CFET SRAM DTCO, Interconnect Guideline, and Benchmark for CMOS Scaling

Hsiao-Hsuan Liu, *Graduate Student Member, IEEE*, Shairfe M. Salahuddin, Boon Teik Chan, Pieter Schuddinck, Yang Xiang, Geert Hellings, *Senior Member, IEEE*, Pieter Weckx, Julien Ryckaert, and Francky Catthoor, *Fellow, IEEE*

***Abstract*— This article explores and evaluates six-transistor static random access memory (SRAM) bitcell design options for sequential and monolithic complementary field-effect transistors (CFET) in 5-Å-compatible (A5) and 3-Å-compatible (A3) technology. A5 CFET offers up to 55% and 40% SRAM bitcell area scaling due to stacked architecture as compared to 14-Å-compatible (A14) nanosheet (NS) technology and 10-Å-compatible (A10) forksheet (FS) technology counterparts, respectively. A dielectric isolation wall (DIW) between gates is introduced in A3 CFET SRAM as a scaling booster. Replacement of gate-cuts with DIW results in up to 17% bitcell area scaling in A3 as compared to A5 CFET SRAM. However, aggressive area scaling introduces routing complexity and limits the node-to-node power and performance (PP) gain. Thus, the interconnect design guidelines are provided to overcome these challenges for power, performance, and area (PPA) enhancements of high-density (HD) SRAM.**

***Index Terms*— CFET, DTCO, scaling, SRAM.**

I. Introduction

STATIC random access memory (SRAM) is the most common embedded memory and a major building block in CMOS ICs. A large portion of the total power, performance, and area (PPA) are occupied by SRAM in many system-on-chips (SoCs) [1], [2], [3]. Increasing last-level cache (LLC) size and the need for improved on-die SRAM density puts enormous pressure on the delay and energy consumption metrics. However, an economically attractive scaled SRAM proposal necessitates overall PPA enhancements of SoCs with evolving technology [3]. Therefore, the continued scaling of SRAM toward 5-/3-Å-compatible technology (A5/A3) associated with design technology co-optimization (DTCO) and global PPA trade-off analysis is explored in this work.

Manuscript received 6 December 2022; accepted 4 January 2023. Date of publication 19 January 2023; date of current version 24 February 2023. The review of this article was arranged by Editor B. K. Kaushik. *(Corresponding author: Hsiao-Hsuan Liu.)*

Hsiao-Hsuan Liu and Francky Catthoor are with IMEC, 3001 Leuven, Belgium, and also with the Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven (KU Leuven), 3000 Leuven, Belgium (e-mail: samantha.liu@imec.be).

Shairfe M. Salahuddin, Boon Teik Chan, Pieter Schuddinck, Yang Xiang, Geert Hellings, Pieter Weckx, and Julien Ryckaert are with IMEC, 3001 Leuven, Belgium.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2023.3235701.

Digital Object Identifier 10.1109/TED.2023.3235701

The complementary field-effect transistor (CFET) [4], [5] is one of the most promising candidates to continue CMOS technology scaling in sub-3 nm nodes. Owing to the vertically stacked architecture, CFET has the potential to scale standard cell and SRAM bitcell area beyond traditional scaling [4]. However, the cell height scaling of CFET SRAM beyond A5 is limited by the large gate-cut requirements for the following reasons. The gate-cut is supposed to be decreased with technology scaling, but it is increased in mono CFET due to the technological challenge which has been identified as the common gate connection [6] compared to its counterparts presented in Fig. 1 [7], [8], [9], [10], [11], [12], [13]. Fig. 3 illustrates that the gate-cut (marked in orange circle) occupies up to 44% of cell height in A5 CFET SRAM. To resolve this scaling bottleneck, replacing the gate-cut with a dielectric isolation wall (DIW) in A3 SRAM is proposed. This design choice enables aggressive SRAM cell height scaling from A5 to A3.

Area scaling of SRAM is always accompanied by the demands of redeveloping suitable interconnect routing designs. Existing routing solutions are mostly proposed to open-up sufficient space at frontside BEOL MINT layer for widening bitlines (BL/BLB). For instance, buried power rails (BPRs) (move the power rails from frontside to backside BEOL [7]) and spacer merge (move cross-coupled formation from BEOL to MEOL [4]) are used for A5 CFET SRAM design. However, those are inadequate for A3 counterparts due to the extensively squeezed cell height. Hence, the area-efficient A3 CFET SRAM DTCO is first approached by solving the interconnect challenges that follow aggressive area scaling. Secondly, it is evaluated with PPA results compared to 14-Å-compatible technology (A14) nanosheet (NS) FET, 10-Å-compatible technology (A10) forksheet (FS) FET, and A5 CFET SRAM. The challenging node-to-node trade-offs between PP and A due to the problematically increased resistance and capacitance (R-C) observed for A3 CFET SRAM. Finally, the BEOL R-C lowering technique is proven to be indispensable for achieving the full PPA benefit of an ultra-scaled SRAM design.

II. SRAM Bitcell Design and Process Flow Challenges

The PMOS and NMOS devices are assumed to be on the bottom and top tiers [5], respectively, in a CFET configuration for maintaining high stress on the p-channel which is

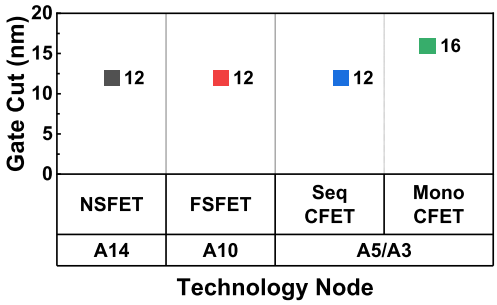

Fig. 1. Gate-cut scaling trend from A14 to A3 [6-13]. The gate-cut assumption is increased in mono CFET due to the common gate connection of the top and bottom tiers [6]. (NSFET: nanosheet FET; FSFET: forksheet FET).

