A 38.1Mb/mm2 SRAM in 2nm CMOS Nanosheet Technology for High Density and Energy Efficient Compute Applications

Nandini Kumawat

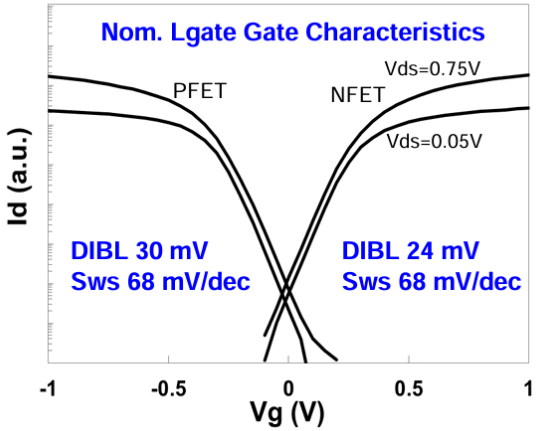
UNIVERSITY OF MINNESOTA
Driven to Discover®

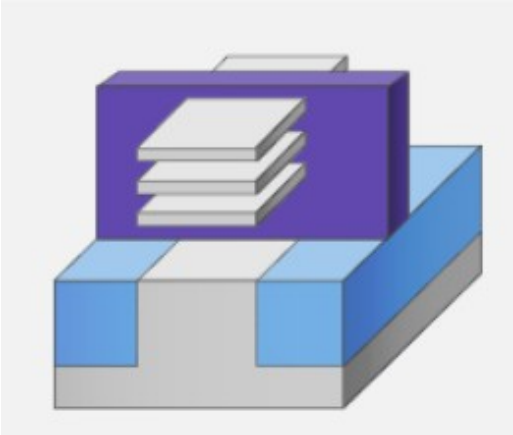
Outline

- 2nm Technology Highlight
- SRAM Bit Cell and scaling trend
- SRAM design techniques
  - Maximize bit cell array efficiency
  - Far-End (FE) - Write Assist
  - FE-Pre-Charger
  - Double Pump SRAM Design for HPC
- Silicon results
- Summary

2nm Technology Overview

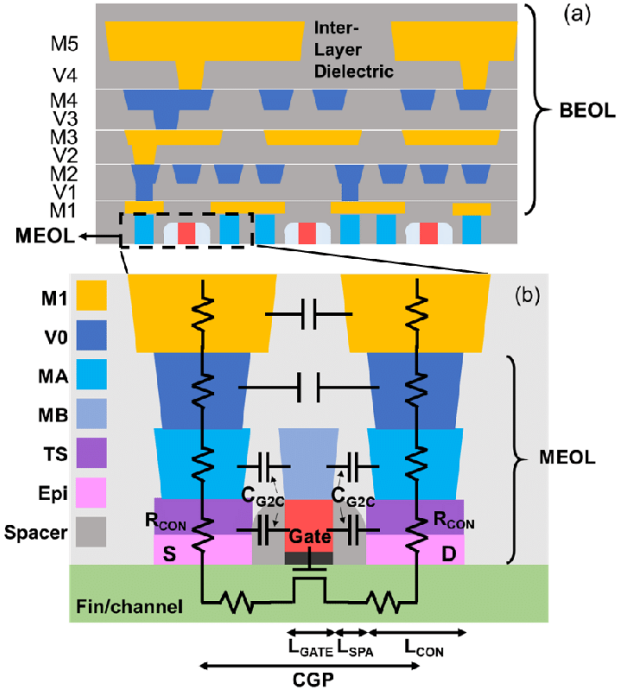
- Successfully transitioned to Nanosheet (NS) era
- Defined and optimized with 3DIC for energy efficiency compute and system integration.
- Energy efficient NS transistors with excellent drives, DIBL and Subthreshold Swing



