



WELCOME To

ISSCC 2014

SESSION 13

ADVANCED

EMBEDDED MEMORY



A 1Gb 2GHz Embedded DRAM in 22nm Tri-Gate CMOS Technology

Fatih Hamzaoglu, Umut Arslan, Nabhendra Bisnik,
Swaroop Ghosh, Manoj B. Lal, Nick Lindert, Mesut
Meterelliyoz, Randy B. Osborne, Joodong Park,
Shigeki Tomishima, Yih Wang, Kevin Zhang

Intel
Hillsboro, OR

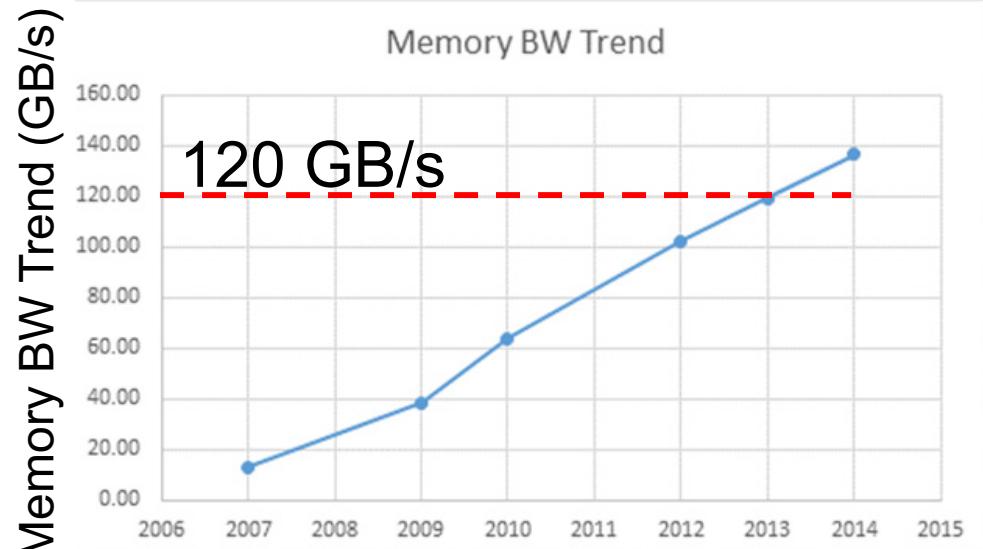
Outline

- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- Integrated Charge Pumps
- Die Architecture and Protocol
- Silicon Data
- Conclusions

Motivation



Intel IvyTown Xeon®, ISSCC’14, Paper#5.4



Shih-Lien Lu, ISSCC’14, Forum2, 2-socket Dell DP Systems

- Need for High Capacity and Low Energy/bit high-BW Memory increases
 - SRAM is not dense enough, and DDR Energy/bit is high
 - A high-density, high-BW In Package Memory is needed

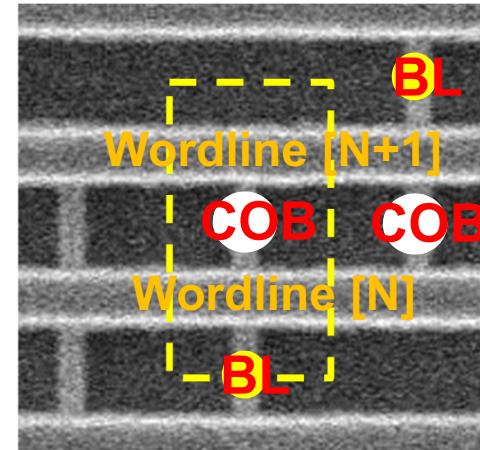
Outline

- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- Integrated Charge Pumps
- Die Architecture and Protocol
- Silicon Data
- Conclusions

eDRAM Process in 22nm Tri-Gate

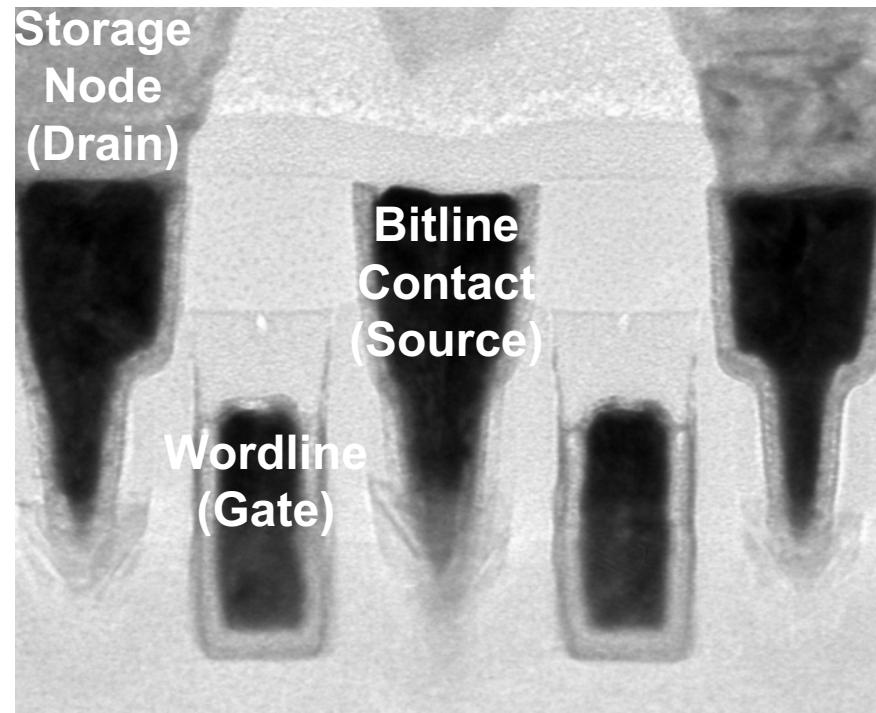
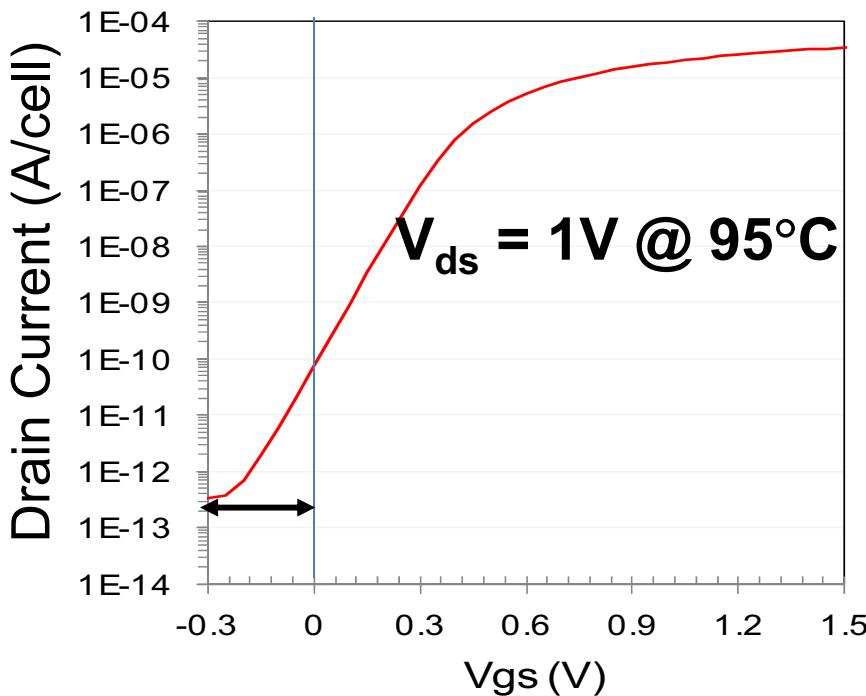
Key features of eDRAM Tech.:

