DTCO Launches Moore's Law Over the Feature Scaling Wall

V. Moroz[1], X.-W. Lin[1], P. Asenov[2], D. Sherlekar[1], M. Choi[1], L. Sponton[3],
L. S. Melvin III[4], J. Lee[2], B. Cheng[2], A. Nannipieri[3], J. Huang[1], and S. Jones[5]
[1]Synopsys, Inc., Mountain View, CA, USA, [2]Synopsys Northern Europe, Glasgow, UK,
[3]Synopsys Switzerland LLC, Zurich, Switzerland, [4]Synopsys, Inc., Hillsboro, OR, USA,
[5]IC Knowledge LLC, Georgetown, MA, USA, email: victor.moroz@synopsys.com

2020 IEEE International Electron Devices Meeting (IEDM) | 978-1-7281-8888-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IEDM13553.2020.9372010

***Abstract*—** Instead of marching from one crisply defined technology node to the next with an uncertain timeline, industry is transitioning toward annual technology updates driven by a schedule, but with an uncertain transistor density increase. Full node updates are expected every other year, with "half-node" updates in between. Pitch scaling began slowing after the 10nm node and is expected to practically cease by the 1nm node. Despite that, transistor density is expected to continue increasing at a similar pace of 45% density increase per node (or 20% per year) through the 1nm node, fueled by increasingly sophisticated Design-Technology Co-Optimization (DTCO) and Electronic Design Automation (EDA) advances.

I. Introduction

For many decades, transistor density has approximately doubled with every technology generation. Most of the density increase came from pitch scaling, with the remainder from DTCO and improvements in design tools and algorithms [1, 2]. However, after 10nm node, pitch scaling has slowed and at the 5nm node, about half of density improvements come from DTCO and advances in EDA [3] (Fig. 1). Despite disappearing feature scaling, transistor density keeps increasing by 46% per node since switching from planar MOSFETs to the FinFETs [3] (Fig. 2). Another important observation from Fig. 2 is that SRAM scaling in the post-planar era is consistently slower than logic scaling.

This work focuses on technology evolution toward 2028 – 2030, which roughly corresponds to 1nm node. Considering that SRAM scaling is falling behind logic, SRAM design and technology are the focus.

II. CFET Logic and SRAM Architecture

Complementary FET (CFET) technology represents the first step in stacking transistors vertically and can increase transistor density for a given feature size [4]. The top tier of transistors must be narrow to allow access to the lower tier of transistors. Therefore, a wide GAA (Gate All Around) transistor was chosen for the lower tier, and a stack of 2 narrow GAA channels for the upper tier (Fig. 3).

PMOS was placed at the lower tier as epitaxial SiGe source/drain (S/D) stress engineering is easier to arrange there. A wide single channel strained PMOS and 2 stacked narrow NMOS have comparable driving strengths, creating a balanced logic architecture. Combined with Buried Power Rails (BPR), this architecture enables a 4-track tall logic library with 4 routing tracks or a 5-track tall logic library with 5 routing tracks (Fig. 3a).

For CFET SRAM, the same width upper and lower tier transistors were considered, bringing the NMOS to PMOS drive strength ratio close to 2:1, which is desirable for an SRAM bitcell (Fig. 3b). Such CFET SRAM cell has about 2x higher transistor density than CFET logic, which is limited by pin access and routing. This is achieved by creating a densely packed bitcell volume with a dummy PMOS under the Pass Gate (PG) NMOS, which is tri-gate rather than GAA, a Word Line (WL) placed on Metal 0 (M0), and a power supply placed on the BPR (Fig. 4). Using tri-gate PG enables a tighter bitcell design and also helps to get desirable Pull-Down (PD) vs PG transistor strength ratio. This work is based on a 39 nm Contacted Poly Pitch (CPP) and a 19 nm Metal Pitch (MP), resulting in a transistor density of just over 1 billion transistors per square millimeter (Fig. 5).