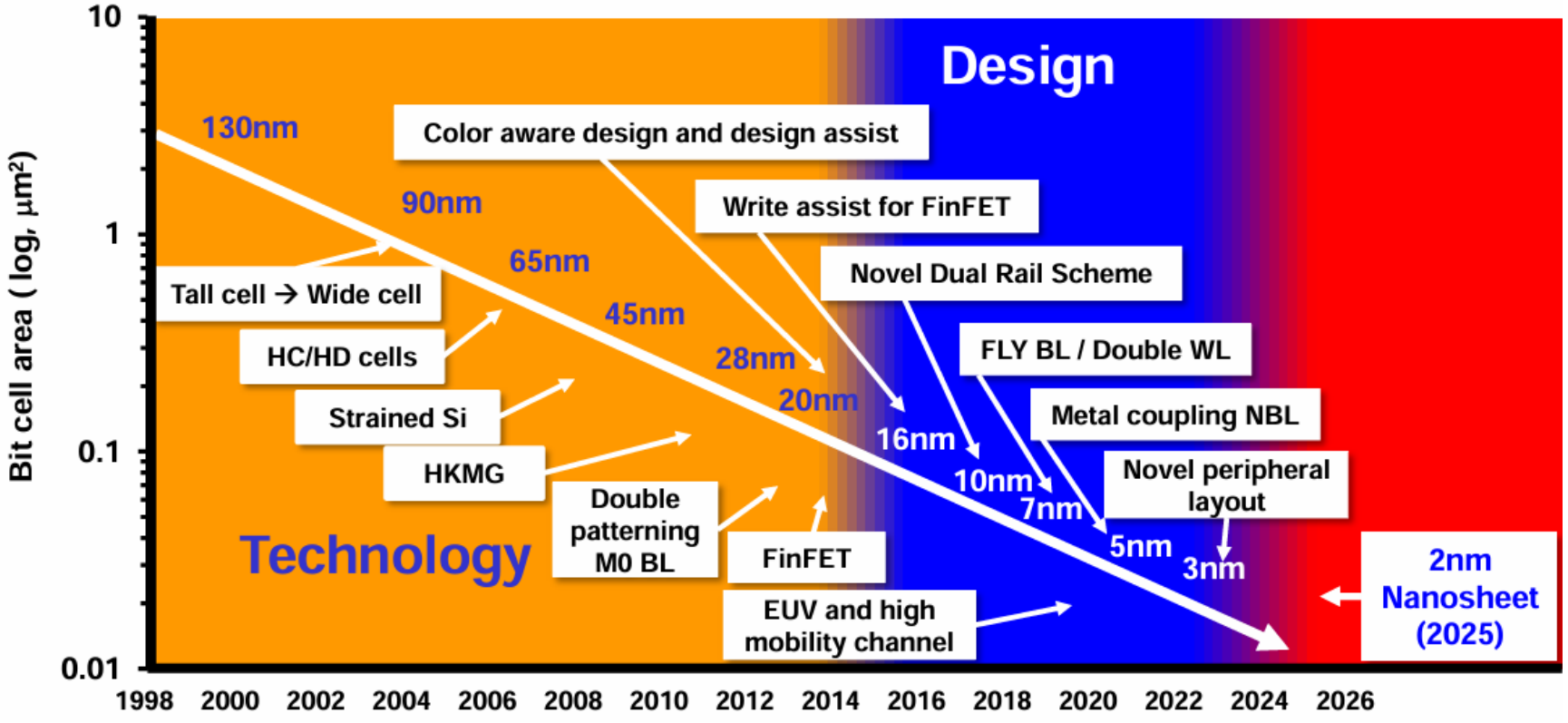


2nm Technology Overview

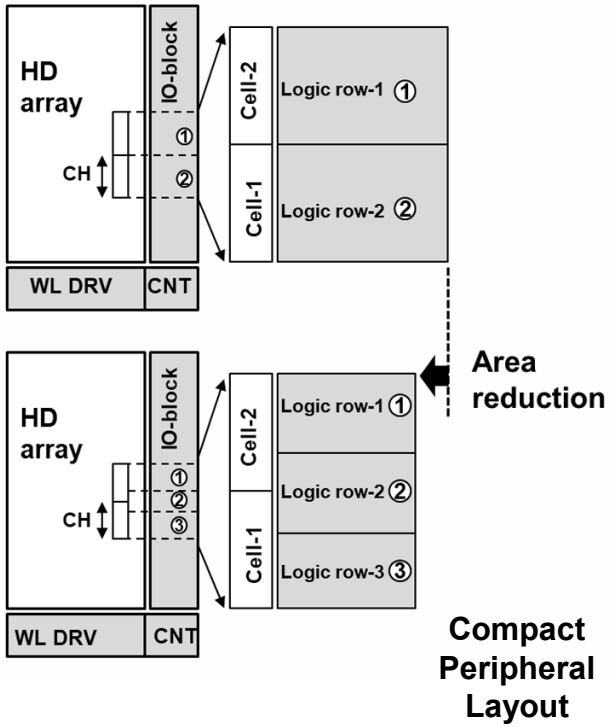
- MEOL/BEOL offers 20% RC reduction and optimized design rules for densest design



