First Experimental Demonstration of Self-aligned Flip FET (FFET): a Breakthrough Stacked Transistor Technology with 2.5T Design, Dual-side Active and Interconnects

Haoran Lu, Yandong Ge, Xun Jiang, Jiacheng Sun, Wanyue Peng, Rui Guo, Ming Li, Yibo Lin, Runsheng Wang, Heng Wu*, Ru Huang
School of Integrated Circuits, Peking University, Beijing, China, *Email: hengwu@pku.edu.cn

Abstract

In this work, for the first time, the Flip FET (FFET), a novel stacked transistor technology with self-aligned active and interconnects on both sides of wafer, is proposed and experimentally demonstrated. Two layers of transistors are formed on the same active and back-to-back stacked, featuring a much more manufacturing-friendly process flow with lower aspect ratio (AR) than CFET. Standard cell (STC) libraries with minimum 2.5 track height (2.5T) design are established, proving further scaling possibility and better routability over CFET. Benefited from the dual-side process, FFET also has better design flexibility with no restriction on N/P polarity for each transistor layer, enabling a bipolar SRAM with further 12% area reduction over CFET SRAM and at least 35.9% over FinFET SRAM, respectively. Meanwhile, based on fins, FFET outperforms CFET with 21.5% higher frequency at iso-power and 45.0% lower power at iso-frequency. Furthermore, for nanosheet-based ones, FFET shows extra benefits over CFET with larger nanosheet width (W_{NS}), with 14.5% higher frequency at iso-power at the same footprint. New concepts of dual-side interconnects are introduced and the P&R result of a RISCV32I core further validates the superiority of FFET with more than 31.3% area reduction compared with CFET.

Introduction

As the conventional effective scaling coming to an end [1], hyper scaling with DTCO emerges as a key enabler for future logic technology. Methods like fin depopulation and track reduction [2][3] are widely used, as shown in Fig. 1(c). Meanwhile, BSPDN [4] and stacked transistors such as CFET [5], also attract hot discussions recently. Unfortunately, due to the limitation of pin access, CFET cell height reduction ends at 3T design with 3 frontside-only signal tracks [5] and 2 backside-only buried power rails (BPR) [6]. The CFET also faces great manufacturing challenges due to the high AR processes.

For here, to maximize the potential of wafer backside and overcome the process complexity of CFET, we propose a breakthrough stacked transistor technology with dual-side signal/power capability: self-aligned Flip FET (Fig. 1 (a-b)), named after its wafer flipping process. Key FFET processes, including the wafer bonding, thinning and self-aligned active, were successfully developed. From our comprehensive study in the aspects of process complexity, STC design flexibility, SRAM scalability, PPA analysis, and P&R assessment, the FFET is a great candidate for the next-generation logic transistors, as shown in Fig. 1(c).

Process Flow and Device Fabrication

Fig. 2(a) summarizes the critical steps of the FFET flow, taking 3.5T dual-fin STC as an example. The fin is etched first and then half-buried in STI. After the frontside (FS) FEOL & BEOL formation, a carrier wafer is bonded to it and flipped. Then, the Si substrate of the active wafer is thinned down with CMP stopping on STI. The backside (BS) fin is then revealed by STI recess in a self-aligned manner, followed by the BS FEOL & BEOL formation. Most importantly, the processes of both sides follow the standard FinFET/NSFET flow except those for inter-side connections. Its process implementation is far easier than Mono. CFET [5], as validated by the lower AR in Fig. 2(b).

Fabrication of fin-based FFET was successfully conducted. Fig. 3(a) shows the SEM image of the FFET FS FEOL & BEOL. Fig. 3(b) depicts the partial backside thinning down with some Si remaining. Fig. 3(c) shows the precise CMP stop on STI in cross gate and fin directions. The tapered fin profile can be improved by a novel optional amorphization process as shown in Fig. 3(d). BS STI recess by wet etch to reveal the BS active is shown in Fig. 3(e). The Common Gate (CG) is shown in Fig. 3(f). Unique in FFET, the Split Gate (SG) can be easily realized, as shown in Fig. 3(g). Further treatment can also be used to optimize the BS fin profile to be fully symmetrical to FS fin, as given in Fig. 3(h).

