin 2018, p. 141-144, doi: 10.1109/ICICDT.2018.8399776.
- [45] D. Gitlin, M. Vinet, et F. Clermidy, « Cost model for monolithic 3D integrated circuits », in *2016 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, oct. 2016, p. 1-2, doi: 10.1109/S3S.2016.7804408.
- [46] A. Mallik *et al.*, « The impact of sequential-3D integration on semiconductor scaling roadmap », in *2017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 32.1.1-31.1.4, doi: 10.1109/IEDM.2017.8268483.
- [47] P. Coudrain *et al.*, « Experimental Insights Into Thermal Dissipation in TSV-Based 3-D Integrated Circuits », *IEEE Des. Test*, vol. 33, n° 3, p. 21-36, juin 2016, doi: 10.1109/MDAT.2015.2506678.
- [48] C. Jacoboni, C. Canali, G. Ottaviani, et A. Alberigi Quaranta, « A review of some charge transport properties of silicon », *Solid-State Electron.*, vol. 20, n° 2, p. 77-89, févr. 1977, doi: 10.1016/0038-1101(77)90054-5.
- [49] A. Bar-Cohen et P. Wang, « Thermal Management of On-Chip Hot Spot », *J. Heat Transf.*, vol. 134, n° 5, mai 2012, doi: 10.1115/1.4005708.
- [50] D. Choudhury, « 3D integration technologies for emerging microsystems », in *2010 IEEE MTT-S International Microwave Symposium*, mai 2010, p. 1-4, doi: 10.1109/MWSYM.2010.5514747.
- [51] C. Santos *et al.*, « Thermal performance of CoolCube™ monolithic and TSV-based 3D integration processes », in *2016 IEEE International 3D Systems Integration Conference (3DIC)*, 2016, p. 1-5.
- [52] A. Garnier, A. Jouve, R. Franiatte, et S. Cheramy, « Underfilling techniques comparison in 3D CtW stacking approach », in *2014 IEEE 64th Electronic Components and Technology Conference (ECTC)*, mai 2014, p. 906-912, doi: 10.1109/ECTC.2014.6897395.
- [53] S. K. Samal, S. Panth, K. Samadi, M. Saedi, Y. Du, et S. K. Lim, « Fast and accurate thermal modeling and optimization for monolithic 3D ICs », in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, juin 2014, p. 1-6, doi: 10.1145/2593069.2593140.
- [54] M. Brocard *et al.*, « Transistor temperature deviation analysis in monolithic 3D standard cells », in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2017, p. 539-544.
- [55] W.-L. Hung, G. M. Link, Yuan Xie, N. Vijaykrishnan, et M. J. Irwin, « Interconnect and thermal-aware floorplanning for 3D microprocessors », in *7th International Symposium on Quality Electronic Design (ISQED'06)*, mars 2006, p. 6 pp. - 104, doi: 10.1109/ISQED.2006.77.
-

-
- [56] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, et D. Tarjan, « Temperature-aware microarchitecture », *ACM SIGARCH Comput. Archit. News*, vol. 31, n° 2, p. 2-13, mai 2003, doi: 10.1145/871656.859620.
- [57] J. Cong, Jie Wei, et Yan Zhang, « A thermal-driven floorplanning algorithm for 3D ICs », in *IEEE/ACM International Conference on Computer Aided Design, 2004. ICCAD-2004.*, nov. 2004, p. 306-313, doi: 10.1109/ICCAD.2004.1382591.
- [58] P. Falkenstern, Yuan Xie, Yao-Wen Chang, et Yu Wang, « Three-dimensional integrated circuits (3D IC) Floorplan and Power/Ground Network Co-synthesis », in *2010 15th Asia and South Pacific Design Automation Conference (ASP-DAC)*, janv. 2010, p. 169-174, doi: 10.1109/ASPDAC.2010.5419899.
- [59] H. Wei, T. F. Wu, D. Sekar, B. Cronquist, R. F. Pease, et S. Mitra, « Cooling three-dimensional integrated circuits using power delivery networks », in *2012 International Electron Devices Meeting*, déc. 2012, p. 14.2.1-14.2.4, doi: 10.1109/IEDM.2012.6479040.
- [60] S. K. Samal, K. Samadi, P. Kamal, Y. Du, et S. K. Lim, « Full chip impact study of power delivery network designs in monolithic 3D ICs », in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, San Jose, CA, USA, nov. 2014, p. 565-572, doi: 10.1109/ICCAD.2014.7001406.
- [61] P. D. Franzon, W. Rhett Davis, et T. Thorolffson, « Creating 3D specific systems: Architecture, design and CAD », in *2010 Design, Automation Test in Europe Conference Exhibition (DATE 2010)*, mars 2010, p. 1684-1688, doi: 10.1109/DATE.2010.5457086.
- [62] L. Brunet *et al.*, « First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers », in *2016 IEEE Symposium on VLSI Technology*, juin 2016, p. 1-2, doi: 10.1109/VLSIT.2016.7573428.
- [63] P. Batude *et al.*, « 3D sequential integration opportunities and technology optimization », in *IEEE International Interconnect Technology Conference*, mai 2014, p. 373-376, doi: 10.1109/IITC.2014.6831837.
- [64] F. Andrieu *et al.*, « Design technology co-optimization of 3D-monolithic standard cells and SRAM exploiting dynamic back-bias for ultra-low-voltage operation », in *2017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 20.3.1-20.3.4, doi: 10.1109/IEDM.2017.8268428.
- [65] A. Vandooren *et al.*, « 3D technologies for analog/RF applications », in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, oct. 2017, p. 1-3, doi: 10.1109/S3S.2017.8308746.
- [66] P. Vivet *et al.*, « Advanced 3D Technologies and Architectures for 3D Smart Image Sensors », in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, mars 2019, p. 674-679, doi: 10.23919/DATE.2019.8714886.
- [67] O. Turkyilmaz, G. Cibrario, O. Rozeau, P. Batude, et F. Clermidy, « 3D FPGA using high-density interconnect Monolithic Integration », in *2014 Design, Automation Test in Europe Conference Exhibition (DATE)*, mars 2014, p. 1-4, doi: 10.7873/DATE.2014.351.
- [68] O. Thomas, M. Vinet, O. Rozeau, P. Batude, et A. Valentian, « Compact 6T SRAM cell with robust read/write stabilizing design in 45nm Monolithic 3D IC technology », in *2009 IEEE International Conference on IC Design and Technology*, mai 2009, p. 195-198, doi: 10.1109/ICICDT.2009.5166294.
- [69] R. Boumchedda *et al.*, « Energy-Efficient 4T SRAM Bitcell with 2T Read-Port for Ultra-Low-Voltage Operations in 28 nm 3D Monolithic CoolCube™ Technology », in *2018 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, juill. 2018, p. 1-7.
- [70] C. Liu et S. K. Lim, « Ultra-high density 3D SRAM cell designs for monolithic 3D integration », in *2012 IEEE International Interconnect Technology Conference*, juin 2012, p. 1-3, doi: 10.1109/IITC.2012.6251581.
- [71] S. Srinivasa *et al.*, « ROBIN: Monolithic-3D SRAM for Enhanced Robustness with In-Memory Computation Support », *IEEE Trans. Circuits Syst. Regul. Pap.*, vol. 66, n° 7, p. 2533-2545, juill. 2019, doi: 10.1109/TCSI.2019.2897497.
- [72] S. Thuries *et al.*, « M3D-ADTCO: Monolithic 3D Architecture, Design and Technology Co-Optimization for High Energy Efficient 3D IC », in *2020 Design, Automation Test in Europe*
-

-
- Conference Exhibition (DATE)*, mars 2020, p. 1740-1745, doi: 10.23919/DATE48585.2020.9116293.
- [73] A. Ayres *et al.*, « Guidelines on 3DVLSI design regarding the intermediate BEOL process influence », in *2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Rohnert Park, CA, USA, oct. 2015, p. 1-2, doi: 10.1109/S3S.2015.7333540.
- [74] O. Billoint *et al.*, « Merging PDKs to Build a Design Environment for 3D Circuits: Methodology, Challenges and Limitations », in *2019 International 3D Systems Integration Conference (3DIC)*, oct. 2019, p. 1-5, doi: 10.1109/3DIC48104.2019.9058793.
- [75] T. Poiroux *et al.*, « Leti-UTSOI2.1: A Compact Model for UTBB-FDSOI Technologies—Part I: Interface Potentials Analytical Model », *IEEE Trans. Electron Devices*, vol. 62, n° 9, p. 2751-2759, sept. 2015, doi: 10.1109/TED.2015.2458339.
- [76] T. Poiroux *et al.*, « Leti-UTSOI2.1: A Compact Model for UTBB-FDSOI Technologies—Part II: DC and AC Model Description », *IEEE Trans. Electron Devices*, vol. 62, n° 9, p. 2760-2768, sept. 2015, doi: 10.1109/TED.2015.2458336.
- [77] O. Weber *et al.*, « 14nm FDSOI upgraded device performance for ultra-low voltage operation », in *2015 Symposium on VLSI Technology (VLSI Technology)*, juin 2015, p. T168-T169, doi: 10.1109/VLSIT.2015.7223664.
- [78] D. Bosch *et al.*, « Novel fine-grain back-bias assist techniques for 3D-monolithic 14 nm FDSOI top-tier SRAMs », *Solid-State Electron.*, vol. 168, p. 107720, juin 2020, doi: 10.1016/j.sse.2019.107720.
- [79] P. Batude *et al.*, « 3DVLSI with CoolCube process: An alternative path to scaling », in *2015 Symposium on VLSI Technology (VLSI Technology)*, juin 2015, p. T48-T49, doi: 10.1109/VLSIT.2015.7223698.
- [80] A. A. de Sousa, « 3D Monolithic Integration : performance, Power and Area Evaluation for 14nm and beyond », phdthesis, Université Grenoble Alpes, 2017.
- [81] E. Seevinck, F. J. List, et J. Lohstroh, « Static-noise margin analysis of MOS SRAM cells », *IEEE J. Solid-State Circuits*, vol. 22, n° 5, p. 748-754, oct. 1987, doi: 10.1109/JSSC.1987.1052809.
- [82] Z. Guo, A. Carlson, L.-T. Pang, K. T. Duong, T.-J. K. Liu, et B. Nikolic, « Large-Scale SRAM Variability Characterization in 45 nm CMOS », *IEEE J. Solid-State Circuits*, vol. 44, n° 11, p. 3174-3192, nov. 2009, doi: 10.1109/JSSC.2009.2032698.
- [83] A. Gupta *et al.*, « High-Aspect-Ratio Ruthenium Lines for Buried Power Rail », in *2018 IEEE International Interconnect Technology Conference (IITC)*, juin 2018, p. 4-6, doi: 10.1109/IITC.2018.8430415.
- [84] P. Debacker *et al.*, « DTCO exploration for efficient standard cell power rails », mars 2018, p. 10, doi: 10.1117/12.2293500.
- [85] D. Prasad *et al.*, « Buried Power Rails and Back-side Power Grids: Arm® CPU Power Delivery Network Design Beyond 5nm », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 19.1.1-19.1.4, doi: 10.1109/IEDM19573.2019.8993617.
- [86] S. M. Salahuddin *et al.*, « SRAM With Buried Power Distribution to Improve Write Margin and Performance in Advanced Technology Nodes », *IEEE Electron Device Lett.*, vol. 40, n° 8, p. 1261-1264, août 2019, doi: 10.1109/LED.2019.2921209.
- [87] L. Zhu, Y. Badr, S. Wang, S. Iyer, et P. Gupta, « Assessing Benefits of a Buried Interconnect Layer in Digital Designs », *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 36, n° 2, p. 346-350, févr. 2017, doi: 10.1109/TCAD.2016.2572144.
- [88] C. Auth *et al.*, « A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects », in *2017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 29.1.1-29.1.4, doi: 10.1109/IEDM.2017.8268472.
- [89] D. H. Kim *et al.*, « Borderless contact leakage induced standby current failure on sub-0.15 μm CMOS device », in *Proceedings of the 2001 8th International Symposium on the Physical and Failure Analysis of Integrated Circuits. IPFA 2001 (Cat. No.01TH8548)*, Singapore, 2001, p. 165-168, doi: 10.1109/IPFA.2001.941478.
-

-
- [90]Minghao Lin, Heming Sun, et S. Kimura, « Power-efficient and slew-aware three dimensional gated clock tree synthesis », in *2016 IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC)*, sept. 2016, p. 1-6, doi: 10.1109/VLSI-SoC.2016.7753535.
- [91]K. Cho, C. Jang, J. Song, S. Kim, et J. Chong, « Thermal aware clock tree optimization with balanced clock skew in 3D ICs », in *The 18th IEEE International Symposium on Consumer Electronics (ISCE 2014)*, juin 2014, p. 1-2, doi: 10.1109/ISCE.2014.6884330.
- [92]J. Minz, Xin Zhao, et Sung Kyu Lim, « Buffered clock tree synthesis for 3D ICs under thermal variations », in *2008 Asia and South Pacific Design Automation Conference*, mars 2008, p. 504-509, doi: 10.1109/ASPDAC.2008.4484003.
- [93]M. M. Navidi et G.-S. Byun, « Comparative analysis of clock distribution networks for TSV-based 3D IC designs », in *Fifteenth International Symposium on Quality Electronic Design*, mars 2014, p. 184-188, doi: 10.1109/ISQED.2014.6783323.
- [94]K. Domanski, « Latch-up in FinFET technologies », in *2018 IEEE International Reliability Physics Symposium (IRPS)*, mars 2018, p. 2C.4-1-2C.4-5, doi: 10.1109/IRPS.2018.8353550.
- [95]O. Weber *et al.*, « 14nm FDSOI technology for high speed and energy efficient applications », in *2014 Symposium on VLSI Technology (VLSI-Technology): Digest of Technical Papers*, 2014, p. 1-2.
- [96]E. Seevinck, F. J. List, et J. Lohstroh, « Static-noise margin analysis of MOS SRAM cells », *IEEE J. Solid-State Circuits*, vol. 22, n° 5, p. 748-754, oct. 1987, doi: 10.1109/JSSC.1987.1052809.
- [97]B. Zimmer *et al.*, « SRAM Assist Techniques for Operation in a Wide Voltage Range in 28-nm CMOS », *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 59, n° 12, p. 853-857, déc. 2012, doi: 10.1109/TCSII.2012.2231015.
- [98]O. Thomas *et al.*, « Dynamic single-p-well SRAM bitcell characterization with back-bias adjustment for optimized wide-voltage-range SRAM operation in 28nm UTBB FD-SOI », in *2014 IEEE International Electron Devices Meeting*, San Francisco, CA, USA, déc. 2014, p. 3.4.1-3.4.4, doi: 10.1109/IEDM.2014.7046973.
- [99]D. Bosch *et al.*, « Back-bias impact on variability and BTI for 3D-monolithic 14nm FDSOI SRAMs applications », in *2019 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, avr. 2019, p. 1-4, doi: 10.1109/EUROSOI-ULIS45800.2019.9041890.
- [100] T.M. Mak, « Is CMOS more reliable with scaling? », 2002, doi: 10.13140/RG.2.1.4684.2003.
- [101] A. Tsiara, « Electrical characterization & modeling of the trapping phenomena impacting the reliability of nanowire transistors for sub 10nm nodes », thesis, Grenoble Alpes, 2019.
- [102] F. Chen et M. Shinosky, « Addressing Cu/Low- ϵ Dielectric TDDB-Reliability Challenges for Advanced CMOS Technologies », *IEEE Trans. Electron Devices*, vol. 56, n° 1, p. 2-12, janv. 2009, doi: 10.1109/TED.2008.2008680.
- [103] A. Bansal, R. Rao, J.-J. Kim, S. Zafar, J. H. Stathis, et C.-T. Chuang, « Impacts of NBTI and PBTI on SRAM static/dynamic noise margins and cell failure probability », *Microelectron. Reliab.*, vol. 49, n° 6, p. 642-649, juin 2009, doi: 10.1016/j.microrel.2009.03.016.
- [104] H. Reisinger, O. Blank, W. Heinrigs, A. Muhlhoff, W. Gustin, et C. Schlunder, « Analysis of NBTI Degradation- and Recovery-Behavior Based on Ultra Fast VT-Measurements », in *2006 IEEE International Reliability Physics Symposium Proceedings*, San Jose, CA, USA, 2006, p. 448-453, doi: 10.1109/RELPHY.2006.251260.
- [105] W.-G. Ho, Z. Zheng, K.-S. Chong, et B.-H. Gwee, « A Comparative Analysis of 65nm CMOS SRAM and Commercial SRAMs in Security Vulnerability Evaluation », in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, nov. 2018, p. 1-5, doi: 10.1109/ICDSP.2018.8631874.
- [106] D. Bosch *et al.*, « Back-bias impact on variability and BTI for 3D-monolithic 14nm FDSOI SRAMs applications », in *2019 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, avr. 2019, p. 1-4, doi: 10.1109/EUROSOI-ULIS45800.2019.9041890.
- [107] J. El Hussein *et al.*, « A Complete Characterization and Modeling of the BTI-Induced Dynamic Variability of SRAM Arrays in 28-nm FD-SOI Technology », *IEEE Trans. Electron Devices*, vol. 61, n° 12, p. 3991-3999, déc. 2014, doi: 10.1109/TED.2014.2361954.
-

