Evaluation Of Nanosheet and Forksheet Width Modulation For Digital IC Design In The Sub-3nm Era

Giuliano Sisto, *Student Member, IEEE,* Odysseas Zografos, Bilal Chehab, Naveen Kakarla, Yang Xiang, *Member, IEEE,* Dragomir Milojevic, Pieter Weckx Geert Hellings, *Senior Member, IEEE,* and Julien Ryckaert

***Abstract*—In this paper we provide a comprehensive evaluation of width modulation capabilities of both nanosheet and forksheet devices, going from device level to a block level implementation. The main innovation introduced by the forksheet consists of a dielectric wall added between the P- and NMOS transistors. Leveraging this feature, forksheet shows approximately the same current behavior as nanosheet, considered as state-of-the-art reference, but reduced parasitic capacitance thanks to its fewer but wider stacked sheets. At block level, an area reduction up to 12% is observed with forksheet, alongside a 13% power reduction and 10% frequency increase. Following the device comparison, the potential of sheet width modulation as additional PPA optimization technique during synthesis and place and route is investigated. A description of the specific steps required to enable this knob in a conventional EDA framework is provided. As demonstrated by the obtained experimental results, the same frequency of the single-width implementation can be achieved using mixed libraries with lower power consumption (13% and 16% for nanosheet and forksheet respectively), leading to improved energy efficiency. Further, it is shown how designs implemented using forksheet benefit more from this type of optimization than the ones using nanosheet, with a 12%-15% energy reduction compared to the 8.5%-14% obtained with nanosheet.**

***Index Terms*—Forksheet, Advanced sub-3nm nodes design, CMOS scaling, Design-Technology Co-Optimization.**

I. Introduction

THE quest for Power, Performance, Area and Cost (PPAC) improvements has characterized the semiconductor industry for several years. This improvement is achieved with every new CMOS generation, thanks to smaller devices (Area improvement) allowing more transistor per unit silicon area (Cost improvement), while guaranteeing also higher drives for less power (Power and Performance improvement).

However, for advanced nodes (sub-10nm), the scaling trend has slowed down considerably due to growing process complexity [1], [2]. New methods to achieve area reduction are required to overcome the limits of traditional contacted gate pitch (CGP) and metal pitch (MP) downsizing. An efficient alternative is provided by Design-Technology Co-Optimization (DTCO) [3]–[7]. This approach aims at answering process concerns with enhancements at higher design levels, as shown already through examples like the Buried Power Rail (BPR) [8]–[10].

Similarly, the advent of FinFET devices not only allowed scaling below 22 nm [11], [12], but it also enabled cell area shrinking through fin de-population, since higher drive can be obtained also by increasing fin height. Therefore, as the number of fins per cell is reduced in order to save area, performance is maintained by thinner and taller fins [13].

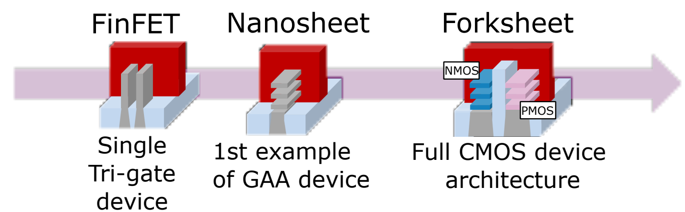


Fig. 1: Roadmap illustrating the device evolution after FinFET, going from the design of a single device to that of the complete CMOS architecture.

FinFETs limitations are highlighted by continued scaling, as desired performance is not achieved by devices with a single tall fin. While a solution, from the device perspective, is provided by Gate All Around (GAA) Nanosheet (NS) FETs [14]–[16], further cell height scaling will be limited by the PN separation. For this reason the novel forksheet (FS) device architecture is proposed [17]–[19]. In this case the space between the P and N devices can be considerably reduced, leveraging the addition of a dielectric wall in-between them. The benefits enabled by this new type of device propagate to the higher levels of design as well with new opportunities enabled at Standard Cell (SDC), circuit , and at physical design level [20]. The CMOS roadmap for advanced post-FinFET technology nodes is showed in Figure 1.

In this paper a comparison between the NS and FS is provided including both their device characteristics and their performance at block level, to highlight the greater benefits of the latter. Different cell libraries, distinguished by the channel width, are presented, showcasing various optimizations options for both devices. Finally, a methodology to perform width-modulation, *i.e.* combining cells with different sheet width in the same physical design, is introduced in an attempt to maximize the quality of PPA results. This work provides first proof that device width of sheet-based technologies can be leveraged as additional optimization parameter for the physical implementation of logic designs, akin to threshold voltage optimization. Furthermore, it also provides insights on how different types of advanced CMOS devices are impacted by this specific optimization technique.

© 2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. https://doi.org/10.1109/TVLSI.2022.3190080

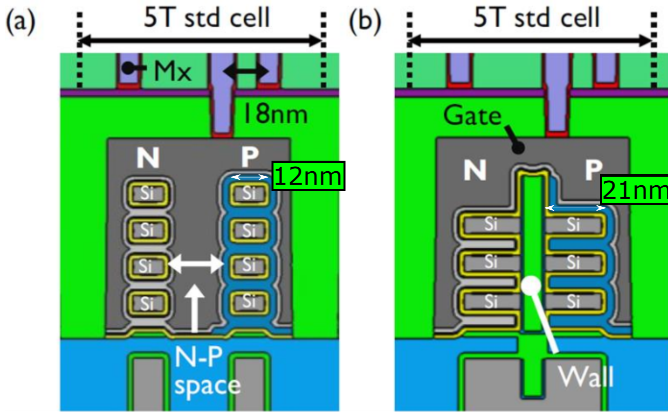


Fig. 2: Cross section comparison between the NS device (a) and FS (b) as shown in [19].

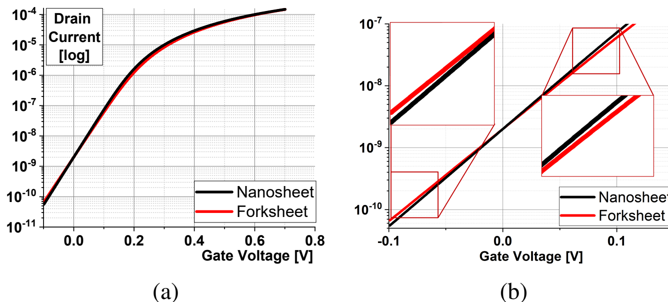
The manuscript is organized as follows. An overview of the device under analysis is provided in the second section. In the third section, the potential optimization opportunities at SDC level are explained. The impact of these performance boosters on a block level implementation is shown in the fourth section. Conclusions based upon the experimental simulation results are drawn in the fifth and last section.

II. Device Overview

An introduction of the state of the art devices for advanced technology nodes is provided in this section. This is instrumental to have a better understanding of the following block-level evaluation.

GAA NS devices have been accepted as main architecture to continue scaling beyond the reach of their competitors in view of the following three advantages over FinFET [21]–[23]. To begin with, the NS devices have superior gate electrostatic control thanks to their GAA architecture compared to the tri-gate FinFET. Second, NS can achieve better tradeoff between W_{eff}[1] and device parasitics at scaled dimension, *i.e.* SDC with less than two fins. Finally, unlike conventional FinFET, NS are not bound to the limitation of having a discrete number of fins, as the sheet width is only subject to the PN separation

[1] W_{eff} is obtained by summing all sheet edges that are contacted to the gate (4 for NS and 3 for FS for both the P- and NMOS device).