Ultra-scaled Standard Cell and SRAM

With 2 signal tracks and 1 power rail shared with adjacent STCs placed in a mirror configuration at both sides, 2.5T FFET has 1 more signal track than 3T CFET [5]. With simpler processes like the non-stacked FET at each side, S/D and gate pins of FFET can be connected to any signals flexibly, providing wider design space than CFET. Although the polarity of FS and BS transistors are switchable, we use nFET at FS and pFET at BS for the thermal budget constraint of pFET [7][8]. Fig. 4(a) shows the layout of 2.5T FFET inverter (INV). CG is realized by merging the SG (Fig. 1(b)). Thanks to the separated gate process, FFET can support critical sequential logic cells such as: transmission gate and C^2MOS (Fig. 4(b)) without area penalty, while 3T CG CFET or 3.5T SG CFET [9] have to waste CPPs or 0.5T. Fig. 4(d-f) give DFF layouts with folded 2-row design for the three stacked FETs above and FFET is the smallest due to reduced cell height and the efficient split-gate design. The area of key cells in our STC libraries are compared in Fig. 4(c), further proving the area benefits of FFET.

As FFET can support both nFET and pFET at each side, unipolar SRAM (Uni-SRAM) and bipolar SRAM (Bi-SRAM) were designed, as shown in Fig. 5(a-b). FFET SRAM is capable of both N-PG (this work) or P-PG while CFET only allows P-PG [5][10]. The tall via connecting FS source to BS power can be removed in Bi-SRAM thanks to its dual-side power. As a result, its area is further reduced to 0.0135 um^2, 12% smaller than the Uni-SRAM and CFET SRAM [10] at the same design rule and 35.9% over FinFET SRAM [11]. Furthermore, about 45% area reduction over FinFET SRAM can be achieved with the same push rule as in ref [11]. Fig. 5(c) shows the cross-couple enabled by the Gate Merge in Bi-SRAM. The RC trade-off was also studied. By placing WL & BL on both sides, 31.5% BL metal line resistance and 55% WL metal line resistance are reduced in Bi-SRAM as given in Fig. 5(d), thanks to the wider BL and the smaller cell height with the two WL tracks running in parallel at FS and BS, respectively. Uni-SRAM also has larger WL pin resistance (Fig. 5(e)) for its shared V0 and VG. However, due to the doubled WL track and the wider BL, WL & BL capacitance increases by 21.5% and 24.1% in Bi-SRAM, as shown in Fig. 5(f).

PPA Analysis

A 15-stage ring oscillator with FO3 and typical BEOL loads [12] was used in the PPA evaluation. The device models were calibrated to ref [11, 13-15]. We assume the same intrinsic transistor performance for all the architectures studied.

Note that for fair footprint benchmark, we compared the 4.5T single-fin FinFET, the 4T dual-fin CG CFET and the 3.5T dual-fin FFET by using the same M0 pitch for all the stacked transistors studied, which results in smaller Mx pitch for CG CFET. As shown in Fig. 6(a), at iso-power @VDD=0.7V, frequency of FFET exceeds FinFET by 4.9% (W/ BEOL load) and exceeds CFET by 21.5% (W/ BEOL load) and 26.4% (W/O BEOL load) due to smaller parasitic RC. FFET INV also has 70.5% reduced source resistance (Fig6. (b)) due to the far longer VD to BPR for top device in CFET [5]. Furthermore, FFET has a wider gate cut thus shorter gate extension because FFET MOL is formed after the gate cut while the bottom tier of CFET [5] isn't. This leads to 34.2% reduction in input capacitance (Fig. 6(c)). By inserting more DTCO knobs, 5.0% frequency gain at iso-power @VDD=0.7V can be realized in FFET (inset in Fig. 6(a)).

3.5T nanosheet-based FFET was also studied and compared with CG CFET at the same footprint. As given in Fig. 7(a), with relaxed restriction of space from vias to active, nanosheet in FFET can be much wider than that in CFET (46nm vs 34nm), resulting in a performance gain of 14.5% at iso-power at respective max W_{NS}. It also outperforms fin-based FFET and nanosheet-based CFET by 10.8% and 10.5% at iso-power and iso active width, as shown in Fig. 7(b).

Dual-side Interconnects

For logic gates, output pin is typically composed of common S/D of a pair of nFET and pFET. Similarly, in FFET STC, each output pin is linked by the Drain Merge (the via connecting FS and BS S/D, shown in Fig. 1(b)) and gets connected to FS and BS M0 to drive the next stage cell with input pins on FS or BS. Fig. 8(a) shows an FFET INV with dual-side output (FOUT & BOUT) and input at either side (FIN or BIN). A typical FFET circuit consisted of an INV driving the next stage INV and NAND is illustrated in Fig. 8(b) and the dual-side interconnects layout and the 3D schematic of routing is given explicitly in Fig. 8(c-d). Signals can also reach the other side (inter-side interconnects) in none-active region by using the same processes as Drain Merge, as given in Fig. 8(e).