-
- [108] A. Karel, M. Comte, J.-M. Galliere, F. Azais, et M. Renovell, « Impact of VT and Body-Biasing on Resistive Short Detection in 28nm UTBB FDSOI -- LVT and RVT Configurations », in *2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, Pittsburgh, PA, USA, juill. 2016, p. 164-169, doi: 10.1109/ISVLSI.2016.102.
- [109] D. Bosch *et al.*, « Novel Fine-Grain Back-Bias Assist Techniques for 14nm FDSOI Top-Tier SRAMs integrated in 3D-Monolithic », in *2019 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, avr. 2019, p. 1-2, doi: 10.1109/VLSI-TSA.2019.8804649.
- [110] J. Wang, S. Nalam, et B. H. Calhoun, « Analyzing static and dynamic write margin for nanometer SRAMs », in *Proceeding of the 13th international symposium on Low power electronics and design (ISLPED '08)*, août 2008, p. 129-134, doi: 10.1145/1393921.1393954.
- [111] K. Lofstrom, W. R. Daasch, et D. Taylor, « IC identification circuit using device mismatch », in *2000 IEEE International Solid-State Circuits Conference. Digest of Technical Papers (Cat. No.00CH37056)*, févr. 2000, p. 372-373, doi: 10.1109/ISSCC.2000.839821.
- [112] Y. Su, J. Holleman, et B. Otis, « A 1.6pJ/bit 96% Stable Chip-ID Generating Circuit using Process Variations », in *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, févr. 2007, p. 406-611, doi: 10.1109/ISSCC.2007.373466.
- [113] D. E. Holcomb, W. P. Burleson, et K. Fu, « Power-Up SRAM State as an Identifying Fingerprint and Source of True Random Numbers », *IEEE Trans. Comput.*, vol. 58, n° 9, p. 1198-1210, sept. 2009, doi: 10.1109/TC.2008.212.
- [114] G. Selimis *et al.*, « Evaluation of 90nm 6T-SRAM as Physical Unclonable Function for secure key generation in wireless sensor nodes », in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, mai 2011, p. 567-570, doi: 10.1109/ISCAS.2011.5937628.
- [115] Y. Dodis, R. Ostrovsky, L. Reyzin, et A. Smith, « Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data », *SIAM J. Comput.*, vol. 38, n° 1, p. 97-139, janv. 2008, doi: 10.1137/060651380.
- [116] S. O'uchi *et al.*, « Robust and compact key generator using physically unclonable function based on logic-transistor-compatible poly-crystalline-Si channel FinFET technology », in *2015 IEEE International Electron Devices Meeting (IEDM)*, déc. 2015, p. 25.6.1-25.6.4, doi: 10.1109/IEDM.2015.7409767.
- [117] A. Garg et T. T. Kim, « Design of SRAM PUF with improved uniformity and reliability utilizing device aging effect », in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, juin 2014, p. 1941-1944, doi: 10.1109/ISCAS.2014.6865541.
- [118] M. Pierre, R. Wacquez, X. Jehl, M. Sanquer, M. Vinet, et O. Cueto, « Single-donor ionization energies in a nanoscale CMOS channel », *Nat. Nanotechnol.*, vol. 5, n° 2, Art. n° 2, févr. 2010, doi: 10.1038/nnano.2009.373.
- [119] Y. Ono, K. Nishiguchi, A. Fujiwara, H. Yamaguchi, H. Inokawa, et Y. Takahashi, « Conductance modulation by individual acceptors in Si nanoscale field-effect transistors », *Appl. Phys. Lett.*, vol. 90, n° 10, p. 102106, mars 2007, doi: 10.1063/1.2679254.
- [120] E. Hamid, D. Moraru, T. Mizuno, et M. Tabe, « Single-electron transport through a single donor at elevated temperatures », in *2012 IEEE Silicon Nanoelectronics Workshop (SNW)*, juin 2012, p. 1-2, doi: 10.1109/SNW.2012.6243293.
- [121] W. Kohn et J. M. Luttinger, « Theory of Donor States in Silicon », *Phys. Rev.*, vol. 98, n° 4, p. 915-922, mai 1955, doi: 10.1103/PhysRev.98.915.
- [122] M. Pierre *et al.*, « Dielectric confinement and fluctuations of the local density of state in the source and drain of an ultra scaled SOI NMOS transistor », in *2010 Silicon Nanoelectronics Workshop*, juin 2010, p. 1-2, doi: 10.1109/SNW.2010.5562598.
- [123] E. I. Vatajelu, G. Di Natale, et P. Prinetto, « Towards a highly reliable SRAM-based PUFs », in *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*, mars 2016, p. 273-276.
- [124] A. T. Elshafiey, P. Zarkesh-Ha, et J. Trujillo, « The effect of power supply ramp time on SRAM PUFs », in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, août 2017, p. 946-949, doi: 10.1109/MWSCAS.2017.8053081.
- [125] W. Wang, A. Singh, U. Guin, et A. Chatterjee, « Exploiting power supply ramp rate for calibrating cell strength in SRAM PUFs », in *2018 IEEE 19th Latin-American Test Symposium (LATS)*, mars 2018, p. 1-6, doi: 10.1109/LATW.2018.8349685.
-

-
- [126] F. Andrieu *et al.*, « A review on opportunities brought by 3D-monolithic integration for CMOS device and digital circuit », in *2018 International Conference on IC Design Technology (ICICDT)*, juin 2018, p. 141-144, doi: 10.1109/ICICDT.2018.8399776.
- [127] P. Batude *et al.*, « 3D monolithic integration », in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, mai 2011, p. 2233-2236, doi: 10.1109/ISCAS.2011.5938045.
- [128] C. Fenouillet-Beranger *et al.*, « Recent advances in 3D VLSI integration », in *2016 International Conference on IC Design and Technology (ICICDT)*, juin 2016, p. 1-4, doi: 10.1109/ICICDT.2016.7542069.
- [129] C. Fenouillet-Beranger *et al.*, « FDSOI bottom MOSFETs stability versus top transistor thermal budget featuring 3D monolithic integration », in *2014 44th European Solid State Device Research Conference (ESSDERC)*, sept. 2014, p. 110-113, doi: 10.1109/ESSDERC.2014.6948770.
- [130] P. Rodriguez *et al.*, « Contacts for monolithic 3D architecture: Study of Ni_{0.9}Co_{0.1} silicide formation », in *2016 IEEE International Interconnect Technology Conference / Advanced Metallization Conference (IITC/AMC)*, mai 2016, p. 72-74, doi: 10.1109/IITC-AMC.2016.7507685.
- [131] C. Fenouillet-Beranger *et al.*, « New insights on bottom layer thermal stability and laser annealing promises for high performance 3D VLSI », déc. 2014, p. 27.5.1-27.5.4, doi: 10.1109/IEDM.2014.7047121.
- [132] L. Brunet *et al.*, « First demonstration of a CMOS over CMOS 3D VLSI CoolCube™ integration on 300mm wafers », in *2016 IEEE Symposium on VLSI Technology*, juin 2016, p. 1-2, doi: 10.1109/VLSIT.2016.7573428.
- [133] T. Naito *et al.*, « World's first monolithic 3D-FPGA with TFT SRAM over 90nm 9 layer Cu CMOS », in *2010 Symposium on VLSI Technology*, juin 2010, p. 219-220, doi: 10.1109/VLSIT.2010.5556234.
- [134] C. Yang *et al.*, « Location-controlled-grain Technique for Monolithic 3D BEOL FinFET Circuits », in *2018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 11.3.1-11.3.4, doi: 10.1109/IEDM.2018.8614708.
- [135] Y. Liu *et al.*, « Variability Analysis of Scaled Crystal Channel and Poly-Si Channel FinFETs », *IEEE Trans. Electron Devices*, vol. 59, n° 3, p. 573-581, mars 2012, doi: 10.1109/TED.2011.2178850.
- [136] R. Ishihara, M. R. T. Mofrad, J. Derakhshandeh, N. Golshani, et C. I. M. Beenakker, « Monolithic 3D-ICs with single grain Si thin film transistors », in *2012 IEEE 11th International Conference on Solid-State and Integrated Circuit Technology*, oct. 2012, p. 1-4, doi: 10.1109/ICSICT.2012.6467714.
- [137] Batude, « Sequential Integration », IEDM, juill. 12, 2019.
- [138] S.-Jung *et al.*, « High Speed and Highly Cost effective 72M bit density S3 SRAM Technology with Doubly Stacked Si Layers, Peripheral only CoSix layers and Tungsten Shunt W/L Scheme for Standalone and Embedded Memory », in *2007 IEEE Symposium on VLSI Technology*, juin 2007, p. 82-83, doi: 10.1109/VLSIT.2007.4339736.
- [139] Y.-H. Son *et al.*, « Laser-induced Epitaxial Growth (LEG) Technology for High Density 3-D Stacked Memory with High Productivity », juin 2007, p. 80-81, doi: 10.1109/VLSIT.2007.4339735.
- [140] P. Batude *et al.*, « 3DVLSI with CoolCube process: An alternative path to scaling », in *2015 Symposium on VLSI Technology (VLSI Technology)*, juin 2015, p. T48-T49, doi: 10.1109/VLSIT.2015.7223698.
- [141] I. Radu, B.-Y. Nguyen, G. Gaudin, et C. Mazure, « 3D monolithic integration: Stacking technology and applications », in *2015 International Conference on IC Design Technology (ICICDT)*, juin 2015, p. 1-3, doi: 10.1109/ICICDT.2015.7165915.
- [142] L. Brunet *et al.*, « Breakthroughs in 3D Sequential technology », in *2018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 7.2.1-7.2.4, doi: 10.1109/IEDM.2018.8614653.
- [143] P. Batude *et al.*, « GeOI and SOI 3D monolithic cell integrations for high density applications », in *2009 Symposium on VLSI Technology*, juin 2009, p. 166-167.
- [144] W. Rachmady *et al.*, « 300mm Heterogeneous 3D Integration of Record Performance Layer Transfer Germanium PMOS with Silicon NMOS for Low Power High Performance Logic
-

- Applications », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 29.7.1-29.7.4, doi: 10.1109/IEDM19573.2019.8993626.
- [145] H. W. Then *et al.*, « 3D heterogeneous integration of high performance high-K metal gate GaN NMOS and Si PMOS transistors on 300mm high-resistivity Si substrate for energy-efficient and compact power delivery, RF (5G and beyond) and SoC applications », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 17.3.1-17.3.4, doi: 10.1109/IEDM19573.2019.8993583.
- [146] C. Cheng *et al.*, « Monolithic Heterogeneous Integration of BEOL Power Gating Transistors of Carbon Nanotube Networks with FEOL Si Ring Oscillator Circuits », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 19.2.1-19.2.4, doi: 10.1109/IEDM19573.2019.8993593.
- [147] P. S. Kanhaiya, Y. Stein, W. Lu, J. A. del Alamo, et M. M. Shulaker, « X3D: Heterogeneous Monolithic 3D Integration of “X” (Arbitrary) Nanowires: Silicon, III–V, and Carbon Nanotubes », *IEEE Trans. Nanotechnol.*, vol. 18, p. 270-273, 2019, doi: 10.1109/TNANO.2019.2902114.
- [148] D. BOSCH *et al.*, « Laser Processing For 3D Junctionless Transistor Fabrication », in *2019 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, oct. 2019, p. 1-2.
- [149] J. Micout *et al.*, « High performance low temperature FinFET with DSPER, gate last and Self Aligned Contact for 3D sequential integration », in *2017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 32.2.1-32.2.4, doi: 10.1109/IEDM.2017.8268484.
- [150] A. Vandooren *et al.*, « First Demonstration of 3D stacked Finfets at a 45nm fin pitch and 110nm gate pitch technology on 300mm wafers », in *2018 IEEE International Electron Devices Meeting (IEDM)*, déc. 2018, p. 7.1.1-7.1.4, doi: 10.1109/IEDM.2018.8614654.
- [151] J. P. Colinge *et al.*, « SOI gated resistor: CMOS without junctions », in *2009 IEEE International SOI Conference*, oct. 2009, p. 1-2, doi: 10.1109/SOI.2009.5318737.
- [152] A. Vandooren *et al.*, « 3-D Sequential Stacked Planar Devices Featuring Low-Temperature Replacement Metal Gate Junctionless Top Devices With Improved Reliability », *IEEE Trans. Electron Devices*, vol. 65, n° 11, p. 5165-5171, nov. 2018, doi: 10.1109/TED.2018.2871265.
- [153] A. Vandooren *et al.*, « Buried metal line compatible with 3D sequential integration for top tier planar devices dynamic V_{th} tuning and RF shielding applications », in *2019 Symposium on VLSI Technology*, juin 2019, p. T56-T57, doi: 10.23919/VLSIT.2019.8776490.
- [154] S. Barraud *et al.*, « Scaling of Trigate Junctionless Nanowire MOSFET With Gate Length Down to 13 nm », *IEEE Electron Device Lett.*, vol. 33, n° 9, p. 1225-1227, sept. 2012, doi: 10.1109/LED.2012.2203091.
- [155] B.-H. Lee *et al.*, « A Vertically Integrated Junctionless Nanowire Transistor », *Nano Lett.*, vol. 16, n° 3, p. 1840-1847, mars 2016, doi: 10.1021/acs.nanolett.5b04926.
- [156] « Impact of crystalline damage on a vertically integrated junctionless nanowire transistor | Request PDF », *ResearchGate*. https://www.researchgate.net/publication/309654844_Impact_of_crystalline_damage_on_a_vertically_integrated_junctionless_nanowire_transistor (consulté le mars 09, 2020).
- [157] « Junction-less stackable SONOS memory realized on vertical-Si-nanowire for 3-D application - IEEE Conference Publication ». <https://ieeexplore.ieee.org/document/5872271> (consulté le mars 09, 2020).
- [158] I.-H. Wong, Y.-T. Chen, S.-H. Huang, W.-H. Tu, Y.-S. Chen, et C. W. Liu, « Junctionless Gate-All-Around pFETs Using In-situ Boron-Doped Ge Channel on Si », *IEEE Trans. Nanotechnol.*, vol. 14, n° 5, p. 878-882, sept. 2015, doi: 10.1109/TNANO.2015.2456182.
- [159] H. Gamble *et al.*, « Germanium Processing », in *Semiconductor-On-Insulator Materials for Nanoelectronics Applications*, A. Nazarov, J.-P. Colinge, F. Balestra, J.-P. Raskin, F. Gamiz, et V. S. Lysenko, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, p. 3-29.
- [160] I-Hsieh Wong *et al.*, « In-situ doped and tensily strained ge junctionless gate-all-around nFETs on SOI featuring Ion = 828 $\mu\text{A}/\mu\text{m}$, Ion/Ioff $\sim 1 \times 10^5$, DIBL = 16–54 mV/V, and 1.4X external strain enhancement », in *2014 IEEE International Electron Devices Meeting*, déc. 2014, p. 9.6.1-9.6.4, doi: 10.1109/IEDM.2014.7047019.
-

