29.1 A 38.1Mb/mm^2 SRAM in a 2nm-CMOS-Nanosheet Technology for High-Density and Energy-Efficient Compute

Tsung-Yung Jonathan Chang, Yen-Huei Chen, K. Venkateswara Reddy, Nikhil Puri, Teja Masina, Kuo-Cheng Lin, Po-Sheng Wang, Yangsyu Lin, Chih-Yu Lin, Yi-Hsin Nien, Hidehiro Fujiwara, Ku-Feng Lin, Ming-Hung Chang, Ching Wei Wu, Robin Lee, Yih Wang, Hung-Jen Liao, Quincy Li, Ping Wei Wang, Geoffrey Yeap

TSMC, Hsinchu, Taiwan

Embedded memories are crucial SoC design components; among these, SRAM plays a vital role in enhancing system performance across various applications. The ongoing demand for high-capacity on-die SRAM necessitates optimal-density scaling as we transition technology nodes. In mature technology nodes, reducing the bit cell area significantly contributed to SRAM scaling. However, as we move into more advanced technology nodes, scaling the cell area becomes increasingly challenging. Design-technology co-optimization (DTCO) becomes essential to achieving further area scaling at the chip level. We focus on optimizing both cell and peripheral designs to improve memory density. Our SRAM design leverages the unique characteristics of 2nm nanosheet technology. We also explore various design domains, such as SRAM macro architecture, design assist techniques, and floor planning. The primary objective of this paper is to minimize the periphery while maximizing the bit cell array size. This is achieved by increasing the number of bit cells per BL, as the 2nm nanosheet technology improves the cell's on-to-off current ratio. This advancement allows for a 2× increase in the maximum BL loading compared to the previous technology. Additionally, we implement special logic rules, for the peripheral logic, to further optimize its area, as illustrated in Fig. 29.1.1(a).

To address memory density, a high-density (HD) SRAM design using 2nm nanosheet technology is presented; the presented SRAM macro has a capacity of 580kb (4096×145) using cells with a size of 0.021µm^2. By using DTCO the overall SRAM density is improved by 10% compared to the previous technology node: resulting in a 38.1Mb/mm^2 density, as shown in Fig. 29.1.1(b). To achieve a lower minimum-write-voltage (V_{MIN}) operation, the negative BL (NBL) technique is employed for write assist [1-5]. Figure 29.1.2(a) illustrates a traditional SRAM macro design using a FinFET technology. With FinFETs, the maximum number of cells per BL is limited to 256. In contrast, the 2nm nanosheet technology allows an increase to 512 cells per BL due to the improved on-current to off-current ratio of the bit cell. This enhancement significantly boosts the cell efficiency of the SRAM macro. Additionally, by increasing the BL capacity to 512 cells and adopting the flying BL (FBL) architecture, the array efficiency improved. Figure 29.1.2(b) shows the FBL macro architecture, which features 512 rows of cells, for both the top and bottom banks. This is a substantial improvement to a conventional design, which can accommodate only for 256 rows using a FBL architecture. The top-bank BLs are connected to the main IO (MIO) block via the FBL metal 2 over the bottom bank, creating a 1024 pseudo-row architecture. However, increasing the number of cells per BL and adding the FBL over the bottom bank results in a higher BL resistance and capacitance for both the top and bottom banks. Implementing the 1024 pseudo-row architecture, with 512 rows per BL and 512 rows on the FBL, presents several key challenges due to the significant increase in BL resistance and capacitance: (1) a greater NBL loss at the far-end of the BL due to the higher BL resistance; (2) an increased BL boost capacitance; (3) a longer BL pre-charge time. To address these challenges, we propose placing the write assist and BL pre-charge blocks at the far side of the array. This enhances the writability and pre-charge strength for the far-end cells.