(a) (b)

Fig. 3: Id-Vg characteristics for both Nanosheet and Forksheet in Logarithmic units (a) and a highlight of the slope in sub-threshold regime (b).

Device Parameter	Unit	Nanosheet(NS)	Forksheet(FS)
Sheet width	[nm]	12	21
Sheet thickness	[nm]	5	5
Number of stacked sheets	[nm]	4	3
Effective width (W_{eff})	[nm]	136	141
Vertical sheet pitch	[nm]	15	15
Drawn gate length	[nm]	14	14
Spacer thickness	[nm]	5	5
Equivalent oxide thickness	[nm]	1.1	1.1
Contacted gate pitch	[nm]	42	42
Metal Pitch	[nm]	18	18
PN separation	[nm]	27	9

TABLE I: Compact model parameters for both NS and FS devices.

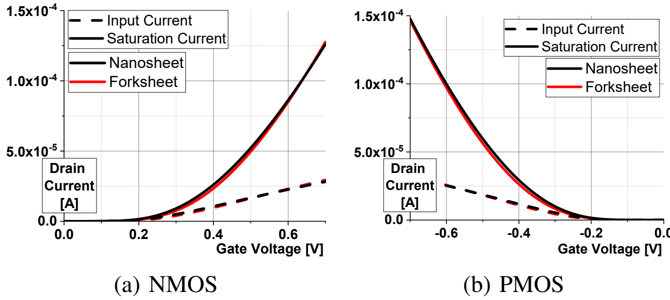
and can be thus varied in a continuous fashion. This enables W_{eff} maximization while also increasing freedom in terms of device drive strength tuning. This is particularly advantageous considering that, for very advanced nodes, only single fin devices can be eligible.

However, a structure with many stacked sheets suffers greatly from parasitic capacitance, growing proportionally to the number of layers and limiting the maximum achievable performance. Moreover, similarly to FinFET, the PN separation of NS is constrained by gate etching limitations. As a result, further cell area scaling in the sub-3 nm regime remains challenging to NS as well.

The FS device is introduced as the next stage in the device scaling roadmap, to overcome the aforementioned limitations of both FinFET and NS devices. It consists of vertically stacked lateral sheets controlled by a forked gate structure, containing a dielectric wall between the P- and NMOS (Figure 2). The insertion of the wall allows the gate edge to be self-aligned with the device, thus circumventing the overlay margin and allowing for patterning simplification. With respect to the traditional GAA NS, only minor updates to the process flow are required. In fact, the dielectric wall can be formed after patterning the Si/SiGe superlattice by spacer deposition and etch back. Therefore the forked gate represent an enticing way to overcome the scaling limitations of NS, with limited process complexity overhead.

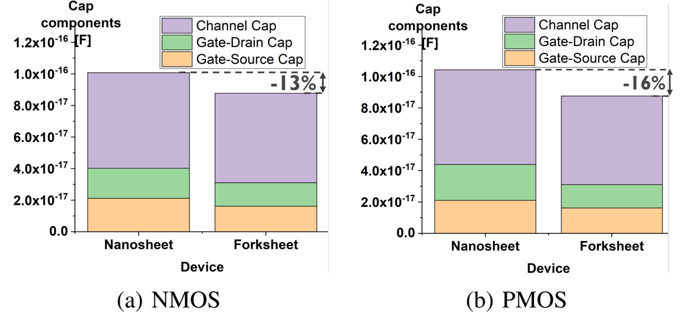
Predictive SPICE compact models are developed for both the NS and the FS devices. The models build upon TCAD-derived device physics (e.g., quantum effects, transport) at ultra-scaled dimensions and further incorporate full Front-End Of-Line (FEOL) device- and Middle-Of-Line (MOL) intra-cell parasitic resistance and capacitance based on finite-element full-field solutions [24], [25]. The essential model assumptions on the devices are listed in Table I. All devices are designed in such a way to have same quiescent current (I_{ddq}=$2nA$), at the nominal voltage supply of $0.7V$.

Despite being a non-GAA configuration, in terms of electrical properties, FS devices exhibit limited sub-threshold slope degradation relative to iso-width NS (Figure 3), thanks to the channel recess treatment that is assumed on the top sheet [17]. Nevertheless, a slightly lower I_{on} in FS arises from the 3-stacked sheets used instead of NS 4-stacked ones (Figure 4). The key advantage of having one fewer sheet is a significant gate-to-source/-drain parasitic capacitance reduction, resulting



(a) NMOS (b) PMOS

Fig. 4: Id-Vg characteristics for both Nanosheet and Forksheet showcasing slightly lower current in saturation region f



(a) NMOS (b) PMOS

Fig. 5: Device capacitance breakdown for both devices, showing both NMOS cap (a) and PMOS's (b)

in 13% less total capacitance compared to NS, as verified in Figure 5.

The discussed capacitance reduction and the potential for area scaling, both enabled by the PN separation reduction induced by the wall insertion, is where the main benefits from the FS architecture lie. With a PN separation as close as 9 nm, no coupling effects have been observed when using SiN as dielectric material for the wall ($k = 20$), but a wide variety of alternative dielectric materials could be possible for processing.

III. Width Modulation

This section provides an introductory overview of the different SDC variants available for both NS and FS. Furthermore, the concept of width modulation and the EDA methodology to enable it in physical design are introduced.

Width modulation represents a DTCO scaling booster, typical of sheet-based devices (*e.g.* NS and FS), relying on the possibility to vary geometrical device parameters (*i.e.* the sheet width) to optimize the circuit performance, as shown in [26] concerning the SRAM design. We aim at applying this technique to the physical implementation of a large CPU block, by combining high-drive (HD) strength and low-capacitance (LC) cells, with wide and narrow sheets respectively. In order to achieve this, a custom PDK, equipped with a set of ground rules and standard cell libraries, targeted for the 2 nm CMOS technology node is developed. Concerning the SDC library, it is composed of 80 cells in total, including Flip-Flops, repeaters (*i.e.* buffers and inverters), and gates for both simple (*i.e.* AND, OR, etc.), and more complex logic functions (*e.g.* And-Or-Inverter, Or-And-Inverter, etc.). Its technical details are summarized in Table II. Keeping all other parameters as detailed in Tables I and II, a total of four device variants and the corresponding cell libraries are created for this work changing only the sheet width. The various libraries used are listed below:

- 5TNS08 - NS with 8nm sheet width (LC);
- 5TNS12 - NS with 12nm sheet width (HD);
- 5TFS14 - FS with 14nm sheet width (LC);
- 5TFS21 - FS with 21nm sheet width (HD).

A W_{eff} reduction is seen in the LC cases, from $136nm$ to $104nm$ in NS, and from $141nm$ to $99nm$ in FS, as a natural consequence of the sheets shrinking.

The main constraint for the sheet width variation is represented by the size of the PN separation, determined by lithography limits and usually constant for a specific device. While multiple widths can be possible for both NS and FS, the dielectric wall inserted in the latter can be leveraged to greatly increase the channel flexibility in the cell design. As a result, with the PN separation scaled to a minimum (half metal pitch), additional space is gained in the cell to enable much wider sheets than in the NS, for both LC and HD cells. The absence of the forked gate structure limits therefore the freedom concerning the size of the channel, due to NS process and cell design constraints.