-
- [161] S. Ren *et al.*, « Total Ionizing Dose (TID) Effects in Ultra-Thin Body Ge-on-Insulator (GOI) Junctionless CMOSFETs With Recessed Source/Drain and Channel », *IEEE Trans. Nucl. Sci.*, vol. 64, n° 1, p. 176-180, janv. 2017, doi: 10.1109/TNS.2016.2624294.
- [162] A. Kranti *et al.*, « Junctionless nanowire transistor (JNT): Properties and design guidelines », in *2010 Proceedings of the European Solid State Device Research Conference*, sept. 2010, p. 357-360, doi: 10.1109/ESSDERC.2010.5618216.
- [163] T. K. Kim *et al.*, « First Demonstration of Junctionless Accumulation-Mode Bulk FinFETs With Robust Junction Isolation », *IEEE Electron Device Lett.*, vol. 34, n° 12, p. 1479-1481, déc. 2013, doi: 10.1109/LED.2013.2283291.
- [164] Y. Cheng *et al.*, « Performance enhancement of a novel P-type junctionless transistor using a hybrid poly-Si fin channel », in *2014 IEEE International Electron Devices Meeting*, déc. 2014, p. 26.7.1-26.7.4, doi: 10.1109/IEDM.2014.7047116.
- [165] Y. Cheng *et al.*, « Characteristics of a Novel Poly-Si P-Channel Junctionless Thin-Film Transistor With Hybrid P/N-Substrate », *IEEE Electron Device Lett.*, vol. 36, n° 2, p. 159-161, févr. 2015, doi: 10.1109/LED.2014.2379673.
- [166] Y. Lin *et al.*, « Hybrid P-Channel/N-Substrate Poly-Si Nanosheet Junctionless Field-Effect Transistors With Trench and Gate-All-Around Structure », *IEEE Trans. Nanotechnol.*, vol. 17, n° 5, p. 1014-1019, sept. 2018, doi: 10.1109/TNANO.2018.2848283.
- [167] H.-H. Li, Y.-R. Lin, Y.-C. Wu, Y.-H. Lin, et J.-J. Yu, « Multi-stacking hybrid P/N/P nanosheet layers junctionless field-effect transistors », in *2018 7th International Symposium on Next Generation Electronics (ISNE)*, Taipei, mai 2018, p. 1-3, doi: 10.1109/ISNE.2018.8394662.
- [168] J.P. Colinge, « Grain Size and Resistivity of LPCVD Polycrystalline Silicon Films », *J. Electrochem. Soc.*, vol. 128, n° 9, p. 2009, 1981, doi: 10.1149/1.2127785.
- [169] J.- Colinge, H. Morel, et J.- Chante, « Field effect in large grain polycrystalline silicon », *IEEE Trans. Electron Devices*, vol. 30, n° 3, p. 197-201, mars 1983, doi: 10.1109/T-ED.1983.21099.
- [170] C. Su, T. Tsai, Y. Liou, Z. Lin, H. Lin, et T. Chao, « Gate-All-Around Junctionless Transistors With Heavily Doped Polysilicon Nanowire Channels », *IEEE Electron Device Lett.*, vol. 32, n° 4, p. 521-523, avr. 2011, doi: 10.1109/LED.2011.2107498.
- [171] T.-Y. Liu, F.-M. Pan, et J.-T. Sheu, « Characteristics of Gate-All-Around Junctionless Polysilicon Nanowire Transistors With Twin 20-nm Gates », *IEEE J. Electron Devices Soc.*, vol. 3, n° 5, p. 405-409, sept. 2015, doi: 10.1109/JEDS.2015.2441736.
- [172] Po-Yi Kuo, Yi-Hsien Lu, et Tien-Sheng Chao, « High-Performance GAA Sidewall-Damascened Sub-10-nm In Situ n⁺-Doped Poly-Si NWs Channels Junctionless FETs », *IEEE Trans. Electron Devices*, vol. 61, n° 11, p. 3821-3826, nov. 2014, doi: 10.1109/TED.2014.2354436.
- [173] D. Hsieh, J. Lin, P. Kuo, et T. Chao, « Comprehensive Analysis on Electrical Characteristics of Pi-Gate Poly-Si Junctionless FETs », *IEEE Trans. Electron Devices*, vol. 64, n° 7, p. 2992-2998, juill. 2017, doi: 10.1109/TED.2017.2704933.
- [174] D. Hsieh, J. Lin, P. Kuo, et T. Chao, « High-Performance Pi-Gate Poly-Si Junctionless and Inversion Mode FET », *IEEE Trans. Electron Devices*, vol. 63, n° 11, p. 4179-4184, nov. 2016, doi: 10.1109/TED.2016.2611021.
- [175] W. C.-Y. Ma, J.-Y. Wang, H.-C. Wang, Y.-J. Huang, et L.-W. Yu, « Dependence of Sub-Thermionic Swing on Channel Thickness and Drain Bias of Poly-Si Junctionless Thin-Film Transistor », *IEEE Electron Device Lett.*, vol. 39, n° 8, p. 1122-1125, août 2018, doi: 10.1109/LED.2018.2850974.
- [176] M.-S. Yeh, Y.-C. Wu, M.-H. Wu, M.-H. Chung, Y.-R. Jhan, et M.-F. Hung, « Characterizing the Electrical Properties of a Novel Junctionless Poly-Si Ultrathin-Body Field-Effect Transistor Using a Trench Structure », *IEEE Electron Device Lett.*, vol. 36, n° 2, p. 150-152, févr. 2015, doi: 10.1109/LED.2014.2378785.
- [177] H.-C. Lin, C.-I. Lin, et T.-Y. Huang, « Characteristics of n-Type Junctionless Poly-Si Thin-Film Transistors With an Ultrathin Channel », *IEEE Electron Device Lett.*, vol. 33, n° 1, p. 53-55, janv. 2012, doi: 10.1109/LED.2011.2171914.
- [178] Y. Kamimuta, K. Ikeda, K. Furuse, T. Irisawa, et T. Tezuka, « Short channel poly-Ge junctionless p-type FinFETs for BEOL transistors », in *2013 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, avr. 2013, p. 1-2, doi: 10.1109/VLSI-TSA.2013.6545620.
-

-
- [179] K. Usuda, Y. Kamata, Y. Kamimuta, T. Mori, M. Koike, et T. Tezuka, « High-performance tri-gate poly-Ge junction-less p- and n-MOSFETs fabricated by flash lamp annealing process », in *2014 IEEE International Electron Devices Meeting*, déc. 2014, p. 16.6.1-16.6.4, doi: 10.1109/IEDM.2014.7047066.
- [180] W.-H. Huang *et al.*, « Enabling n-type polycrystalline Ge junctionless FinFET of low thermal budget by in situ doping of channel and visible pulsed laser annealing », *Appl. Phys. Express*, vol. 10, n° 2, p. 026502, févr. 2017, doi: 10.7567/APEX.10.026502.
- [181] V. Djara *et al.*, « Junctionless InGaAs MOSFETs with InAlAs barrier isolation and channel thinning by digital wet etching », in *71st Device Research Conference*, juin 2013, p. 131-132, doi: 10.1109/DRC.2013.6633828.
- [182] V. Djara *et al.*, « Tri-gate In_{0.53}Ga_{0.47}As-on-insulator junctionless field effect transistors », in *EUROSOI-ULIS 2015: 2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon*, janv. 2015, p. 97-100, doi: 10.1109/ULIS.2015.7063782.
- [183] W. Cao, J. Kang, et K. Banerjee, « Junction-Less Monolayer MoS₂ FETs », *arXiv*, p. arXiv:1509.00561, sept. 2015.
- [184] M. A. Barik, M. K. Sarma, et J. C. Dutta, « Traditional graphene and junctionless carbon nanotube field effect transistor for cholesterol sensing », in *2014 IEEE 2nd International Conference on Emerging Electronics (ICEE)*, déc. 2014, p. 1-4, doi: 10.1109/ICEmElec.2014.7151205.
- [185] J. Jiang, Q. Wan, et Q. Zhang, « Transparent Junctionless Electric-Double-Layer Transistors Gated by a Reinforced Chitosan-Based Biopolymer Electrolyte », *IEEE Trans. Electron Devices*, vol. 60, n° 6, p. 1951-1957, juin 2013, doi: 10.1109/TED.2013.2258922.
- [186] G.-D. Wu, J. Zhang, et X. Wan, « Junctionless Coplanar-Gate Oxide-Based Thin-Film Transistors Gated by Al₂O₃ Proton Conducting Films on Paper Substrates », *Chin. Phys. Lett.*, vol. 31, n° 10, p. 108505, oct. 2014, doi: 10.1088/0256-307X/31/10/108505.
- [187] « TCAD ». <https://www.synopsys.com/silicon/tcad.html> (consulté le mars 09, 2020).
- [188] « Sentaurus Device ». <https://www.synopsys.com/silicon/tcad/device-simulation/sentaurus-device.html> (consulté le mars 09, 2020).
- [189] C.-W. Lee *et al.*, « Influence of gate misalignment on the electrical characteristics of MuGFETS », *Solid-State Electron.*, vol. 54, n° 3, p. 226-230, mars 2010, doi: 10.1016/j.sse.2009.09.001.
- [190] T. Morimoto *et al.*, « Self-aligned nickel-mono-silicide technology for high-speed deep submicrometer logic CMOS ULSI », *IEEE Trans. Electron Devices*, vol. 42, n° 5, p. 915-922, mai 1995, doi: 10.1109/16.381988.
- [191] D. J. Roulston, N. D. Arora, et S. G. Chamberlain, « Modeling and measurement of minority-carrier lifetime versus doping in diffused layers of n+-p silicon diodes », *IEEE Trans. Electron Devices*, vol. 29, n° 2, p. 284-291, févr. 1982, doi: 10.1109/T-ED.1982.20697.
- [192] J. W. Slotboom et H. C. de Graaff, « Bandgap narrowing in silicon bipolar transistors », *IEEE Trans. Electron Devices*, vol. 24, n° 8, p. 1123-1125, août 1977, doi: 10.1109/T-ED.1977.18889.
- [193] D. B. M. Klaassen, « A unified mobility model for device simulation—I. Model equations and concentration dependence », *Solid-State Electron.*, vol. 35, n° 7, p. 953-959, juill. 1992, doi: 10.1016/0038-1101(92)90325-7.
- [194] J. P. Colinge *et al.*, « Junctionless Transistors: Physics and Properties », in *Semiconductor-On-Insulator Materials for Nanoelectronics Applications*, A. Nazarov, J.-P. Colinge, F. Balestra, J.-P. Raskin, F. Gamiz, et V. S. Lysenko, Éd. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, p. 187-200.
- [195] R. Trevisoli, R. T. Doria, M. de Souza, S. Barraud, M. Vinet, et M. A. Pavanello, « Analytical Model for the Dynamic Behavior of Triple-Gate Junctionless Nanowire Transistors », *IEEE Trans. Electron Devices*, vol. 63, n° 2, p. 856-863, févr. 2016, doi: 10.1109/TED.2015.2507571.
- [196] R. Trevisoli, R. Doria, M. De Souza, et M. Pavanello, « Accounting for Series Resistance in the Compact Model of Triple-Gate Junctionless Nanowire Transistors », in *2018 33rd Symposium on Microelectronics Technology and Devices (SBMicro)*, Bento Gonçalves, août 2018, p. 1-4, doi: 10.1109/SBMicro.2018.8511376.
-

-
- [197] J.-M. Sallese, F. Jazaeri, L. Barbut, N. Chevillon, et C. Lallement, « A Common Core Model for Junctionless Nanowires and Symmetric Double-Gate FETs », *IEEE Trans. Electron Devices*, vol. 60, n° 12, p. 4277-4280, déc. 2013, doi: 10.1109/TED.2013.2287528.
- [198] D. Gola, B. Singh, J. Singh, S. Jit, et P. K. Tiwari, « Static and Quasi-Static Drain Current Modeling of Tri-Gate Junctionless Transistor With Substrate Bias-Induced Effects », *IEEE Trans. Electron Devices*, vol. 66, n° 7, p. 2876-2883, juill. 2019, doi: 10.1109/TED.2019.2915294.
- [199] T.-K. Chiang, « A Short-Channel-Effect-Degraded Noise Margin Model for Junctionless Double-Gate MOSFET Working on Subthreshold CMOS Logic Gates », *IEEE Trans. Electron Devices*, vol. 63, n° 8, p. 3354-3359, août 2016, doi: 10.1109/TED.2016.2581826.
- [200] Y. Xiao, B. Zhang, H. Lou, L. Zhang, et X. Lin, « A Compact Model of Subthreshold Current With Source/Drain Depletion Effect for the Short-Channel Junctionless Cylindrical Surrounding-Gate MOSFETs », *IEEE Trans. Electron Devices*, vol. 63, n° 5, p. 2176-2181, mai 2016, doi: 10.1109/TED.2016.2535247.
- [201] T.-K. Chiang, « A New Subthreshold Current Model for Junctionless Trigate MOSFETs to Examine Interface-Trapped Charge Effects », *IEEE Trans. Electron Devices*, vol. 62, n° 9, p. 2745-2750, sept. 2015, doi: 10.1109/TED.2015.2456040.
- [202] A. Yesayan, F. Jazaeri, et J.-M. Sallese, « Charge-Based Modeling of Double-Gate and Nanowire Junctionless FETs Including Interface-Trapped Charges », *IEEE Trans. Electron Devices*, vol. 63, n° 3, p. 1368-1374, mars 2016, doi: 10.1109/TED.2016.2521359.
- [203] F. Jazaeri et J.-M. Sallese, « Modeling Channel Thermal Noise and Induced Gate Noise in Junctionless FETs », *IEEE Trans. Electron Devices*, vol. 62, n° 8, p. 2593-2597, août 2015, doi: 10.1109/TED.2015.2437954.
- [204] D. Jeon, S. J. Park, M. Mouis, S. Barraud, G. Kim, et G. Ghibaudo, « A Simple Method for Estimation of Silicon Film Thickness in Tri-Gate Junctionless Transistors », *IEEE Electron Device Lett.*, vol. 39, n° 9, p. 1282-1285, sept. 2018, doi: 10.1109/LED.2018.2857623.
- [205] J.-P. Colinge *et al.*, « A Simulation Comparison between Junctionless and Inversion-Mode MuGFETs », Montreal, QC, Canada, 2011, p. 63-72, doi: 10.1149/1.3570778.
- [206] R. Trevisoli, R. T. Doria, M. de Souza, et M. A. Pavanello, « Lateral spacers influence on the effective channel length of junctionless nanowire transistors », in *2017 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, oct. 2017, p. 1-3, doi: 10.1109/S3S.2017.8309260.
- [207] R. Trevisoli, R. T. Doria, M. de Souza, et M. A. Pavanello, « Effective channel length in Junctionless Nanowire Transistors », in *2015 30th Symposium on Microelectronics Technology and Devices (SBMicro)*, août 2015, p. 1-4, doi: 10.1109/SBMicro.2015.7298144.
- [208] C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, et J.-P. Colinge, « Junctionless multigate field-effect transistor », *Appl. Phys. Lett.*, vol. 94, n° 5, p. 053511, févr. 2009, doi: 10.1063/1.3079411.
- [209] S. Gundapaneni, S. Ganguly, et A. Kottantharayil, « Enhanced Electrostatic Integrity of Short-Channel Junctionless Transistor With High- κ Spacers », *IEEE Electron Device Lett.*, vol. 32, n° 10, p. 1325-1327, oct. 2011, doi: 10.1109/LED.2011.2162309.
- [210] G. Saini et S. Choudhary, « Improving the subthreshold performance of junctionless transistor using spacer engineering », *Microelectron. J.*, vol. 59, p. 55-58, janv. 2017, doi: 10.1016/j.mejo.2016.11.012.
- [211] J. Hur *et al.*, « Comprehensive Analysis of Gate-Induced Drain Leakage in Vertically Stacked Nanowire FETs: Inversion-Mode Versus Junctionless Mode », *IEEE Electron Device Lett.*, vol. 37, n° 5, p. 541-544, mai 2016, doi: 10.1109/LED.2016.2540645.
- [212] S. Takagi, A. Toriumi, M. Iwase, et H. Tango, « On the universality of inversion layer mobility in Si MOSFET's: Part I-effects of substrate impurity concentration », *IEEE Trans. Electron Devices*, vol. 41, n° 12, p. 2357-2362, déc. 1994, doi: 10.1109/16.337449.
- [213] C. Lee *et al.*, « High-Temperature Performance of Silicon Junctionless MOSFETs », *IEEE Trans. Electron Devices*, vol. 57, n° 3, p. 620-625, mars 2010, doi: 10.1109/TED.2009.2039093.
- [214] T. Rudenko *et al.*, « Mobility enhancement effect in heavily doped junctionless nanowire silicon-on-insulator metal-oxide-semiconductor field-effect transistors », *Appl. Phys. Lett.*, vol. 101, n° 21, p. 213502, nov. 2012, doi: 10.1063/1.4767353.
-

