Design enablement of CFET devices for sub-2nm CMOS nodes

Odysseas Zografos, Bilal Chehab, Pieter Schuddinck, Gioele Mirabelli, Naveen Kakarla, Yang Xiang, Pieter Weckx, Julien Ryckaert
imec, Kapeldreef 75, 3001 Leuven, Belgium
email: Odysseas.Zografos@imec.be

***Abstract*— Novel devices that optimize their structure in a three-dimensional fashion and offer significant area gains by reducing standard cell track height are adopted to scale silicon technologies beyond the 5nm node. Such a device is the Complementary FET (CFET), which consists of an n-type channel stacked vertically over a p-type channel. In this paper we review the significant benefits of CFET devices as well as the challenges that arise with their use. More specifically, we focus on the standard cell design challenges as well as the physical implementation ones. We show that to fully exploit the area benefits of the CFET devices, one must carefully select the metal stack used for the physical implementation of a large design.**

I. Introduction

The FinFET has been a remarkably efficient solution for the silicon industry, enabling the scale down from planar transistors to 5nm technology node. The transition to FinFETs has essentially been driven by mitigating short-channel control using multi-gate structures [1]. Short-channel electrostatic control is critical for technology scaling since it allows for shorter channel lengths and lower operating voltages. Only by scaling these intrinsic device electrical properties, can a technology node proposal hope to meet the industrial demands for smaller footprint and higher performance at constant power. Moreover, one also needs to consider the scalability of the structure when embedded into a functional block such as a standard cell or an SRAM. These are indeed the core primitives of logic implementations and as such set the area and performance targets [1].

Advanced technology nodes have utilized FinFET extensively and today's most compact standard cells feature 2 fins per device in a 6-track (6T) cell [2]. However, as scaling continues, two key problems arise limiting the scalability of the FinFET architecture. First, at scaled gate length, the FinFET structure fails to provide sufficient electrostatic control even at reduced fin width. This phenomenon imposes significant challenges below 15nm effective gate length with 5nm fin [3]. The second problem is related to the drive strength for smaller track height in FinFET-based standard cells. A further reduction to a single fin will reduce the effective width by half, which results in significant performance degradation. Moreover, process variability in single fin device will heavily affect performance [1]. Scaling further than the 5nm node new device architectures will be required. Based on the co-optimization of both device and technology aspects, imec has proposed the scaling roadmap shown in Fig. 1, which features several new device architectures. The common thread among these architectures is the tendency to optimize the device in a three-dimensional manner, organizing the active part and internal interconnects in a vertical fashion. Therefore, the proposed structures are more amenable to reduced standard cell track height, w.r.t FinFET.

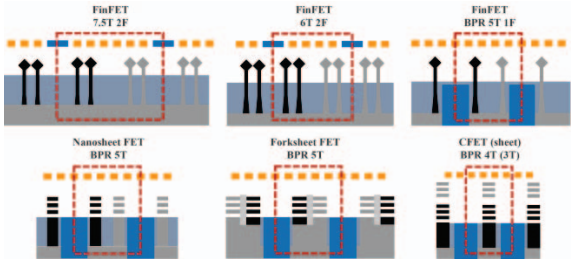

Fig. 1. Various device architectures beyond 5nm. From FinFET to NanoSheet with buried *power* rails, ForkSheet and CFET [JulienIEDM19].

The ultimate scaling solution proposed in the roadmap of Fig. 1 is the Complementary FET (CFET) structure which was introduced by Ryckaert et al., in [4] and consists of an n-type channel stacked vertically over a p-type channel. The folding of both channels over a single footprint allows for extremely compact standard cell designs of 4T or even 3T track height.

The main contributions of this paper are the exploration of standard cell design challenges of ultra-low track height as well as the physical implementation optimization for such standard cells utilizing the CFET device. We show that to fully exploit the area benefits of the CFET devices, one must carefully select the metal stack used for the physical implementation of a large design.

The remainder of this paper is organized as follows. In Section II, we introduce in further detail the CFET concept as well as a more specific focus is given on standard cell design analysis of CFET. Physical implementation and BEOL optimization thereof are shown in Sections III and IV respectively, followed by conclusions in Section V.