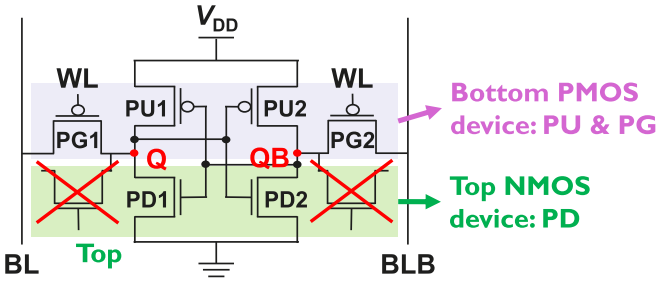

Fig. 2. Schematic view of 6T CFET SRAM.

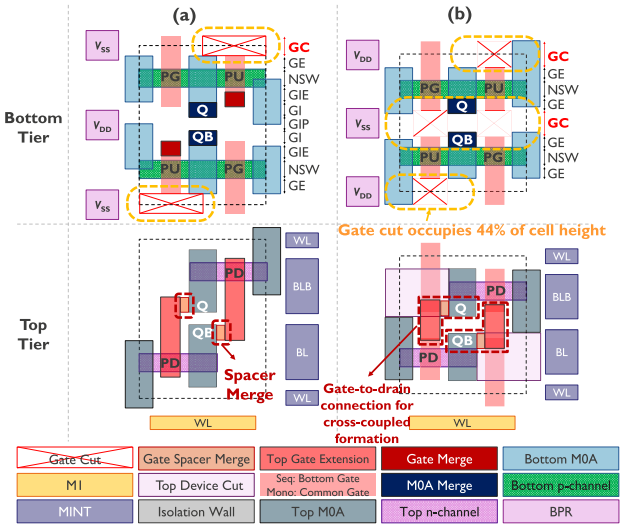

Fig. 3. A5 (a) Seq and (b) mono NS-on-NS CFET SRAM bitcell layout design. Gate-cut occupies up to 44% of cell height among stacked fin and NS CFET SRAM.

preferable for logic applications [13]. Unlike conventional six-transistor (6T) SRAM, a CFET SRAM bitcell consists of four PMOS and two NMOS devices (see Fig. 2). The fin or NS width ratio of pull-up (PU):pass-gate (PG):pull-down (PD) devices is 1:1:1 for high-density (HD) SRAM bitcell design. Table I. lists the 6T SRAM design rules and other parameters from A14 to A3.

Detailed CFET process comparisons between seq (e.g., independent gate connections and separate top-bottom tier fabrications, etc.) and mono (e.g., inherent top-bottom self-alignment, common gate connections, and high aspect ratio processes, etc.) are reported in [4], [6], [11], and [13]. The corresponding process considerations on the proposed CFET SRAM are described in the following Sections II-A and II-B.

TABLE I
6T SRAM DESIGN RULES AND OTHER PARAMETERS FROM A14 TO A3

Parameters		A5 & A3	A14 & A10
Transistor Type		Fin- & NS-based Seq & Mono CFET	NS (A14) FS (A10)
Gate Length (L_g)		12 nm	15 nm
Metal Pitch (MP)		16 nm	18 nm
Contacted Poly Pitch (CPP)		39 nm	42 nm
M0A/MINT Spacing		7 nm	9 nm
M1 Spacing		10 nm	14 nm
Fin/Nanosheet Width (FW/NSW)		5/11 nm	
Fin/Nanosheet Height (H_{fin}/H_{NS})		35/5 nm	
Number of Stacked Nanosheet (*NNS*)		2 (A5) 3 (A3)	4 (A14) 3 (A10)
Gate-Cut (GC)		12 nm (Seq) 16 nm (Mono)	12 nm
Gate Extension (GE)		7.5 nm (A5) 9 nm (A3)	7.5 nm
Gate to Internal Node Contact (GI) Gate to Internal Node Extension (GIE) Gate to Internal Node Peak (GIP)		9 nm (A5)	--
Isolation Wall Width (ISO)		7	--
Nominal Supply Voltage (V_{DD})		0.7 V	
Pass-gate (PG) Devices		PMOS	NMOS
N-channel Stress (S_N)	Seq/Mono Fin	0.962 GPa	--
	Mono NS	0.923 GPa	--
	Seq NS	0.8936 GPa	--
	NS/FS	--	0.694 GPa
P-channel Stress (S_P)	Seq/Mono Fin	1.850 GPa	--
	Mono NS	1.7412 GPa	--
	Seq NS	1.820 GPa	--
	NS/FS	--	1.7 GPa

A. Parasitic Top Device Removal

The redundant parasitic transistors on top of the bottom PG devices in a CFET SRAM should be removed to maintain functionality [4]. In seq CFET, these top devices can be removed by active and gate removal processes without impacting the devices in the bottom tier since the top and bottom tiers are fabricated separately. The top and bottom gates are tied together in a mono CFET device. Thus, the top device removal should be achieved by controlled time-etch. The consequently freed spaces on top of the bottom PG devices (in both mono and seq CFET) are used to implement cross-coupled feedback. Hence, the cell height does not require to be enlarged for building up the cross-coupled connections.

B. M0A Merge and Gate Merge Process Assumptions

The top and bottom M0A merge (dark blue) and gate merge (dark red) in A5 seq CFET SRAM [see Fig. 3(a)] are non-self-aligned, requiring additional spacings from the active region for lithography. The imprecise lithography may have the risk of damage to peripheral surroundings. To reduce up to 21% of cell height occupied by these spacings and the risk of damage, the self-aligned process should be implemented for area scaling in A3. Thus, M0A merge and gate merge are proposed to be abutted to the active region for A3 seq CFET SRAM [see Fig. 4(a)] at the cost of routing challenges.

C. DIW as Gate-Cut Replacement

To mitigate limited cell height reduction due to large gate-cut requirements, the DIW is introduced (see Fig. 4). The

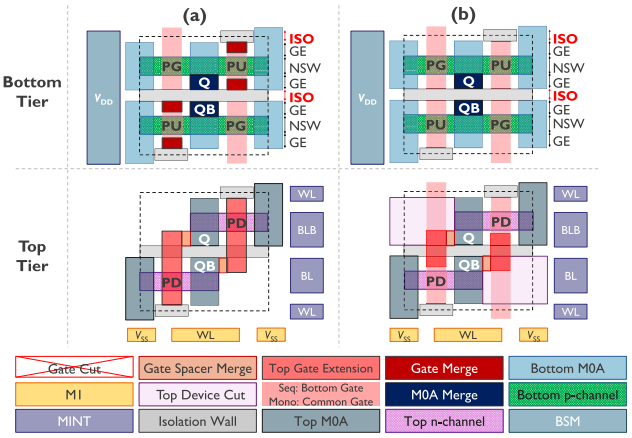

Fig. 4. A3 (a) Seq and (b) mono NS-on-NS CFET SRAM bitcell layout design with DIW.