III. Patterning Challenges

The CFET SRAM has lithographically standard shrink effects except for the M0 and cross-couple (CC) layers. The M0 layer combines wide and narrow wires (Fig. 5) that can be patterned with a 13.5 nm wavelength (EUV) 0.33 numerical aperture (NA) system. NA quantifies the imaging information collected from the mask by the scanner optics. This information forms the wafer pattern. The Rayleigh criteria describes this image formation [5]:

$$CD_{min} = k_1 \frac{\lambda}{NA} \qquad (1)$$

$$Depth\ of\ Focus = \frac{\lambda}{NA^2} \qquad (2)$$

CD_{min} is the minimum resolvable pitch for a given system (nm), λ is the wavelength of light (nm), and k_1 is the process factor, typically 0.5. Depth of Focus (DoF) describes the total variation from the focal plane where a CD_{min} feature images. To increase features on a wafer, CD_{min} is pushed smaller, but to reduce manufacturing costs, DoF must increase, therefore equations 1 and 2 represent competing considerations.

The CC layer angle (Fig. 5) and the 8 nm space between the line pair presents a difficult imaging problem. EUV systems use reflective optics, which require mask illumination at an angle [6]. The illumination angle imposes preferential printing in the X and Y directions and degrades imaging for intermediate directions including the CC layer angle.

Simulation results for 0.55 NA EUV imagining of the CC layer using a single exposure (SE) with 2 Sub-Resolution Assist Features (SRAF) per side results in a Normalized Image Log Slope (NILS) of approximately 1.4 (Figs. 6 and 7). NILS is a measure of manufacturability where a minimum value of 2.0 is regarded as acceptable with sufficient manufacturing process margin [7]. Using additional lithography techniques such as OPC and SMO, Hi NA EUV should image CC using a single exposure, if the Hi NA system is released within target 2025 to 2027 release window [8].

IV. Stress Analysis

Growing epitaxial SiGe S/D from 3 seeds (Si wafer at the bottom and 2 adjacent PMOS channels on the sides), produces beneficial compressive channel stress of -1 GPa. However, it is difficult to avoid dislocations and {111} stacking faults with the epitaxy merging from several sides, and a pair of {111} stacking faults reduces stress to -720 MPa (Fig. 8). This is the stress level used for SRAM performance analysis. One important implication of a pair of stacking faults shifting the stress by about 280 MPa is that it creates a new transistor performance variability mechanism quantized in ~15% Idsat increments, but is likely limited to 2 or 3 values.

Growing SiGe epitaxy from the sides only gives detrimental tensile channel stress of +450 MPa, mainly from sacrificial SiGe layers surrounding Si channel (Figs. 9 and 10).

V. SRAM Bitcell and SRAM Array Analysis

Area reduction is the major CFET benefit in the SRAM bitcell, However, it also must achieve high yield and match or exceed state of the art performance. This was analyzed with Boltzmann transport model to evaluate transistor behavior at the nominal conditions and at the process corners [9]. Static Noise Margin (SNM) analysis was performed at the Fast/Slow (FS) process corner at 125°C and included 6σ local variability offsets to verify high-sigma yield at the worst-case Process-Voltage-Temperature (PVT) conditions. Analysis of SNM as a function of core (VDDC) and periphery (VDDP) voltages showed the cell functions down to 0.57V (Fig. 11), and Fig. 12 shows SNM curves with and without local variability offsets. For this CFET SRAM cell, writability was not found as a limiting factor due to the 1:2 PU:PG ratio.

Array-level behavior was also considered, with read-time simulations performed at a range of different configurations ranging from 32x32 cells, to 256x256 cells. The read analysis showed that both the word line length (*ncolumns*) and the bitline length (*nrows*) have similar impact on read delay (Fig. 13). An array of 256x256 cells is viable, with access times of 0.3ns. For reference this was compared with a 3nm BPR design [10], where similar speed was achieved for a smaller array with half *nrows*.

VI. Transistor Density and Cost Analysis

Moore's Law is driven by reducing cost per transistor as an absolute requirement for moving to the next technology generation. Therefore, technology cost analysis was performed based on the methodology from IC Knowledge LLC (ICK) [11]. ICK data on transistor density is shown on Fig. 14. A 0.7x scaling is used to establish a consistent node progression. The logic transistor density was calculated based on 2-input NAND gate and Scanned Flip Flop (SFF) standard cells weighted 60% and 40% respectively. A, B and C represent the three industry leaders. An important observation is that transistor density grows at 45% per node, including 1.25nm node.