Based on the routing principles above, a block-level routability evaluation on FFET and CG & SG CFET W/ BSPDN was conducted by doing P&R with a RISCV32I core. FFET has much less design rule violation (DRV) than CG & SG CFET with the same chip area (Fig. 9(a)) as the input pin number is halved at each side for FFET. With less area and better routability, FFET has 44% and 31.3% core area reduction against CG & SG CFET respectively, as in Fig. 9(b).

Other Crucial Aspects

Due to the shared sub-fin region of FFET, the electrical coupling from both sides cannot be ignored, especially in the OFF-state. Thus, the RO IDDQ was studied by TCAD mixed mode simulation in Fig. 10(a) for different well profiles. The IDDQ extracted (Fig. 10(b)) can be effectively reduced by optimizing the well doping and matched to the target without coupling as shown in Fig. 10(c). Note that the coupling can be further reduced by MDI as in CFET [16].

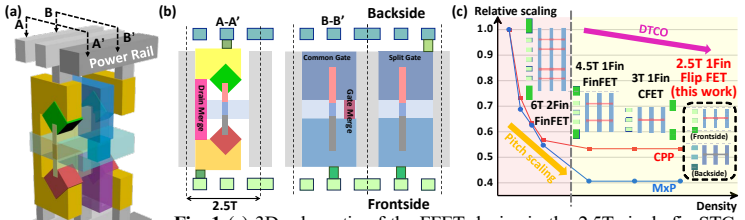
Lastly, for even better thermal budget and variability control, we further propose the Fully-aligned Flip FET (F3ET), with both active and gate self-aligned. By bridging the separated dummy gates after stripping the sacrificial material under the active Si, self-aligned RMG can be formed after BS epitaxy.

Conclusion

Table 1 benchmarks reported stacked FETs with the FFET technology demonstrated in work in all key aspects. With self-aligned active and gate, lower AR processes, higher design flexibility, smaller cell size and better routability, FFET shows great potential. The experimental demonstration and DTCO framework in this work provide new insights for future logic technology.

Reference

[1] M. Radosavljević, et al., *IEDM 2021*. [2] S. -Y. Wu, et al., *IEDM 2022*. [3] J. Cai, *VLSI 2021* SC. [4] M. Kobrinsky, et al., *IEDM 2023*. [5] J. Ryckaert, et al., *VLSI 2018*. [6] A. Veloso, et al., *IEDM 2023*. [7] J. Wang, et al., *VLSI 2021*. [8] A. Vandooren, et al., *IEDM 2018*. [9] L. Liebmann, et al., *IEDM 2021*. [10] H. -H. Liu, et al., *TED, p.883, 2023*. [11] G. Yeap, et al., *IEDM 2019*. [12] S. Yang, et al., *IEDM 2023*. [13] J. Jeong, et al., *VLSI 2023*. [14] R. Li, et al., *www.techinsights.com/blog/qualcomm-snapdragon-8-gen1-soc, 2022*. [15] Y. Lee, et al., *IEEE Access, p. 80695, 2022*. [16] S. Liao, et al., *IEDM 2023*. [17] B. Chehab, et al., *SPIE 2021*. [18] W. Rachmady, et al., *IEDM 2019*. [19] L. Brunet, et al., *VLSI 2016*. [20] J. Park, et al., *IEDM 2023*.



CPP	Fin Pitch	Mx Pitch
50nm	28nm	30nm

Fig. 1 (a) 3D schematic of the FFET device in the 2.5T single-fin STC. Design rule assumptions [14] are listed below. (b) X-sections along the S/D (A-A') and gate region (B-B'), which shows the Drain Merge and the Gate Merge for the CG. The STC right next to the B-B' shows the SG. (c) Area scaling roadmap with FFET as a breakthrough scaling device beyond CFET.

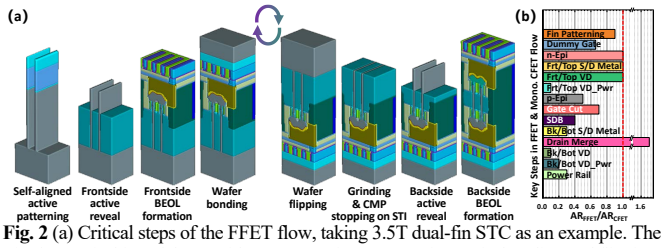


Fig. 2 (a) Critical steps of the FFET flow, taking 3.5T dual-fin STC as an example. The Si substrate is thinned down with CMP stopping on STI after wafer bonding and flipping. Then the BS fin is revealed by STI recess in a self-aligned manner. Most of the processes of both sides follow the standard FinFET/NSFET flow, resulting in (b) manufacturing-friendly processes with lower AR compared with Mono. CFET.