A comparison showing a sample cell layout for the 4 different device flavours is shown in Figure 6. All NS and FS SDC layouts are 5T height cell with equal BPR width and via to ensure a realistic comparison. A key distinction is represented by the absence of the gate cut in the NS (a,b) due to tight PN separation requirements, resulting in an area penalty in all SDCs that include transmission gate sub-circuits (e.g. mux, flops, etc.). Additionally, given the large space between the devices, the M1 layer (vertical) is used to perform these north-south connections, as directly contacting the gate would cause a steep increase in coupling capacitance. The presence of the dielectric wall in the FS (c,d), not only allows for the PN separation to be vastly reduced increasing active area, but also allows for self-aligned gate-cut

PVT Corner			
Process Corner	Typical-Typical		
Voltage Supply	0.7V		
Temperature	25C		
BEOL			
13 metal layers (6Mx, 7Mx)			
Metal layer	Pitch [nm]	Width [nm]	Spacing [nm]
M0 & M2	18	9	9
M1 to M5	28	14	14
M6 to M12	80	40	40
Standard Cell Library			
# Cells	80		
# Flip-Flops	6	DFF and Scan-DFF - with Set/Reset options	
# Buffers	5	Drive Strengths: D1, D2, D4, D8, and D16	
# Inverters	5	Drive Strengths: D1, D2, D4, D5, and D8	
# Simple Logic	15	Drive Strengths: D1, D2, and D4	
# Complex Logic	24	Drive Strengths: D1 and D2	

TABLE II: imec $iN3$ (2 nm) PDK details

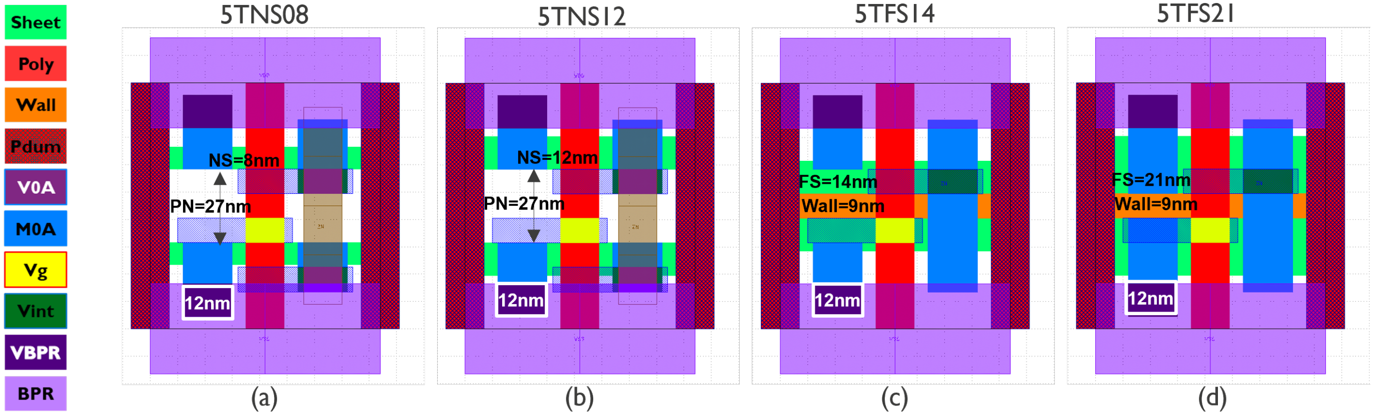


Fig. 6: Overview of all SDC layouts available for both devices. An increased sheet width is shown in (b) and (d), which are in fact the high-performance cells for NS and FS respectively. (a) and (c) are instead optimized towards low power consumption.

and independent gate pick-up. Also, north-south contact can be formed using only MOL metals. This contributes in providing an area reduction in cells with cross-coupled devices, as well as in freeing intra-cell routing resources in the M1 layer [17]. An overview of the I-V curves for the four cell flavours is shown in Figure 7, highlighting the current reduction $\sim 5\%$ induced by the smaller sheets.

The various libraries can all be used individually for the physical implementation of any digital IC. Each library relies, in fact, on its own layout abstract, provided to the different tools in the LEF file default format, and power/timing characterization, read by the tools also as a conventional liberty (LIB) file. However, the main goal of this paper is that of combining both types of cells in the same physical design implementation, so that each can help mitigating the downside of the other at block level, resulting into overall better PPA results. This process, also referred to as w-mixing, can be applied equally to NS and FS, even though larger benefits are expected from the FS as explained prior. This can be enabled in

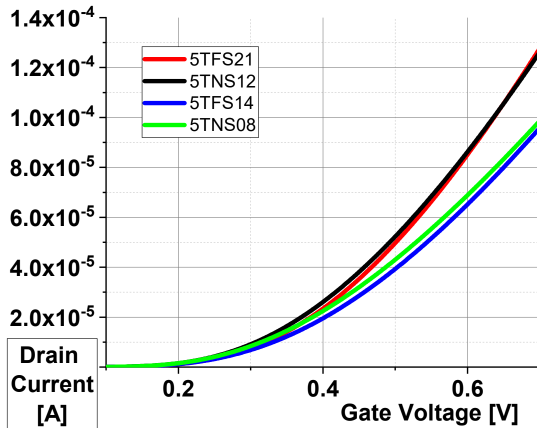


Fig. 7: Id-Vg characteristics for all the different devices, highlighting the higher saturation current obtained with wider sheets. All 4 devices are targeting same leakage current of $2nA$

a conventional PNR flow by combining the contents of the LIB and LEF files belonging to the HD cells with their counterparts from LC cells, obtaining a set of files containing gates with mixed sheet width. Since the cell names are normally the same between different libraries, in the w-mixing LIB and LEF the different cells are differentiated with a suffix according to their width size, $i.e.$ cells such as BUFFLVTD1_14W and BUFFLVTD1_21W will coexist in the same file. Two mixed libraries, namely $in3_5tnswmix$ for NS, and $in3_5tfswmix$ for FS, are developed and used as part of this study.

Leveraging the naming differences adopted in the w-mixing libraries to distinguish between LC and HD cells, they are added to different optimization groups, in order for the tools to fully understand them and use them for optimization accordingly to their optimized properties. This is made possible, for both physical synthesis and Place and Route (PNR), through a specific feature developed for the Innovus® tool. The various cells are assigned to the corresponding groups during physical synthesis, so that the output netlist is already optimized with both cell flavours. For the design back-end, during placement the tool can choose between the two kinds of cells, since the groups are defined among the pre-placement options. These specific settings are retained throughout the rest of the implementation flow, therefore even during post-route optimization, both LC and HD cells can be added ($e.g.$ for buffer insertion or hold fixing). An overview of the entire PNR flow used for width modulation is showed in Figure 8, calling attention to the steps where the w-mixing settings are added to the process.

IV. Experimental Evaluation

An extensive analysis of the proposed devices as well their corresponding width modulation is presented in this section.