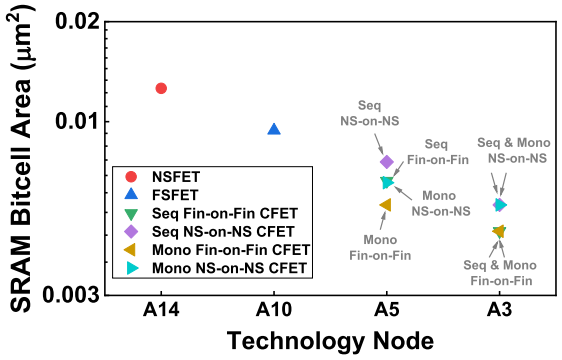

Fig. 5. HD SRAM bitcell area scaling benchmark from A14 to A3.

litho-defined gate-cuts are 12 and 16 nm for seq and mono CFET SRAM, respectively. By replacing these litho-defined gate-cuts with the DIW, the tip-to-tip gate spacing is reduced to 7 nm (critical dimension in A3) by forming the DIW before gate patterning. However, the gate extension is consequently increased to sustain double work function formation. Thus, cell height scaling of up to 8% and 17% are obtained for seq and mono CFET SRAM, respectively, coming purely from DIW implementation. The routing challenges are resolved with the following interconnect design and process flow exploration.

D. SRAM Bitcell Area Benchmark

Fig. 5 shows the HD SRAM bitcell benchmark from A14 to A3. The area scaling of A5 CFET SRAM is up to 55% as compared to A14 NSFET and A10 FSFET SRAM designs. By introducing DIW and self-aligned M0A/gate merge, up to 29% area scaling is predicted in A3 CFET SRAM as compared to the A5 CFET SRAM.

E. Interconnect Challenge and Solution

Fig. 6(a) shows the power delivery network (PDN) solution for A5 CFET SRAM. Both the supply (V_{DD}) and ground (V_{SS}) voltage are connected to the backside metal (BSM) through BPR and via BPR (VBPR) [14]. However, aggressive cell height scaling in A3 CFET SRAM introduces routing challenges. There is an insufficient spacing between the active region and VBPR (V_{SP}) to place V_{SS} as BPR at the backside

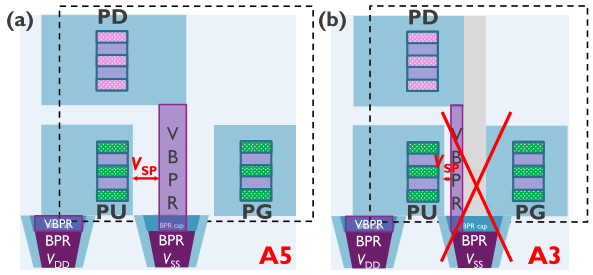

Fig. 6. Schematic views of (a) previous PDN solution (BPRs are implemented on both V_{DD} and V_{SS}) in A5 and (b) PDN challenge (BPR is only applied for V_{DD} since there is insufficient V_{SP} to place BPR for V_{SS}) in A3 for NS-on-NS CFET SRAM. (V_{SP}: spacing between the active region and VBPR.)

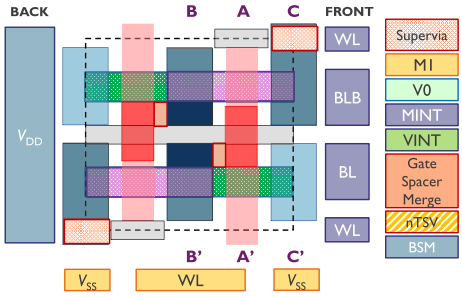

Fig. 7. Top view of interconnect solution for A3 mono NS-on-NS CFET SRAM bitcell design.

BEOL as illustrated in Fig. 6(b). The V_{SS} interconnect thus must be placed at the frontside BEOL in A3 CFET SRAM.

Figs. 7 and 8 illustrate the mono NS-on-NS CFET SRAM bitcell layout and a cross-sectional view, respectively, with the proposed interconnect solution for PDN and signals. Note that the layout designs (see Figs. 3, 4, 7, 8) work for both Fin-on-Fin and NS-on-NS CFET SRAM by replacing the NS with Fin. In A3 design, V_{DD} as a BSM layer can be connected by either BPR or nanoscale through-silicon via (nTSV) [15], [16], and nTSV is assumed in this work. Regarding the insufficient space at the backside BEOL and frontside MINT layer, the V_{SS} is therefore placed on the frontside M1 layer directly connected to the top PD device by a supervia. Unlike the conventional BEOL connection sequence (VINT-MINT-V0-M1), the supervia from M0A to M1 must be implemented here for V_{SS} connection to avoid using extra MINT rails. The cross-coupled inverter is formed by a gate-to-drain merge by locally removing the gate spacer (see Fig. 3) instead of using two MINT tracks [4]. Hence, wider BL and BLB are obtained by using the saved area in the MINT layer. This proposed solution mitigates at least part of the *R-C* challenges of these ultra-scaled SRAM layouts.