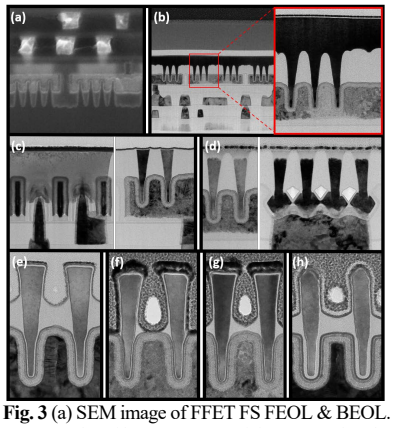


Fig. 3 (a) SEM image of FFET FS FEOL & BEOL. TEM results of key processes of the FFET: (b) The partial substrate thinning down with some Si remaining. (c) CMP stops on STI (in cross gate and fin directions). (d) Tapered fin profile improved by amorphization process (in gate and S/D regions). (e) STI recess by wet etch. (f) Common Gate. (g) Split Gate. (h) Split Gate with further fin smoothing.

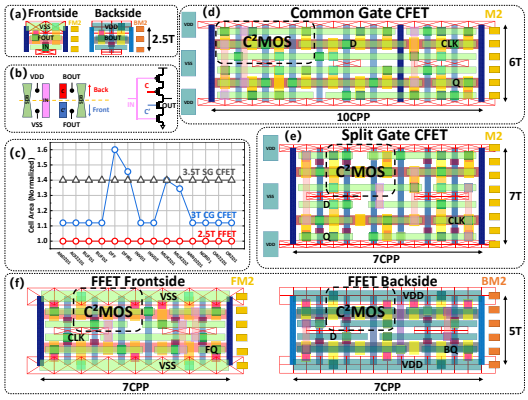


Fig. 4 FFET STC layout design. (a) 2.5T Inverter with 2+2 signal tracks. FFET can support transmission gate and (b) C²MOS without extra area penalty by SG. (c) STC area comparison between 2.5T FFET, 3T CG CFET and 3.5T SG CFET. (d-f) DFF layouts with folded 2-row design for the three stacked FETs. FFET has fewer CPPs than CG CFET and smaller cell height than CG & SG CFET.

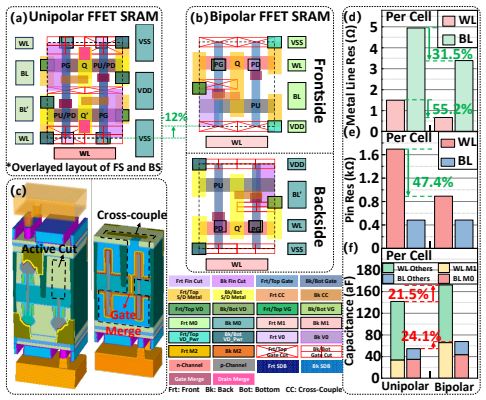


Fig. 5 FFET SRAM comparison between (a) Uni-SRAM and (b) Bi-SRAM. (c) 3D illustration of a Bi-SRAM and the cross-couple enabled by Gate Merge. (d-f) Parasitic RC trade-off. Bi-SRAM has 55.2% and 47.4% lower WL metal line and pin resistance and 31.5% lower BL metal line resistance while WL & BL capacitance increase by 21.5% and 24.1% in Bi-SRAM.

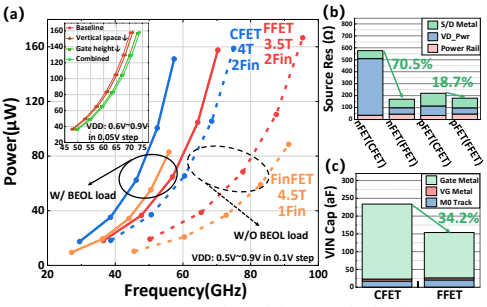


Fig. 6 (a) Power-frequency plots of the 4.5T single-fin FinFET, 4T dual-fin CFET and 3.5T dual-fin FFET with the same M0 pitch at similar footprint. DTCO knobs in the top-left corner includes reduction of vertical space, which lowers the resistance of Drain Merge and Gate Merge, and reduction of gate height, which reduces the gate-to-drain capacitance. (b) Reduced source resistance of FFET due to the shorter VD to power. (c) Reduced input capacitance of FFET due to shorter gate extension.