Figure 29.1.3(a) illustrates the proposed far-end write-assist (FE-WA) and far-end pre-charge (FE-PRE) schemes, aiming to extend the number of cells per bit line to 512. To mitigate far-end cell write degradation, the FE-WA and FE-PRE blocks are placed on top of the top and bottom banks, respectively. The top-bank BL uses metal 2 (FBL) to fly over the bottom bank to drive the MIO block; the MIO's write driver is uses NBL for write assist. Typically, the NBL boost capacitor is a MOS capacitor, which is used to generate the coupling voltage with negative bias (NVSS). The NBL bias signal, generated in the MIO, must travel over the bottom bank to reach the FE-WA block for both top and bottom banks. Metal 4 tracks are used as the metal coupling capacitance to transfer the NVSS voltage to the FE-WA blocks. In the FE-WA block, a pair of NMOS write drivers is controlled by DT and DC, which is the data to be written and its compliment. The write driver's source terminals are connected to NVSS to propagate the negative bias to the cells. Another NMOS pair, in series with the write driver's drain, serves as the write column-multiplexing selectors. The gate of this NMOS pair is controlled by the column address (Y[0], …Y[n]) to enable the selected write columns. During write-0, the write driver in the MIO block is triggered by DT = 0 and DC = 1 write data signals and the WPB-selection signal to start pulling down BL[0]. The FE-WA block is also activated by DT = 0 and DC = 1 signals to assist in discharging the far-side BL[0] to achieve the required negative BL level. Next, the NBL_ENB signal activates NBL boost signal to couple the MOS and metal-4 capacitance to generate the negative bias signal NVSS. This negative bias signal (NVSS) propagates to both near-side and far-side through the NMOS write driver pairs to the selected BL. After write completion, the BL is pre charged to V_{DD}, terminating the write cycle. To enhance the write cycle time, the FE-PRE block, equipped with a pair of PMOS pre-charge and equalizing transistors, assists in restoring the BL back to V_{DD}. Figure 29.1.3(b) shows a diagram of the global signals that control the FE-WA and FE-PRE blocks. To activate the FE-WA block, column-selection signals are transmitted using metal-4 tracks from the control block (CNT) to the FE-WA using local buffers to aid signal reconstruction. Additionally, the write-data signals (DT and DC), which are the latched data-to-be-written signals, also travel over the arrays using metal 4 to the FE-WA. The FE-PRE block is enabled by the BL pre-charge signal (BLPRE), which also uses metal 4 to propagate to the FE-PRE block.

Figure 29.1.4(a) shows the simulated waveforms, when the FE-WA and FE-PRE blocks are disabled. Due to the high BL time constant, the far-side BL fails to reach the required NBL voltage, when only the near-side NBL is activated. Consequently, the far-side cells experience write failures. Additionally, the high BL time constant increases the pre-charge time needed to restore the BL to V_{DD}. In contrast, Fig. 29.1.4(b) presents simulation waveforms, with the FE-WA and FE-PRE blocks enabled: the NBL bias signal (NVSS) is able to propagate to the far-side cells; thus, these cells achieve the required NBL voltage required for a successful write operation. Moreover, enabling the FE-PRE block aids in a faster BL recovery to V_{DD}, resulting in a ~2× improvement in BL pre-charge time.

In addition to high-density SRAM, a double-pumped SRAM with a high-current (HC) cell is another critical enabler for high-performance computing (HPC) applications. To improve energy efficiency, a dual-tracking scheme, illustrated in Fig. 29.1.5, is implemented to reduce active power and boost speed. At a lower V_{DD} the tracking scheme ensures a sufficient read margin (RM) for the SRAM to operate at V_{MIN}. For the nominal V_{DD} range, the design switches to TURBO mode, bypassing the tracking scheme and using the pure logic-delay path. This TUBO mode switch improves maximum operation frequency (f_{MAX}) and prevents excessive RMs when operating from a nominal V_{DD}. The proposed dual-tracking scheme enables the double-pumped SRAM to achieve a 6.3% speed increase, and a 11.5% reduction in active power, compared to its 3nm counterpart; this results in a 20% energy improvement.