F. Coventor Process Flow

Fig. 9 shows the proposed process flow by Coventor simulation of A3 mono NS-on-NS CFET SRAM giving a brief overview of the fabrication sequence. This flow also highlights the remaining challenges in the process integration stage, which is not the main focus of this article. These have to be addressed in the future, and they include DIW formation, redundant top device removal, and cross-coupled contacts formation. Both the bottom PMOS and top NMOS devices

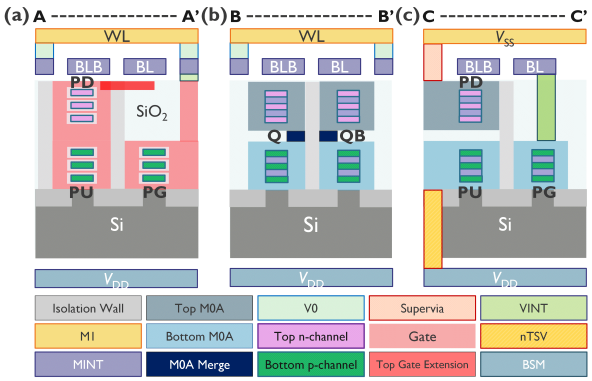

Fig. 8. Cross-sectional views of interconnect solution along the (a) gate, (b) internal nodes (across WL), and (c) edge (across V_{SS}) for A3 mono NS-on-NS CFET SRAM bitcell design. (a) Cross-coupled feedback is formed by extending the top gate to reach the other side of the internal node (M0A) and connecting via spacer merge. (b) Internal nodes (drains of PU and PD) are merged between the top and bottom M0A. (c) Tall via (VINT) connection from BL to bottom PG, supervia connection from V_{SS} (M1) to top PD, and nTSV connection from backside V_{DD} (BSM) to bottom PG.

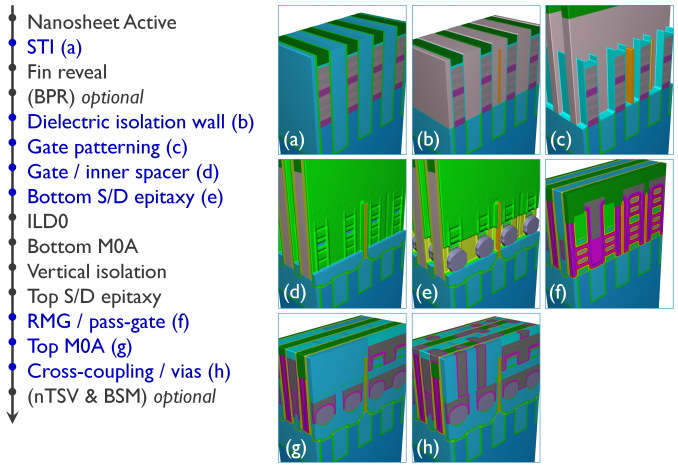

Fig. 9. A3 mono NS-on-NS CFET SRAM process flow description with Coventor simulation. (a) Shallow trench isolation (STI). (b) Dielectric isolation wall (DIW). (c) Gate patterning. (d) Gate and inner spacer formation. (e) Bottom source/drain (S/D) epitaxy. (f) Replacement metal gate (RMG) and redundant pass-gate removal. (g) Top S/D metal (M0A) fabrication. (h) Cross-coupling inverter and vias formation.

are fabricated with vertically stacked lateral NS followed by shallow trench isolation (STI) formation [see Fig. 9(a)]. BPR as backside connections can be optionally fabricated after fin reveal. DIWs are formed in between NSs [see Fig. 9(b)]. Next, gate patterning, gate and inner spacer formation, and bottom source/drain (S/D) epitaxy processes are performed [see Fig. 9(c)–9(e)]. An interlayer dielectric (ILD0) is deposited and followed by bottom S/D metal (M0A) formation. To isolate the bottom and top tiers, vertical isolation is fabricated before the top S/D epitaxy. After fabricating the replacement metal gate (RMG) process, the redundant top PG devices are removed to form the bottom PG device [see Fig. 9(f)], and the top S/D M0A is fabricated [see Fig. 9(g)]. The final frontside process of cross-coupling inverter (L-shaped region) and vias formation are then performed [see Fig. 9(h)]. nTSV and BSM can be optionally fabricated after the frontside process as backside PDN. Note that the concept of fabricating DIW before gate patterning is the same for seq counterparts.

III. SRAM DTCO and PPA Benchmark for HD Option

A. SRAM Parasitics

In this section, all the results are attained considering the reported logic dimensions of fin height (H_{fin}) and number of stacked NS (NNS) [13] as listed in Table I. Fig. 10 shows the parasitic resistance (R_{BL} and R_{WL}) and capacitance (C_{BL} and C_{WL}) of BL and WL in SRAM bitcell extracted by Cadence (for FEOL intrinsic capacitance) and QuickCap (for BEOL resistance and BEOL + MEOL + FEOL extrinsic capacitance) [17], as they dominate the performance and energy in the subarray. The cell height scaling has a direct impact on BL width resulting in increased R_{BL} and the demand for R_{BL} reduction techniques. In this work, although BPR is assumed in both A14 [7] and A10 [9], R_{BL} from A14 to A10 still dramatically increases due to scaling. Therefore, the spacer merge technique is introduced in A5 (already with BPR assumption) to reduce R_{BL} by decreasing the MINT metal track number. However, R_{BL} again increases from A5 to

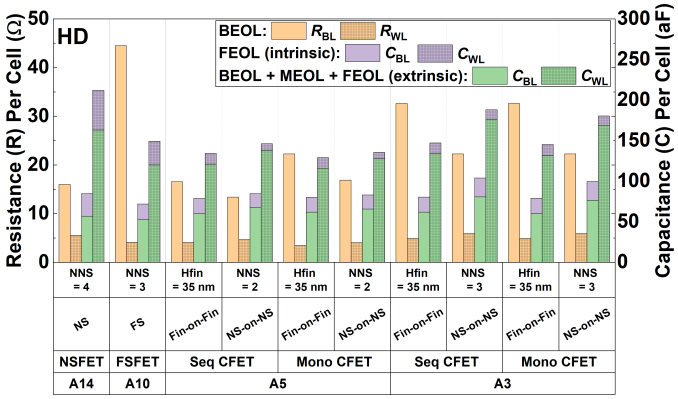

Fig. 10. Parasitic *R-C* per cell of BL and WL for HD A14 NS FET, A10 FS FET, and A5 and A3 CFET SRAM.

A3 implying the necessity of BEOL optimization for SRAM scaling. CFET SRAM exhibits larger BEOL capacitance than NSFET and FSFET counterparts due to the tall via (VINT) connection from BL/WL to the bottom PG. Those increased R-C in A3 CFET SRAM serves as a bottleneck of node-to-node power and performance (PP) gain.