Wafer cost analysis shows a consistent increase of 13% per node (Fig. 15). This was based on ICK Strategic Cost and Price Model except for 1.25nm node where ICK Cost Explorer Model is applied due to its compatibility with Synopsys Process Explorer [11]. Wafer cost was calculated for the core transistors through Metal 1 only, no other devices or higher-level interconnects are considered. Because mask usage and amortization vary so much between foundry and IDM products, mask set amortization is not included. These omissions make for a consistent comparison but flatten the cost curve versus full process flows. The curve is also flattened for the 5nm node due to increased EUV usage reducing costs versus multi-patterning. In each case fab locations and capacities were selected consistent with typical production fabs and the companies being modeled.

Combining these data on transistor density and wafer cost, a consistent transistor cost reduction of 32% per node was found, including the 1.25nm node based on this work (Fig. 16). Figure 16 does not include design costs, which require high volume manufacturing to achieve profits [1]. Note the reported increases of transistor density and reduction of transistor cost occur despite disappearing feature scaling due to the combined efforts of DTCO and EDA (Fig. 17).

VII. Conclusions

In this work, traditional Power-Performance-Area (PPA) analysis was combined with the wafer and transistor cost analysis, making it PPA+C. As an example of DTCO extending transistor density gains towards 1nm node, a CFET architecture for logic and SRAM was characterized. The CFET exhibits an SRAM transistor density of 1 billion transistors per square millimeter and continues the pace of 32% reduction of cost per transistor through the 1nm node that corresponds to 2028 – 2030 timeline.

References

[1] G. Yeric, IEDM Tech. Dig., pp. 1-4, 2015.
[2] A. de Geus, Semicon West keynote, 2020.
[3] S.-Y. Wu, IEDM Tech. Dig., pp. 864-867, 2019.
[4] J. Smith, SPCC Proceedings, 2019.
[5] L. Rayleigh, J. Royal Microscopical Soc., v. 23, pp. 447-473, 1903.
[6] A. Hawryluk et al., Solid State Tech., v. 40, no. 7, p. 151, 1997.
[7] C. Mack, www.lithoguru.com/scientist/litho_tutor/TUTOR32, 2001.
[8] J. van Schoot et al., SPIE Proceedings, 1132307, 2020.
[9] S.C. Song et al., VLSI Tech. Proceedings, pp. 206-207, 2019.
[10] S. Salahuddin et al., VLSI Tech. Proceedings, 2020.
[11] www.icknowledge.com

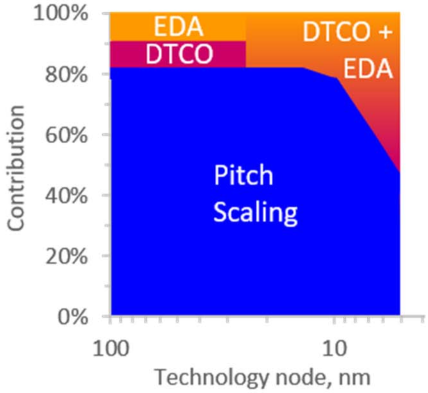
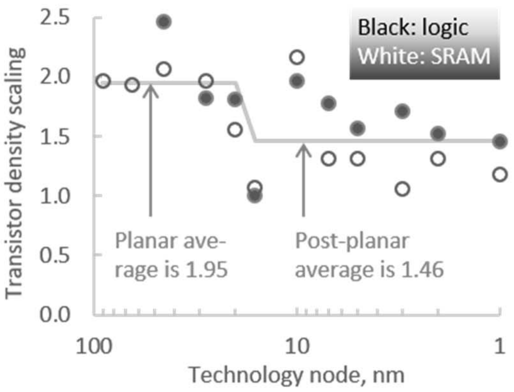


Fig. 1. Evolution of the key vehicles of transistor density scaling.