II. Background on CFET

In this section, we introduce the structure of a CFET device, review the previous work reported and we elaborate on the standard cell design using a CFET.

A. Structure & Previous work

The CFET concept originates from the complementary nature of CMOS logic where both nFET and pFET are controlled by the same gate, see Fig. 2. Two pairs of stacked S/D electrodes are used to provide access to device pins forming a 5-terminal structure. By construction, CFET requires two levels of S/D contact which can be connected both to the buried power rail (BPR) or to the first routing layer (M0). This configuration assumes a fin-on-fin construction; while other can be envisioned as well (e.g., wire-on-wire, sheet-on-sheet).

Significant amount of work has been already done on the CFET device. Since its introduction in [4], a complete overview of the process assumptions, device characteristics, and circuit evaluation (including SRAM bit-cell assessment) was presented in [5]. In that work, a 4T CFET was shown to outperform in the power-performance-area (PPA) product a

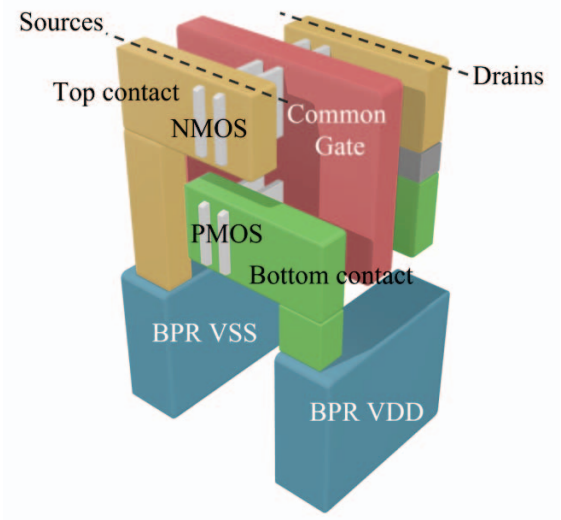

Fig. 4. CFET architecture forming a stacked p-n CMOS primitive structure with 2-level local interconnects, based on [1].

5T FinFET design. Subramanian et al. [6] demonstrate the monolithic integration of CFET devices on 300mm wafers using imec's N14 platform. Furthermore, standard cell design was analysed in [7] and in [8].

B. Standard cell analysis

In terms of standard cell design, using CFET brings up two interesting aspects. First is the connectivity of pull-up and pull-down networks. In a traditional 2D scheme (e.g., FinFET) a vertical back end of line (BEOL) layer – usually M1 – is required to provide intra-cell S/D connectivity. However, the 3D nature of CFET allows for the use of only one horizontal middle of line (MOL) layer (M0) since the S/D of devices are stacked on top of each other and can be interconnected in a 3D fashion. This feature, at a standard cell library level, produces a significant reduction of M1 usage inside standard cells and therefore *should* reduce congestion at the block-level.

Second interesting aspect that arises from the standard cell design with CFETs is the common gate terminal for both channels. As shown in Fig. 2, the CFET structure proposed assumes that both the nFET and pFET channels are controlled by a common gate[1]. This is important because it penalizes the design efficiency of a transmission gate (TG). Transmission gates (shown in Fig. 3a) are commonly used in sequential standard cells, such as flops, clock gates, mux-es, and others. In a traditional 2D design, to minimize the area, a "gate cut" is needed over the PN separation area, so that all S/D are shared and only three gate pitches are used to complete the design (Fig. 3b). Such a gate cut is not available in the monolithic integration of CFET, and therefore dummy gates are required to complete the TG. This increases the width to five gate pitches (Fig. 3c).