Concerning the latter, the analysis is performed both at Ring Oscillator (RO) and block level to obtain a complete understanding of the mixing procedure. As a reference design for block level, we utilize the logic part of a 64-bit ARM® CPU ($\sim 500K$ cells with no SRAM caches) to guarantee a relevant BEOL load for the technology. All the different

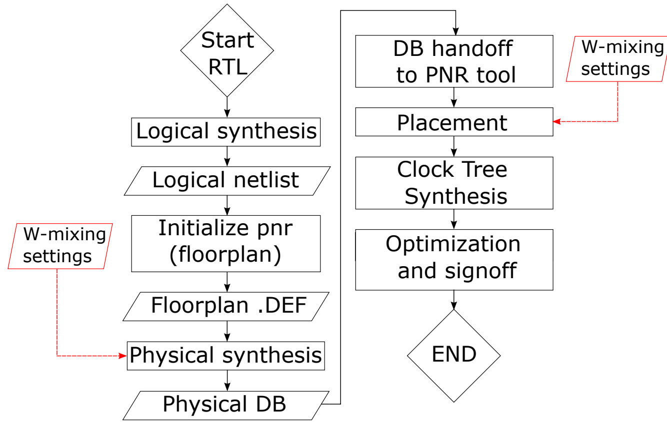


Fig. 8: Complete flow description from RTL to signoff optimization. All the steps belong to the standard Digital IC design and signoff flow, as the w-mixing features are fully integrated with the conventional algorithms.

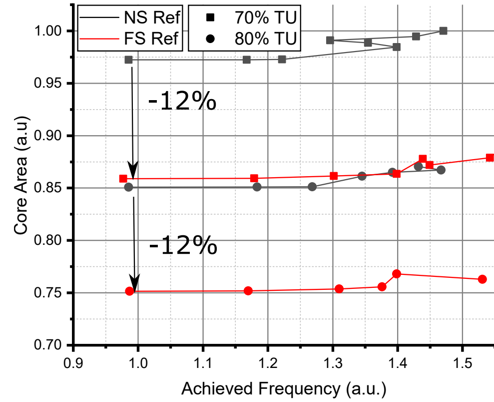


Fig. 9: Core area comparison between FS and NS. A fixed offset is present in both cases between the two different target utilization used.

libraries have been developed as part of imec 2nm PDK (iN3), using 42 nm CGP and 18 nm MP. The BPR is assumed as default constituent of all iN3 libraries, as it is required to obtain 5-Tracks tall SDC (90 nm cell height). All block-level results generated based on the 64-bit ARM® CPU are reported in relative (arbitrary) units honouring the confidentiality agreement between imec and ARM®, based on which the block design was shared.

A. Device comparison

In this subsection, a PPA comparison at block level is carried out between the FS and NS reference libraries, corresponding to the HD ones (*i.e.* 5TFS21 and 5TNS12). A frequency sweep is performed for both technologies, to identify their maximum achievable frequency, in different design utilization scenarios, namely 70% and 80% target. The design area is tightly coupled with the cell area by the formula $A_{design} = \#Cells \cdot A_{cell}$. In this case both libraries share same SDC height and width, however, a key distinction is represented by the absence of the gate cut in the NS due the tight PN separation requirements. Such difference results in area penalty in all SDCs that include transmission gate sub-circuts (*e.g.* Mux, flops, etc.). On the other hand, in the FS, the dielectric wall behaves as a physical separation between the N and the P gate, which can compensate the area penalty seen in the NS as well as enabling additional intra-cell routing resources in the M1 layer. As a consequence of these improvements, a 12% lower area is observed with the FS compared to the NS after PNR (Figure 9).

As previously explained in the device section, roughly the same I_{eff} (slightly lower for FS) is exhibited by both devices, but with a smaller amount of stacked sheets in the FS (3 as opposed to the 4 of the NS). This key difference between the devices translates at block level into a 13% lower power with FS, justified by their lower parasitic capacitance compared to NS. The comparison is performed using the total design power, composed by (i) internal power, (ii) switching power, and (iii) leakage power. The first two contribute respectively to $\sim 53\%$ and $\sim 46\%$ of the total power. The specific impact of leakage, whose contribution is $< 1\%$, can therefore be neglected. In addition to the better power consumption in iso-frequency condition, the combination of improved parasitic and cell layout allows the FS to also reach a 11% higher maximum frequency. Both power and frequency results are shown in Figure 10.

Lastly, to combine the improvements just described, an evaluation of the two technologies in the Energy-Performance design space is performed. From the various data points reported in Figure 11, two pareto curves are plotted for each device. Comparing points in iso-power conditions, a $9-14\%$ energy reduction is observed with FS, alongside a $10-20\%$ frequency increase.

B. W-mixing

1) Ring Oscillator level: A RO study is performed in order to model the critical path behavior of NS and FS under width modulation, comparing it with the HD-only case. The analyzed circuit is built with the HD cells (5TFS21 and 5TNS12 for FS and NS respectively), while the additional fanout cells are modeled as LC cells (5FS14 and 5TNS08 for FS and NS

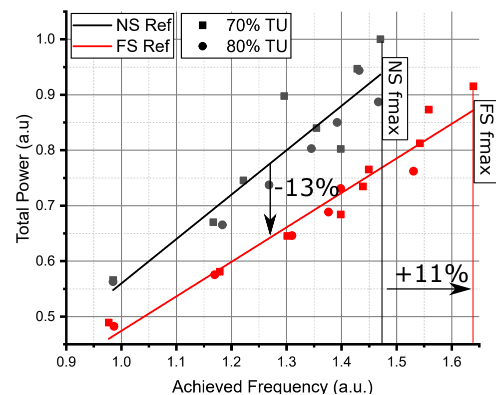


Fig. 10: Power comparison between FS and NS. The maximum frequency reached by both devices is also highlighted.

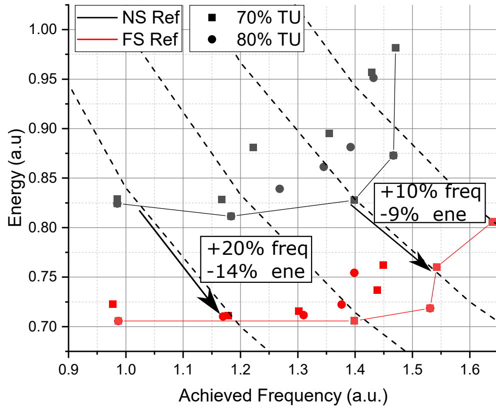


Fig. 11: FS and NS comparison in the Energy-Performance design space. Iso-power dotted lines have been drawn to compare points from the pareto curves corresponding to the devices.

respectively). This RO configuration represents the scenario where all the cells in the critical path fanout to LC cells in the non-critical paths. For the reference simulation, all the cells (critical and fanout) are kept as HD. The entire RO is comprised of 19 stages, and the circuit schematic of a single stage is shown in Figure 12.

As cell under test a HD INVD1 cells with a fanout of 3 is used. The interconnect load is modeled as a generic metal with a resistance of $400\ \Omega/\mu m$ and capacitance of $200\ aF/\mu m$ which is the projection for scaled nodes. 3-pi RC network is employed to model the metal resistance and capacitance. As observed in Figure 13, when considering the interconnect loading, performance benefits induced by width modulation result very small. For a total metal length of $15\ \mu m$ ($5\ \mu m$ per fanout leg), the performance gain corresponds to 1% and 0.8% for FS and NS respectively. Even the pin capacitance difference becomes increasingly less relevant as the metal length increases. As a means of comparison, in an ideal case, assuming no BEOL load, w-mixing provides a 12% and 10% performance increase due to the 8% and 7% reduction in pin capacitance for FS and NS respectively. This demonstrates that width modulation achieves very similar performance to that of HD libraries under majority of design and BEOL load conditions.