B. SRAM Subarray Circuitry and Simulation Methodology

The simulation circuits include address/timing control, precharge circuits, a row decoder, a column multiplexer (4 to 1; M4), sense amplifiers, write drivers, and the SRAM array with the size of 288 columns (WL) and 256 rows (BL) (1024×72M4). PPA analysis is performed with the worst case bitcell, which is the furthest bitcell from the row decoder, sense amplifier, and write driver. Detailed read and write operation steps and bias conditions for PMOS-PG-based SRAM (A5 and A3) are listed in Table II (operations with charging BL), whereas NMOS-PG-based SRAM (A14 and A10) follows conventional procedures (operations with discharging BL) [7]. The read and write operations for the proposed CFET SRAM design start with pre-discharging BLs to the ground due to PMOS PG. During WL period, one of the BL is charging for sensing the voltage difference or writing the internal node

TABLE II
READ AND WRITE OPERATION STEPS AND BIAS CONDITIONS FOR 6T CFET SRAM WITH PMOS PG IN A5 AND A3

Read Operation (path: V_{DD}, PU, PG)	
1. Set initial condition	Q = V_{DD}; QB = 0
2. Pre-discharging BL & BLB to ground (WL is off)	$V_{BL} = V_{BLB} = 0$; $V_{WL} = V_{DD}$
3. WL activation	$V_{WL} = 0$
4. Charging BL to V_{DD} (through PU and PG devices)	$V_{BL} = V_{DD}$; $V_{BLB} = 0$; $V_{WL} = 0$
5. Sensing voltage difference (+ΔV=150mV) between BL & BLB, then activate sense amplifiers	
6. Output data '1' & '0'	
7. WL deactivation	$V_{WL} = V_{DD}$
Write Operation (path: PG, PD, V_{SS})	
1. Set initial condition	Q = 0; QB = V_{DD}
2. Pre-discharging BL & BLB to ground (WL is off)	$V_{BL} = V_{BLB} = 0$; $V_{WL} = V_{DD}$
3. Asserting data to BL and BLB from the write driver during WL activation	$V_{WL} = 0$; $V_{WD} > V_{DD}$ for V_{BL}
4. Internal node data flipping	Q = V_{DD}; QB = 0
5. WL deactivation	$V_{WL} = V_{DD}$

Q and QB. The read delay is defined as the timing from half of the clock signal to half of the data output from the sense amplifier when reaching 150 mV voltage difference between BLs. An artificial write-assist write driver voltage (V_{WD}) [18] with a positive ramp-up slope from zero to a voltage larger than V_{DD} is applied on BL to ensure successful write operation without specific peripheral circuits. The write margin is defined as V_{DD} minus minimum V_{WD} which confirms the data flip [18]. After obtaining this write margin value, it is set as a constant V_{WD} for both write delay and write energy worst case simulations. A more negative V_{WD} implies stronger write assistance which may decrease the write time and increase the energy consumption. The read and write energy are derived by integrating the active power of the full subarray circuit over the operating time. The compact models used in this work for A14 [19] and A5/A3 [13] are reported, and the same effective off current (~0.02 nA) assumptions are applied for each technology.

C. Impacts of P/NMOS Epitaxy Placements and Stress

To compensate for the difference in intrinsic mobility and thereby achieve comparable transistors' drive strength between PMOS and NMOS, PMOS and NMOS devices are assumed to be placed on the bottom and top tiers, respectively [5], [13]. Placing PMOS (PU and PG) devices on the bottom tier may attain more stress on the p-channel from the substrate resulting in PMOS strength enhancement. By contrast, placing NMOS (PD) devices on the top tier may cause the risk of lack of stress on the n-channel due to the absence of substrate-induced strain and consequently weaken NMOS strength. The NMOS and PMOS stress conditions may affect the read stability and write ability in SRAM. As a result, the influences on read static noise margin (RSNM) (see Fig. 11) and write margin (see Fig. 12) are investigated with varying NMOS/PMOS stress before moving to the PP analysis.

Fig. 11 shows the lack of NMOS stress has negligible impacts on RSNM since the NMOS devices are not in the read

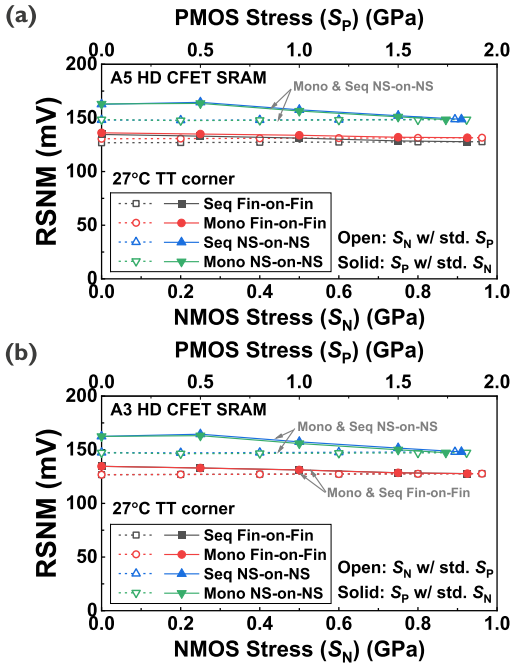

Fig. 11. RSNM with varying NMOS and PMOS stress of (a) A5 and (b) A3 HD CFET SRAM.

path (see Table II). However, increasing PMOS stress for both PU and PG devices slightly degrades the RSNM due to more read disturbance. RSNM is determined by the transistor's drive strength between PG and PU devices and the V_{DD} resistance (R_{VDD}) caused by wires and connected vias in the read path. Only the impact of R_{VDD} on RSNM is discussed here since PG and PU devices use the same PMOS model. The larger R_{VDD} would degrade the RSNM. In both A5 and A3, the larger R_{VDD} of the Fin-on-Fin cases leads to worse RSNM compared to the NS-on-NS counterparts due to the longer VBPR/nTSV connection. In A5, the mono Fin-on-Fin case has slightly less R_{VDD} than the seq counterpart resulting in better RSNM due to the wider VBPR assumptions at the edge. In A3, similar RSNM between seq and mono cases are obtained since they have the same V_{DD} and nTSV widths.