Fig. 2. Evolution of transistor density scaling between nodes for logic and SRAM [3]. Averages are shown for a 50/50 logic/SRAM mix



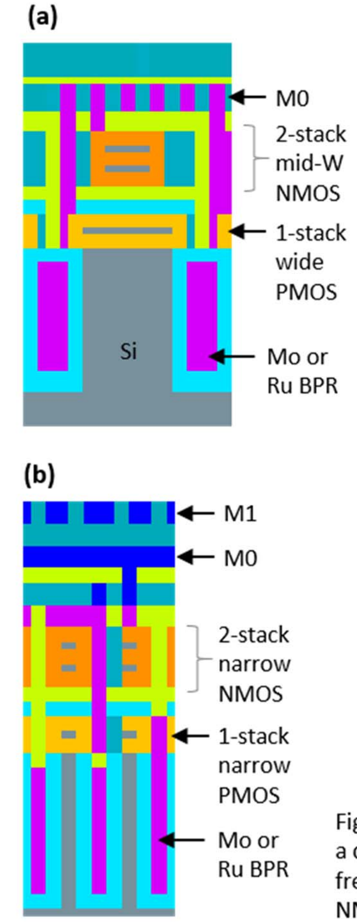


Fig. 3. CFET architecture for logic (a) and SRAM (b) that enables 4 track logic and 4 track SRAM bitcell while balancing both with ~1:1 NMOS/PMOS strength ratio for logic and desirable 2:1 N/P ratio for the SRAM bitcell.

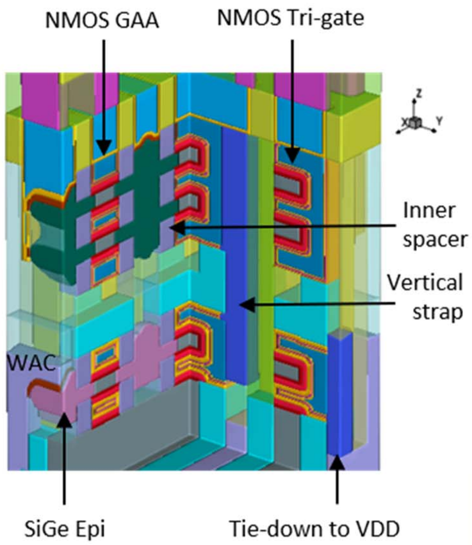


Fig. 4. CFET SRAM bitcell architecture with a densely populated volume, where stress-free NMOS is above strained PMOS. The NMOS/PMOS pair occupies footprint of 39 nm by 38 nm (1 CPP by 2 MP).

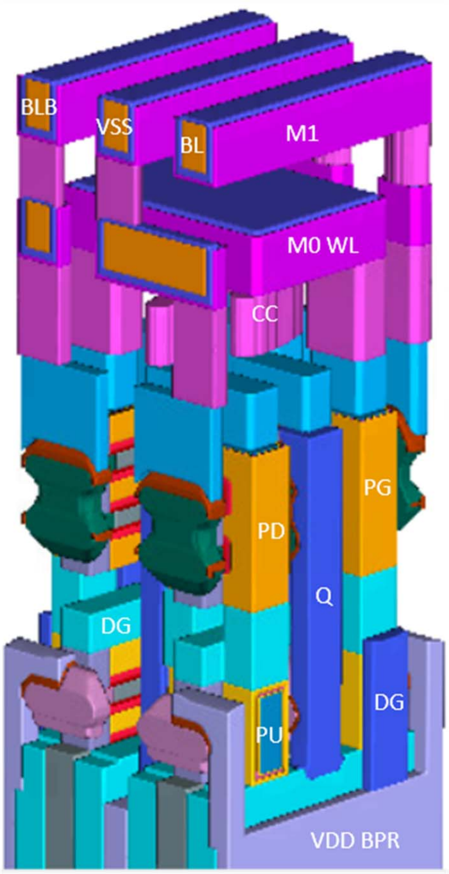


Fig. 5. SRAM bitcell with 1 billion transistors per square millimeter at CPP/MP of 39nm/19nm.