- 0.029 μm^2 eDRAM bitcell
 - 3.2x smaller than SRAM (ISSCC'12)
- 17.5Mb/mm² density for a 128Mb Macro
- High-k, Metal Tri-Gate transistors supporting both low leakage and high performance
- 9 Metal Layers and 7-layers self-aligned Vias with ULK ILD
- MIM Capacitor-Over-Bitline (COB) embedded in interconnect stack



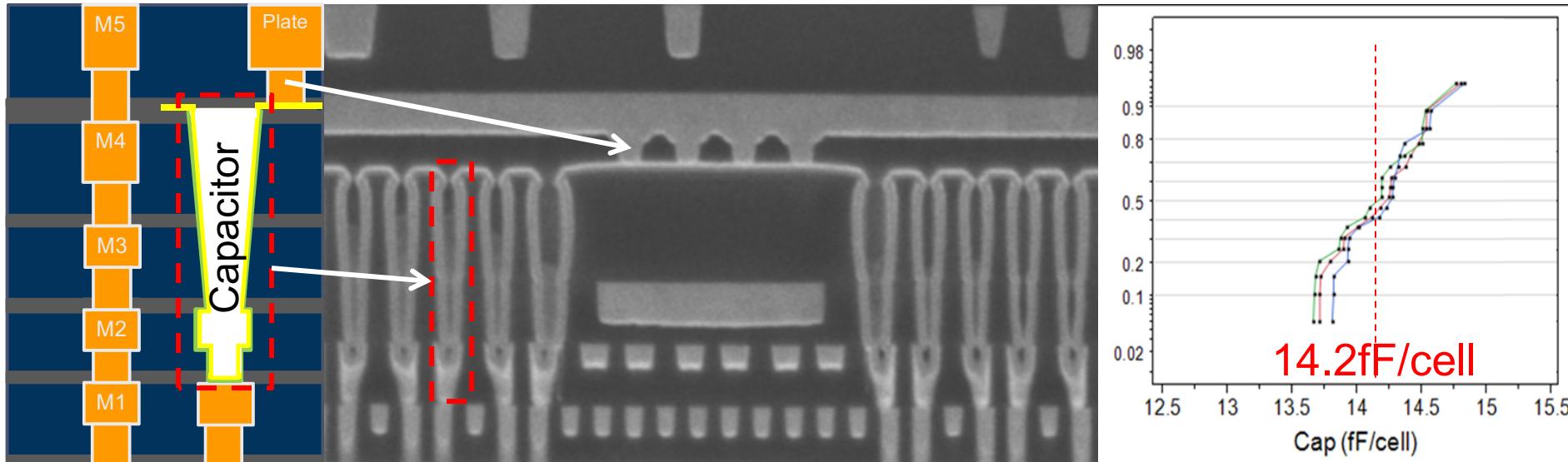
1T-1C eDRAM: 0.029 μm^2

eDRAM Access Transistor



- Optimized Tri-Gate transistors for Latency, BW and Power
 - <<1pA/cell leakage with Gate Underdrive
 - Fifth-generation strained Si