-
- [215] K.-I. Goto, T.-H. Yu, J. Wu, C. H. Diaz, et J. P. Colinge, « Mobility and screening effect in heavily doped accumulation-mode metal-oxide-semiconductor field-effect transistors », *Appl. Phys. Lett.*, vol. 101, n° 7, p. 073503, août 2012, doi: 10.1063/1.4745604.
- [216] R. T. Doria, R. Trevisoli, M. de Souza, et M. A. Pavanello, « Effective mobility analysis of n- and p-types SOI junctionless nanowire transistors », in *2014 29th Symposium on Microelectronics Technology and Devices (SBMicro)*, Aracaju, Brazil, sept. 2014, p. 1-4, doi: 10.1109/SBMicro.2014.6940108.
- [217] B. Van Zeghbroeck, « Principles of semiconductor devices, 2004 », *Colo. Univ. Ed.*, 2007.
- [218] J.-P. Colinge *et al.*, « Reduced electric field in junctionless transistors », *Appl. Phys. Lett.*, vol. 96, n° 7, p. 073510, févr. 2010, doi: 10.1063/1.3299014.
- [219] S. J. Park *et al.*, « Less mobility degradation induced by transverse electric-field in junctionless transistors », *Appl. Phys. Lett.*, vol. 105, n° 21, p. 213504, nov. 2014, doi: 10.1063/1.4902549.
- [220] D.-Y. Jeon *et al.*, « Separation of surface accumulation and bulk neutral channel in junctionless transistors », *Appl. Phys. Lett.*, vol. 104, n° 26, p. 263510, juin 2014, doi: 10.1063/1.4886139.
- [221] M. S. Parihar *et al.*, « Back-gate effects and detailed characterization of junctionless transistor », in *2015 45th European Solid State Device Research Conference (ESSDERC)*, Graz, Austria, sept. 2015, p. 282-285, doi: 10.1109/ESSDERC.2015.7324769.
- [222] G. Mariniello, R. T. Doria, M. de Souza, M. A. Pavanello, et R. D. Trevisoli, « Analysis of gate capacitance of n-type junctionless transistors using three-dimensional device simulations », in *2012 8th International Caribbean Conference on Devices, Circuits and Systems (ICCDACS)*, Playa del Carmen, Mexico, mars 2012, p. 1-4, doi: 10.1109/ICCDACS.2012.6188946.
- [223] A. Kranti et G. A. Armstrong, « Design and Optimization of FinFETs for Ultra-Low-Voltage Analog Applications », *IEEE Trans. Electron Devices*, vol. 54, n° 12, p. 3308-3316, déc. 2007, doi: 10.1109/TED.2007.908596.
- [224] S. R. Nassif, « Process variability at the 65nm node and beyond », in *2008 IEEE Custom Integrated Circuits Conference*, sept. 2008, p. 1-8, doi: 10.1109/CICC.2008.4672005.
- [225] M. J. M. Pelgrom, A. C. J. Duinmaijer, et A. P. G. Welbers, « Matching properties of MOS transistors », *IEEE J. Solid-State Circuits*, vol. 24, n° 5, p. 1433-1439, oct. 1989, doi: 10.1109/JSSC.1989.572629.
- [226] O. Weber *et al.*, « High immunity to threshold voltage variability in undoped ultra-thin FDSOI MOSFETs and its physical understanding », in *2008 IEEE International Electron Devices Meeting*, déc. 2008, p. 1-4, doi: 10.1109/IEDM.2008.4796663.
- [227] M. Aldegunde, A. Martinez, et J. R. Barker, « Study of Discrete Doping-Induced Variability in Junctionless Nanowire MOSFETs Using Dissipative Quantum Transport Simulations », *IEEE Electron Device Lett.*, vol. 33, n° 2, p. 194-196, févr. 2012, doi: 10.1109/LED.2011.2177634.
- [228] H. F. Dadgour, K. Endo, V. K. De, et K. Banerjee, « Grain-Oriented Induced Work Function Variation in Nanoscale Metal-Gate Transistors—Part I: Modeling, Analysis, and Experimental Validation », *IEEE Trans. Electron Devices*, vol. 57, n° 10, p. 2504-2514, oct. 2010, doi: 10.1109/TED.2010.2063191.
- [229] S. M. Nawaz et A. Mallik, « Effects of Device Scaling on the Performance of Junctionless FinFETs Due to Gate-Metal Work Function Variability and Random Dopant Fluctuations », *IEEE Electron Device Lett.*, vol. 37, n° 8, p. 958-961, août 2016, doi: 10.1109/LED.2016.2578349.
- [230] X. Wang, A. R. Brown, Binjie Cheng, et A. Asenov, « Statistical variability and reliability in nanoscale FinFETs », in *2011 International Electron Devices Meeting*, déc. 2011, p. 5.4.1-5.4.4, doi: 10.1109/IEDM.2011.6131494.
- [231] « FD-SOI - Soitec - Soitec ». <https://www.soitec.com/fr/produits/fd-soi> (consulté le juin 29, 2020).
- [232] K. J. Kuhn *et al.*, « Process Technology Variation », *IEEE Trans. Electron Devices*, vol. 58, n° 8, p. 2197-2208, août 2011, doi: 10.1109/TED.2011.2121913.
- [233] A. Asenov, S. Kaya, et J. H. Davies, « Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations », *IEEE Trans. Electron Devices*, vol. 49, n° 1, p. 112-119, janv. 2002, doi: 10.1109/16.974757.
-

-
- [234] A. Vandooren *et al.*, « 3-D Sequential Stacked Planar Devices Featuring Low-Temperature Replacement Metal Gate Junctionless Top Devices With Improved Reliability », *IEEE Trans. Electron Devices*, vol. 65, n° 11, p. 5165-5171, nov. 2018, doi: 10.1109/TED.2018.2871265.
- [235] J.-T. Park, J. Kim, et J. Colinge, « Negative-bias-temperature-instability and hot carrier effects in nanowire junctionless p-channel multigate transistors », *Appl. Phys. Lett.*, vol. 100, févr. 2012, doi: 10.1063/1.3688245.
- [236] M. Cho *et al.*, « On and off state hot carrier reliability in junctionless high-K MG gate-all-around nanowires », déc. 2015, p. 14.5.1-14.5.4, doi: 10.1109/IEDM.2015.7409697.
- [237] M. Toledano-Luque *et al.*, « Superior Reliability of Junctionless pFinFETs by Reduced Oxide Electric Field », *Electron Device Lett. IEEE*, vol. 35, p. 1179-1181, déc. 2014, doi: 10.1109/LED.2014.2361769.
- [238] Theodorou, « Low frequency noise in advanced CMOS/SOI nanoscale multi-gate devices and noise models for applications in electronic circuits », ARISTOTLE UNIVERSITY OF THESSALONIKI, 2013.
- [239] T. Hida, H.-H. Kuo, J. Potthoff, et W. Streit, *White Noise: An Infinite Dimensional Calculus*. Springer Netherlands, 1993.
- [240] Y. Tsvividis et C. McAndrew, « Operation and modeling of the MOS transistor », *CERN Document Server*, 2011. <https://cds.cern.ch/record/1546736> (consulté le mars 17, 2020).
- [241] A. Van Der Ziel et E. R. Chenette, « Noise in Solid State Devices », in *Advances in Electronics and Electron Physics*, vol. 46, L. Marton, Éd. Academic Press, 1978, p. 313-383.
- [242] M. J. Kirton, M. J. Uren, S. Collins, M. Schulz, A. Karmann, et K. Scheffer, « Individual defects at the Si:SiO₂ interface », *Semicond. Sci. Technol.*, vol. 4, n° 12, p. 1116–1126, déc. 1989, doi: 10.1088/0268-1242/4/12/013.
- [243] R. Jayaraman et C. G. Sodini, « A 1/f noise technique to extract the oxide trap density near the conduction band edge of silicon », *IEEE Trans. Electron Devices*, vol. 36, n° 9, p. 1773-1782, sept. 1989, doi: 10.1109/16.34242.
- [244] N. Opondo, S. Ramadurgam, C. Yang, et S. Mohammadi, « Trap studies in silicon nanowire junctionless transistors using low-frequency noise », *J. Vac. Sci. Technol. B*, vol. 34, janv. 2016, doi: 10.1116/1.4939787.
- [245] D.-Y. Jeon, S. J. Park, M. Mouis, S. Barraud, G.-T. Kim, et G. Ghibaudo, « Low-frequency noise behavior of junctionless transistors compared to inversion-mode transistors », *Solid-State Electron.*, vol. 81, p. 101–104, mars 2013, doi: 10.1016/j.sse.2012.12.003.
- [246] D. Jang *et al.*, « Low-frequency noise in junctionless multigate transistors », *Appl. Phys. Lett.*, vol. 98, p. 133502, mars 2011, doi: 10.1063/1.3569724.
- [247] R. T. Doria, R. Trevisoli, M. de Souza, et M. A. Pavanello, « Low-frequency noise and effective trap density of short channel p- and n-types junctionless nanowire transistors », *Solid-State Electron.*, vol. 96, p. 22-26, juin 2014, doi: 10.1016/j.sse.2014.04.019.
- [248] R. T. Doria *et al.*, « Analysis of the substrate bias effect on the interface trapped charges in junctionless nanowire transistors through low-frequency noise characterization », *Microelectron. Eng.*, vol. 178, p. 17-20, juin 2017, doi: 10.1016/j.mee.2017.04.014.
- [249] T.-I. Tsai *et al.*, « Low-Operating-Voltage Ultrathin Junctionless Poly-Si Thin-Film Transistor Technology for RF Applications », *IEEE Electron Device Lett.*, vol. 33, n° 11, p. 1565-1567, nov. 2012, doi: 10.1109/LED.2012.2212174.
- [250] D. Ghosh, M. S. Parihar, G. A. Armstrong, et A. Kranti, « High-Performance Junctionless MOSFETs for Ultralow-Power Analog/RF Applications », *IEEE Electron Device Lett.*, vol. 33, n° 10, p. 1477-1479, oct. 2012, doi: 10.1109/LED.2012.2210535.
- [251] J.-P. Colinge, *Silicon-on-Insulator Technology: Materials to VLSI: Materials to VLSI*. Springer Science & Business Media, 2012.
- [252] « p-n Junctions ». http://ece.colorado.edu/~bart/book/book/chapter4/ch4_2.htm (consulté le mars 18, 2020).
- [253] C.-C. Yang *et al.*, « High Gamma Value 3D-Stackable HK/MG-Stacked Tri-Gate Nanowire Poly-Si FETs With Embedded Source/Drain and Back Gate Using Low Thermal Budget Green Nanosecond Laser Crystallization Technology », *IEEE Electron Device Lett.*, vol. 37, n° 5, p. 533-536, mai 2016, doi: 10.1109/LED.2016.2537381.
-

-
- [254] J.-Y. Lin, P.-Y. Kuo, K.-L. Lin, C.-C. Chin, et T.-S. Chao, « Junctionless Poly-Si Nanowire Transistors With Low-Temperature Trimming Process for Monolithic 3-D IC Application », *IEEE Trans. Electron Devices*, vol. 63, n° 12, p. 4998-5003, déc. 2016, doi: 10.1109/TED.2016.2615805.
- [255] M. Mandurah, « Phosphorus Doping of Low Pressure Chemically Vapor-Deposited Silicon Films », *J. Electrochem. Soc. - J ELECTROCHEM SOC*, vol. 126, janv. 1979, doi: 10.1149/1.2129167.
- [256] C. Fenouillet-Beranger *et al.*, « 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)*, oct. 2016, p. 1-2, doi: 10.1109/S3S.2016.7804375.
- [257] P. Acosta Alba *et al.*, « Nanosecond Laser Annealing for Phosphorous Activation in Ultra-Thin Implanted Silicon-On-Insulator Substrates », in *2016 21st International Conference on Ion Implantation Technology (IIT)*, sept. 2016, p. 1-4, doi: 10.1109/IIT.2016.7882896.
- [258] H. Kuriyama *et al.*, « Comprehensive Study of Lateral Grain Growth in Poly-Si Films by Excimer Laser Annealing and Its Application to Thin Film Transistors », *Jpn. J. Appl. Phys.*, vol. 33, p. 5657, oct. 1994, doi: 10.1143/JJAP.33.5657.
- [259] J. Y. W. Seto, « The electrical properties of polycrystalline silicon films », *J. Appl. Phys.*, vol. 46, n° 12, p. 5247-5254, déc. 1975, doi: 10.1063/1.321593.
- [260] H. Kuriyama *et al.*, « Enlargement of poly-Si film grain size by excimer laser annealing and its application to high-performance poly-Si thin film transistor », *Jpn. J. Appl. Phys.*, vol. 30, n° 12S, p. 3700, 1991.
- [261] D. A. Buchanan, « Scaling the gate dielectric: Materials, integration, and reliability », *IBM J. Res. Dev.*, vol. 43, n° 3, p. 245-264, mai 1999, doi: 10.1147/rd.433.0245.
- [262] P. Kumar, « Impact of 14/28nm FDSOI high-k metal gate stack processes on reliability and electrostatic control through combined electrical and physicochemical characterization techniques », phdthesis, Université Grenoble Alpes, 2018.
- [263] K. Shiraishi *et al.*, « Physics in Fermi level pinning at the polySi/Hf-based high-k oxide interface », juill. 2004, vol. 43, p. 108-109, doi: 10.1109/VLSIT.2004.1345421.
- [264] M. T. Bohr, R. S. Chau, T. Ghani, et K. Mistry, « The High-k Solution », *IEEE Spectr.*, vol. 44, n° 10, p. 29-35, oct. 2007, doi: 10.1109/MSPEC.2007.4337663.
- [265] O. Weber *et al.*, « Work-function engineering in gate first technology for multi-VT dual-gate FDSOI CMOS on UTBOX », *IEDM Tech Dig*, déc. 2010, doi: 10.1109/IEDM.2010.5703289.
- [266] E. P. Gusev, V. Narayanan, et M. M. Frank, « Advanced high- κ dielectric stacks with polySi and metal gates: Recent progress and current challenges », *IBM J. Res. Dev.*, vol. 50, n° 4.5, p. 387-410, juill. 2006, doi: 10.1147/rd.504.0387.
- [267] J.-B. Henry, « Contribution à l'étude expérimentale des résistances d'accès dans les transistors de dimensions deca-nanométrique des technologies CMOS FD-SOI », thesis, Grenoble Alpes, 2018.
- [268] S. E. Thompson et S. Parthasarathy, « Moore's law: the future of Si microelectronics », *Mater. Today*, vol. 9, n° 6, p. 20-25, juin 2006, doi: 10.1016/S1369-7021(06)71539-5.
- [269] L. Pasini *et al.*, « High performance low temperature activated devices and optimization guidelines for 3D VLSI integration of FD, TriGate, FinFET on insulator », in *2015 Symposium on VLSI Technology (VLSI Technology)*, juin 2015, p. T50-T51, doi: 10.1109/VLSIT.2015.7223699.
- [270] L. Pasini, « Low temperature devices (FDSOI, TriGate) junction optimization for 3D sequential integration », phdthesis, Université Grenoble Alpes, 2016.
- [271] J. Micout, « Fabrication et caractérisation de transistor réalisée à basse température pour l'intégration 3D séquentielle », phdthesis, Université Grenoble Alpes, 2019.
- [272] F. Cristiano, « Ion Implantation-Induced extended defects: structural investigations and impact on Ultra-Shallow Junction properties », thesis, Université Paul Sabatier - Toulouse III, 2013.
- [273] S. U. Campisano, « Impurity and concentration dependence of growth rate during solid epitaxy of implanted Si », *Appl. Phys. A*, vol. 29, n° 3, p. 147-149, nov. 1982, doi: 10.1007/BF00617771.
- [274] J. S. Williams, « Solid phase epitaxial regrowth phenomena in silicon », *Nucl. Instrum. Methods Phys. Res.*, vol. 209-210, p. 219-228, mai 1983, doi: 10.1016/0167-5087(83)90803-7.
-