Silicon test-chip results for the 2Mb HD SRAM's V_{MIN} are shown in Fig. 29.1.6(a) at 25°C: including four 580kb SRAM macros that are configured as 4096×145 mux-4 with pseudo-1024cells/BL. Figure 29.1.6(b) shows the 256Mb HD SRAM's V_{MIN} at 25°C, using 2048 SRAM macros configured as 4096×32 mux-16 with 256cells/BL. Applying write-assist improves both 2Mb and 256Mb SRAM V_{MIN} by 300mV at the 95th percentile, in comparison to without write-assist. Figure 29.1.6(c) illustrates the frequency Shmoo for the double-pumped 32kb SRAM at 25°C, which is configured as 512×64 mux-4. The proposed dual-tracking scheme enables f_{MAX} = 4.2GHz using a 1.05V supply.

Figure 29.1.7 shows the SRAM test chips, along with their respective key-summary information. One test chip incorporates four 580kb SRAM macros (configured as 1024 pseudo cells/BL), features a far-end write-assist and pre-charging schemes for post-silicon tuning. The test chip has a 2Mb total capacity. The other test chip includes 2048 SRAM macros, each with a 128kb capacity (configured as 256cells/BL), with a 256Mb total capacity. This chip is equipped with redundancy and write assist options. Both test chips were manufactured using 2nm CMOS nanosheet technology.

Acknowledgement:
The authors thank the RD team for wafer manufacturing, and the test department for chip measurements for this work.

2025 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3315-4101-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISSCC49661.2025.10904759

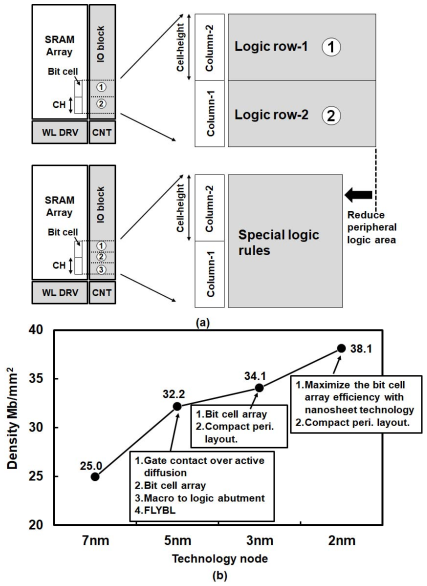


Figure 29.1.1: (a) Special logic rules are used for peripheral logic-area reduction, and (b) technology trendline for HD SRAM bit density.

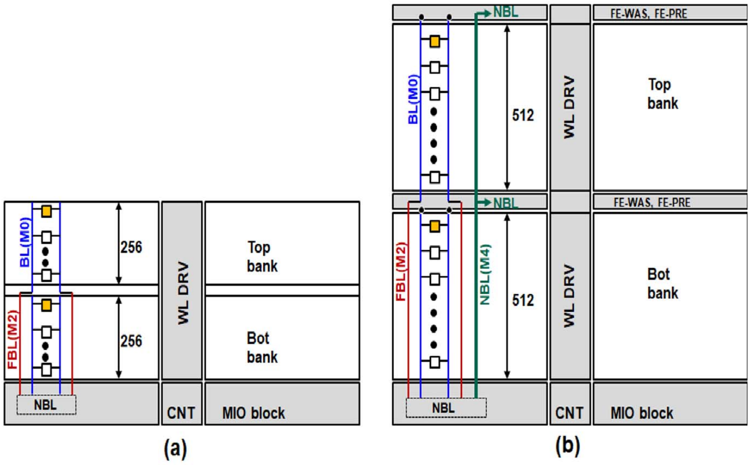


Figure 29.1.2: (a) A typical FinFET SRAM macro design, and (b) the FBL macro architecture that features 512-row top and bottom banks.