eDRAM Capacitor



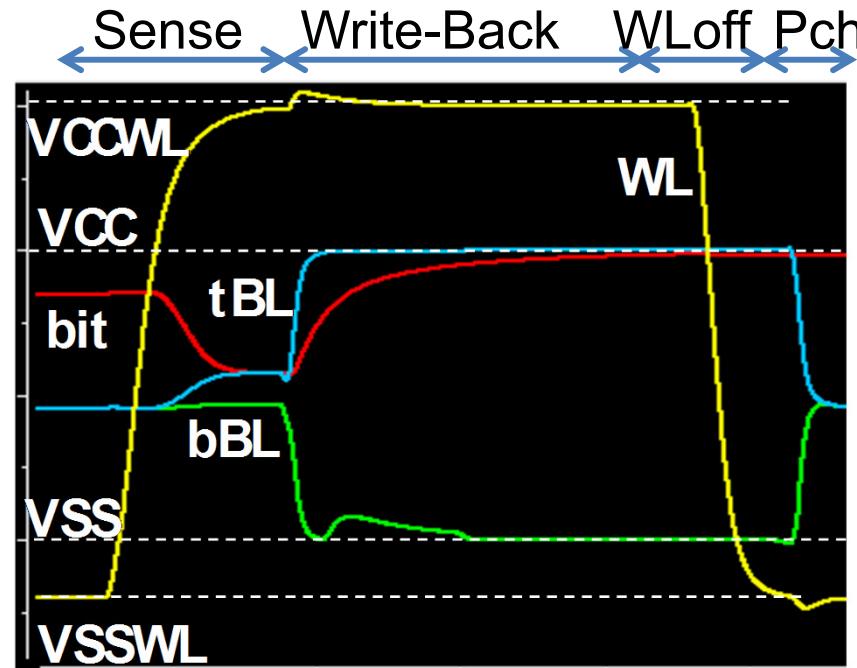
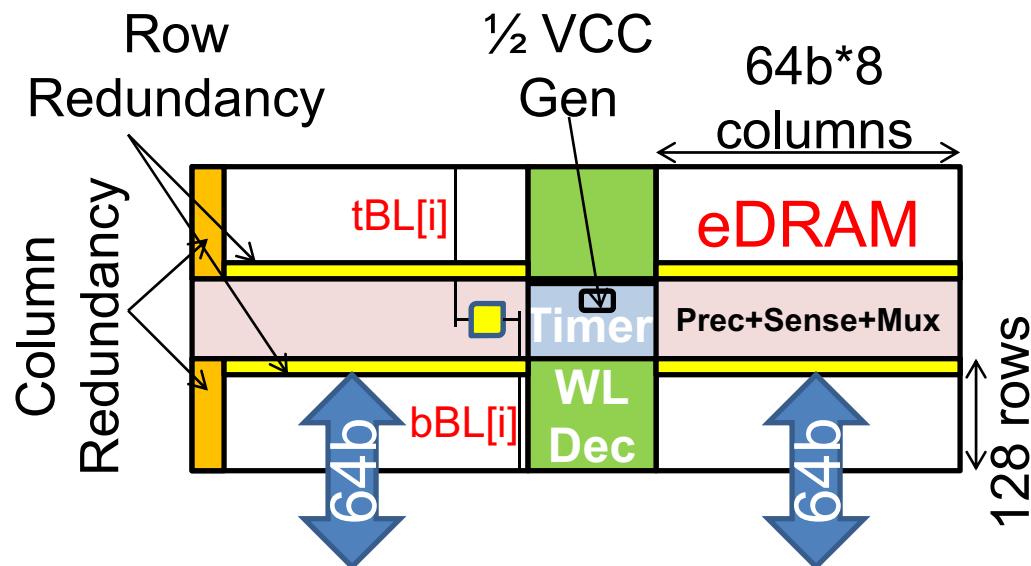
- ❑ Fully integrated MIM Trench Capacitor
 - Aggressive aspect ratio and compatible with logic process
 - Median 14.2fF/cell, <<0.1pA/cell

Outline

- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- Integrated Charge Pumps
- Die Architecture and Protocol
- Silicon Data
- Conclusions

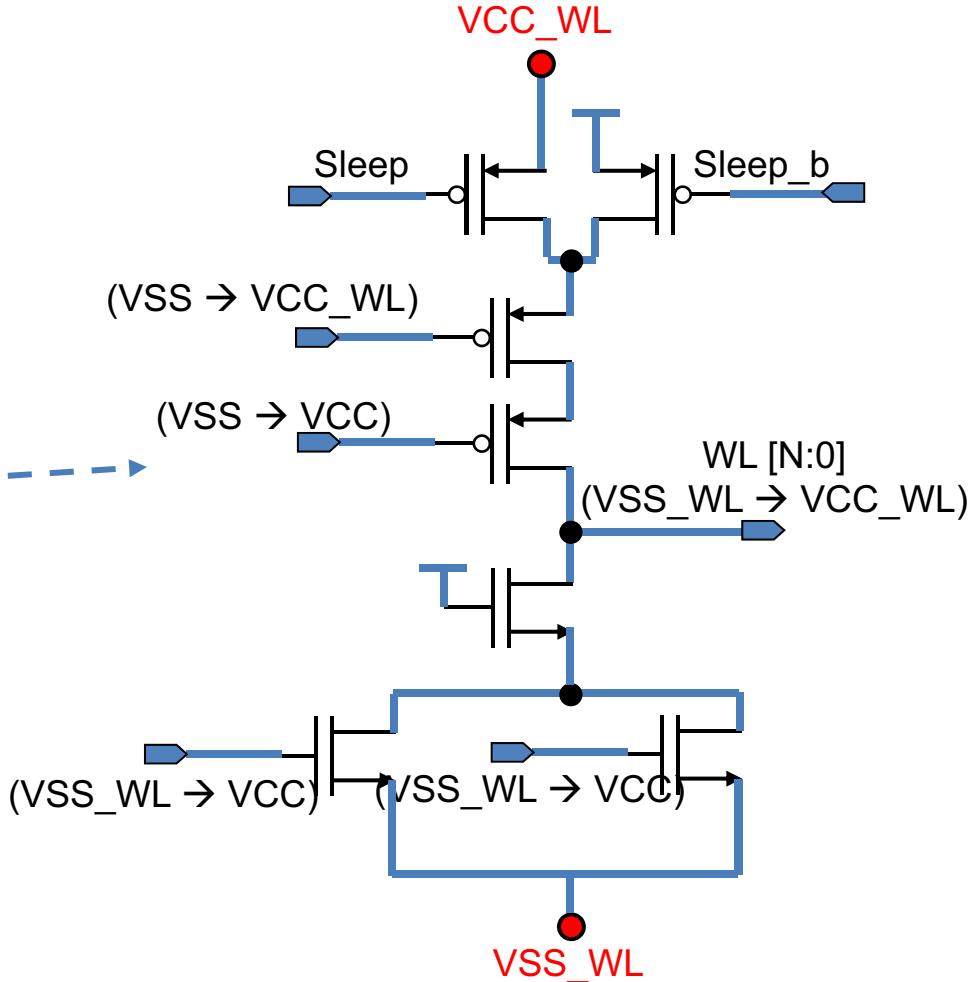
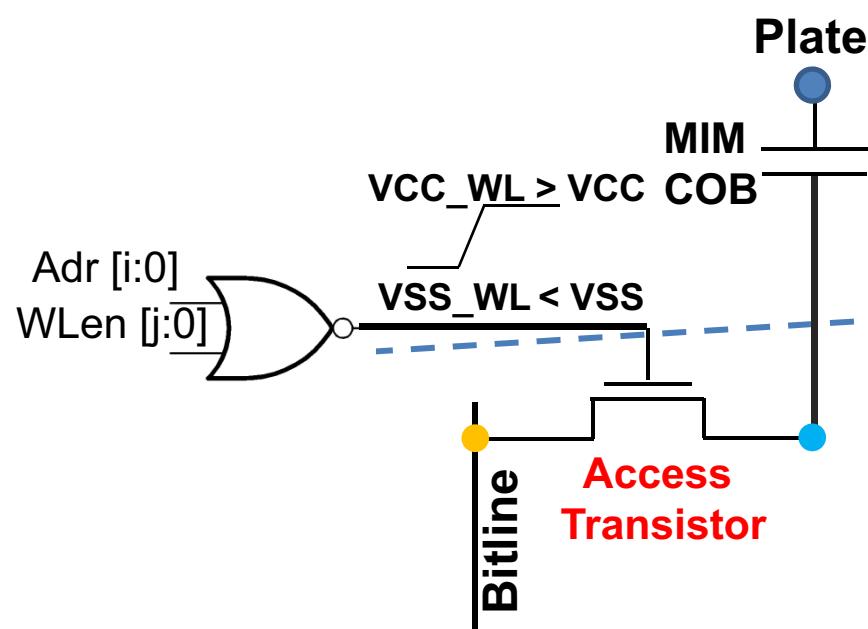
Subarray Architecture and Operation