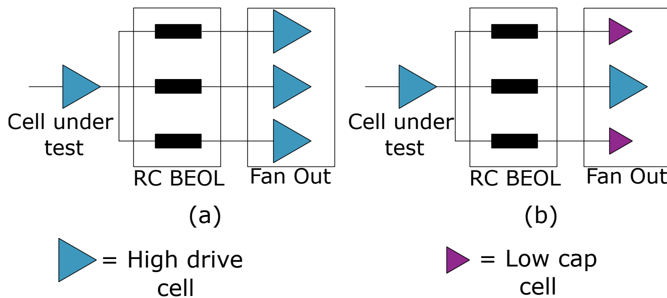
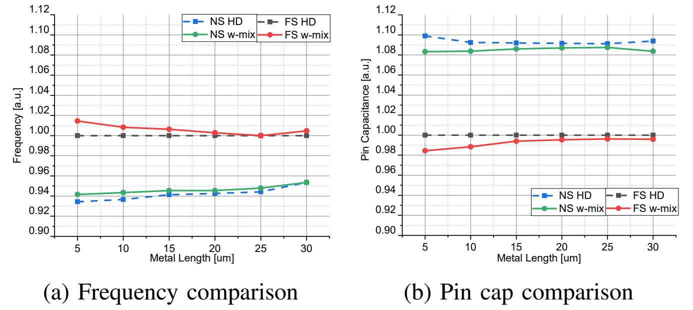


Fig. 12: Schematic of a single stage of the RO circuit used to simulate the reference case with all HD cells (a), and the width modulation case with LC cells in the fan-out legs (b).



(a) Frequency comparison

(b) Pin cap comparison

Fig. 13: Variation of achieved frequency (a) and total pin capacitance (b) with respect to the total metal length, as resulting from RO circuit simulation. All data points corresponding to different metal lengths are normalized with respect to the FS HD results.

2) Block level: The previously described design methodology for width-modulation has been integrated in the same PNR flow used for the analysis of the standalone libraries. The updated framework is employed for the evaluation of the two mixed libraries mentioned prior ($i.e. in3_5tnswmix$ and $in3_5tfswmix$). The same frequency sweep as seen in the device comparison section is performed, pairing the reference libraries for LC and HD for each technology with their w-mixing counterpart. All PNR runs are performed in single-mode-single-corner analysis mode, using the same voltage supply of 0.7 V.

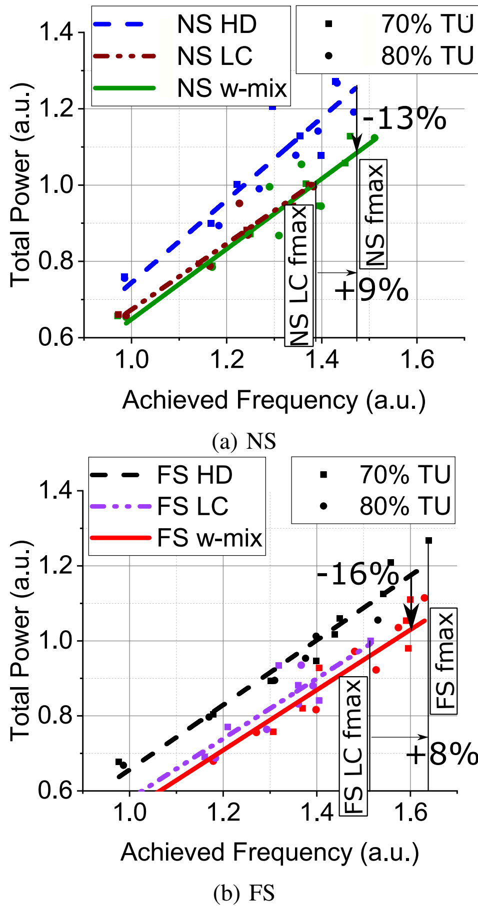
For each w-mixing run, the two optimization groups corresponding to LC cells (5TFS14 and 5TNS08 for FS and NS respectively) and HD cells (5FS21 and 5TNS12 for FS and NS respectively) are defined both during synthesis and PNR. Consequently, the cell distribution for both parts of the design flow captures the same trend in terms of relative usage of HD cells, in correlation with the target frequency. As visible in Table III, the HD-to-LC cell ratio increases proportionally to the design target frequency, showing a good understanding of the two libraries by the synthesis and PNR tools. Furthermore, it is interesting to notice how even for very high target frequency values, the cell ratio is still in favor of LC cells. This is confirmed by the relative percentage of HD cells in the maximum frequency points, which reaches approximately 15% for FS, after PNR, and 19% for NS.

Having a vast majority of LC cells in the design, even in high-frequency conditions, translates into a reduction of the total block power (Figure 14). A 16% lower power is observed for FS as opposed to 13% for NS, driven by the superior width modulation capability of the first. More or less the same power is consumed by the w-mixing and LC libraries, with a slightly higher average value in the latter (1% and 3% for NS and FS respectively).

Even though power is reduced thanks to w-mixing, the design performance is not affected by the process, as both technologies can achieve the same maximum frequency values. The design is instead always sped-up by $8-9\%$ compared to the LC implementation. This is in line with expectations, and it is motivated by the fact that in all analyzed cases the critical paths of the design consist only of HD cells. In fact,

	HD to LC Cell Ratio	Target Frequency [a.u.]									
Nanosheet											
		1.00	1.20	1.33	1.43	1.50	1.58	1.67	1.76	1.88	2.00
Physical Synthesis	70% TU	0.01	0.03	0.04	0.06	0.07	0.09	0.11	0.13	0.14	0.18
	80% TU	0.01	0.02	0.03	0.06	0.06	0.08	0.10	0.11	0.13	0.17
Physical Implement.	70% TU	0.01	0.02	0.04	0.05	0.07	0.08	0.09	0.13	0.19	0.25
	80% TU	0.01	0.02	0.03	0.04	0.06	0.07	0.09	0.12	0.16	0.22
Forksheet											
	HD to LC Cell Ratio	Target Frequency [a.u.]									
		1.00	1.20	1.33	1.43	1.50	1.58	1.67	1.76	1.88	2.00
Physical Synthesis	70% TU	0.01	0.03	0.05	0.06	0.08	0.09	0.11	0.12	0.15	0.19
	80% TU	0.01	0.02	0.04	0.05	0.07	0.08	0.10	0.11	0.14	0.18
Physical Implement.	70% TU	0.01	0.03	0.04	0.06	0.07	0.09	0.11	0.13	0.14	0.18
	80% TU	0.01	0.02	0.03	0.05	0.05	0.06	0.09	0.10	0.14	0.17

TABLE III: Ratio between HD and LC cells for different target frequencies and utilization, for both NS and FS after physical synthesis and after PNR. While the same trend is observed with both devices, a better correlation between synthesis and PNR is showed by th FS. For NS on the other hand, for high target values, the ratio registered at PNR exceeds the corresponding synthesis value, due to the buffer insertion occurring in the Clock Tree Synthesis (CTS) and hold fixing steps.