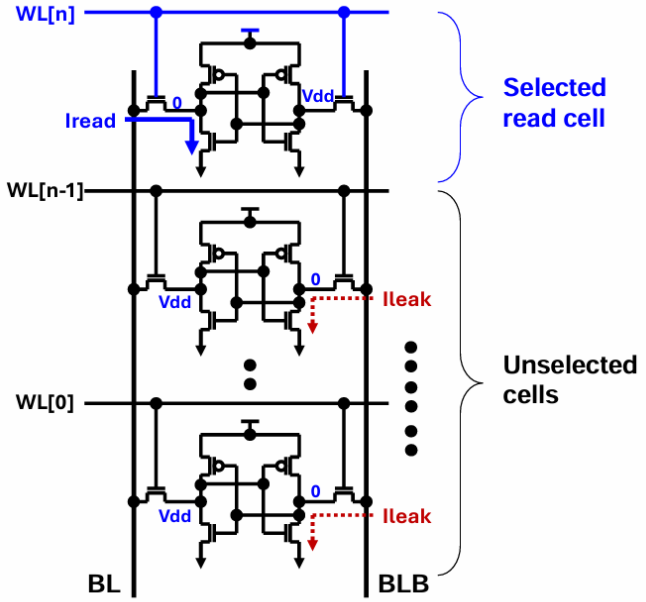
SRAM Bit Cell Scaling Trend



Peripheral Layout for HD Cell



- Cell-area scaling becomes harder at advanced nodes, demanding DTCO.
- Need co-optimization of bitcell and periphery to maintain density improvements.
- 2 nm nanosheets offer higher Ion/Ioff, allowing longer bitlines without margin loss.
- This work reduces periphery area and increases array utilization.
- Supports up to **2 x larger BL loading** than 3 nm while keeping timing stable.
- Requires special logic design rules to handle increased BL length.



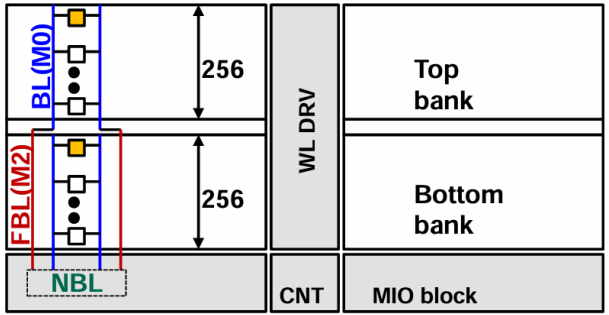
$$SN\ ratio = \frac{I_{on}}{I_{off}} = \frac{I_{read}}{n * I_{leak}}$$

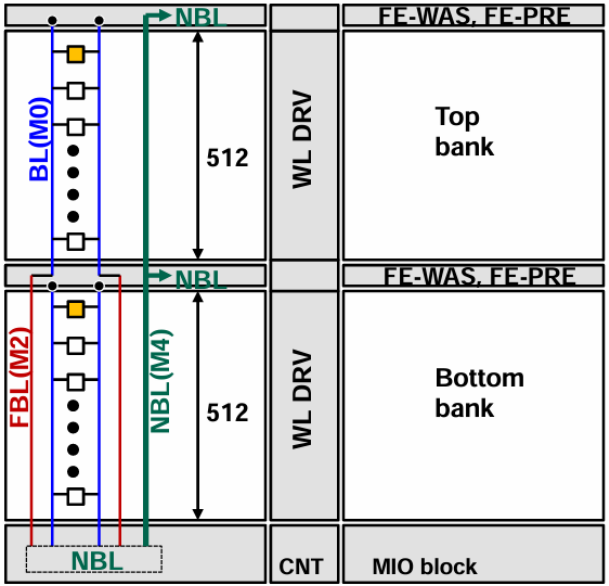
FBL Macro Architecture

- FinFET SRAM limits bitline length to **256 cells per BL** due to weaker Ion/Ioff and higher leakage.
- 2 nm nanosheet SRAM supports **512 cells per BL.**
- Higher Ion improves BL discharge speed, and lower Ioff limits leakage from unselected cells, enabling longer BLs without margin loss.

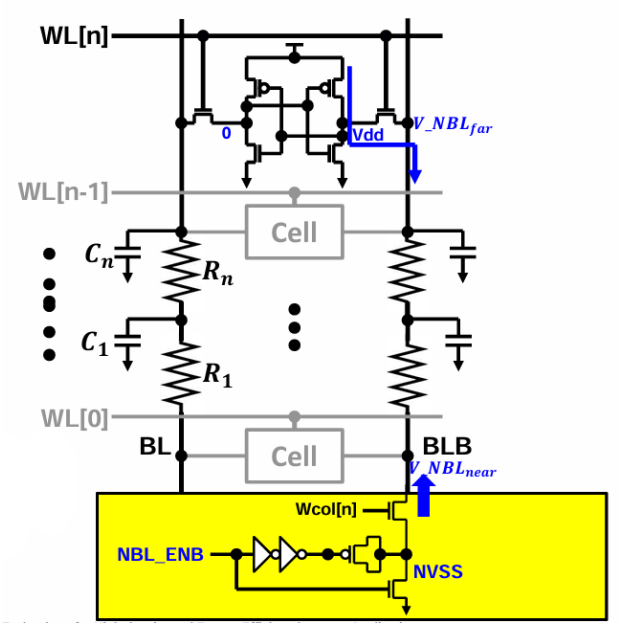
SRAM Macro: 580 kb
Cell size: 0.021 μm²
Density: 38.1 Mb/mm²

- Increasing BL capacity to 512 cells boosts **array efficiency.**
- Using **flying bitline (FBL)** routing, the top bank's bitlines connect to the Main IO (MIO) block using Metal-2 over the bottom bank.
- This creates a **1024-row pseudo-BL** structure, improving density and simplifying global routing.