256Kbit Subarray



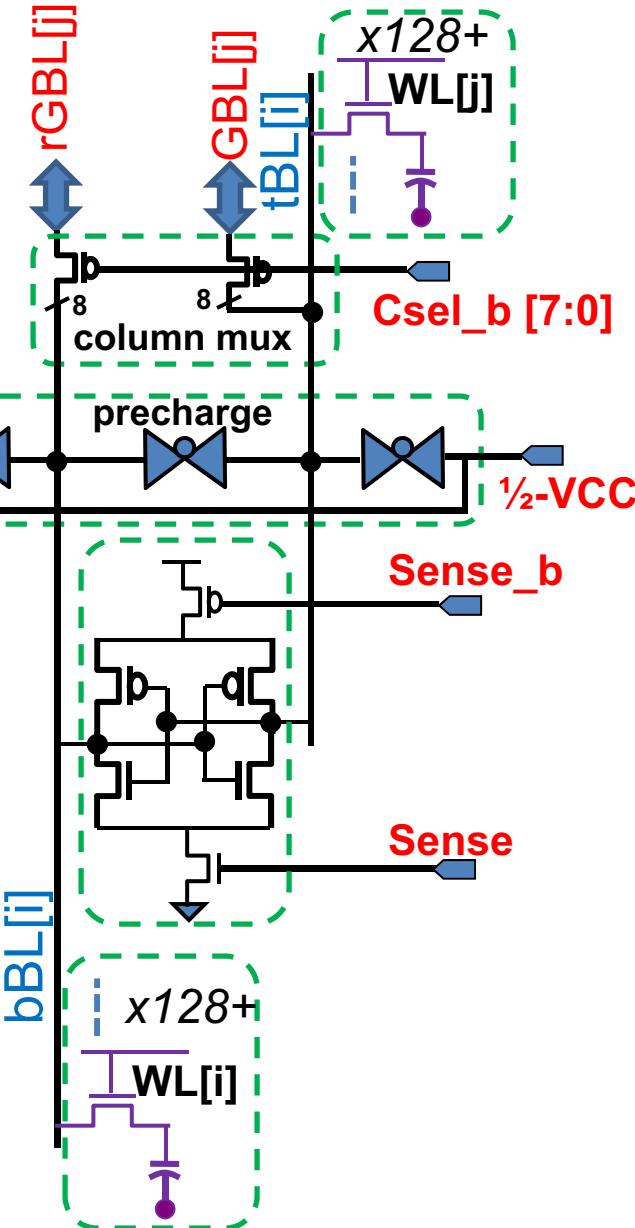
- 65% Array efficiency is achieved at 256Kbit Subarray
- Subarray also includes Row and Column Redundancies, as well as local $\frac{1}{2}$ -VCC Generator
- Random Cycle Time of 6 Array-Clock

WordLine Driver for Power and Vmax



- NOR Wordline Driver is used for simplicity and Reliability
 - None of the transistors' gate is exposed to $>\text{VCC}_\text{WL}$
 - Dynamic Sleep/Wakeup between $\text{VCC}/\text{VCC}_\text{WL}$