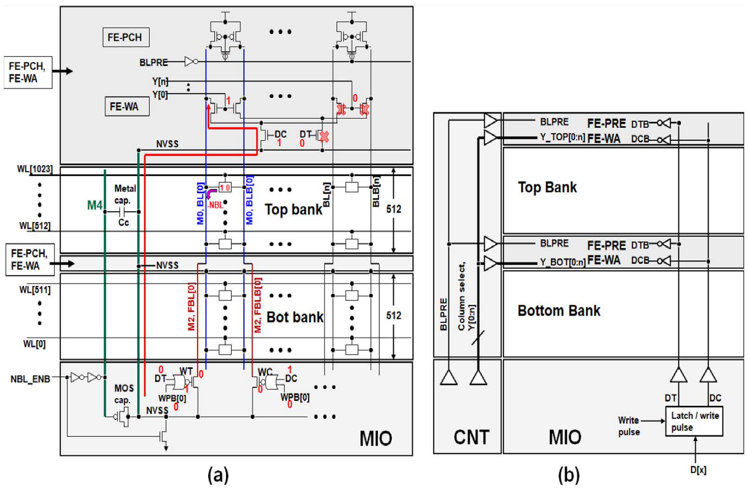


Figure 29.1.3: (a) The proposed far-end write-assist (FE-WA) and far-end pre-charge (FE-PRE) schemes that enable increasing the number of cells per bit line to 512. (b) Block diagram showing the global signals, which control FE-WA and FE-PRE blocks.

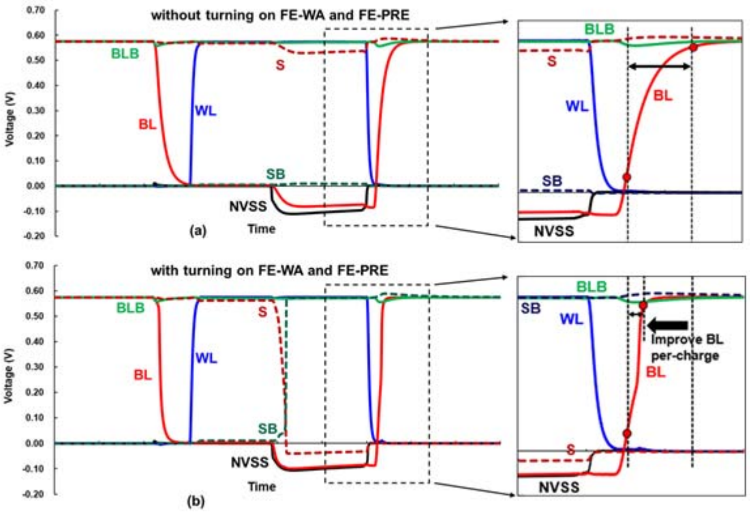


Figure 29.1.4: (a) Simulated waveforms with the FE-WA and FE-PRE disabled, and (b) with the FE-WA and FE-PRE enabled.