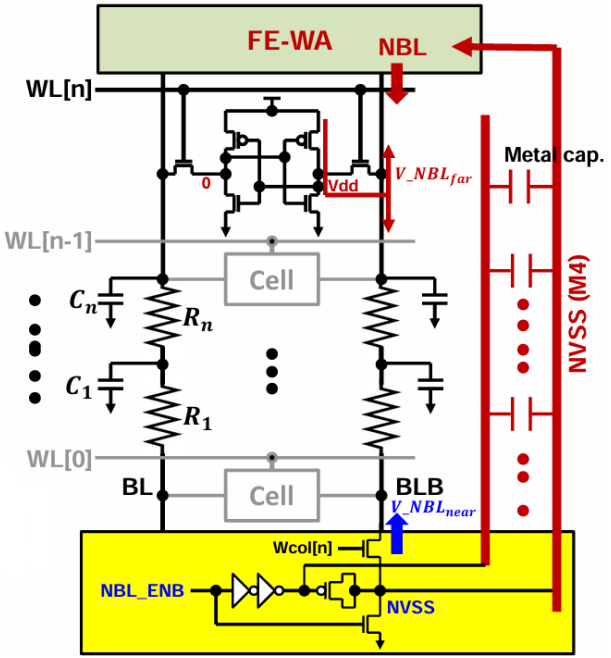


Challenge



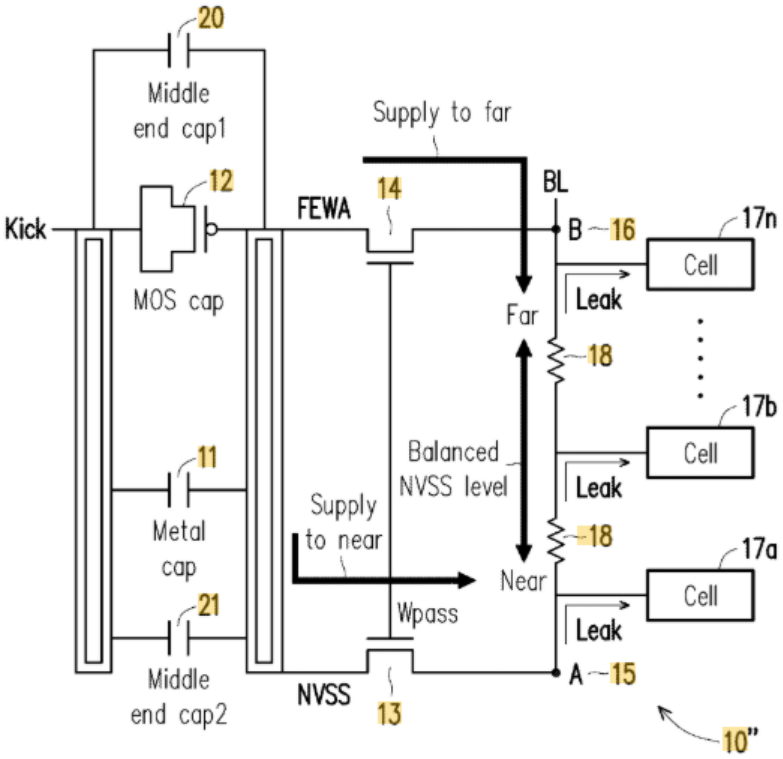
$$|V\_NBL_{far}| \ll |V\_NBL_{near}|$$

Mitigation



$$|V\_NBL_{far}| \cong |V\_NBL_{near}|$$

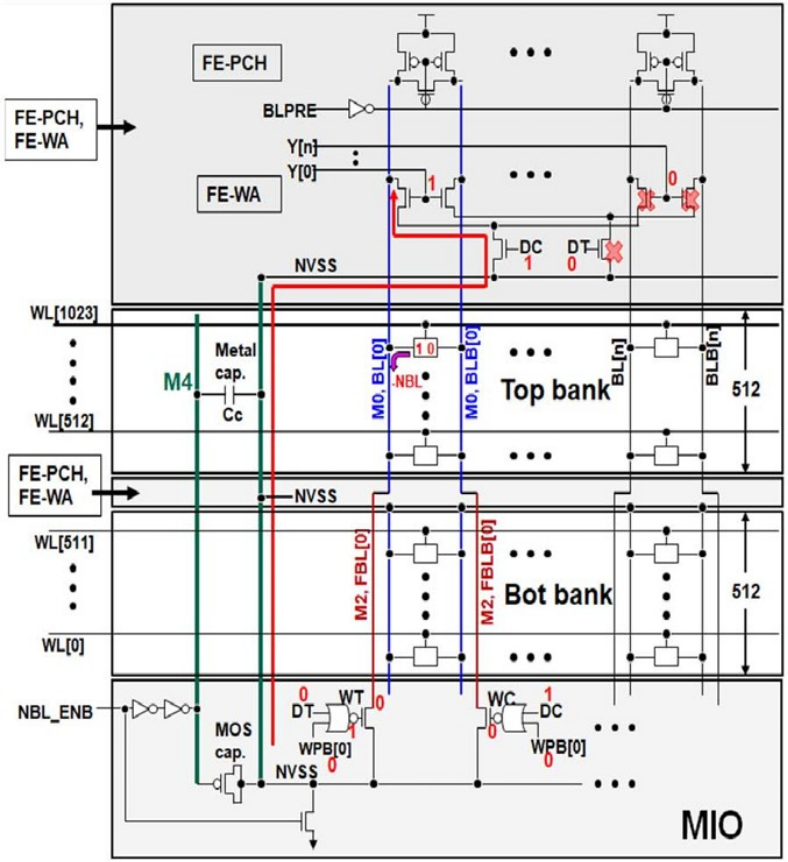
Balanced Negative Bitline Voltage for a Write Assist Circuit