-
- [275] P. Pichler, *Intrinsic Point Defects, Impurities, and Their Diffusion in Silicon*. Wien: Springer-Verlag, 2004.
- [276] L. Ehouarne, « Métallisation des mémoires Flash à base de NiSi et d'éléments d'alliage », thesis, Aix-Marseille 3, 2008.
- [277] R. Trevisoli, R. T. Doria, M. de Souza, et M. A. Pavanello, « Improved analog operation of junctionless nanowire transistors using back bias », in *EUROSOI-ULIS 2015: 2015 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon*, janv. 2015, p. 265-268, doi: 10.1109/ULIS.2015.7063824.
- [278] G. Ghibaudo, O. Roux, et J. Brini, « Modeling of conductance fluctuations in small area metal-oxide-semiconductor transistors », *Phys. Status Solidi A*, vol. 127, n° 1, p. 281-294, 1991, doi: 10.1002/pssa.2211270132.
- [279] E. G. Ioannidis, S. Haendler, C. G. Theodorou, S. Lasserre, C. A. Dimitriadis, et G. Ghibaudo, « Evolution of low frequency noise and noise variability through CMOS bulk technology nodes from 0.5 μ m down to 20nm », *Solid-State Electron.*, vol. 95, p. 28-31, mai 2014, doi: 10.1016/j.sse.2014.03.002.
- [280] C. Diaz-Llorente *et al.*, « Impact of Low-Temperature Coolcube™ Process on the Performance of FDSOI Tunnel FETs », oct. 2018, p. 1-3, doi: 10.1109/S3S.2018.8640190.
- [281] M. Horowitz, « 1.1 Computing's energy problem (and what we can do about it) », in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, févr. 2014, p. 10-14, doi: 10.1109/ISSCC.2014.6757323.
- [282] H. Esmaeilzadeh, E. Blem, R. S. Amant, K. Sankaralingam, et D. Burger, « Dark Silicon and the End of Multicore Scaling », *IEEE Micro*, vol. 32, n° 3, p. 122-134, mai 2012, doi: 10.1109/MM.2012.17.
- [283] S. Hamdioui *et al.*, « Memristor for computing: Myth or reality? », in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, mars 2017, p. 722-731, doi: 10.23919/DATE.2017.7927083.
- [284] I. Vourkas et G. C. Sirakoulis, « A Novel Design and Modeling Paradigm for Memristor-Based Crossbar Circuits », *IEEE Trans. Nanotechnol.*, vol. 11, n° 6, p. 1151-1159, nov. 2012, doi: 10.1109/TNANO.2012.2217153.
- [285] S. Kvatinsky, N. Wald, G. Satat, A. Kolodny, U. C. Weiser, et E. G. Friedman, « MRL — Memristor Ratioed Logic », in *2012 13th International Workshop on Cellular Nanoscale Networks and their Applications*, août 2012, p. 1-6, doi: 10.1109/CNNA.2012.6331426.
- [286] S. Kvatinsky *et al.*, « MAGIC—Memristor-Aided Logic », *IEEE Trans. Circuits Syst. II Express Briefs*, vol. 61, n° 11, p. 895-899, nov. 2014, doi: 10.1109/TCSII.2014.2357292.
- [287] Lei Xie, Hoang Anh Du Nguyen, M. Taouil, S. Hamdioui, et K. Bertels, « Fast boolean logic mapped on memristor crossbar », in *2015 33rd IEEE International Conference on Computer Design (ICCD)*, oct. 2015, p. 335-342, doi: 10.1109/ICCD.2015.7357122.
- [288] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, et R. S. Williams, « 'Memristive' switches enable 'stateful' logic operations via material implication », *Nature*, vol. 464, n° 7290, p. 873-876, avr. 2010, doi: 10.1038/nature08940.
- [289] S. Kvatinsky, G. Satat, N. Wald, E. G. Friedman, A. Kolodny, et U. C. Weiser, « Memristor-Based Material Implication (IMPLY) Logic: Design Principles and Methodologies », *IEEE Trans. Very Large Scale Integr. VLSI Syst.*, vol. 22, n° 10, p. 2054-2066, oct. 2014, doi: 10.1109/TVLSI.2013.2282132.
- [290] L. E. R. R. T. S. B. U, et W. R., « Beyond von Neumann--logic operations in passive crossbar arrays alongside memory operations », *Nanotechnology*, mars 08, 2012. <https://pubmed.ncbi.nlm.nih.gov/22782173/> (consulté le juill. 31, 2020).
- [291] L. Gao, F. Alibart, et D. B. Strukov, « Programmable CMOS/Memristor Threshold Logic », *IEEE Trans. Nanotechnol.*, vol. 12, n° 2, p. 115-119, mars 2013, doi: 10.1109/TNANO.2013.2241075.
- [292] G. S. Rose, J. Rajendran, H. Manem, R. Karri, et R. E. Pino, « Leveraging Memristive Systems in the Construction of Digital Logic Circuits », *Proc. IEEE*, vol. 100, n° 6, p. 2033-2049, juin 2012, doi: 10.1109/JPROC.2011.2167489.
-

-
- [293] S. Li, C. Xu, Q. Zou, J. Zhao, Y. Lu, et Y. Xie, « Pinatubo: A processing-in-memory architecture for bulk bitwise operations in emerging non-volatile memories », in *2016 53rd ACM/EDAC/IEEE Design Automation Conference (DAC)*, juin 2016, p. 1-6, doi: 10.1145/2897937.2898064.
- [294] L. Xie *et al.*, « Scouting Logic: A Novel Memristor-Based Logic Design for Resistive Computing », in *2017 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, juill. 2017, p. 176-181, doi: 10.1109/ISVLSI.2017.39.
- [295] C.-X. Xue *et al.*, « Embedded 1-Mb ReRAM-Based Computing-in-Memory Macro With Multibit Input and Weight for CNN-Based AI Edge Processors », *IEEE J. Solid-State Circuits*, vol. 55, n° 1, p. 203-215, janv. 2020, doi: 10.1109/JSSC.2019.2951363.
- [296] W.-H. Chen *et al.*, « A 65nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16ns multiply-and-accumulate for binary DNN AI edge processors », in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, févr. 2018, p. 494-496, doi: 10.1109/ISSCC.2018.8310400.
- [297] W.-H. Chen *et al.*, « A 16Mb dual-mode ReRAM macro with sub-14ns computing-in-memory and memory functions enabled by self-write termination scheme », in *2017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 28.2.1-28.2.4, doi: 10.1109/IEDM.2017.8268468.
- [298] R. Mochida *et al.*, « A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture », in *2018 IEEE Symposium on VLSI Technology*, juin 2018, p. 175-176, doi: 10.1109/VLSIT.2018.8510676.
- [299] Q. Luo *et al.*, « 8-Layers 3D vertical RRAM with excellent scalability towards storage class memory applications », in *2017 IEEE International Electron Devices Meeting (IEDM)*, déc. 2017, p. 2.7.1-2.7.4, doi: 10.1109/IEDM.2017.8268315.
- [300] D. B. Strukov, G. S. Snider, D. R. Stewart, et R. S. Williams, « The missing memristor found », *Nature*, vol. 453, n° 7191, p. 80-83, mai 2008, doi: 10.1038/nature06932.
- [301] M. Suri *et al.*, « Impact of PCM resistance-drift in neuromorphic systems and drift-mitigation strategy », New York, United States, juill. 2013, doi: 10.1109/NanoArch.2013.6623059.
- [302] F. Xiong *et al.*, « Self-Aligned Nanotube–Nanowire Phase Change Memory », *Nano Lett.*, vol. 13, n° 2, p. 464-469, févr. 2013, doi: 10.1021/nl3038097.
- [303] H. Jeong et L. Shi, « Memristor devices for neural networks », *J. Phys. Appl. Phys.*, vol. 52, n° 2, p. 023003, oct. 2018, doi: 10.1088/1361-6463/aae223.
- [304] L. Perniola *et al.*, « Universal Signatures from Non-Universal Memories: Clues for the Future... », in *2016 IEEE 8th International Memory Workshop (IMW)*, mai 2016, p. 1-3, doi: 10.1109/IMW.2016.7495295.
- [305] H. Y. Lee *et al.*, « Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO₂ based RRAM », in *2008 IEEE International Electron Devices Meeting*, déc. 2008, p. 1-4, doi: 10.1109/IEDM.2008.4796677.
- [306] D. Kumar, R. Aluguri, U. Chand, et T. Y. Tseng, « Metal oxide resistive switching memory: Materials, properties and switching mechanisms », *Ceram. Int.*, vol. 43, p. S547-S556, août 2017, doi: 10.1016/j.ceramint.2017.05.289.
- [307] « Superior filament formation control in HfO₂ based RRAM for high-performance low-power operation of 1 μ A to 20 μ A at \pm 1V | Request PDF », *ResearchGate*. https://www.researchgate.net/publication/254036362_Superior_filament_formation_control_in_HfO2_based_RRAM_for_high-performance_low-power_operation_of_1_A_to_20_A_at_-1V (consulté le août 03, 2020).
- [308] U. Chand, K. Huang, C. Huang, et T. Tseng, « Mechanism of Nonlinear Switching in HfO₂-Based Crossbar RRAM With Inserting Large Bandgap Tunneling Barrier Layer », *IEEE Trans. Electron Devices*, vol. 62, n° 11, p. 3665-3670, nov. 2015, doi: 10.1109/TED.2015.2471835.
- [309] E. Vianello *et al.*, « Back-end 3D integration of HfO₂-based RRAMs for low-voltage advanced IC digital design », in *Proceedings of 2013 International Conference on IC Design & Technology (ICICDT)*, Pavia, Italy, mai 2013, p. 235-238, doi: 10.1109/ICICDT.2013.6563344.
- [310] « (PDF) Resistive switching memories based on metal oxides: Mechanisms, reliability and scaling », *ResearchGate*.
-

- https://www.researchgate.net/publication/303291282_Resistive_switching_memories_based_on_metal_oxides_Mechanisms_reliability_and_scaling (consulté le août 03, 2020).
- [311] T. Dalgaty *et al.*, « Hybrid neuromorphic circuits exploiting non-conventional properties of RRAM for massively parallel local plasticity mechanisms », *APL Mater.*, vol. 7, n° 8, p. 081125, août 2019, doi: 10.1063/1.5108663.
- [312] A. Levisse, B. Giraud, J. P. Noël, M. Moreau, et J. M. Portal, « SneakPath compensation circuit for programming and read operations in RRAM-based CrossPoint architectures », in *2015 15th Non-Volatile Memory Technology Symposium (NVMTS)*, oct. 2015, p. 1-4, doi: 10.1109/NVMTS.2015.7457426.
- [313] J. Sandrini *et al.*, « OxRAM for embedded solutions on advanced node: scaling perspectives considering statistical reliability and design constraints », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 30.5.1-30.5.4, doi: 10.1109/IEDM19573.2019.8993484.
- [314] E. Vianello *et al.*, « Resistive Memories for Ultra-Low-Power embedded computing design », in *2014 IEEE International Electron Devices Meeting*, déc. 2014, p. 6.3.1-6.3.4, doi: 10.1109/IEDM.2014.7046995.
- [315] A. Grossi *et al.*, « Fundamental variability limits of filament-based RRAM », in *2016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 4.7.1-4.7.4, doi: 10.1109/IEDM.2016.7838348.
- [316] G. W. Burr *et al.*, « Access devices for 3D crosspoint memory », *J. Vac. Sci. Technol. B*, vol. 32, n° 4, p. 040802, juill. 2014, doi: 10.1116/1.4889999.
- [317] Y. Sasago *et al.*, « Cross-point phase change memory with 4F2 cell size driven by low-contact-resistivity poly-Si diode », in *2009 Symposium on VLSI Technology*, juin 2009, p. 24-25.
- [318] M. Lee *et al.*, « 2-stack 1D-1R Cross-point Structure with Oxide Diodes as Switch Elements for High Density Resistance RAM Applications », in *2007 IEEE International Electron Devices Meeting*, déc. 2007, p. 771-774, doi: 10.1109/IEDM.2007.4419061.
- [319] Qianqian Huang *et al.*, « Self-depleted T-gate Schottky barrier tunneling FET with low average subthreshold slope and high ION/IOFF by gate configuration and barrier modulation », in *2011 International Electron Devices Meeting*, déc. 2011, p. 16.2.1-16.2.4, doi: 10.1109/IEDM.2011.6131564.
- [320] W. Y. Park *et al.*, « A Pt/TiO₂/Ti Schottky-type selection diode for alleviating the sneak current in resistance switching memory arrays », *Nanotechnology*, vol. 21, n° 19, p. 195201, avr. 2010, doi: 10.1088/0957-4484/21/19/195201.
- [321] DerChang Kau *et al.*, « A stackable cross point Phase Change Memory », in *2009 IEEE International Electron Devices Meeting (IEDM)*, déc. 2009, p. 1-4, doi: 10.1109/IEDM.2009.5424263.
- [322] K. Gopalakrishnan *et al.*, « Highly-scalable novel access device based on Mixed Ionic Electronic conduction (MIEC) materials for high density phase change memory (PCM) arrays », in *2010 Symposium on VLSI Technology*, juin 2010, p. 205-206, doi: 10.1109/VLSIT.2010.5556229.
- [323] S. H. Jo, T. Kumar, S. Narayanan, et H. Nazarian, « Cross-Point Resistive RAM Based on Field-Assisted Superlinear Threshold Selector », *IEEE Trans. Electron Devices*, vol. 62, n° 11, p. 3477-3481, nov. 2015, doi: 10.1109/TED.2015.2426717.
- [324] Sung Hyun Jo, T. Kumar, S. Narayanan, W. D. Lu, et H. Nazarian, « 3D-stackable crossbar resistive memory based on Field Assisted Superlinear Threshold (FAST) selector », in *2014 IEEE International Electron Devices Meeting*, déc. 2014, p. 6.7.1-6.7.4, doi: 10.1109/IEDM.2014.7046999.
- [325] S. Kim *et al.*, « Ultrathin (<10nm) Nb₂O₅/NbO₂ hybrid memory with both memory and selector characteristics for high density 3D vertically stackable RRAM applications », in *2012 Symposium on VLSI Technology (VLSIT)*, juin 2012, p. 155-156, doi: 10.1109/VLSIT.2012.6242508.
- [326] C. Ho *et al.*, « Threshold Vacuum Switch (TVS) on 3D-stackable and 4F2 cross-point bipolar and unipolar resistive random access memory », in *2012 International Electron Devices Meeting*, déc. 2012, p. 2.8.1-2.8.4, doi: 10.1109/IEDM.2012.6478968.
- [327] X. P. Wang *et al.*, « Highly compact 1T-1R architecture (4F2 footprint) involving fully CMOS compatible vertical GAA nano-pillar transistors and oxide-based RRAM cells exhibiting excellent