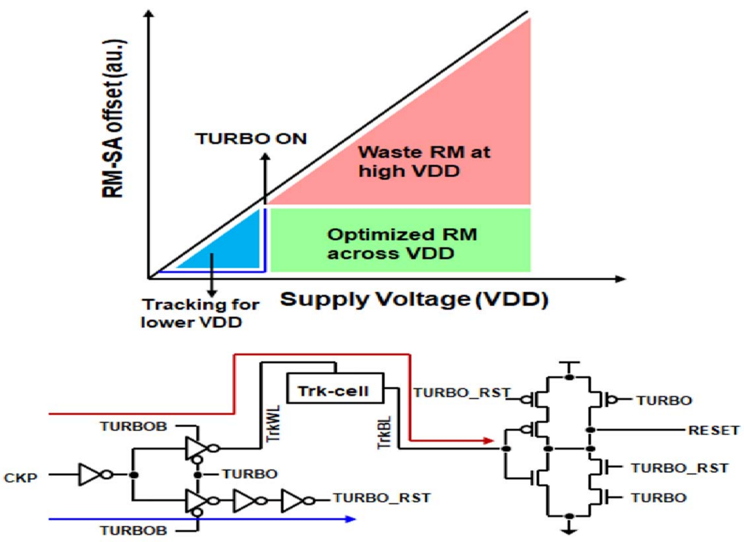
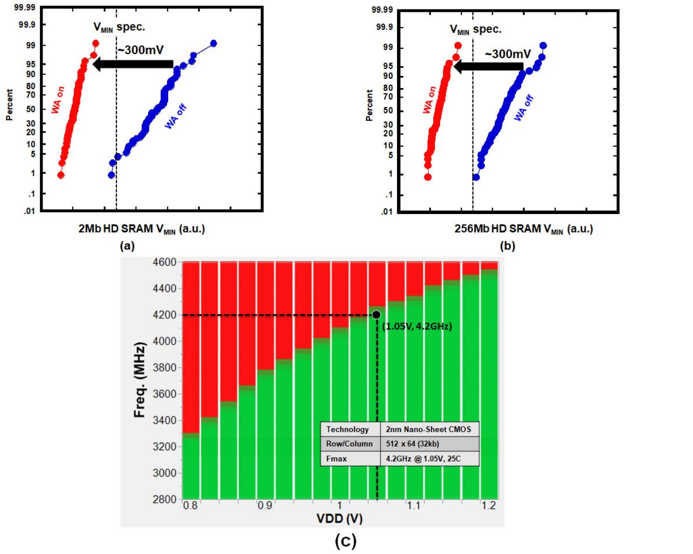
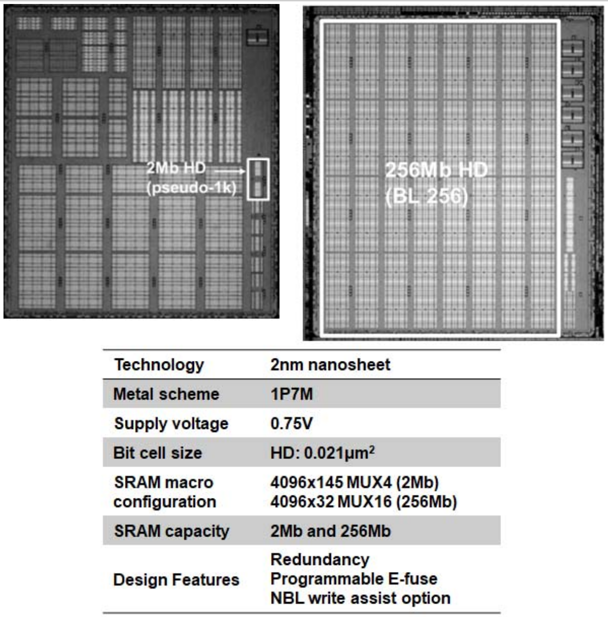


Figure 29.1.5: The dual-tracking scheme for double-pumped SRAM.



Technology	2nm Nano-Sheet CMOS
Row/Column	512 x 64 (32kb)
Fmax	4.2GHz @ 1.05V, 25C

Figure 29.1.6: (a) 2Mb HD-SRAM V_{MIN} 25°C silicon results for a 1024 pseudo-cell architecture. (b) 256Mb HD-SRAM V_{MIN} cumulative distribution for conventional 256 cells/BL, and (c) frequency/voltage Shmoo for double pumped SRAM.



Technology	2nm nanosheet
Metal scheme	1P7M
Supply voltage	0.75V
Bit cell size	HD: 0.021μm^2
SRAM macro configuration	4096x145 MUX4 (2Mb) 4096x32 MUX16 (256Mb)
SRAM capacity	2Mb and 256Mb
Design Features	Redundancy Programmable E-fuse NBL write assist option

Figure 29.1.7: Test-chip micrograph and key-metric summary table.

References:

[1] J. Chang et al., "A 3nm 256Mb SRAM in FinFET Technology with New Array Banking Architecture and Write-Assist Circuitry Scheme for High-Density and Low-VMIN Applications", *IEEE VLSI Symp.*, 2023.

[2] J. Chang et al., "A 5nm 135Mb SRAM in EUV and High-Mobility-Channel FinFET Technology with Metal Coupling and Charge-Sharing Write-Assist Circuitry Schemes for High-Density and Low-VMIN Applications," *ISSCC*, pp. 238-239, 2020.

[3] J. Chang et al., "A 7nm 256Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-VMIN Applications", *ISSCC*, pp. 206-207, 2017.

[4] T. Song et al., "A 7nm FinFET SRAM using EUV lithography with dual write-driver assist circuitry for low-voltage applications", *ISSCC*, pp. 198-200, 2018.

[5] Y. Kim et al., "Energy-Efficient High Bandwidth 6T SRAM Design on Intel 4 CMOS Technology", *IEEE VLSI Symp.*, pp. 212-213, 2022.