Area Efficient Column-IO

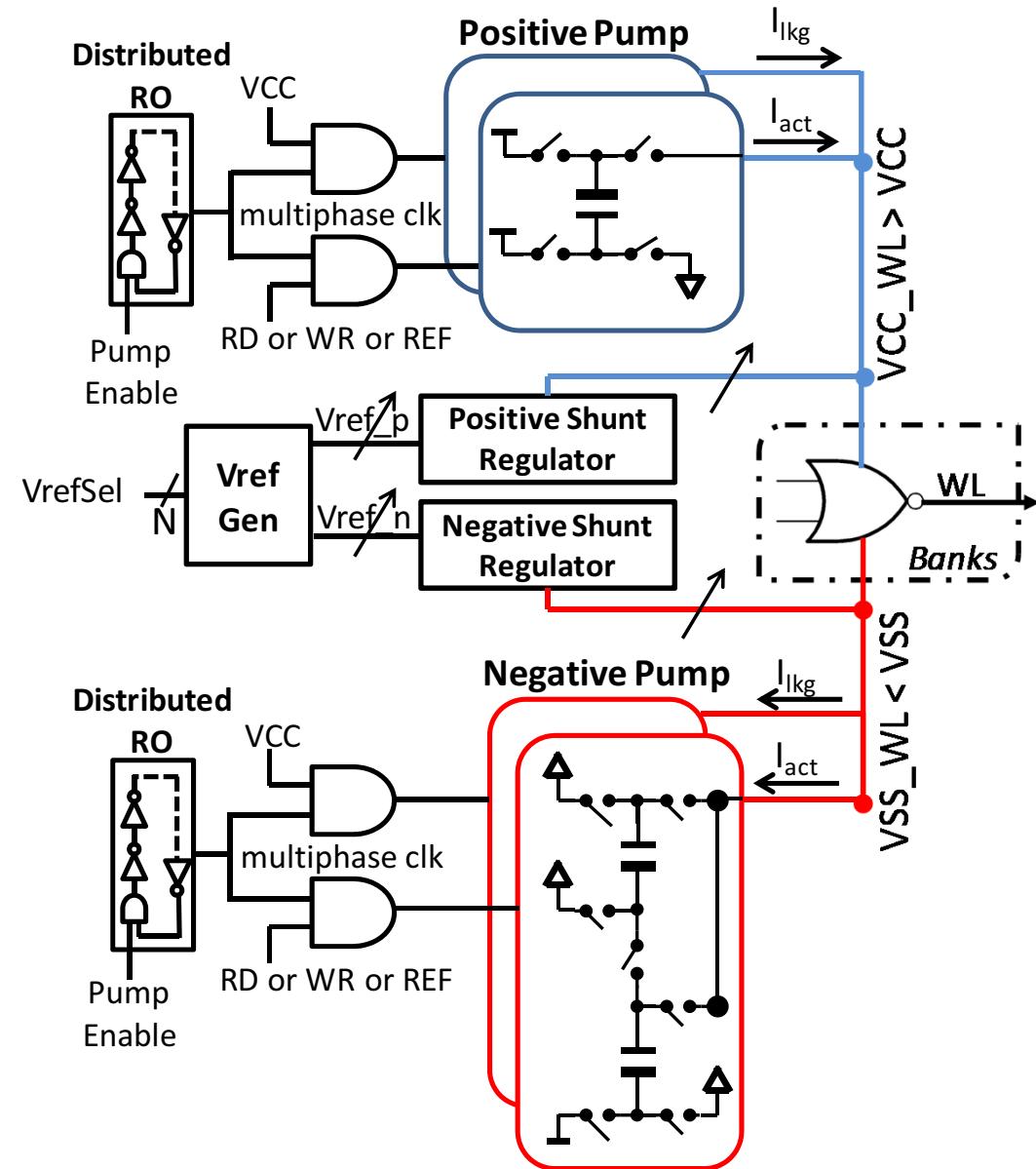


- $\frac{1}{2}$ -VCC Small Signal Sensing at Local Bitline
- 8:1 Column-Mux to Read/Write 128bits per Subarray
- 2nd Stage Small Signal Sense at Global Bitline
 - Shared among Subarrays

Outline

- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- **Integrated Charge Pumps**
- Die Architecture and Protocol
- Silicon Data
- Conclusions

Integrated Programmable Charge Pumps

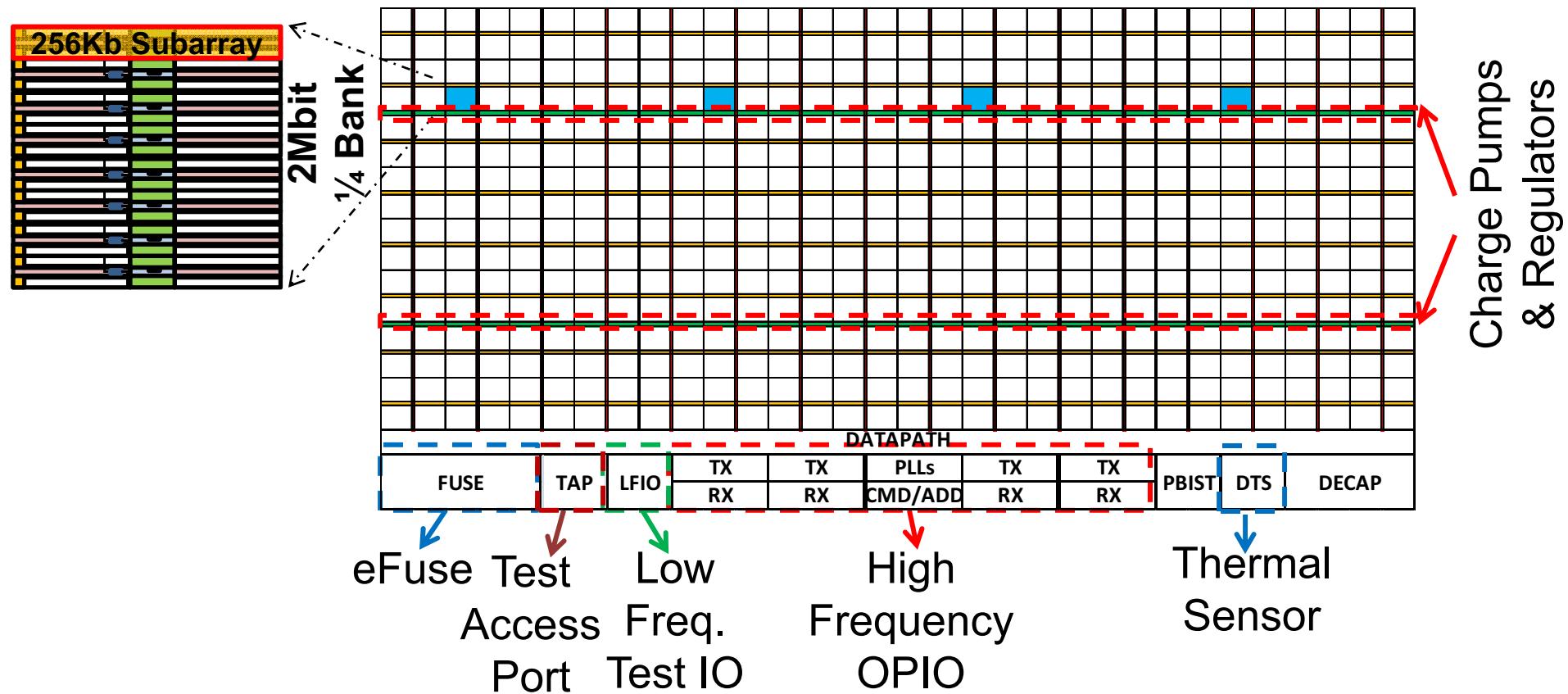


- WL-ON Voltage is generated through Voltage Doubler
- WL-OFF Voltage is driven from $-VCC/2$
- Separate Pumps for Leakage and Activity Charge Loss
- Both Voltages are Regulated for Vmax and GIDL
 - Programmable