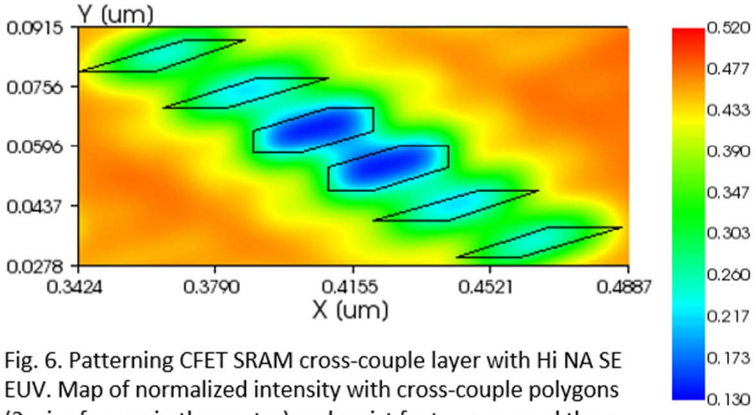


Fig. 6. Patterning CFET SRAM cross-couple layer with Hi NA SE EUV. Map of normalized intensity with cross-couple polygons (2 wire frames in the center) and assist features around them.

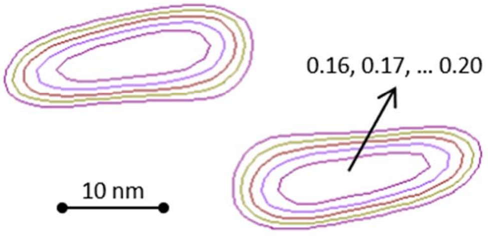
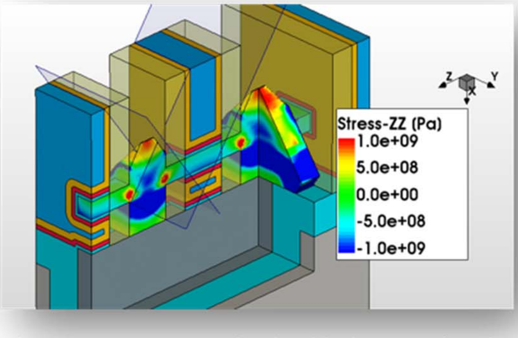


Fig. 7. Zoom into cross-couple light intensity contours that define photo-resist shapes in the layout plane.



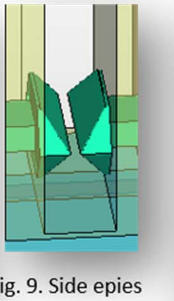


Fig. 9. Side epies before they merge.

Fig. 8. PMOS stress map for the side+bottom S/D epitaxy with 2 {111} stacking faults. Compressive channel stress is -720 MPa.

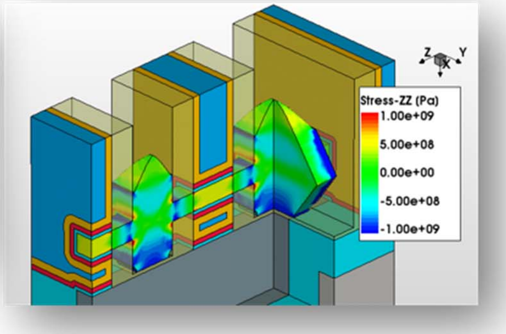


Fig. 10. PMOS stress map for the side-only S/D epitaxy. Tensile channel stress is +450 MPa.

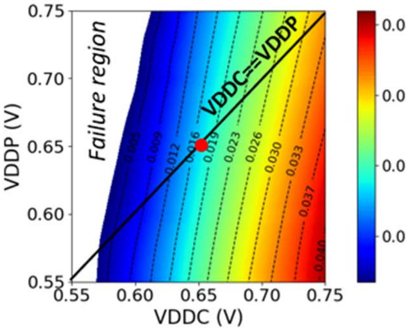


Fig. 11. SNM at FS, 125C of a 6σ cell. VDDC is bitcell voltage, VDDP is peri. voltage.