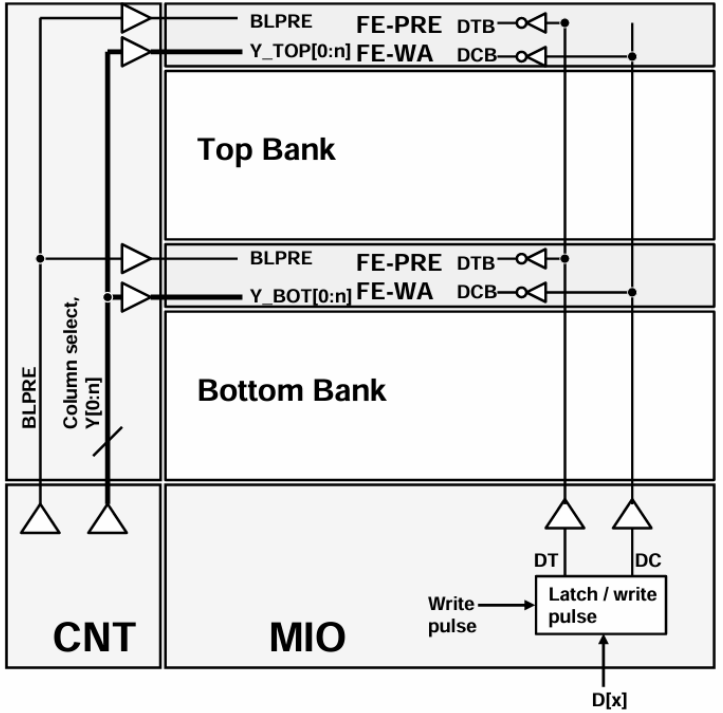


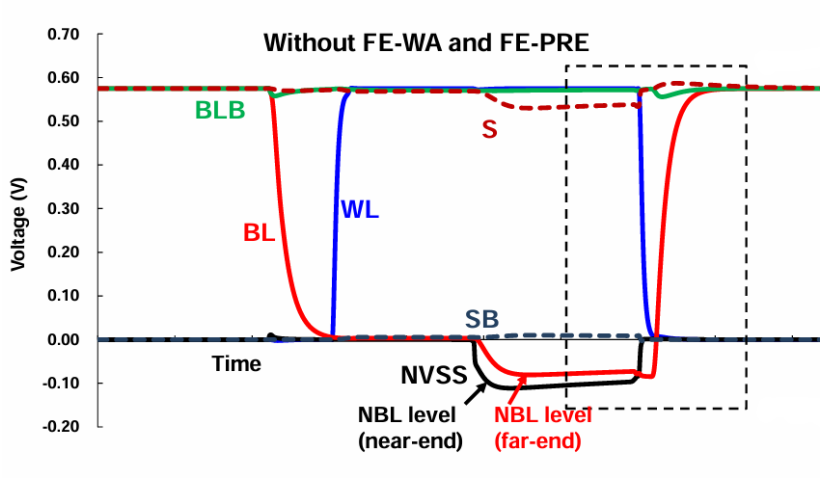
- You tie a MOS capacitor and a metal capacitor in parallel to boost total capacitance.
- You connect their plates to the bitline through switch transistors at the near end and far end.
- You apply a falling negative kick to the first plates of both capacitors.

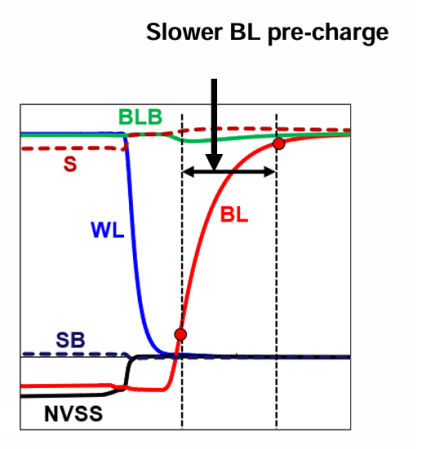
Source: https://patents.google.com/patent/US20210201977A1/en