- NVM properties and ultra-low power operation », in *2012 International Electron Devices Meeting*, déc. 2012, p. 20.6.1-20.6.4, doi: 10.1109/IEDM.2012.6479082.
- [328] M. Ezzadeen *et al.*, « Ultrahigh-Density 3-D Vertical RRAM With Stacked Junctionless Nanowires for in-Memory-Computing Applications », présenté à *IEEE Transactions on Electron Devices*, in press 2020.
- [329] S. Barraud *et al.*, « Vertically stacked-NanoWires MOSFETs in a replacement metal gate process with inner spacer and SiGe source/drain », in *2016 IEEE International Electron Devices Meeting (IEDM)*, déc. 2016, p. 17.6.1-17.6.4, doi: 10.1109/IEDM.2016.7838441.
- [330] N. Loubet *et al.*, « A Novel Dry Selective Etch of SiGe for the Enablement of High Performance Logic Stacked Gate-All-Around NanoSheet Devices », in *2019 IEEE International Electron Devices Meeting (IEDM)*, déc. 2019, p. 11.4.1-11.4.4, doi: 10.1109/IEDM19573.2019.8993615.
- [331] D. BOSCH *et al.*, « All-operation-regime characterization and modeling of drain current variability in junctionless and inversion-mode FDSOI transistors », présenté à *Symposium on VLSI Technology (VLSI Technology)*, 2020.
- [332] T. Karatsori, « Caractérisation et modélisation de UTBB MOSFET sur SOI pour les technologies CMOS avancées et applications en simulations circuits », thesis, Grenoble Alpes, 2017.
- [333] L. Rahhal *et al.*, « New methodology for drain current local variability characterization using Y function method », in *2013 IEEE International Conference on Microelectronic Test Structures (ICMTS)*, Osaka, Japan, mars 2013, p. 99-103, doi: 10.1109/ICMTS.2013.6528153.
- [334] T. A. Karatsori, C. G. Theodorou, E. Josse, C. A. Dimitriadis, et G. Ghibaudo, « All Operation Region Characterization and Modeling of Drain and Gate Current Mismatch in 14-nm Fully Depleted SOI MOSFETs », *IEEE Trans. Electron Devices*, vol. 64, n° 5, p. 2080-2085, mai 2017, doi: 10.1109/TED.2017.2686381.
- [335] E. G. Ioannidis, C. G. Theodorou, S. Haendler, E. Josse, C. A. Dimitriadis, et G. Ghibaudo, « Impact of Source-Drain Series Resistance on Drain Current Mismatch in Advanced Fully Depleted SOI n-MOSFETs », *IEEE Electron Device Lett.*, vol. 36, n° 5, p. 433-435, mai 2015, doi: 10.1109/LED.2015.2411289.
- [336] V. P. Georgiev *et al.*, « Variability study of high current junctionless silicon nanowire transistors », in *2017 IEEE 12th Nanotechnology Materials and Devices Conference (NMDC)*, oct. 2017, p. 87-88, doi: 10.1109/NMDC.2017.8350514.
- [337] M. Denais, « ETUDE DES PHENOMENES DE DEGRADATION DE TYPE
NEGATIVE BIAS TEMPERATURE INSTABILITY (NBTI)
DANS LES TRANSISTORS MOS SUBMICRONIQUES DES
FILIERES CMOS AVANCEES », phdthesis, Université de Provence - Aix-Marseille I, 2005.
- [338] E. H. Snow, A. S. Grove, B. E. Deal, et C. T. Sah, « Ion Transport Phenomena in Insulating Films », *J. Appl. Phys.*, vol. 36, n° 5, p. 1664-1673, mai 1965, doi: 10.1063/1.1703105.
- [339] A. S. Foster, F. Lopez Gejo, A. L. Shluger, et R. M. Nieminen, « Vacancy and interstitial defects in hafnia », *Phys. Rev. B*, vol. 65, n° 17, p. 174117, mai 2002, doi: 10.1103/PhysRevB.65.174117.
- [340] Y. T. Yeow, D. R. Lamb, et S. D. Brotherton, « An investigation of the influence of low-temperature annealing treatments on the interface state density at the Si-SiO₂ », *J. Phys. Appl. Phys.*, vol. 8, n° 13, p. 1495-1506, sept. 1975, doi: 10.1088/0022-3727/8/13/011.
- [341] D. M. Fleetwood, « Border traps and bias-temperature instabilities in MOS devices », *Microelectron. Reliab.*, vol. 80, p. 266-277, janv. 2018, doi: 10.1016/j.microrel.2017.11.007.
- [342] D. M. Fleetwood, M. R. Shaneyfelt, et J. R. Schwank, « Estimating oxide-trap, interface-trap, and border-trap charge densities in metal-oxide-semiconductor transistors », *Appl. Phys. Lett.*, vol. 64, n° 15, p. 1965-1967, avr. 1994, doi: 10.1063/1.111757.
- [343] X. Garros, « Reliability of 3D Architectures (Finfets, Nanowires) », VLSI-TSA, 2020.
- [344] Kueing-Long Chen, S. A. Saller, I. A. Groves, et D. B. Scott, « Reliability effects on MOS transistors due to hot-carrier injection », *IEEE Trans. Electron Devices*, vol. 32, n° 2, p. 386-393, févr. 1985, doi: 10.1109/T-ED.1985.21953.
- [345] M. A. Alam et S. Mahapatra, « A comprehensive model of PMOS NBTI degradation », *Microelectron. Reliab.*, vol. 45, n° 1, p. 71-81, janv. 2005, doi: 10.1016/j.microrel.2004.03.019.

- [346] A. Laurent, « Etude des mécanismes physiques de fiabilité sur transistors Trigate/Nanowire », thesis, Grenoble Alpes, 2018.
- [347] et al. S. Barraud, « 3D RRAMs with Gate-All-Around Stacked Nanosheet Transistors for In-Memory-Computing », 2020.

Chapter V: General conclusions and perspectives

1- Conclusion

To pursue Moore's law, the transistor dimensions have shrunk geometrically, causing the apparition of undesirable parasitic effects for scaled dimensions. To mitigate them, and to scale down the technological nodes further, new architectures have risen to improve the electrostatic control of the gate on the transistor channel. However, performance in nowadays circuits are no longer dictated by the intrinsic transistor delay but rather by interconnections delay. To follow Moore's trend, 3D monolithic integration proposes to stack active layers on top of the other with a lithographic alignment. This particular integration scheme allows to integrate in the same silicon footprint more devices with higher interconnections resources. If we focus on power rather than energy, at a larger scale, half of the energy can be wasted during the memory access, *i.e.* when data are transferred back and forth between memory and computational units. To overcome this so-called "Memory wall", solutions like multi-core processors are already implemented but to handle power density, part of the chip (so-called "dark silicon") cannot be used. A solution called In-Memory Computing (IMC) gathers memory and computational blocks to process directly the data into the memory block and avoid transfers. In this PhD manuscript both directions which are complementary have been explored. Concerning 3D monolithic integration, the main question was how and to what extent can the circuit designers use the freedom enabled by such a vertical integration. Can stacking provide different levers to optimize a layout in terms of wire congestion, performance or area? And if the gains are substantial enough, how to integrate devices on top of the others without degrading the performances? As far as IMC is concerned, there was a desire to combine junctionless nanowires and resistive memory to create an ultra dense cube. Can such a structure enable IMC and perform Boolean operations? If yes, how to create a full 1T1R 3D cube and select the correct materials?

Chapter II investigated the interest of 3D monolithic integration from a circuit designer point of view. In fact, 3D monolithic integration is not just the stacking of two 2D planar circuits but fine-grain interconnections between the two tiers offers unique opportunities. We proposed to share resource (like power rail, clock signal...) between the two tiers to relax the constraint on interconnections to avoid interconnection congestion and shorter connections. In addition, to reduce the cell area and interconnections lengths, NAND gates have been designed to detain their inputs in top-tier and outputs in bottom tiers. Furthermore, 3D monolithic integration enables the integration of local back bias to modulate dynamically the threshold voltage of top transistors. Thanks to this additional degree of freedom for 3D designers, a versatile back-bias assist have been explored for top tier 14nm SRAM. By modulating the transistors power ratio in SRAM, depending of the performed operation (read, write), a reduction of 92mV of the minimum operating voltage is demonstrated. This assist can be integrated without area overhead and with minor design work and paves the way for top-tier low power designs. This demonstrates the attraction of 3D monolithic integration for high performance designs. However, instead of counteracting the SRAM variability, it can be seen as an asset to create physically unclonable functions (PUF) for security applications. In fact, when an SRAM cell is powered up, its state will be initialized either in '0' or '1' depending of the cell skew (due to variability). If the skew is sufficient, the state will be the same for each power-up. A matrix of such devices can define a unique fingerprint based on power-up state of SRAMs. To enhance the skew, we investigated the impact of the presence of a single dopant in the channel. The emulation of single dopant transport indicates that the introduced skew in SRAM cells leads to more reproducible power-up state with a higher tolerance to noise. To conclude, chapter II demonstrated the interest of 3D monolithic integration for high performances circuits. Nevertheless, the manufacturing of a stacked device must be done at low temperature (<500°C

instead of 1050°C) to maintain the bottom tier stability. Low temperature fabrication have been tackled in chapter III.

Chapter III focuses on the fabrication of 3D monolithic devices which must be done at low thermal budget (<500°C, 2 hours). For this, devices without source and drain to channel junctions called “junctionless”, featuring a uniformly doped channel, are envisioned to avoid the thermally costly source and drain implantation annealing. TCAD simulations show good performance of Fully-Depleted SOI junctionless devices compared to standard one “inversion –mode” (IM), especially for RF and low-power applications. In fact, the continuous doping reduces overlap capacitances. To study the impact of channel doping only, devices without temperature constraints were fabricated, highlighted excellent performance of Junctionless-Accumulation Mode transistors, with a maximum operating frequency of 182 Ghz (L=30nm and W=120nm), respectively 165 GHz for IM and an analog gain of 68.8 at $V_B = -10V$. The interest of junctionless devices being pointed out, a 400°C process flow for junctionless devices have been exposed. Several process developments, including low temperature poly-silicon channel creation with laser annealing, silicide and implantation though Solid Phase Epitaxy Regrowth process optimization are presented module by module. To go further in terms of monolithic integration, we propose to take advantage of the verticality provided by stacked nanowires. Our vision consists in integrating memory elements at the drain side of junctionless stacked nanowires. Such a structure is analyzed in depth in the next chapter.

Chapter IV deals with In-Memory Computing. We proposed an ultra-dense low-power cube, combining the emerging technologies of stacked nanowires gate-all-around transistors and Oxide-based RAM. The transistors are chosen junctionless for ease of fabrication. To verify if junctionless transistors The 3D 1T1R cube structure enables Boolean logic operation (such as bitwise AND, OR...) by performing a read operation in the desired cells. This IMC feature has been demonstrated compatible with up to 3 stacked layers by means of TCAD and SPICE simulations. A process flow, as well as a layout, have been proposed to create such a cube. Preliminaries batches are fabricated to size the future devices. An extensive study of mismatch demonstrated that the doped channel introduces more variability in the sub-threshold region due to Coulomb scattering but at higher gate voltages, lower variability is seen. This is attributed to Coulomb scattering screening.

2- Future Work: short term perspectives

Concerning 3D monolithic integration design part the possibilities brought by the introduction of back-bias, which dynamically modulate the threshold voltage of top transistors, are infinite and could be investigated at the cell design level, but also as an additional step of the 3D VLSI design flow.

During this PhD thesis, I initiated the fabrication of junctionless transistors fabricated at 400°C, far below the 500°C state-of-the-art results. Based on future measurements, the behavior of these low-temperature junctionless devices will be compared to high-temperature one. I would suggest to convey variability study on theses devices and the impact of low temperature processing. Also, we fabricated two types of gate shape, the first one being straight and the second one being T-shaped. It would be interesting to analyze their impact, in particular for RF applications.

For the IMC part, the project just begun and several parts must be gathered and are still in development to build the full operational structure. Among them I would propose to consider the peripheral circuit required for the proper operation but also to investigate the energy consumption of the proposed system. From the fabrication point of view, based on on-going studies about transistor and memory element sizing, the full matrix could be processed. In parallel, electrical characterisations can be done to select and optimize carefully the memory element. After, IMC and in particular scouting logic could be demonstrated into the full matrix.

3- Perspectives

3D monolithic integration is foreseen as an alternative to ground rule scaling at the horizon 2028 by the 2018 IRDS roadmap. A main challenge is about system partitioning to take all the benefits from this technology. From my point of view, the fabrication of low-temperature devices featuring high performance is understudied and excellent results are already demonstrated in literature, but there is still a lack of 3D place and route tools to make it appealing for industries. So one general direction would be to tend to develop 3D place and route tools with a better understanding of 3D monolithic integration advantages for semi-custom design. Even better, dedicated tools could offer different levers or functionalities to optimise 3D circuits PPAC.

Concerning IMC, I'm convinced that energy savings brought by this new computation paradigm will make the difference into a world where data are everywhere. It combines state-of-the-art technologies with ground-breaking designs and might be beneficial to others paradigms, like deep learning. On such a vast topic, there are many directions to explore and I expect to be amazed with the future development of IMC.

Annex I: Oxide defects and failure mechanisms in CMOS technology.

In fact, the crystalline structure of silicon is face-centered cubic, whereas the SiO₂ is amorphous. From this lattice discrepancy, some defects are created and can influence the MOSFET electrical performance according to their state (charged or not) [337]. Also, some additional defects are created during device processing, especially during gate oxide annealing.

Some details about oxide defects and their origin are presented below:

- **Fixed oxide charges:** they are near the Si/SiO₂ interface and are related to silicon oxidation step. The oxidization process can be optimized, in particular its temperature to achieved a good quality oxide. The fixed charges are pre-existing defects, having an impact of MOSFET parameters but do not interact with silicon in the channel, thus do not impact the ageing process.
- **Mobile oxide charges:** they are the result of ionic contamination from impurities such as K⁺, Li⁺... They are causing threshold voltage instabilities when positive gate bias is applied [338].
- **Oxide trapped charges:** when the oxide is fabricated, some defects are created in the volume especially in HfO₂ [339]. They can be filled or unfilled when an electrical stress is applied on the gate. The electron traps are distributed through the oxide whereas the hole are located near the Si/SiO₂ interface. These traps can also be generated by the device operation.
- **Interface trapped charges:** after the substrate oxidation, the mechanical strain is relaxed, creating interface traps. The interface trapped charges are created by the dangling bonds at Si/SiO₂ interface. Their density is usually noted N_I or D_{IT} (in cm⁻²eV⁻¹ or cm⁻²) and characterised by its energy level and its capacity to capture and emit mobile charges. They can be generated with device operation, under electrical stress at high field. They can be either acceptor or donor according to their position with respect to the bandgap (upper or lower half). A way to decrease the D_{IT} is to passivate the SiO₂ interface with H₂ [340]. In fact, the hydrogen atom will form a neutral Si-H liaison.
- **Border traps:** it consists of positively charged oxide traps passivated with hydrogen, for instance oxygen (O) vacancies and hydrogen. They are near the interface and can tunneled from the semiconductor to the trap back and force, but also through trap-assisted tunneling or thermal activation [341]. Border traps can be differentiated from interface traps with Charge Pumping, Low Frequency Noise and Time Dependent Defect Spectroscopy measurement [342].

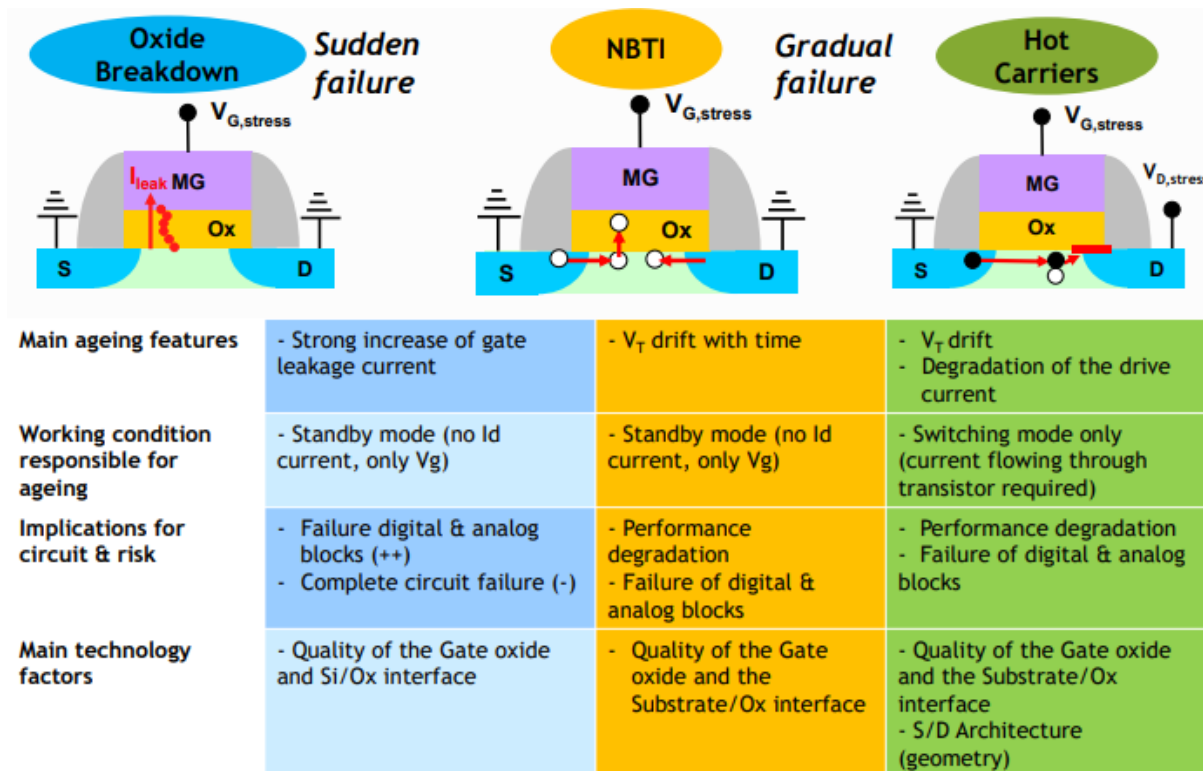


Fig. 275: Summary of the main mechanism of failure for MOS transistors. In this work, the BTI failure mechanism is analysed more in depth. Figure from [343].