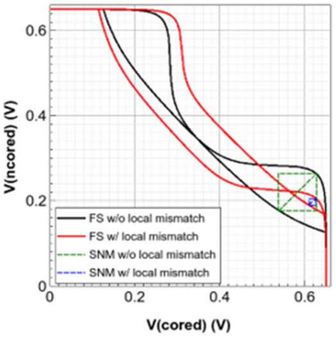


Fig. 12. SNM at FS, 0.65V 125C with and w/o local variability.

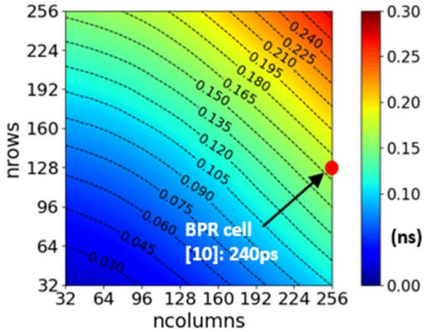


Fig. 13. Read time (wordline to sense-amp) at TT, 0.65V 27C for far bitcell scaling with array size.

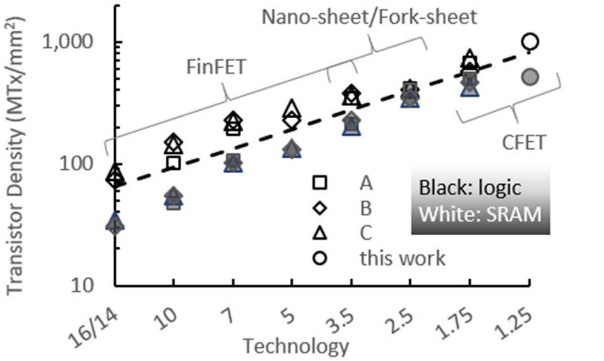


Fig. 14. Evolution of transistor density for logic and HD SRAM (45% increase per node for 50/50 SRAM/logic mix: dashed line).

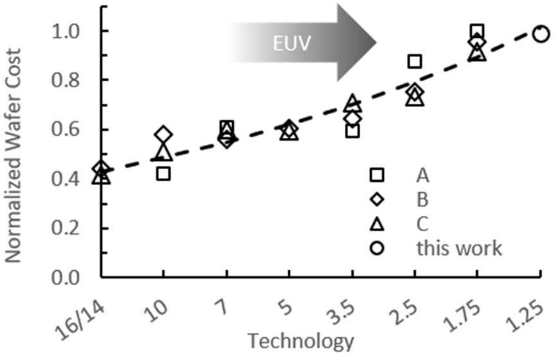


Fig. 15. Evolution of wafer cost (13% increase per node). Wafer cost is calculated for core transistors through M1.

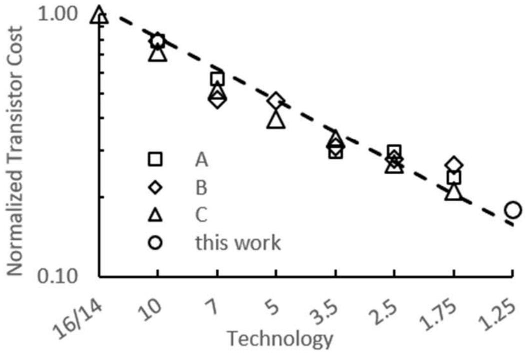


Fig. 16. Evolution of transistor cost for a 50/50 SRAM/logic mix (32% reduction per node).

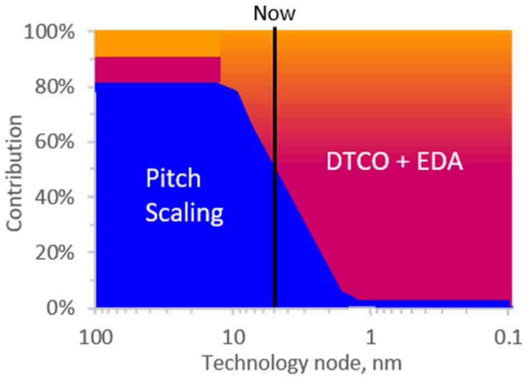


Fig. 17. Evolution of the key vehicles of transistor density scaling in semiconductor industry extended beyond 1nm.