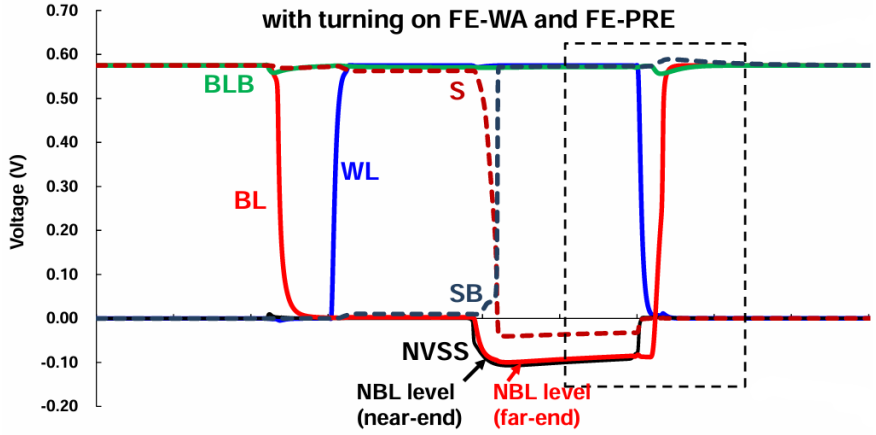
$$V_B = -\frac{C_c}{C_c + C_{par}} \cdot V_{DD}$$

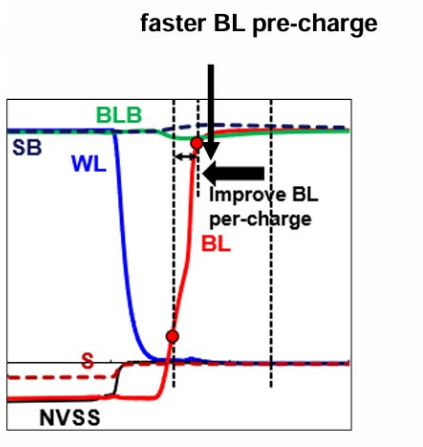

FE-PCH
FE-PCH, FE-WA
BLPRE
Y[n]
Y[0]
FE-WA
NVSS
DC
DT
WL[1023]
WL[512]
Metal cap.
Cc
M4
M0, BL[0]
M0, BLB[0]
NBL
Top bank
BL[n]
BLB[n]
512
FE-PCH, FE-WA
NVSS
WL[511]
WL[0]
M2, FBL[0]
M2, FBLB[0]
Bot bank
512
NBL_ENB
MOS cap.
NVSS
DT
WT
WPB[0]
WC
DC
WPB[0]
MIO


BLPRE
Y_TOP[0:n]
FE-PRE
FE-WA
DTB
DCB
Top Bank
BLPRE
Y_BOT[0:n]
FE-PRE
FE-WA
DTB
DCB
Bottom Bank
BLPRE
Column select, Y[0:n]
DT
DC
Write pulse
Latch / write pulse
D[x]
CNT
MIO


Without FE-WA and FE-PRE
Voltage (V)
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
BLB
S
WL
BL
SB
Time
NVSS
NBL level
(near-end)
NBL level
(far-end)

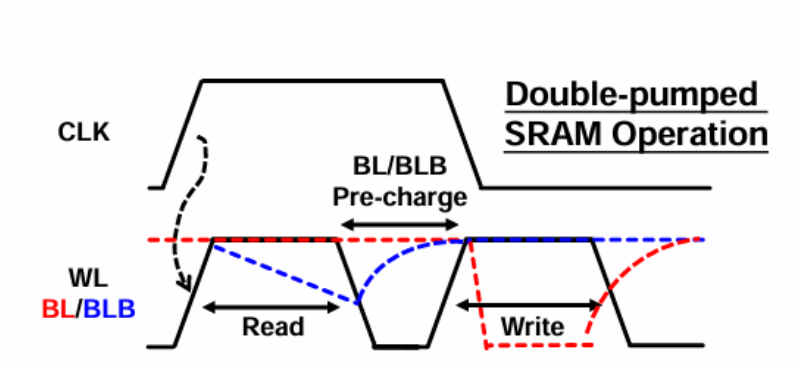

Slower BL pre-charge
BLB
S
WL
BL
SB
NVSS


with turning on FE-WA and FE-PRE
Voltage (V)
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
BLB
S
WL
BL
SB
NVSS
NBL level
(near-end)
NBL level
(far-end)