In [8] a monolithic 4T CFET standard cell library was designed, and Fig. 4 shows the area gains w.r.t a 5T ForkSheet (FS) library at the same ground rules (45nm poly pitch, 21nm metal pitch). We can see that for most standard cell in the library, CFET achieves a 20% area reduction (20% is the ideal

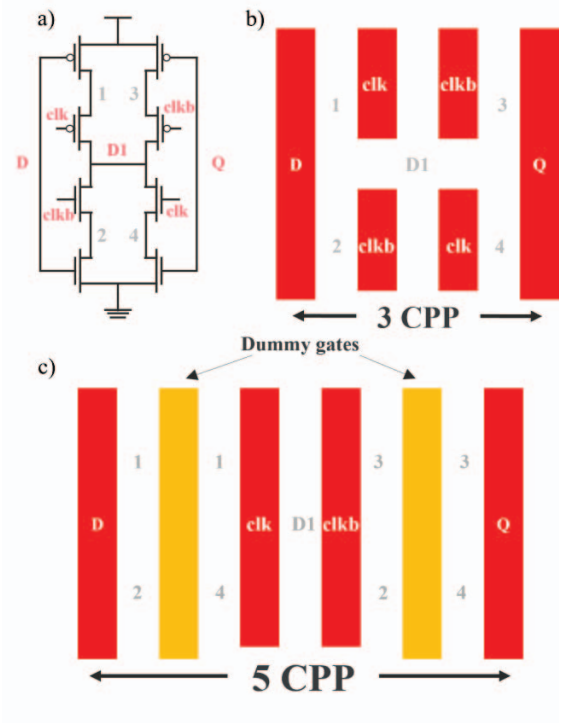

Fig. 3. (a) Typical schematic of a transmission gate, (b) Stick diagram in a 2D layout with gate cut over the PN separation, (c) Stick diagram without a gate cut [8].

area gain when moving from a 5T to a 4T standard cell design in the same ground rules). The exceptions to these gains are the cells that contain a TG structure.

III. Physical Implementation of CFET

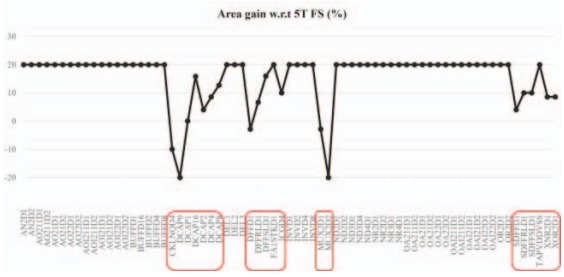

Fig. 2. Area scaling for each cell in the standard library, comparing 4T CFET (monolithic) against a 5T FS library [8].

In this section, we detail the physical implementation setup of the CFET library and present the initial results of the physical implementation experiment. The main purpose of this experiment is to study the effects of the standard cell track height scaling CFET offers (from 5T to 4T) at given set of ground rules.

[1] The common gate approach is part of the monolithic integration strategy. A sequential integration strategy has been proposed as well, where the nFET and pFET devices are controlled by two independent gate terminals. The sequential integration of CFET requires more process steps (increasing the integration costs) and is outside the scope of this paper.

A. Experimental setup

This physical implementation experiment is going to compare two standard cell libraries designed using imec's research process assumptions that correspond to a sub-5nm technology node (see Table I).

TABLE I. NOMINAL PROCESS ASSUMPTIONS

Layers	Pitch (nm)	Patterning
CPP	45	SADP 193i
MINT / M2	21	SALE EUV + 2 EUV cuts
M1 / M3	30	
M4 / M5	48	LELE 193i
M6 – M12	80	SE 193i

The standard cell libraries consist of around 150 cells, one is 5T FS library [9] and one is a 4T CFET library (assuming monolithic CFET integration as presented in Section II). A basic library-level comparison is shown in Table II, where we can observe that the 4T CFET has on average 15.3% smaller area, is 23% more dense (meaning that it has the same amount of cell pins in smaller areas) and has ~10% freer M1 resources compared to 5T FS.

TABLE II. STANDARD CELL LIBRARY COMPARISON

Library averages	Cell area (μm^2)	Pin density (pins/μm^2)	M1 porosity [a]
5T FS	0.0403	124	83%
4T CFET	0.0341	153	91%

[a] Ratio of free M1 tracks over occupied ones for the entire width of a cell

A 64-bit arm© CPU is used as benchmark design and is implemented using an EDA flow with Cadence Genus© for physical synthesis and Cadence Innovus™ implementation system. We utilize only the logic part of the single core CPU (by detaching the memory modules), and the resulting design has around 500k instances. To minimize the physical implementation complexity, we have removed the power delivery network to provide maximum flexibility to the routing engine. Local power delivery is done through the buried power rail. Also, to mitigate for any power-performance discrepancies due to the different devices used in the libraries, we synthesize to very achievable target frequency which should result to a fair comparison plane.