Outline

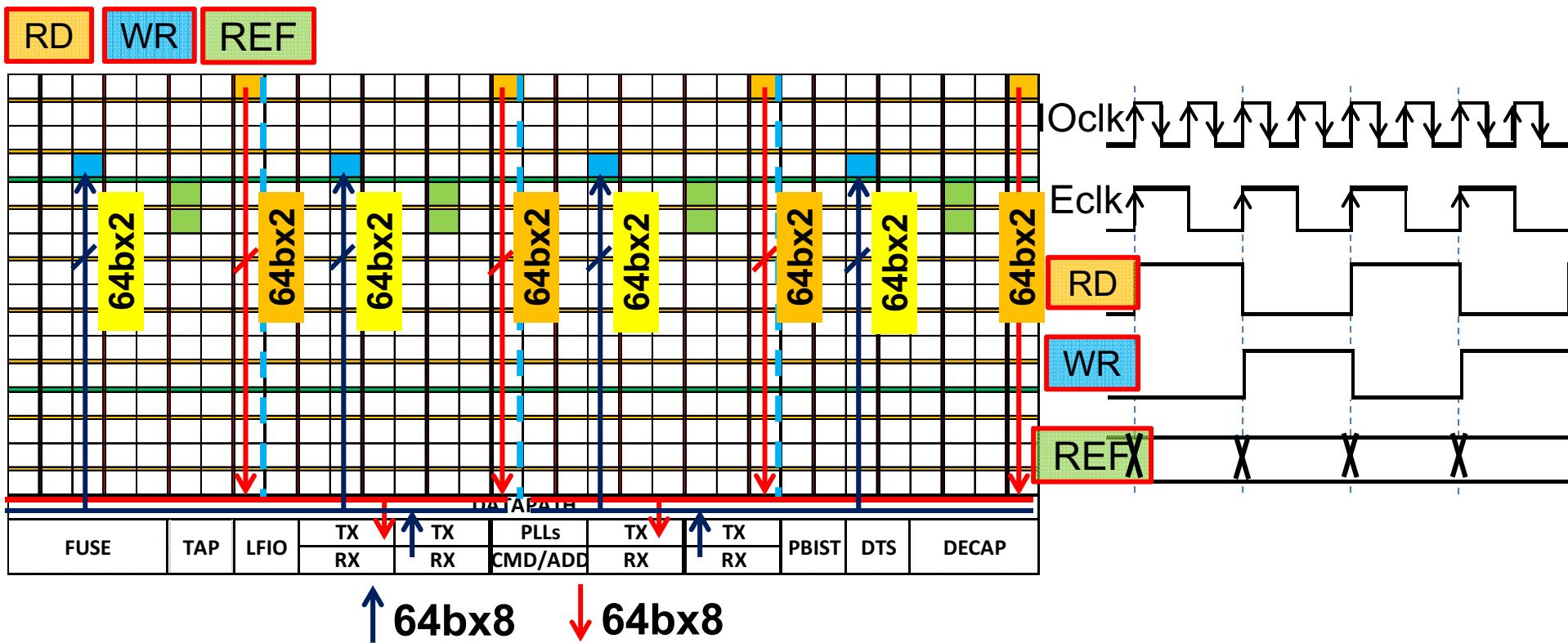
- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- Integrated Charge Pumps
- Die Architecture and Protocol
- Silicon Data
- Conclusions

1Gb Die Floorplan



- 77mm² 1Gbit die, with 128 Independent Banks
- On-die Positive and Negative Charge Pumps with 2% Area Overhead
- Thermal Sensor for T_{jmax}

1Gb Die Operation and Bandwidth

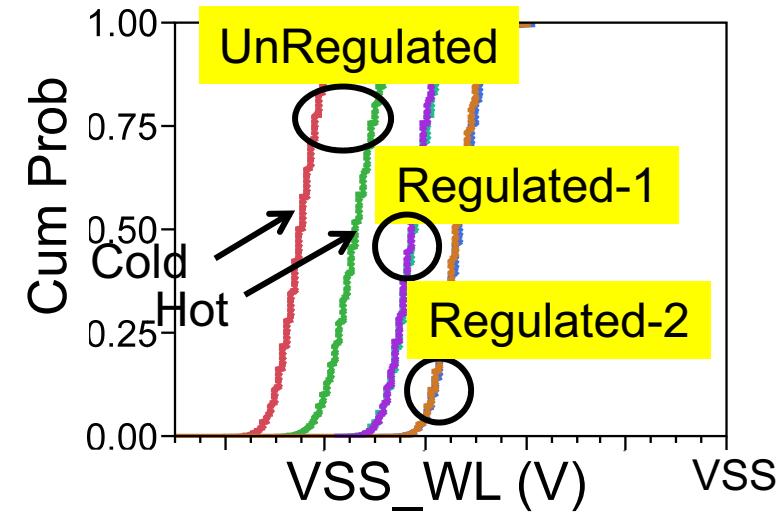
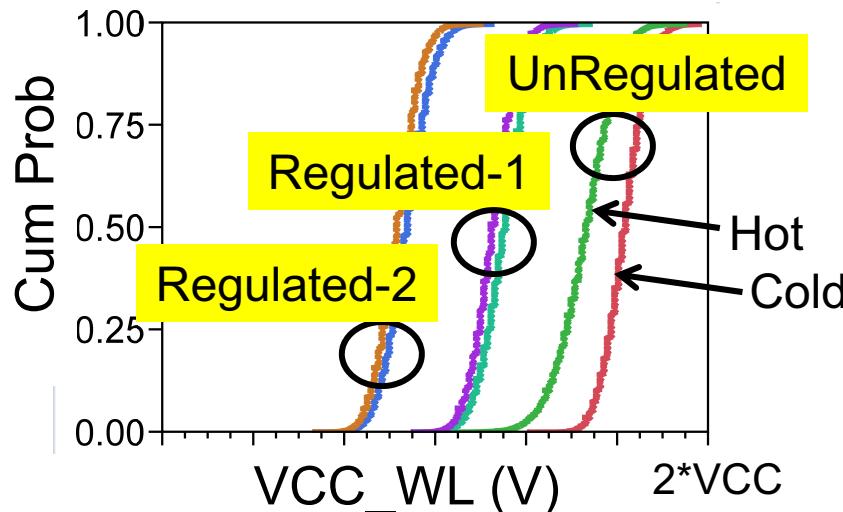


- IO transmission rate is 4x of eDRAM Array to save IO area
 - 2x Frequency and samples at both rising & falling edges of IOclk
- Array supports back to back Read/Write to different Banks
 - Completely decoupled Read and Write Datapaths
- A Refresh can be issued every cycle in parallel

Outline

- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- Integrated Charge Pumps
- Die Architecture and Protocol
- Silicon Data**
- Conclusions