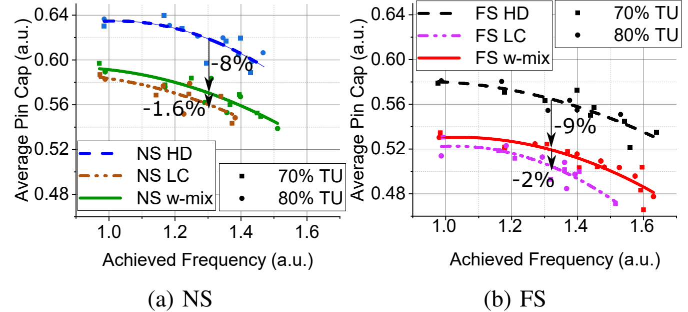
(a) NS

(b) FS

Fig. 14: Total power variation for NS (a) and FS (b) with respect to the achieved frequency, effectively capturing both power and performance behavior.

all different implementations are relying on HD cells being employed in the data paths for which the time is most difficult to meet. Since the overall achieved frequency is determined by these specific paths, no difference in performance is registered with w-mixing compared to the HD standalone implementation after signoff timing optimization.

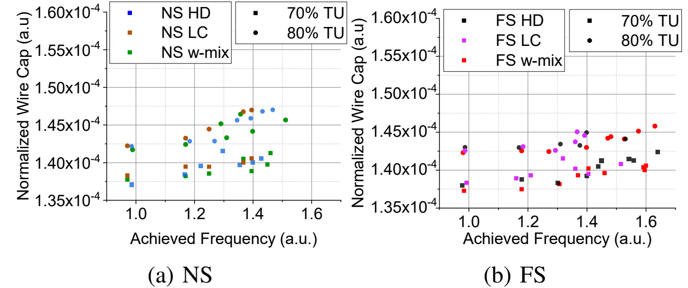
On the other hand, the presence of LC cells in all the non-timing critical paths allows the mixed libraries to greatly reduce their power consumption. This benefit is mainly induced by the lower capacitance provided by the LC cells. To verify this, a capacitance breakdown of the entire design including both pin and wire capacitance is extracted for both FS and NS. The results are shown in Figure 15 concerning the pin capacitance, while wire capacitance is illustrated in Figure 16. The expected capacitance behavior is experimentally con-



(a) NS

(b) FS

Fig. 15: Average pin capacitance for NS (a) and FS (b) for different values of achieved frequency. Values are normalized with respect to the number of instances in the design.



(a) NS

(b) FS

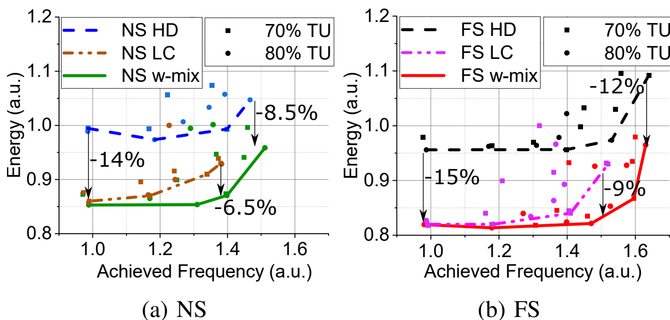
Fig. 16: Wire capacitance results for NS (a) and FS (b) with respect to the achieved frequency. To capture solely the impact of the cap values, all the values are normalized to the total wirelength of the corresponding physical design implementation.

firmed, since the w-mixing pin cap falls in between the other two runs, but closer to the LC (2%), reflecting the majority of this type of cells in the mixed implementation. However lower pin cap values compared with the HD implementation, 8% and 9% for NS and FS respectively, are observed, thus justifying the power reduction. Contrarily, the wire capacitance is only affected by the target utilization and it does not contribute to the power improvement, since roughly the same values are achieved regardless of the width mixing.

The width modulation capability to reduce the total power of the design with no penalty on frequency ultimately entails a better energy efficiency when compared to any of the non-mixed libraries. This is demonstrated by the results in Figure 17. Unlike FS, for which the energy benefits are retained across the entire frequency spectrum, ranging from 15% to 12%, for NS they are considerably reduced, from 14% down to 8.5%, in a high-frequency scenario. This behavior is aligned with the cell distribution trend showed in Table III and it is driven by the incapability of the NS device to reach high frequency targets. To try and overcome this limitation and close timing, an increasingly larger amount of HD cells are added by the PNR tool during post-route optimization, resulting into both the HD-to-LC cell ratio and the energy pareto curve shifting closer to the results of the HD reference implementation.

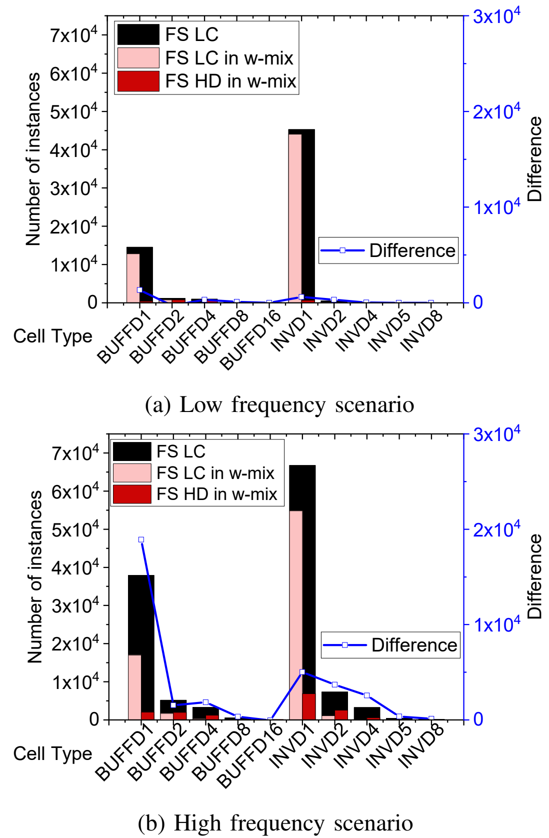
Another effect of this behavior is seen when comparing the mixing results with the LC pareto curve. Compared to the corresponding w-mixing implementations, a larger amount of buffers and inverters is added for both FS14 and NS12 during PNR for high frequency targets. The graphs from Figure 18 are showing the different repeater insertion in low-frequency and high-frequency condition. In the first case, the two implementations show almost the same amount of instances, as shown by the difference line whose peak number is 1340 (BUFFLVTD1) and average is 247. In the scenario with high frequency target, the design using only LC cells has a much larger amount of repeaters, with the difference peaking at 18903 (BUFFLVTD1) and averaging at 3434. The ratio between the total repeater power in the LC-only and w-mixing runs goes from 2%, in the low-frequency case, to 11% in high-frequency, explaining the energy increase with LC cells previously showed in Figure 17.

To provide a comprehensive overview of the entire 2 nm



(a) NS (b) FS

Fig. 17: Assessment of the w-mixing in the Energy-Performance design space for both NS (a) and FS (b).