Fig. 12 shows the impact of decreasing NMOS stress is insignificant since the NMOS strength at zero and standard stress are similar. On the other hand, increasing PMOS stress (stronger PG devices) improves the write margin by facilitating faster charging of data storage nodes. Note that our write margin methodology uses transient simulation to capture the overall impacts of parasitic resistance, capacitance, and PG transistors (reported in [18]). Hence, the write margin is determined by the R_{BL}, C_{BL}, transistor's drive strength between PG and PD devices, and V_{SS} resistance (R_{VSS}) caused by wires and connected vias in the write path. Only the impacts of R_{BL} and R_{VSS} are discussed here since the strength between NMOS PD and PMOS PG devices are comparable and C_{BL} does not differ more than R_{BL}. The lower R_{BL} and the larger R_{VSS} would improve the write margin. Lower R_{BL} can be achieved with wider BL naturally coming with larger cell height (see Fig. 10). In both A5 and A3, the Fin-on-Fin cases have both larger R_{BL} and R_{VSS} compared to the NS-on-NS counterparts due to the smaller cell height and

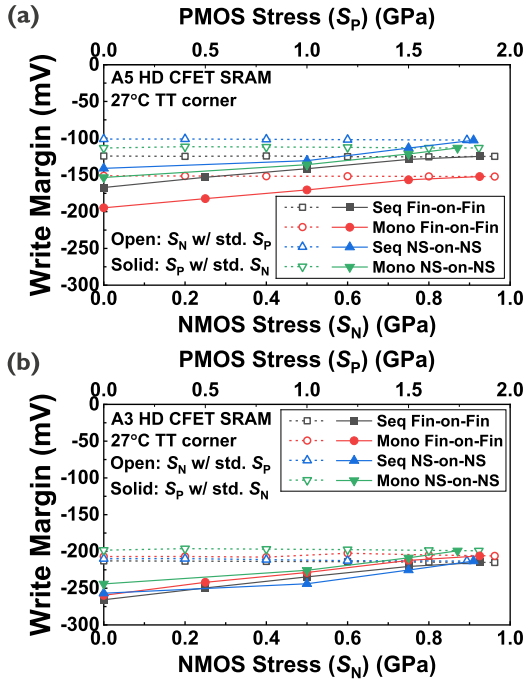

Fig. 12. Write margin with varying NMOS and PMOS stress of (a) A5 and (b) A3 HD CFET SRAM.

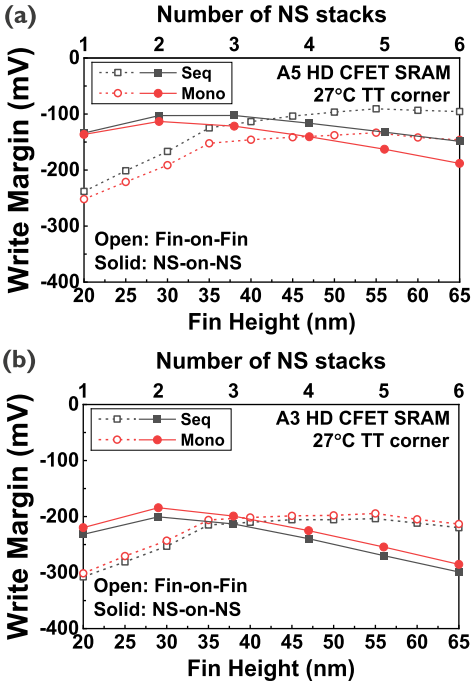

Fig. 13. Write margin with varying H_{fin} and *NNS* of (a) A5 and (b) A3 HD CFET SRAM.

longer VBPR/supervia connection, respectively. Therefore, the differences in write margin between each type of SRAM are obtained by competing with these two factors. For example, the mono Fin-on-Fin case with less R_{VSS} still has the worst write margin among A5 counterparts due to the largest R_{BL}. By contrast, the A5 seq NS-on-NS case has the best write margin since it has the least R_{BL} even with less R_{VSS}. The overall write margins of A5 CFET SRAM are better than A3 counterparts due to the less R_{BL}.

D. Impacts of H_{fin} and NNS

After discussing the stress conditions on both RSNM and write margin, the current logic assumption of NMOS and PMOS epitaxy placements is proven adequate and applicable to the CFET SRAM in this work. Likewise, SRAM dimensions should align with logic (H_{fin} = 35 nm and *NNS* = 2/3 in A5/A3 listed in Fig. 10 [13]) to avoid extra fabrication steps. Therefore, the compatibility between SRAM and logic dimensions is verified by varying H_{fin} or *NNS* on both write margin (see Fig. 13) and read delay (see Fig. 14). In both Figs. 13 and 14, there are two competing phenomena happen simultaneously as the H_{fin} or *NNS* are increased: 1) improvement due to the increased transistor's drive strength and 2) degradation because of a larger voltage drop caused by the increased device capacitance. Note that the current flow to the bitcells also flows through the parasitic capacitors resulting in this voltage drop on BL. As a result, the write margin and read delay are firstly improved then start degrading or saturating at certain points with increasing H_{fin} and *NNS*.

In Fig. 13, a better (more positive) write margin appears at H_{fin} of 35∼55 nm and *NNS* of 2∼3 for Fin-on-Fin and NS-on-NS CFET SRAM, respectively. Detailed write margin analysis and its dominant factors have already been discussed in Section III-C. Fig. 14 shows less read delay at H_{fin} and *NNS* above 35 nm and 2, respectively. The read delay is dominated by R_{WL} and C_{BL}, and the less *R-C* gives the better performance. Considering the logic dimensions, the A5 mono CFET SRAM cases are better than the A5 seq counterparts since it has lower R_{WL} coming from smaller cell height. However, the performances of mono and seq cases are similar in A3 due to the same cell height. Differences between each type of CFET SRAM are mostly determined by the overall impacts of R_{WL} and C_{BL}. Figs. 13 and 14 confirm that the optimal H_{fin} and *NNS* of SRAM are compatible with logic dimensions.