B. Physical implementation results

In Table III we present the cell type breakdown from physical synthesis, targeting a low operating frequency (which is achieved in terms of slack timing by both libraries).

TABLE III. PHYSICAL SYNTHESIS INSTANCE BREAKDOWN

Cell type	Diff (%)
Flops	0%
Clock gates	0%
BUF/INV	+4%
Logic	-1%

The synthesized designs have the same number of flops and clock gates (as expected, since these are defined by the benchmark design and not by power-performance targets), but we can also observe that the differences of BUF/INV cells and logic cells are not significant. This result verifies that, in terms of number of instances, the comparison of physical implementation results from both libraries is fair.

Moving to place and route (P&R) data, the most interesting comparison for this study is the core area scaling shown in Fig. 5.

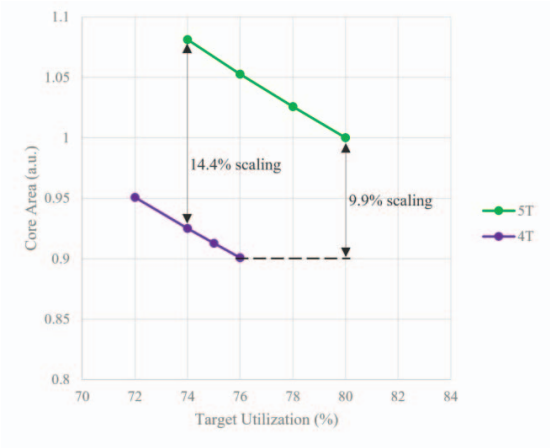

Fig. 5. Post-P&R core area versus target utilization, showing only legal P&R'ed design point. Comparing 5T FS to 4T CFET.

During P&R we iterated over the design to minimize the core area achievable, to do that we swept the target utilization and Fig. 5 presents all the valid P&R designs (where number of violations was insignificant). Comparing the core area at equal target utilization (e.g., 74%), we observe that the 15.3% area gain presented at library level (see Table II), is almost preserved with 4T CFET having 14.4% less core area w.r.t 5T FS. However, when comparing the minimum areas achievable by both libraries we see that the area gains of 4T CFET falls to 9.9%. This gain reduction signifies that the 4T CFET library is more complex to route (since it achieves lower legal utilization) *even though* the CFET cells provide freer M1 resources (as shown in Table II).

To verify the routing complexity of 4T CFET we compare the relative metal layer distribution in Fig. 6. Through this comparison we observe two important facts: a) M1 usage is

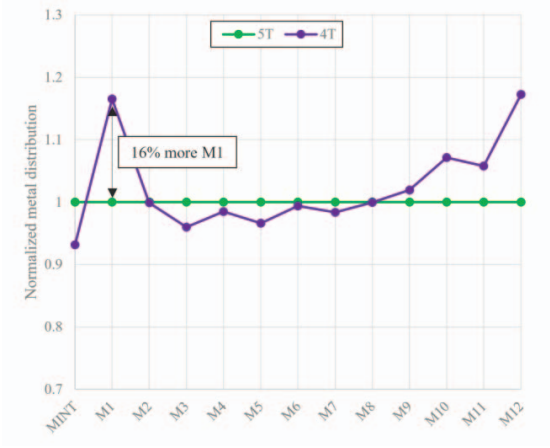

Fig. 6. Normalized metal distribution of 4T CFET w.r.t 5T FS.

16% in 4T CFET; b) M10, M11, M12 usage is significantly higher in 4T CFET which once again supports the notion of routing complexity.