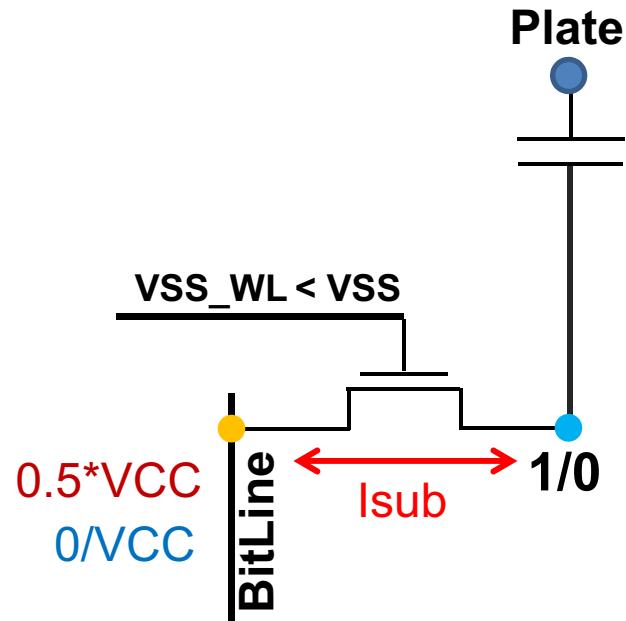
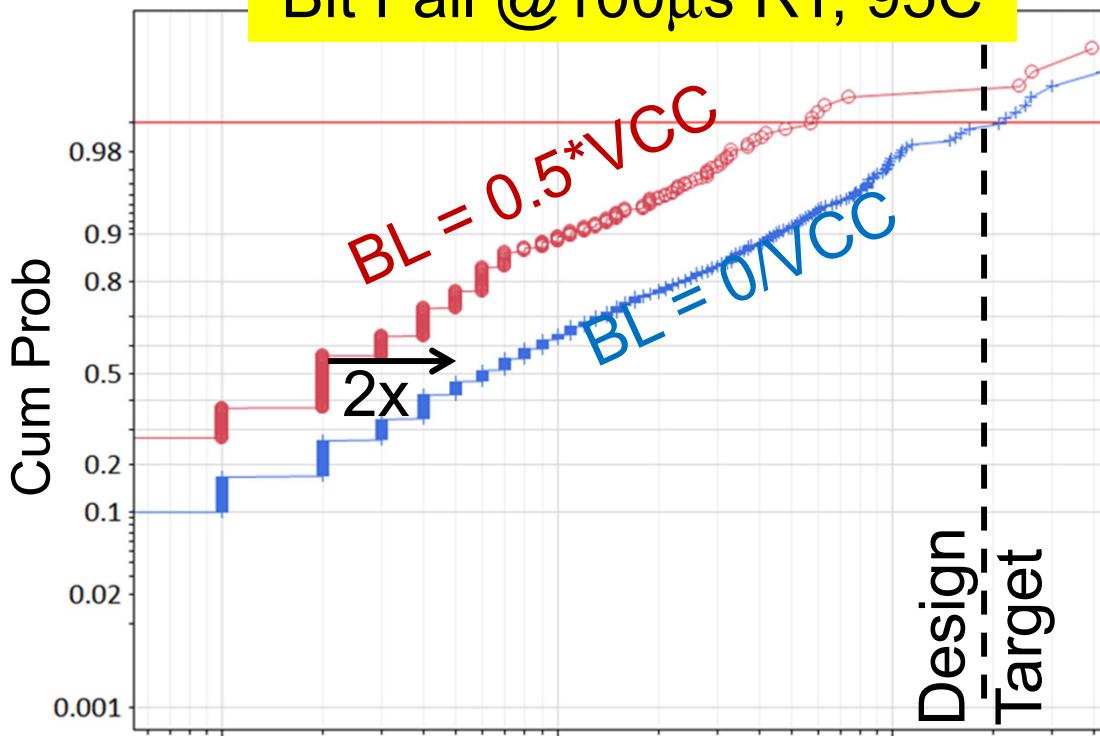
Integrated Programmable Charge Pumps



- Tight V_{CC_WL} control is achieved to avoid V_{CCmax}
- Similarly, tight V_{SS_WL} control helps to achieve good Retention Time
 - Programmability allows optimizing Subthreshold leakage vs. GIDL

Retention Time Testing

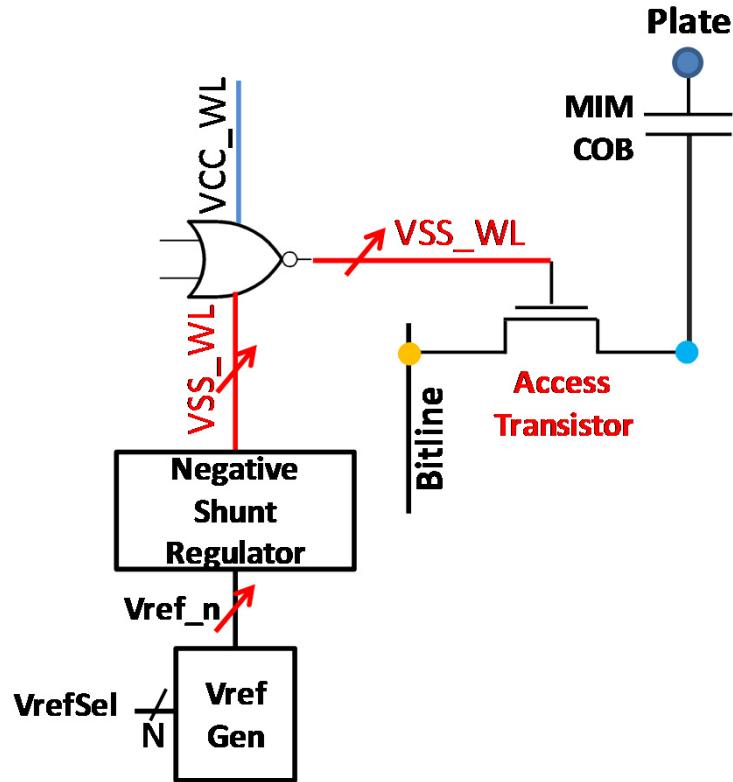
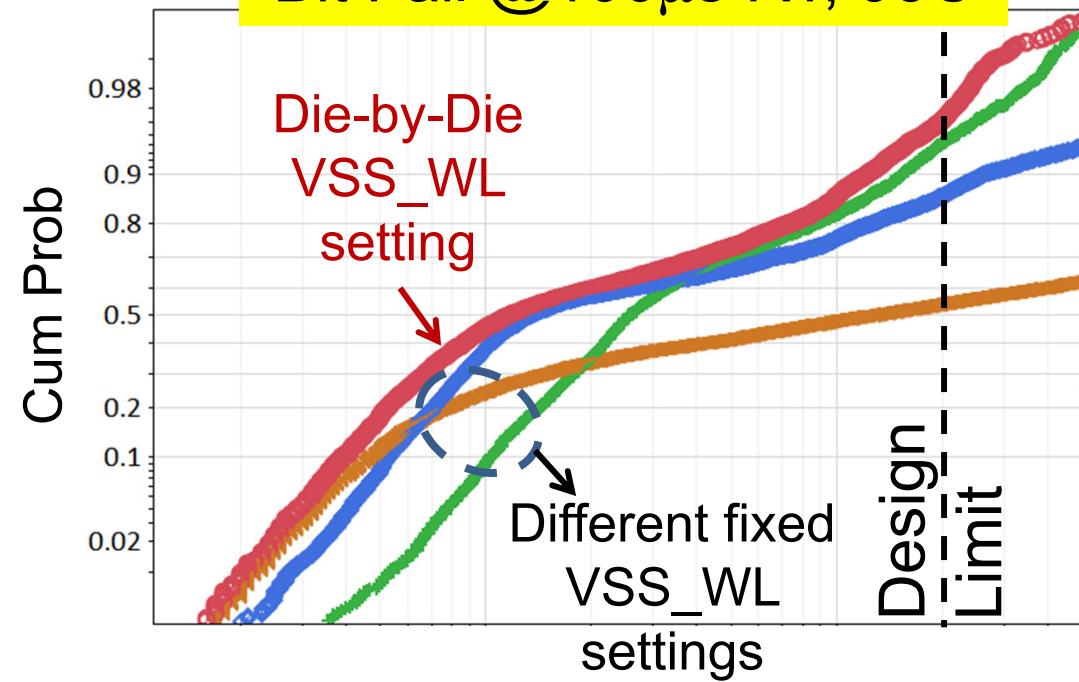
Bit Fail @100 μ s RT, 95C