E. SRAM Benchmark and BEOL Optimization Guideline

Fig. 15 shows the margin, performance, active energy, and energy-delay product (EDP) from A14 to A3 SRAM designs. These results are obtained at nominal temperature (27 °C) and typical corners for a fair comparison. The dominant factors and relationships between each margin and performance for double-stack (CFET) SRAMs have been elaborated on in previous sections. Those concepts are also applicable to single-stack (NSFET and FSFET) SRAMs based on opposite read and write paths due to the NMOS-PG-based structure. Therefore, this section focuses on the differences between single- and double-stack SRAMs.

RSNM of each case is similar due to comparable R_{VDD} and transistors' drive strengths. A10 FS SRAM has a worse write margin than A14 NS SRAM due to a much larger R_{BL} coming from area scaling. The write margins of A5 and A3 CFET SRAMs are improved compared to A10 FS SRAM due to less R_{BL} thanks to the spacer merge technique. However, limited R_{BL} and C_{BL} reduction in A3 CFET SRAMs still cause poor (negative value) write margins. This implies a

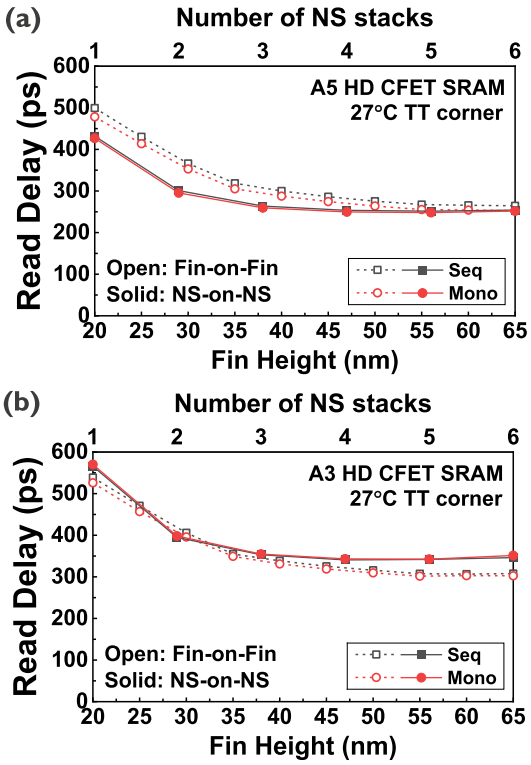

Fig. 14. Read delay with varying H_{fin} and *NNS* of (a) A5 and (b) A3 HD CFET SRAM.

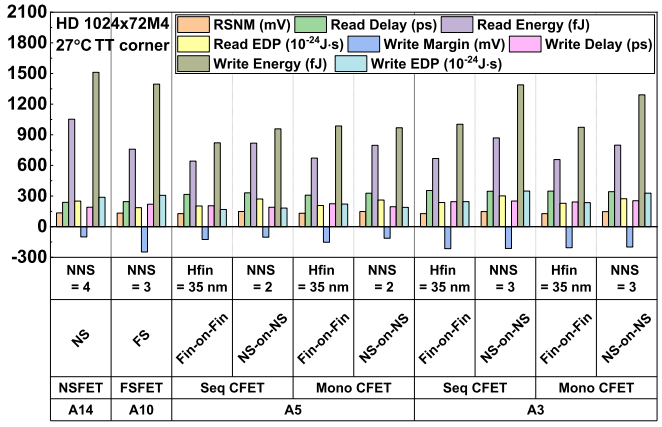

Fig. 15. Benchmark of margin, delay, energy, and EDP for read and write operations in A14 NS, A10 FS, A5 CFET, and A3 CFET SRAM.

high risk of write failure [18], which should be solved by further write margin enhancement strategies such as BEOL optimization and supplementary DTCO in the future. The read delay between A14 NSFET and A10 FSFET SRAM is similar since the latter has less R_{WL} and C_{BL} but a larger R_{BL}. This compensates for the improved or degraded effects on read delay. Although the R_{WL} decreases significantly due to cell height scaling in A5 and A3 CFET SRAMs, the read and write delays do not improve due to larger R_{BL} and C_{BL} caused by the narrower BL and longer via connected from BL to bottom PG, respectively. The energy is mostly dominated by the C_{BL} and the operating time. A10 FS SRAM has less read and write energy compared to A14 NS SRAM due to similar read time but less FEOL C_{BL}. For A5 and A3 CFET SRAMs,

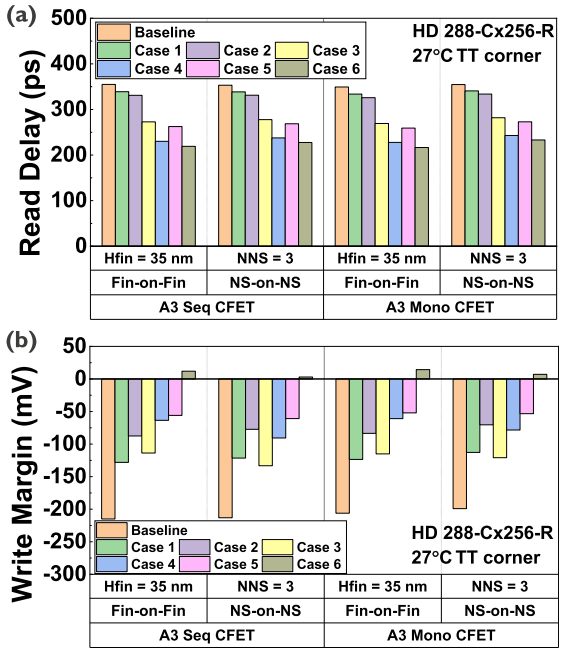

Fig. 16. (a) Read delay and (b) write margin comparison with different BEOL R_{BL} and C_{BL} for A3 CFET SRAM.

TABLE III
TEST CASES OF DIFFERENT BEOL R_{BL} AND C_{BL}

Cases	R_{BL}	C_{BL}
Baseline	1×	1×
Case 1	0.5×	1×
Case 2	0.25×	1×
Case 3	1×	0.5×
Case 4	1×	0.25×
Case 5	0.5×	0.5×
Case 6	0.25×	0.25×

the energies may decrease due to the scaled logic circuits but also increase by large C_{BL}.