Further insight into the routing behaviour of 4T CFET

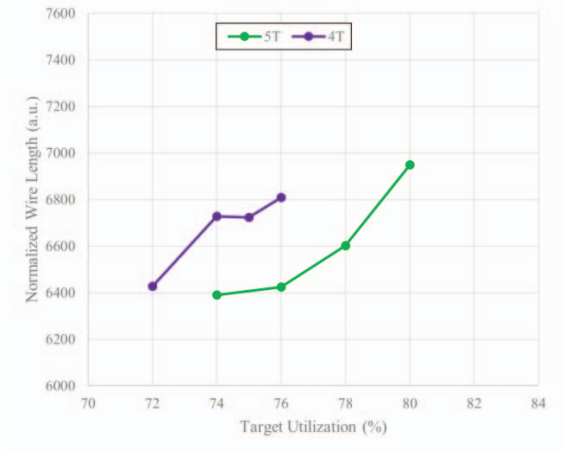

Fig. 7. Normalized wirelength versus target utilization for both 5T FS and 4T CFET libraries.

compared to 5T FS is provided by Fig. 7, where we show the normalized wirelength for each of the libraries (and P&R'ed designs). We define normalized wirelength of a design as the total wirelength divided by the square root of the core area.

The 4T CFET library (due to the routing complexity caused by the low track height) presents a wirelength overhead compared to 5T FS, at iso-utilizations. Based on that, for 4T CFET to achieve higher utilizations (and therefore smaller core areas), somehow the total wirelength should be increased. In the current setup, we are attempting to P&R the same design (meaning same number of nets), with the same BEOL (Table I), in a smaller core area (dictated by the track height scaling of CFET. Therefore, to observe higher values for total wirelength more routing resources need to be provided.

IV. BEOL Optimization for Low-Track Height Nodes

In this section, we analyse the impact of the BEOL metal stack on the two standard cell libraries and the resulting P&R'ed designs.

A. Proposed BEOL stacks

As shown in Section III, more routing resources should be provided to the 4T CFET library to close the utilization gap w.r.t 5T FS and recover as much as possible of the area gain. Towards this, we setup a new round of physical implementation experiments given the BEOL metal stacks shown in Table IV.

TABLE IV. Scaled BEOL Stacks

Layers	BEOL Pitches (nm)		
	Nominal	***m4p30***	***m4m5p30***
CPP	45	45	45
MINT / M2	21	21	21
M1 / M3	30	30	30
M4	48	30	30
M5	48	48	30
M6 – M12	80	80	80

Since we need to keep the comparison in the same technology node as the initial runs (of Section III), we shouldn't scale any of the first 4 routing layers (MINT to M3). So, we try out two different BEOL stacks, one where only M4 is scaled to 30nm (same as M1/M3) pitch, and one where both M4 and M5 are scaled to 30nm pitch.

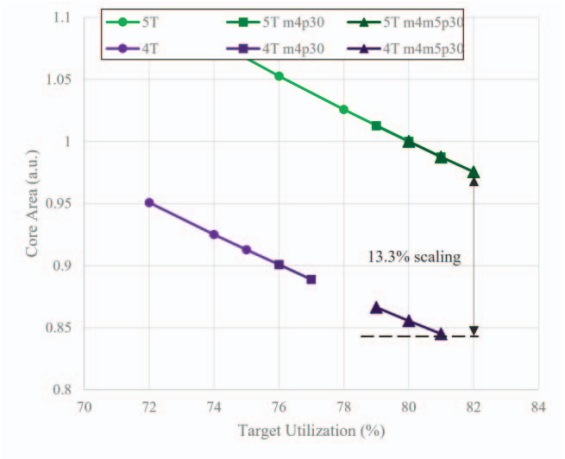

Fig. 9. Post-P&R core area versus target utilization, showing only legal P&R'ed design point. Comparing 5T FS to 4T CFET using the scaled BEOL stacks.

B. Physical implementation results

Given the newly defined BEOL stacks of Table IV, we iterate through the P&R methodology as shown in Section III and compare with the nominal BEOL designs. Fig. 8 shows the area scaling obtained with the scaled BEOL stacks.