With the reduction of transistor dimension, and with a lower supply voltage scaling, the electric field have been increased. In a n-channel MOSFET, this field can accelerate electrons which when stopped by collision events create a electron-holes pair. The generated electrons have enough energy to overcome the oxide potential barrier and be injected into the gate material (see Fig. 275). This electron flow generates interface states shifting the transistor threshold voltage. Different process and design solutions can be done to reduce the impact of HCI [344]. In fact, a lightly doped implantation for transistor is usually done to reduce the electric field near the drain junction edge and thus reduce the emission probability of hot carriers. Also, the gate oxide quality can be improved as far as the size (capture cross section) and density of hot-carrier traps are concerned. Please note that to overcome the oxide potential barrier, an electron (or a hole) must gain a kinetic energy of 3.2eV (4.6eV).

As for as BTI is concerned, the stress applied on the gate will de-passivate the neutral Si-H liaison, creating interface state N_{IT} and shifting the V_T . The time evolution is described by a power-law relationship: $\Delta V_T \sim t^\gamma$ [345]. According to the sign of the applied potential, positive or negative, PBTI (usually for NMOS) or NBTI (usually for PMOS) term is used. The NBTI degradation more critical than PBTI, inducing a threshold voltage degradation 4.5 times higher for NBTI than PBTI [346].

Annex II: Junctionless Threshold voltage analytical expression

Let's consider the case of a single (or planar) gate transistor. The threshold voltage is defined as the gate voltage for which the neutral region disappears in the middle of the channel. The depletion comes from the gate-semiconductor work function difference represented by the flat band voltage V_{FB} . Gate oxide charges are taken into account in VFB expression. The applied gate voltage (V_G) will change the voltage drop across the gate oxide (φ_{ox}) and the surface potential (φ_s).

$$V_G = V_{FB} + \varphi_{ox} + \varphi_s \quad \text{Eq. 30}$$

By using Poisson's law, the potential distribution $\varphi(x)$ is linked to ρ the charge density in the silicon film and ϵ_{si} the permittivity of silicon. ρ can be expressed as $\rho = q \cdot N_D$, N_D being the dopant donor density and q the electronic charge.

$$\frac{\partial^2 \varphi(x)}{\partial x^2} = \frac{-\rho}{\epsilon_{si}} \quad \text{Eq. 31}$$

When integrating the potential distribution (Eq. 31) with respect to x , the electric field distribution across the film can be expressed as:

$$E(x) = \frac{q \cdot N_D \cdot x}{\epsilon_{si}} + cst \quad \text{Eq. 32}$$

A limit condition is that the electric field vanished when the depletion length is reached, *i.e.* $E(x_{dep})=0$. So the expression of the constant is:

$$cst = \frac{-q \cdot N_D \cdot x_{dep}}{\epsilon_{si}} \quad \text{Eq. 33}$$

By integrating Eq. 32 with respect to x between $x=0$ (SiO_2 interface) and x_{dep} , the potential is expressed as:

$$\varphi(x_{dep}) - \varphi(0) = \frac{q \cdot N_D \cdot x_{dep}^2}{2 \cdot \epsilon_{si}} \quad \text{Eq. 34}$$

We also assumed that the potential at x_{dep} is null, so $\varphi(x_{dep}) = 0$ which lead to:

$$\varphi(0) = \varphi_s = \frac{-q \cdot N_D \cdot x_{dep}^2}{2 \cdot \epsilon_{si}} \quad \text{Eq. 35}$$

The electric field at the surface, E_s can be derived:

$$E_s = E(0) = \frac{-q \cdot N_D \cdot x_{dep}}{\epsilon_{si}} \quad \text{Eq. 36}$$

The displacement vector must be continuous at the interface so $E_{ox} \cdot \epsilon_{ox} = E_s \cdot \epsilon_{si}$. Also, assuming a perfect gate oxide, the electric field is constant in the oxide thickness and the voltage drop can be obtained:

$$\varphi_{ox} = E_{ox} t_{ox} = \frac{-q \cdot N_D \cdot x_{dep} t_{ox}}{\epsilon_{ox}} \quad \text{Eq. 37}$$

We have now an analytical expression for V_G . Note that for $V_G=V_T$, x_{dep} is equals for single gate to t_{si} (for a double gate to $t_{si}/2$). We obtain an equation for V_T :

$$V_T = V_{FB} - \frac{q \cdot N_D \cdot t_{si}^2}{2 \cdot \epsilon_{si}} - \frac{q \cdot N_D \cdot t_{ox} \cdot t_{si}}{\epsilon_{ox}} \quad \text{Eq. 38}$$

Annex III: MY-CUBE layout, process flow and limitations

In this annex we will first introduce the layout of MY-CUBE structure and in a second time propose a process flow before ending by Design analysis.

i. MY-CUBE layout:

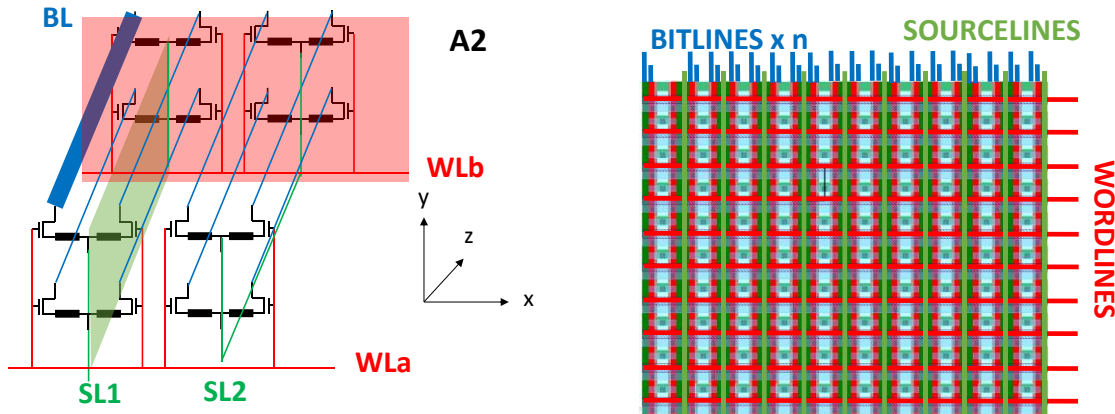


Fig. 276: equivalent electrical schematics of MY-CUBE topology to present the routing between 1T1R pillars.

Fig. 277: layout of MY-CUBE topology. The bitline contact are done in a dedicated area.

We demonstrated previously that up to three layers were compatible with Scouting logic based IMC. As far as the array is concerned, the addressing scheme is quite straightforward and have been presented in part 2-d. In fact, as represented in Fig. 276, to select a particular 1T1R cell in the 3D array, a wordline (WL), a bitline (BL) and a sourceline (SL) are required. For the shake of simplicity, we consider the case of two layers and two pillars. The bitlines rank from one to eight, each selecting one of the four bitlines in the z axis. The number of wordlines and sourceline is for the moment two but can be extended to obtain larger arrays. The main question relies on how to connect with metal one and two wordlines, bitlines and sourcelines. Similarly to an SRAM matrix, the wordline and sourceline, each selecting a row or a column can form a grid whose connections are external to the matrix as illustrated in the layout (Fig. 277). However, for the bitlines, they need to contact each layer independently. For this, in a dedicated area the contact will be done to each layer following a stair scheme. This area is necessary, and depends on the number of layers and wordlines and is not scalable.

Fig. 278 provides a 3D representation of this layout done with Coventor SEMulator 3D software. SEMulator 3D is a process emulation software to perform process variation studies. In the next part, a process flow, modeled with Coventor, to realise MY-CUBE array is presented. For each step, basic input parameters such as selectivity or deposition conformity have been entered to account for cleanroom process. Based on the proposed layout (Fig. 279), the cell size have been evaluated to $(23.9 \times F^2)/n$, F being the minimum feature size (in our custom design kit $F=45\text{nm}$) and n the number of stacked layers. Main limitations to the cell size area the poly-cut width (W_{CT}) mandatory to separate each wordline and metal-via pitch (P_{M1}). We can notice that for $n=6$ the obtained cell size is competitive with crossbar memory density ($4F^2$).

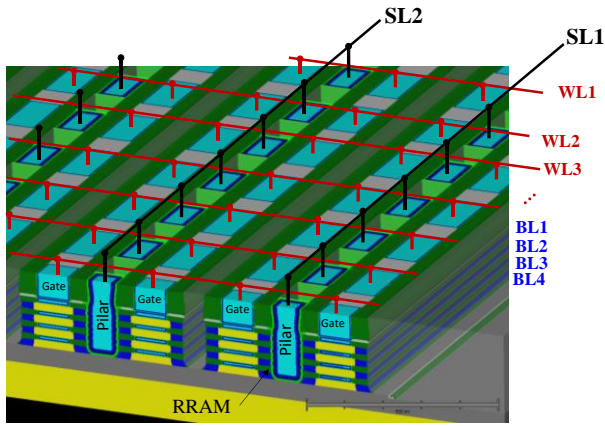


Fig. 278: 3D coventor representation of MY-CUBE from the previous layout.

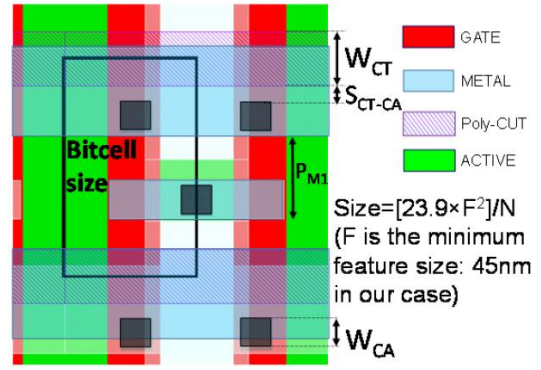
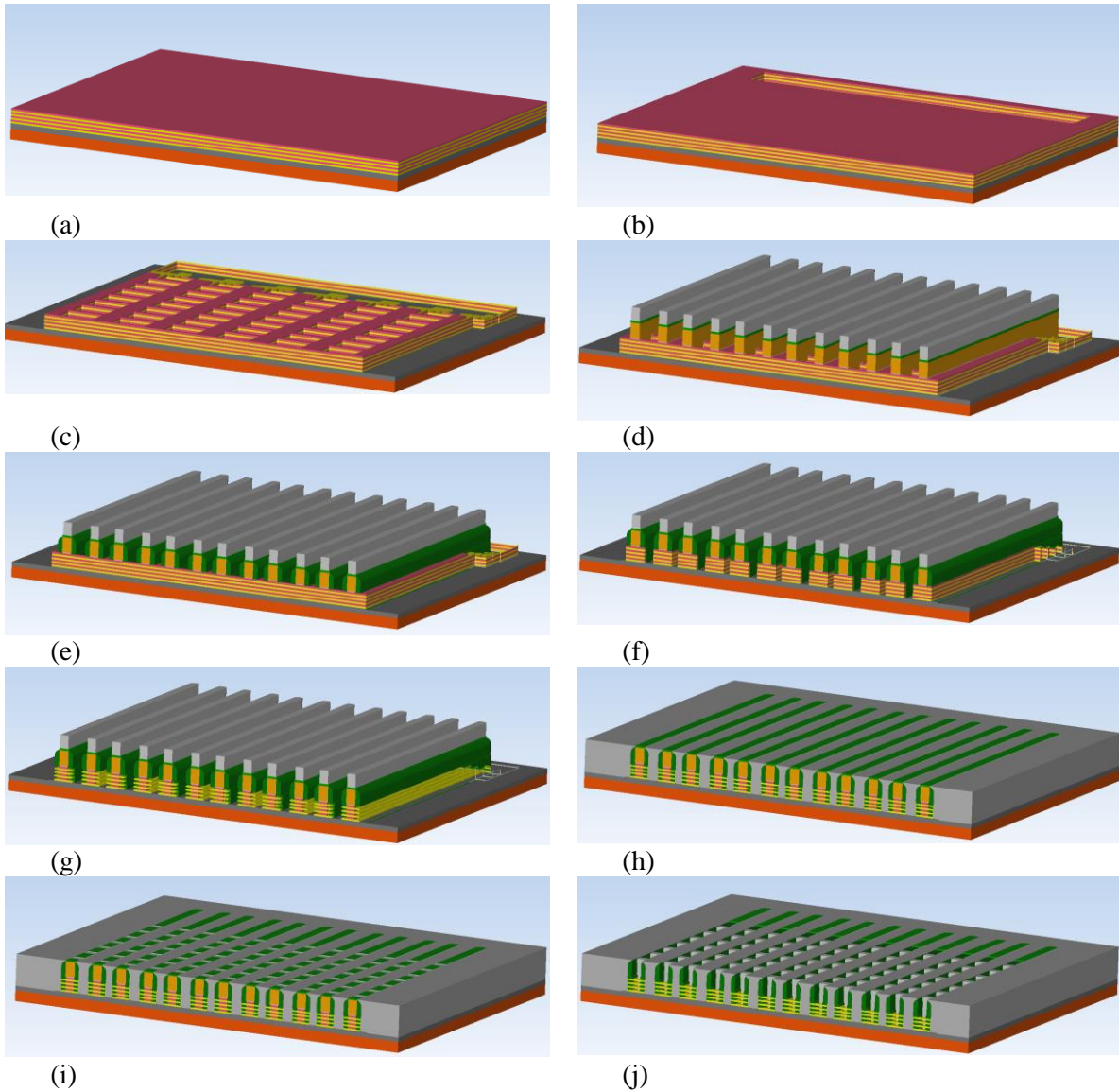


Fig. 279: Evaluation of the cell size. Reproduction from [347].

ii. Proposed process flow for MY-CUBE integration:



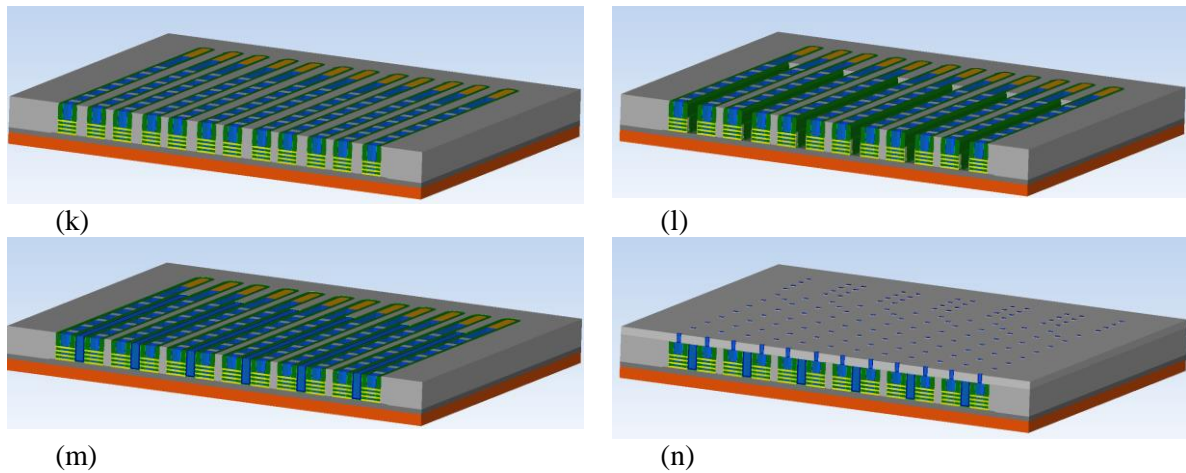


Fig. 280: Presentation of the main process steps to create MY-CUBE structure with four layers. (a) Epitaxy of $(\text{SiGe}_{0.3}/\text{Si:P}) \times 4$ superlattice; the silicon is doped in-situ to provide the future transistor channel material. (b) Scheduling of lithography and etching steps ($\times 4$) to create the bitline contacts (c) Etching of the active area. (d) Deposition TEOS 7nm and polysilicon, planarization. Then, the hard mask is done with SiN and TEOS deposition and etching. The poly-silicon and TEOS are also etched before removing the resist. (e) The IRAD spacer is deposited and etched. (f) the superlattice is etched. (g) As in the process flow of stacked nanowires, the SiGe is etched (h) Formation of inner spacer (IRAD). (i) Separation of the wordlines with the definition and the filling of a cut by oxide. (j) Removal of the dummy gate. (k) Deposition of HfO_2 as a gate oxide and $\text{TiN} + \text{W}$ as a gate metal and planarization. (l) Definition of sourcelines: etching. (m) Sourceline filling by the memory element stack (silicide, Ti, HfO_2 , TiN) and W filling. (n) Contact definition.