As discussed in the previous paragraph, node-to-node area scaling may not guarantee PP benefits due to the increased parasitic *R-C* problem. To achieve more PP benefits in the deeply scaled technology, BEOL optimization for lowering R_{BL} and C_{BL} would be indispensable. Hence, the corresponding BEOL optimization guidelines for improving read delay and write margin based on different BEOL R_{BL} and C_{BL} test cases (see Table III) is provided in Fig. 16. The read speed improvement is dominated by C_{BL} reduction since case 4 and 6 give the best but similar results. On the other hand, reducing both R_{BL} and C_{BL} is crucial to write margin improvement since cases 5 and 6 provide the best results. Several approaches have been reported to reduce *R-C* such as flying BL (which can reduce both R_{BL} and C_{BL}) [20], hybrid-height metal [21], airgap [21], and high aspect ratio [21] techniques. Further developments of BEOL optimization with realistic *R-C* interaction should be tackled in the future by following the *R-C* lowering instructions given in this work.

IV. CONCLUSION

This work evaluates the PPA benchmark of seq and mono stacked fin and NS CFET SRAM bitcell design in A5 and

A3. A3 CFET SRAM offers up to 63%, 50%, and 29 % of area scaling compared to A14 NSFET, A10 FSFET, and A5 CFET counterparts, respectively. However, this aggressive area scaling comes at the cost of routing complexity and exacerbated BEOL parasitic limiting PP gain. Thus, we demonstrated the routing solution and BEOL design instruction to accomplish full PPA benefit for HD CFET SRAM.

REFERENCES

[1] T. Singh et al., "Zen: A next-generation high-performance ×86 core," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 52–53, doi: 10.1109/ISSCC.2017.7870256.

[2] T. Singh et al., "Zen: An energy-efficient high-performance ×86 core," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 102–114, Jan. 2018, doi: 10.1109/JSSC.2017.2752839.

[3] T. Singh et al., "Zen 2: The AMD 7 nm energy-efficient high-performance ×86–64 microprocessor core," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2020, pp. 42–44, doi: 10.1109/ISSCC19947.2020.9063113.

[4] J. Ryckaert et al., "The complementary FET (CFET) for CMOS scaling beyond N3," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2018, pp. 141–142, doi: 10.1109/VLSIT.2018.8510618.

[5] P. Schuddinck et al., "Device-, circuit- & block-level evaluation of CFET in a 4 track library," *Proc. Symp. VLSI Technol.*, Jun. 2019, pp. T204–T205, doi: 10.23919/VLSIT.2019.8776513.

[6] B. Chehab et al., "Design-technology co-optimization of sequential and monolithic CFET as enabler of technology node beyond 2 nm," *Proc. SPIE*, vol. 11614, Apr. 2021, Art. no. 116140D, doi: 10.1117/12.2583395.

[7] S. Salahuddin et al., "Buried power SRAM DTCO and system-level benchmarking in N3," in *Proc. IEEE Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265076.

[8] N. Loubet et al., "Stacked nanosheet gate-all-around transistor to enable scaling beyond FinFET," in *Proc. Symp. VLSI Technol.*, Jun. 2017, pp. T230–T231, doi: 10.23919/VLSIT.2017.7998183.

[9] P. Weckx et al., "Novel forksheet device architecture as ultimate logic scaling device towards 2 nm," in *IEDM Tech. Dig.*, Dec. 2019, p. 36, doi: 10.1109/IEDM19573.2019.8993635.

[10] W. Rachmady et al., "300 mm heterogeneous 3D integration of record performance layer transfer germanium PMOS with silicon NMOS for low power high performance logic applications," in *IEDM Tech. Dig.*, Dec. 2019, p. 29, doi: 10.1109/IEDM19573.2019.8993626.

[11] S. Subramanian et al., "First monolithic integration of 3D complementary FET (CFET) on 300 mm wafers," in *Proc. IEEE Symp. VLSI Technol.*, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265073.

[12] C.-Y. Huang et al., "3-D self-aligned stacked NMOS-on-PMOS nanoribbon transistors for continued Moore's law scaling," in *IEDM Tech. Dig.*, Dec. 2020, p. 20, doi: 10.1109/IEDM13553.2020.9372066.

[13] P. Schuddinck et al., "PPAC of sheet-based CFET configurations for 4 track design with 16 nm metal pitch," in *Proc. IEEE Symp. VLSI Technol. Circuits (VLSI Technol. Circuits)*, Jun. 2022, pp. 365–366, doi: 10.1109/VLSITechnologyandCir46769.2022.9830492.

[14] A. Gupta et al., "Buried power rail metal exploration towards the 1 nm node," in *IEDM Tech. Dig.*, Dec. 2021, p. 22, doi: 10.1109/IEDM19574.2021.9720684.

[15] G. Sisto et al., "IR-drop analysis of hybrid bonded 3D-ICs with backside power delivery and μ-& n-TSVs," in *Proc. IEEE Int. Interconnect Technol. Conf. (IITC)*, Jul. 2021, pp. 1–3, doi: 10.1109/IITC51362.2021.9537541.

[16] A. Veloso et al., "Enabling logic with backside connectivity via n-TSVs and its potential as a scaling booster," in *Proc. Symp. VLSI Technol.*, Jun. 2021, pp. 1–2.

[17] *QuickCap R-2020.09-SP4*, Synopsys, Mountain View, CA, USA, 2020.

[18] H.-H. Liu et al., "Extended methodology to determine SRAM write margin in resistance-dominated technology node," *IEEE Trans. Electron Devices*, vol. 60, no. 6, pp. 3113–3117, Jul. 2022, doi: 10.1109/TED.2022.3165738.

[19] D. Jang et al., "Device exploration of nanosheet transistors for sub-7-nm technology node," *IEEE Trans. Electron Devices*, vol. 64, no. 6, pp. 2707–2713, Jun. 2017, doi: 10.1109/TED.2017.2695455.

[20] J. Chang et al., "12.1 A 7 nm 256 Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-V_{MIN} applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 206–207, doi: 10.1109/ISSCC.2017.7870333.

[21] A. Farokhnejad et al., "Evaluation of BEOL scaling boosters for sub-2 nm using enhanced-RO analysis," in *Proc. IEEE Int. Interconnect Technol. Conf. (IITC)*, Jun. 2022, pp. 136–138, doi: 10.1109/IITC52079.2022.9881286.