(a) Low frequency scenario

(b) High frequency scenario

Fig. 18: Breakdown of the different repeaters between sample FS LC and FS-wmixing runs, one for low frequency target (a) and the other for high-frequencies (b). Same behavior is observed for NS as well.

design space and how width-modulation can play an important role in optimizing PPA results for the physical implementation using sheet-based devices, a complete comparison of all different NS and FS libraries is provided in Figure 19. Considering the FS w-mix library in iso-power comparison with the HD reference, the energy consumption is reduced by 11% and 22%

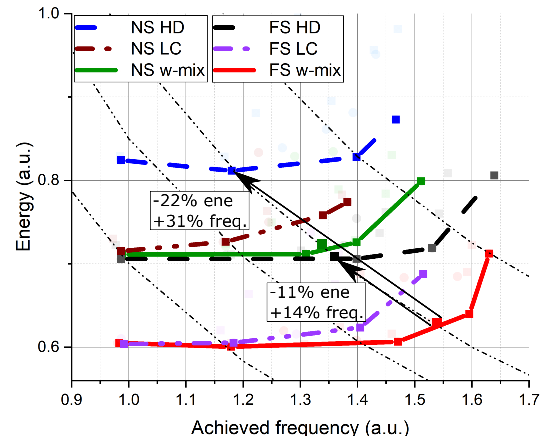


Fig. 19: Complete overview of all different libraries in the design space. Iso-power curves have been used again to compare specific points in this condition.

with respect to the FS and NS respectively. In the same cases, frequency is also improved by 14% compared to FS and 31% compared to NS. Moreover, from the device perspective, NS is brought considerably closer to the FS reference, in terms of energy, thanks to the width modulation, although it continues to show worse performance.

V. Conclusion

A block-level examination of sheet-based devices, $i.e.$ NS and FS, is provided as part of this work, alongside the proposal and evaluation of width modulation as a novel PPA optimization technique. From the perspective of the device architecture, it is shown how FS exhibits lower capacitance values than its counterpart, without sacrificing drive current. The benefits from the improved parasitic are retained at block level, where a 13% power reduction is achieved. In addition to power, area (-12%), frequency ($+11\%$), and, consequently, energy (up to 14%) are also improved with FS thanks to enhancements at cell level, making this device an ideal candidate to overcome the NS limitations.

The flexibility provided to the SDC design represents the key benefit of the FS CMOS architecture. We focus on this aspect to investigate the possibility of further PPA results optimization through the sheet width modulation process. From the RO circuit simulation, the potential performance improvement appears to be strongly limited by the BEOL load.

The block level results, obtained with the developed methodology to mix different sheet widths at PNR, despite confirming no impact on the design achieved frequency, show a significant power and energy reduction in the w-mixing implementations, consequence of the improved global capacitance. Power is reduced on average by 13% and 16% on NS and FS respectively. Similarly, the energy consumption is scaled down up to 14% (NS) and 15% (FS).

This work is the first relying on a full PDK ($i.e.$ device compact models, SDC libraries, etc.) to provide a block-level comparison of a system physically implemented using NS and FS. Additionally, a novel w-mixing approach is proposed to improve the physical design PPA. The obtained results demonstrate that width modulation plays a fundamental role in optimizing power and energy consumption at block level, with no complexity overhead from the technology perspective. Furthermore, while these benefits are shared by both NS and FS, they are more pronounced in the latter as a consequence of their increased freedom in terms of sheet width. All the observed results contribute to strengthen FS position as flagship device for IC design at sub-3nm nodes.

VI. Acknowledgments

The authors would like to acknowledge Gioele Mirabelli and Pieter Schuddinck for their support throughout the revision process.

References

[1] M. T. Bohr and I. A. Young, "Cmos scaling trends and beyond," *IEEE Micro*, vol. 37, no. 6, pp. 20–29, 2017.

[2] S. B. Samavedam *et al.*, "Future logic scaling: Towards atomic channels and deconstructed chips," in *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020.

[3] M. G. Bardon *et al.*, "Extreme scaling enabled by 5 tracks cells: Holistic design-device co-optimization for finfets and lateral nanowires," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016, pp. 28.2.1–28.2.4.

[4] L. W. Liebmann and R. O. Topaloglu, "Design and technology co-optimization near single-digit nodes," in *2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2014.

[5] X. Wang *et al.*, "Design-technology co-optimization of standard cell libraries on intel 10nm process," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018.

[6] L. W. Liebmann *et al.*, "Overcoming scaling barriers through design technology cooptimization," in *2016 IEEE Symposium on VLSI Technology*, 2016.

[7] A. Mocuta *et al.*, "Enabling cmos scaling towards 3nm and beyond," in *2018 IEEE Symposium on VLSI Technology*, 2018.

[8] 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)*, 2019, pp. 19.1.1–19.1.4.

[9] A. Gupta *et al.*, "High-aspect-ratio ruthenium lines for buried power rail," in *2018 IEEE International Interconnect Technology Conference (IITC)*, 2018, pp. 4–6.

[10] ——, "Buried power rail integration with finfets for ultimate cmos scaling," *IEEE Transactions on Electron Devices*, 2020.

[11] H.-J. Lee *et al.*, "Intel 22nm finfet (22ffl) process technology for rf and mm wave applications and circuit design optimization for finfet technology," in *2018 IEEE International Electron Devices Meeting (IEDM)*, 2018.

[12] 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)*, 2012.

[13] J. Ryckaert *et al.*, "Enabling sub-5nm cmos technology scaling thinner and taller!" in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019.

[14] N. Loubet *et al.*, "Stacked nanosheet gate-all-around transistor to enable scaling beyond finfet," in *2017 Symposium on VLSI Technology*, 2017.

[15] A. Veloso *et al.*, "Nanowire nanosheet fets for ultra-scaled, high-density logic and memory applications," in *EUROSOI-ULIS*, 2019.

[16] H. Mertens *et al.*, "Vertically stacked gate-all-around si nanowire cmos transistors with dual work function metal gates," in *2016 IEEE International Electron Devices Meeting (IEDM)*, 2016.

[17] P. Weckx *et al.*, "Novel forksheet device architecture as ultimate logic scaling device towards 2nm," in *2019 IEEE International Electron Devices Meeting (IEDM)*, 2019.

[18] ——, "Stacked nanosheet fork architecture for sram design and device co-optimization toward 3nm," in *2017 IEEE International Electron Devices Meeting (IEDM)*, 2017.

[19] H. Mertens *et al.*, "Forksheet fets for advanced cmos scaling: Forksheet-nanosheet co-integration and dual work function metal gates at 17nm n-p space," in *2021 Symposium on VLSI Technology*, 2021, pp. 1–2.

[20] J. Ryckaert, B. Chehab, D. Jang, G. Mirabelli, S. Salahuddin, P. Schuddinck, O. Zografos, A. Zubair, P. Weckx, and G. Hellings, "From design to system-technology optimization for cmos," in *2021 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)*, 2021, pp. 1–2.

[21] L. Gaben *et al.*, "Stacked-nanowire and finfet transistors: Guideline for the 7 nm node,," in *2015 IEEE International Conference on Solid State Devices and Materials*, 2015.

[22] S.-D. Kim *et al.*, "Performance trade-offs in finfet and gate-all-around device architectures for 7nm-node and beyond," in *2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, 2015.