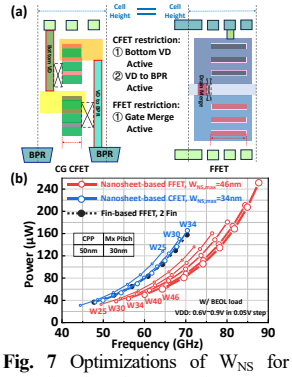


Fig. 7 Optimizations of W_{NS} for nanosheet-based FFET and CFET. (a) Relaxed space constraint enables FFET with wider W_{NS} at the same footprint as CFET. (b) Based on nanosheets, FFET outperforms CFET with 14.5% higher frequency at iso-power at respective maximum W_{NS}.

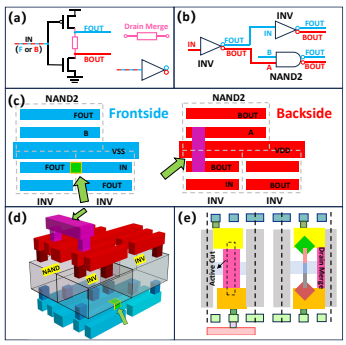


Fig. 8 Principles of dual-side interconnects. Each output is linked by the Drain Merge. (a) FFET INV with dual-side output and either-side input. (b) A typical FFET circuit and its (c) dual-side interconnects layout and (d) 3D schematic of routing. (e) Signals can reach the other side in non-active region by using the same processes as the Drain Merge.

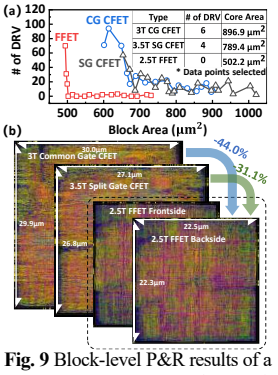


Fig. 9 Block-level P&R results of a RISCV32I core. (a) FFET has much less DRV because input pins are distributed evenly into FS and BS. (b) BEOL layout of CG CFET, SG CFET and FFET with similar DRV correspond to the table inserted in (a). FFET has the smallest core area.

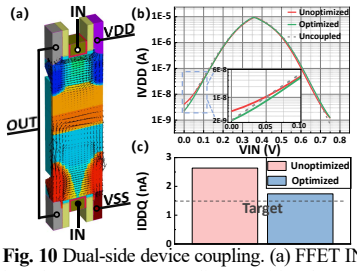


Fig. 10 Dual-side device coupling. (a) FFET INV based on TCAD. Coupling-considered IDDQ, which is extracted by (b) IVDD-VIN curve, (c) can be well reduced by optimizing the well dopants and matched to the target without coupling.

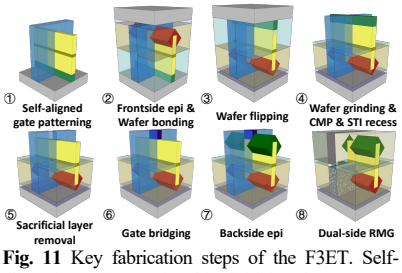


Fig. 11 Key fabrication steps of the F3ET. Self-aligned gate is completed by bridging the separated dummy gates after removing the sacrificial material buried under the self-aligned active. Self-aligned RMG can be formed after BS epitaxy.

	Seq. CFET [8][10][17][18][19]	CG Mono. CFET [4][5][10][16][20]	SG Mono. CFET [9]	FFET	F3ET
Active	None Self-aligned	Self-aligned	Self-aligned	Self-aligned	Self-aligned
Gate	None Self-aligned	Self-aligned	Self-aligned	None Self-aligned	Self-aligned
	Split-gate	Common-gate	Split-gate	Split-gate	Split-gate
Aspect Ratio	Low	High	High	Low	Low
Post-Epi RMG	No(Bottom Tier)	Yes	Yes	No(Frontside)	Yes
Device Polarity for Each Active Layer	Unipolar/ Bipolar	Unipolar	Unipolar	Unipolar/ Bipolar	Unipolar/ Bipolar
SRAM Categories	P-PG	P-PG	P-PG	N-PG/P-PG	N-PG/P-PG
Smallest STC Height	4 T	3 T	3.5 T	2.5 T	2.5 T
Smallest STC Signal Track #	4	3	3	4	4
FEOL/BEOL Categories	Frontside: Active/Signal & PDN Backside: Signal (Optional) & PDN			Frontside: Active/Signal & PDN Backside: Active/Signal & PDN	

Table 1 Benchmark of reported stacked transistor technologies with FFET.