The m4p30 BEOL stack does not offer any significant improvement in terms of recovering the area gain of 4T CFET. Indeed, scaling only one metal resource without providing appropriate access to an adjacent orthogonal resource has very limited efficiency, since very few nets can be routed exclusively on one dimension. Consequently, the m4m5p30

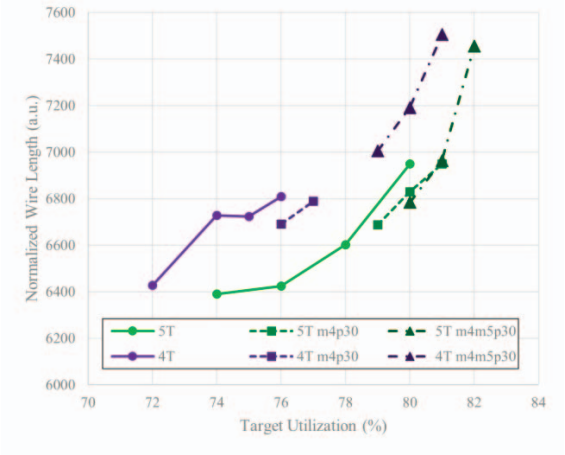

Fig. 8. Normalized wirelength versus target utilization for both 5T FS and 4T CFET libraries using the scaled BEOL stacks.

BEOL stack is much more beneficial for 4T CFET, than m4p30. More specifically, with m4m5p30 the 4T CFET design reaches almost the same maximum utilization as 5T FS, and the area gain is recovered from 9.9% (Fig. 5) to 13.3%. Additionally, it is important to observe that the scaled BEOL stacks do not offer much of area optimization for the 5T FS which means that the number of resources (per unit area) provided by the nominal BEOL stack are adequate for the 5T technologies.

To further substantiate the impact of scaled BEOL stacks we present once again the normalized wirelength of each design in Fig. 9.

As predicted in Section III, providing more routing resources enables both libraries (but especially 4T CFET) to increase significantly their total wirelength and, hence achieving higher utilizations. In fact, with the m4m5p30 BEOL stack both libraries can achieve very close to their absolute minimum core area, since every target utilization above 82% results in placement violations (meaning that the floorplan is too compact to fit all instances of the core design).

V. Conclusions

In this work, we benchmark the area gains of a monolithic 4T CFET library against a 5T ForkSheet one. We show that to maximize these area gains, one must consider that the lower track height of 4T CFET will require more complex routing with longer wirelengths. Therefore, the BEOL stack used for the lower track height library needs to be optimized. This is an important takeaway message for the proposal of any device that offers standard cell track height reduction.

References

[1] J. Ryckaert et al., "Enabling Sub-5nm CMOS Technology Scaling Thinner and Taller!," 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 29.4.1-29.4.4.

[2] X. Wang et al., "Design-Technology Co-Optimization of Standard Cell Libraries on Intel 10nm Process," 2018 IEEE International Electron Devices Meeting (IEDM), 2018, pp. 28.2.1-28.2.4.

[3] D. Jang et al., "Device Exploration of NanoSheet Transistors for Sub-7-nm Technology Node," in IEEE Transactions on Electron Devices, vol. 64, no. 6, pp. 2707-2713, June 2017.

[4] J. Ryckaert et al., "The Complementary FET (CFET) for CMOS scaling beyond N3," 2018 IEEE Symposium on VLSI Technology, 2018, pp. 141-142.

[5] P. Schuddinck et al., "Device-, Circuit- & Block-level evaluation of CFET in a 4 track library," 2019 Symposium on VLSI Technology, 2019, pp. T204-T205.

[6] S. Subramanian et al., "First Monolithic Integration of 3D Complementary FET (CFET) on 300mm Wafers," 2020 IEEE Symposium on VLSI Technology, 2020, pp. 1-2.

[7] S. M. Y. Sherazi, et al., "CFET standard-cell design down to 3Track height for node 3nm and below," Proc. SPIE 10962, Design-Process-Technology Co-optimization for Manufacturability XIII, 1096206 (20 March 2019).

[8] B. Chehab, et al., "Design-technology co-optimization of sequential and monolithic CFET as enabler of technology node beyond 2nm," Proc. SPIE 11614, Design-Process-Technology Co-optimization XV, 116140D (22 April 2021).

[9] P. Weckx et al., "Novel forksheet device architecture as ultimate logic scaling device towards 2nm," 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 36.5.1-36.5.4