faster BL pre-charge
BLB
SB
WL
Improve BL
per-charge
BL
S
NVSS

Double-pumped SRAM Design for AI/HPC

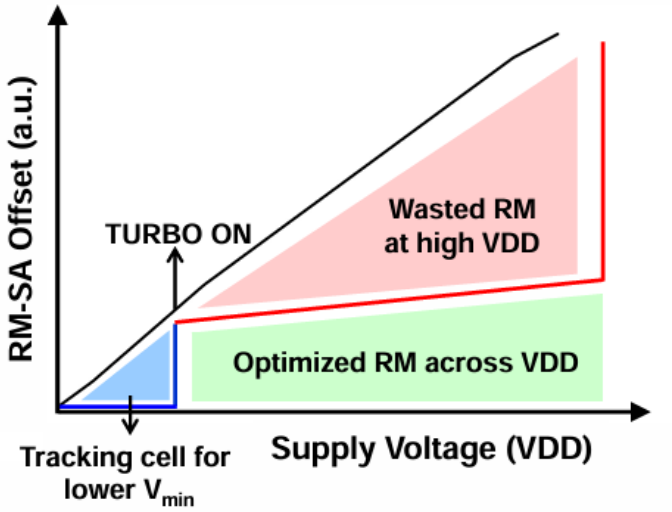
- AI/HPC demands high memory bandwidth: Double-pumped 6T SRAM
- Double-pumped SRAM perform 1-Read and 1-Write (1R1W) per cycle
- Better density than 8T (2-Port) SRAM
- Higher BW and Fmax than Single-Port SRAM for 1R1W operations

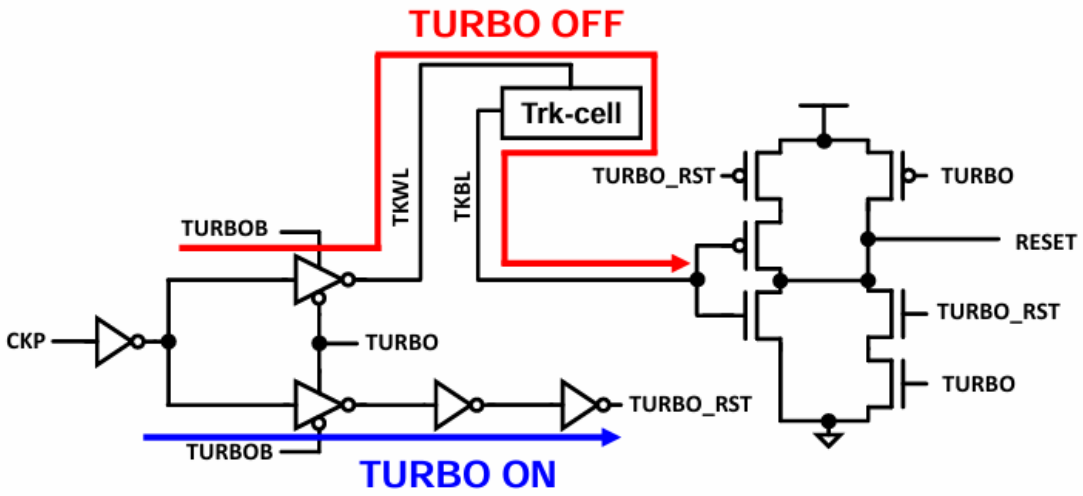


Dual Tracking Scheme in Double-pumped SRAM

TURBO ON: bypass tracking cell to optimize design margin
Fmax: 6.3% improvement while maintaining sufficient RM/WM
Active power: 11.5% reduction, from less BL/BLB swing

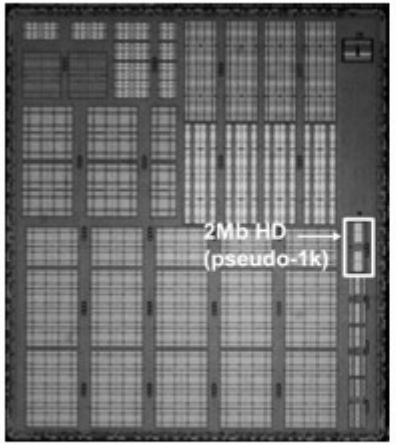
TURBO OFF: tracking cell enables lower Vmin operation