[23] S. Dey *et al.*, "Performance and opportunities of gate-all-around vertically-stacked nanowire transistors at 3nm technology nodes," in *2019 Devices for Integrated Circuit (DevIC)*, 2019.

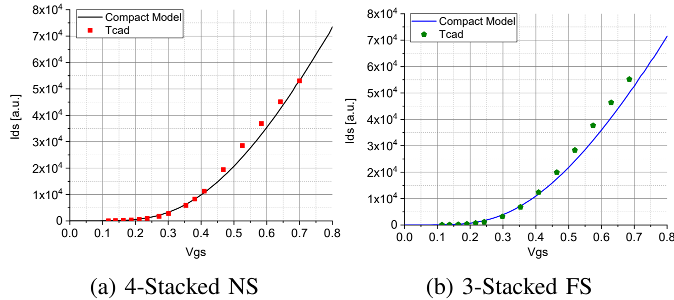
[24] D. Jang *et al.*, "Device exploration of nanosheet transistors for sub-7-nm technology node," *IEEE Transactions on Electron Devices*, vol. 64, no. 6, pp. 2707–2713, 2017.

[25] M. G. Bardon *et al.*, "Power-performance trade-offs for lateral nanosheets on ultra-scaled standard cells," in *2018 IEEE Symposium on VLSI Technology*, 2018, pp. 143–144.

[26] M. K. Gupta *et al.*, "A comprehensive study of nanosheet and forksheet sram for beyond n5 node," *IEEE Transactions on Electron Devices*, vol. 68, no. 8, pp. 3819–3825, 2021.

Appendix A
Compact model and TCAD alignment

As mentioned in the paper, the device compact models, on which the PDKs used in this work are based, are built upon TCAD-derived device physics. This means that a variety of device physics features, *e.g.* ballisticity, electrostatics, and stress effects, are captured by them. The alignment between the models and the TCAD simulations is therefore verified to ensure the validity of the former. Figure 20 is showing the results from this study for both 4-stacked NS and 3-stacked FS NMOS device, displaying an overall good agreement between the two curves.

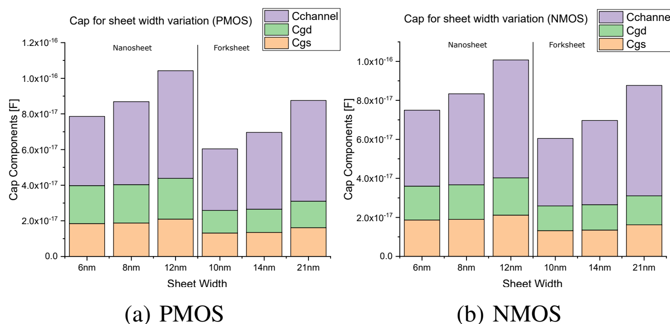


(a) 4-Stacked NS (b) 3-Stacked FS

Fig. 20: Alignment between the compact model and TCAD simulations for NS (a) and FS (b).

Appendix B
Capacitance scaling with sheet width

The width modulation methodology proposed in this work relies on the assumption that the device parasitic capacitance scales with its sheet width. In order for this optimization method to remain relevant for future configurations of sheet-based devices, it is important to verify the linear dependency between the sheet width and the total capacitance. To evaluate this, two additional configurations, namely NS 6nm and FS 10nm, is analyzed and added to the existing ones (NS 8nm and 12nm, and FS 14nm and 21nm). The capacitance results shown in Figure 21 exhibits a mostly linear trend, thus confirming the applicability of the w-mixing techniques also for devices with further scaled sheet width.



(a) PMOS (b) NMOS

Fig. 21: PMOS (a) and NMOS (b) device capacitance variation for different values of sheet width.

Giuliano Sisto received his B.Sc. and M.Sc. from Politecnico di Bari. Since 2018 he is pursuing a PhD degree with Cadence Design Systems, in collaboration with imec. Since 2021 he joined imec as a researcher working on EDA enablement of advanced CMOS technology nodes, and Design and System-Technology Co-Optimization techniques for SoC design.

Odysseas Zografos received his M.Sc. degree in Micro and Nanotechnologies for Integrated Systems from Institut National Polytechnique de Grenoble. He received his PhD degree on Electrical Engineering from KU Leuven. Since 2018, he has joined imec as an R&D engineer focusing on block-level benchmarking of advanced CMOS technology options.

Bilal Chehab received the B.Sc and M.Sc. degree from the University of Pavia-Italy. He joined IMEC in 2014, where he is currently a researcher for CMOS technologies. His main focus is on design and technology co-optimization in advanced node, for Power, Performance, and Area (PPA) scaling enablement.

Naveen Kakarla received his B.Sc. from Rashtreeya Vidyalaya College of engineering, Bangalore and M.S. from Carnegie Mellon University. His experience includes standard cell design and architecture, low power design and implementation and Design-Technology Co-optimization in advanced nodes.

Yang Xiang received his M.Sc. degree in Nanoscience and Nanotechnology (Erasmus Mundus) from KU Leuven, Belgium and Chalmers University of Technology, Sweden. Currently he is pursuing a PhD degree in Electrical Engineering at KU Leuven and imec, Belgium. His research interests include compact modeling of novel logic and memory devices.

Dragomir Milojevic Dragomir Milojevic received his M.S. and PhD degrees in Electrical Engineering from Université libre de Bruxelles (ULB), Belgium, where he holds the position of professor of digital electronics and systems design. In 2004 he joined IMEC working on multi-processor and NOC architectures for low-power multimedia systems. Since 2008 he is working on enablement of 3D stacked ICs, as well as system and design technology co-optimization of advanced technology nodes, and design methodologies for technology aware 3D ICs.

Pieter Weckx received the B.Sc degree in Electronic Engineering, M.Sc. degree in Nanoscience and –technology and Ph.D degree in Engineering from the Katholieke Universiteit Leuven - Belgium, in 2009, 2011 and 2016 respectively. In 2015 he joined imec where he is currently R&D Manager leading the research group for future scaled CMOS technologies, spanning from device to components for systems.

Geert Hellings received the B.S. and M.S. degrees in Electrical Engineering from the KU Leuven, Belgium, in 2007. He obtained a Ph.D. degree at the Electrical Engineering Department (ESAT), Integrated Systems Division (INSYS) of the KU Leuven, and the CMOS Technology Department at imec, Belgium in 2012. In 2011, he joined the Device Reliability and Electrical Characterization group in imec, researching Electrostatic Discharge events in integrated circuitry. Since 2020, he is managing the imec Logic INSITE Research Program.

Julien Ryckaert Julien Ryckaert received the M.Sc. degree in electrical engineering from the University of Brussels (ULB), Belgium, in 2000 and the PhD degree from the Vrije Universiteit Brussel (VUB) in 2007. He joined imec as a mixed-signal designer in 2000 specializing in RF transceivers, ultra-low power circuit techniques and analog-to-digital converters. In 2010, he joined the process technology division in charge of design enablement for 3DIC technology. Since 2013, he is in charge of imec's design- technology co-optimization (DTCO) platform for advanced CMOS technology nodes. In 2018, he became program director focusing on scaling beyond the 3nm technology node as well as the 3D scaling extensions of CMOS. Today, he is vice president logic in charge of compute scaling.