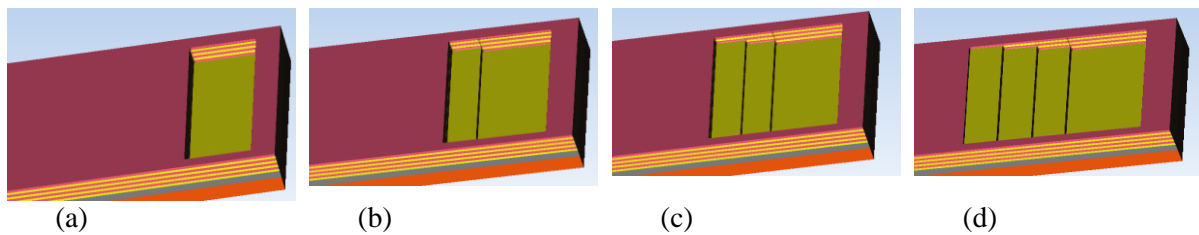


Fig. 281: Presentation of the steps of bitlines contact module.

Fig. 280 presents the main steps of MY-CUBE process flow. Resist deposition, exposure and removal are not indicated but are usually part of an etching step. For more information about the basics process steps, please refer to chapter III section 6-. The masks used to emulate the structure are taken from the layout view of Fig. 277. The first step, depicted in Fig. 280-a, consists in an epitaxial growth of $(\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si:P})$ multilayers. Unlike the nanowire case, the silicon is phosphorus in-situ doped to create the future channel material of junctionless transistor. The sizing of these layers (thickness and doping) will determine the characteristics of the transistor. However, note that there is an interest in having a large level of doping (*i.e* a low resistivity), since the future bitlines will be constituted of the same material. In this representation, four $(\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si:P})$ layers are created to account for a MY-CUBE structure of four levels. The second step is about the bitline module (Fig. 280-b) to dissociate each $(\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si:P})$ layer in order to connect them independently later. The stairs scheme is detailed in Fig. 281. First, the resist is deposited and exposed to allow a rectangle shape etching at the end of matrix. Then the first three $(\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si:P})$ layers are etched as well as the final $\text{Si}_{0.7}\text{Ge}_{0.3}$ to reveal the last Si:P layer. After the resist is removed and the same operation is performed again with a shifted to the left rectangle. This time, the etching process stops at the third Si:P layer (second to last Si:P layer). After two more steps, the final structure (Fig. 281-d) dissociates each level like stairs. Note that the dissociation of bitlines in a same level is not represented on the schematic for the shakes of simplicity. To conclude this bitline module, n lithography steps are needed for n stacked layers. The third step (Fig. 280-c) define the active area according to the layout. The fourth step presented in Fig. 280-d is about the creation of the dummy gate as in the gate-last nanowire flow. For this, a TEOS and polysilicon layers are deposited and planarized. Then, the hard mask is done with SiN and TEOS deposition. The full stack

is then etched to form the sacrificial gate. The fifth step (Fig. 280-e) is about the formation of an IRAD spacer and is realized similarly to the stacked nanowire process flow. The sixth step (Fig. 280-f) the $(\text{Si}_{0.7}\text{Ge}_{0.3}/\text{Si:P})$ multilayers are etched to allow the etching of SiGe extremity layer (Fig. 280-g) and the formation of inner spacer in Fig. 280-h. The seventh step (Fig. 280-i) consists in isolating the gate in the depth since the gate connections must be perpendicular to the sourceline and bitline ones. This is done thanks to a CUT mask where in the spacing between gates an oxide is deposited (a shape is etched and filled) to isolate the gates. After, the gate module will remove the dummy gate (Fig. 280-j) and form the future gate stack (Fig. 280-k) with the HfO_2 and TiN and W deposition. The HfO_2 deposition is conformal and will wrap the silicon nanowires. Then, the sourceline module is executed and a trench at the transistor edge is made and filled with the memory element and W. To finish with, the contact module is performed to connect the wordlines, sourcelines and bitlines according to the addressing schematic.

We presented a process flow proposition for the final structure. However, if the feasibility of scouting logic have been demonstrated in a single pillar, the full matrix was not considered. Especially, depending of the Si:P doping level, the sourcelines resistivity to address a bitcell is quite high and can endanger the good operation of the matrix for far away bitcell. In the next part, we will tackle this issue and analyse if it limits the dimension of the matrix.

iii. Impact of long access line in Si:P

In the matrix, the RRAM are addressed by source lines, made of the same material as transistor channel (so phosphorous doped at N_D), which are running along the array. Based on the previous GDS, we emulated the process with CLEVER (Fig. 282) to compute the sourceline resistances for different doping. Fig. 283 presents the resistance of the sourceline as a function of the distance for various phosphorus doping level lines. Even for short distances (around $0.5\mu\text{m}$) the Si:P induced resistance is of the order of $10\text{k}\Omega$. If we consider such a resistance line in series with a junctionless transistors, the access resistance will be so important that the I_D-V_G curve will be flatten depending on the distance to the bitline contact (Fig. 284). This effect will limit the depth of the array (and thus the number of source lines). However, if the lines are made of tungsten the line resistivity is divided by a factor 100. In this case, the resistivity is of the order of $\text{k}\Omega$ ($4\mu\text{m}$ length), detaining a smaller impact on transistor drive current. That is why, the lines have to be made out of W instead of keeping the Silicon phosphorous doped material to ensure a correct drive current. This metallic bitline module is still in development and validation and remains as a challenge to overcome for future device processing.

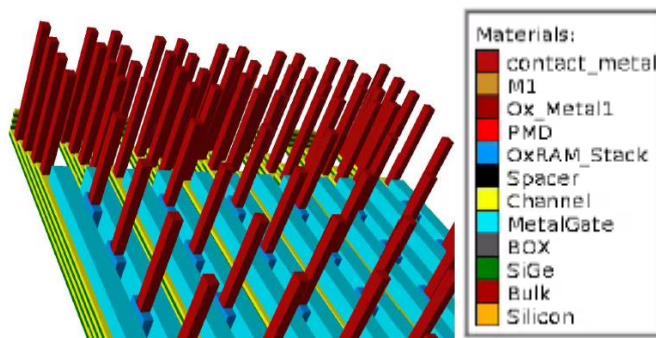


Fig. 282: 3D structure emulated by Clever to compute the access resistances. This simulation have been realized by J. Lacord.

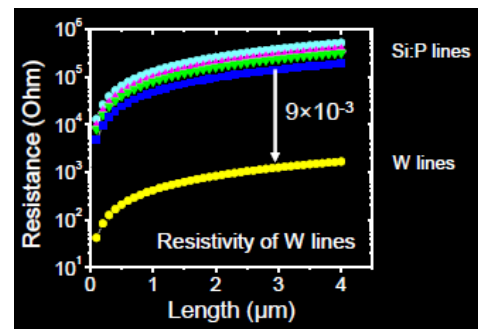


Fig. 283: Resistance of the metal lines as a function of length. This simulation have been realized by J. Lacord.

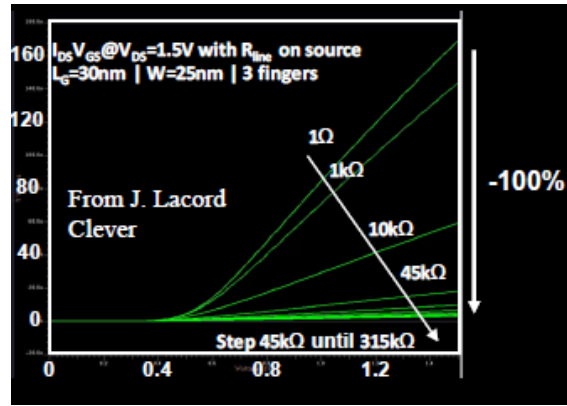


Fig. 284: Impact on I_D - V_G of the additional access resistance.

List of publications

As first author:

- Bosch, D., Garros, X., Makosiej, A., Ciampolini, L., Weber, O., Lacord, J., Cluzel, J., Giraud, B., Berthelon, R., Cibrario, G., Brunet, L., Batude, P., Fenouillet-Béranger, C., Lattard, D., Colinge, J.P., Balestra, F., Andrieu, F., 2019. Novel fine-grain back-bias assist techniques for 3D-monolithic 14 nm FDSOI top-tier SRAMs. *Solid-State Electronics* 107720. <https://doi.org/10.1016/j.sse.2019.107720>
- Bosch, D., Garros, X., Makosiej, A., Ciampolini, L., Weber, O., Lacord, J., Cluzel, J., Giraud, B., Berthelon, R., Cibrario, G., Brunet, L., Batude, P., Fenouillet-Béranger, C., Lattard, D., Colinge, J.P., Balestra, F., Andrieu, F., 2019. Back-bias impact on variability and BTI for 3D-monolithic 14nm FDSOI SRAMs applications, EUROSOI.
- Bosch, D., F., Andrieu, Ciampolini, L., Makosiej, A., Weber, O., Garros, X. Lacord, J., Cluzel, Esmanhotto, E., Rios, M., Lang S., J., Giraud, B., Berthelon, R., Cibrario, G., Brunet, L., Batude, P., Fenouillet-Béranger, C., Lattard, D., Colinge, J.P., Balestra, Vinet, M., Novel Fine-Grain Back-Bias Assist Techniques for 14nm FDSOI Top-Tier SRAMs integrated in 3D-Monolithic, 2019, VLSI-TSA.
- Bosch, D., Acosta P., Kerdiles S., Benevent V., Perrot C., Lassarre J., Richy J., Brunet, L., Batude, P., Fenouillet-Béranger, C., Lattard, D., Colinge J.P. , Balestra, F., Andrieu, 2019, Laser Processing for 3D junctionless transistor fabrication, S3S.
- Bosch, D., Colinge, J.P., Lugo J., Tataridou A., Theodorou C., Garros, X., Barraud S., Lacord, J., Sklenard B, Casse M., Brunet, L., Batude, P., Fenouillet-Béranger, C., Lattard, D., Cluzel, J., Allain F., Youcef R., Hartmann J.M., Vizioz C., Audoit G., Balestra, F., Andrieu, Comparative experimental study of junctionless and inversion-mode nanowire transistors for analog applications, 2020 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA), Hsinchu, Taiwan.
- Bosch, D., Colinge, Ghibaudo, G., C., Garros, X., Barraud S., Lacord, J., Sklenard B, Brunet, L., Batude, P., Fenouillet-Béranger, C., Cluzel, J., Kies, R., Hartmann J.M., Vizioz C., Audoit G., Balestra, F., Andrieu, All-operation-regime characterization and modeling of drain current variability in junctionless and inversion-mode FDSOI transistors, VLSI symposium.

As co-author:

- Fenouillet, C., Brunet, L., Batude, P., Brevard, L., Garros, X., Casse, M., Lugo, J., Mota-Frutososa, T., Lacord J., **Bosch, D.**, Bernard, N., Ribotta, M., Sklenard, B., Milesi, F., Magalhaes-Lucas, A., Kies, R., Romano, G., Acosta-Alba, P., Kerdiles, S., Tavernier, A., Vizzioz, C., Besson, P., Gassilloud, R., Kanyandekwe, J., Cooper, D., Lapras, V., Kim, W., Sasaki, Y., Oh, S., Kang, P., Lee, S., Na, H., Arcamone, J., Andrieu, F., 2020, First demonstration of low temperature ($\leq 500^{\circ}\text{C}$) CMOS devices featuring functional RO and SRAM bitcells toward 3D VLSI integration, VLSI.
- Ezzadeen, M., **Bosch, D.**, Giraud, B., Barraud, S., Noel, J.-P., Lattard, D., Lacord, J., Portal, J.M., Andrieu, F., Ultra-High-density 3D vertical RRAM with stacked JunctionLess nanowires for In-Memory-Computing applications, TED-Brief (accepted).

Award: 2019 VLSI-TSA Best student paper award

Simulation, fabrication and electrical characterization of advanced silicon MOS transistors for 3D-monolithic integration

Nowadays, Microelectronics industry must handle a real “data deluge” and a growing demand of added functionalities due to the new market sector of Internet Of Things, 5G but also Artificial Intelligence... At the same time, energy becomes a major issue and new computation paradigms emerge to break the traditional Von-Neumann architecture. In this context, this PhD manuscript explores both 3D monolithic integration and nano-electronic devices for In-Memory Computing. First, 3D monolithic integration is not seen only as an alternative to Moore’s law historic scaling but also to leverage circuit diversification. The advantages of this integration are analysed in depth and in particular an original top-tier Static Random Access Memories (SRAM) assist is proposed, improving significantly SRAM stability and performances without area overhead. In a second time, an original transistor architecture, called junctionless, suitable for 3D-monolithic integration is studied in detail. Devices are simulated, fabricated and electrically characterised for mixed digital/analog applications. In particular, the impact of channel doping density on mismatch is tackled. Also, low temperature (<500°C) junctionless bricks are developed and device optimization trade-off are discussed. In a third time, an innovative 3D structure combining state of the art devices: junctionless stacked Silicon nanowires and Resistive Random Access Memories (RRAM) is envisioned. This technology is proved to enable In-Memory Boolean operations through a so-called “scouting logic” approach.

Simulation, fabrication et caractérisation de transistors MOS avancés pour une intégration 3D monolithique

De nos jours, l’industrie microélectronique doit maîtriser un véritable « déluge de données » et une demande toujours en croissance de fonctionnalités ajoutées pour les nouveaux secteurs de marchés tels que la 5G, l’internet des objets, l’intelligence artificielle... Par ailleurs, l’énergie et sa gestion est un enjeu majeur au sein des architectures Von-Neumann traditionnelles. Dans ce cadre, ce travail de thèse explore l’intégration 3D monolithique ainsi que des dispositifs pour le calcul dans la mémoire. Premièrement, l’intégration 3D monolithique n’est pas perçue uniquement comme une alternative à la loi de Moore mais permet de diversifier les circuits. Les avantages de cette intégration sont analysés en détails et en particulier, une aide à la stabilité des mémoires SRAM (Static Random Access Memory) est proposée. Cette aide améliore significativement la stabilité ainsi que les performances des SRAM de l’étage supérieur, sans dégrader l’empreinte silicium. Secondement, des transistors sans jonctions (junctionless), compatibles avec une intégration 3D séquentielle sont étudiés. Les dispositifs sont simulés, fabriqués et caractérisés électriquement pour des applications digitales et analogiques. En particulier, l’impact du dopage canal sur la variabilité est analysée. Egalement des briques à basse température (<500°C) sont développées. Troisièmement, une structure 3D innovante combinant des transistors sans jonctions empilées et des mémoires résistives (RRAM).