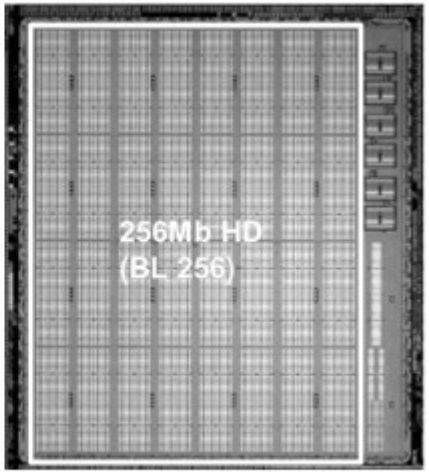




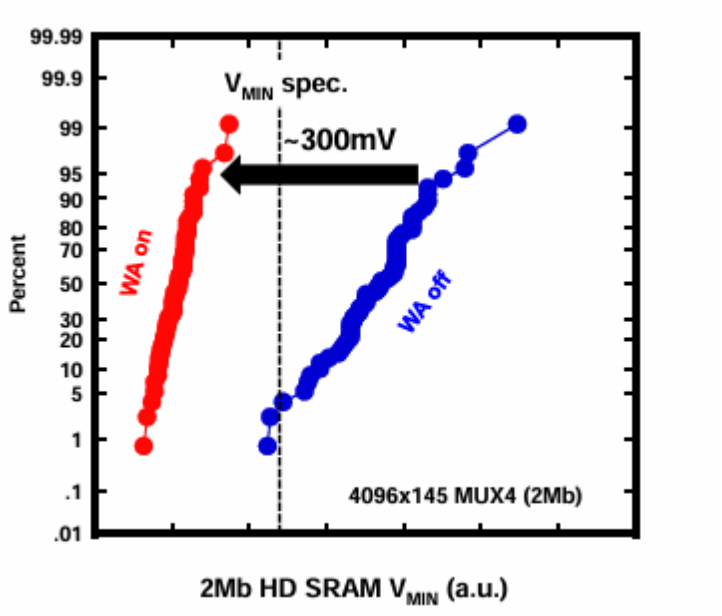
Test Chip

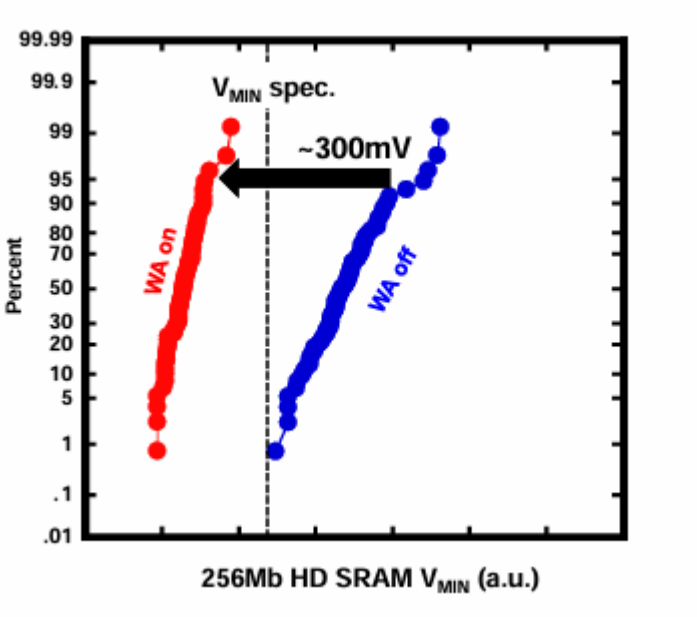
Technology	2nm nanosheet
Metal scheme	1P7M
Supply voltage	0.75V
Bit cell size	HD: 0.021μm^2
SRAM macro configuration	4096x145 MUX4 (2Mb) 4096x32 MUX16 (256Mb)
SRAM capacity	2Mb and 256Mb
Design Features	Redundancy Programmable E-fuse NBL write assist option





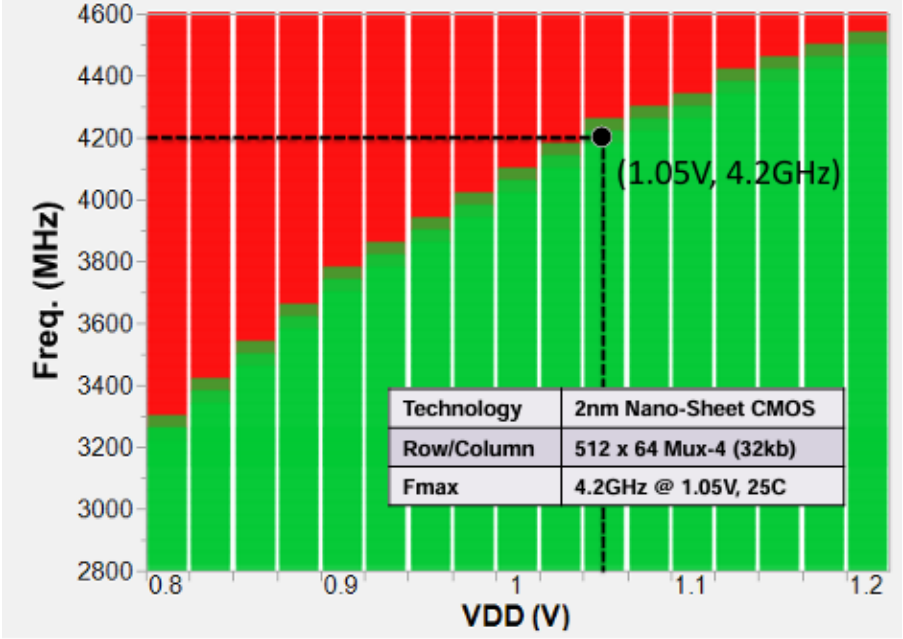
HD SRAM VMIN Cumulative Plot





Fmax Shmoo of Double-Pumped SRAM

Dual-tracking scheme enables 4.2GHz at 1.05V



University of Minnesota

Driven to Discover®

Crookston Duluth Morris Rochester Twin Cities