- Retention Time tested at worst case Bitline Voltage of 0V/1V
- Potential 2x lower Bit Fail with BL = VCC/2 in Self Refresh

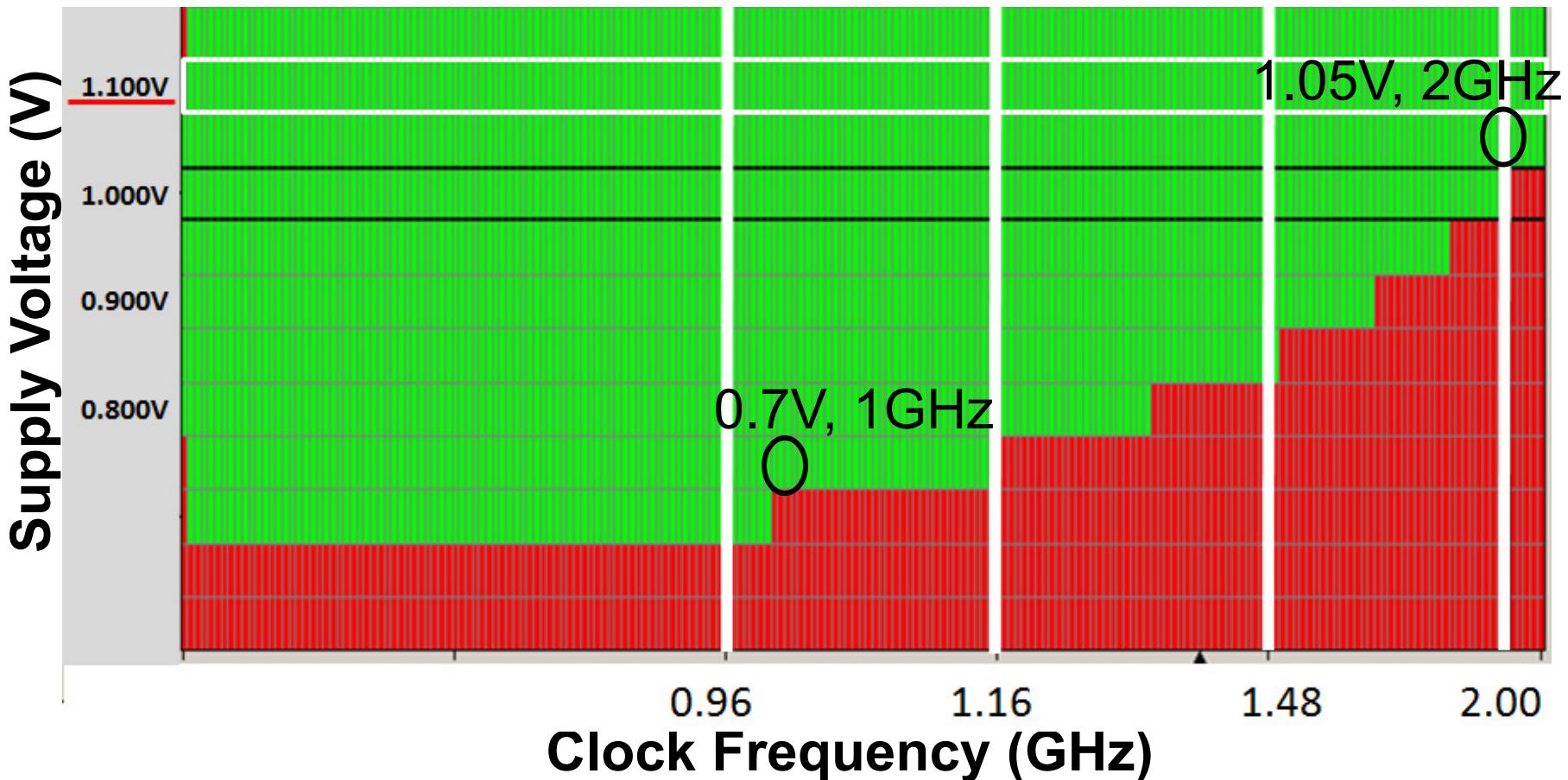
Retention Bit-Fail Optimization

Bit Fail @100 μ s RT, 95C



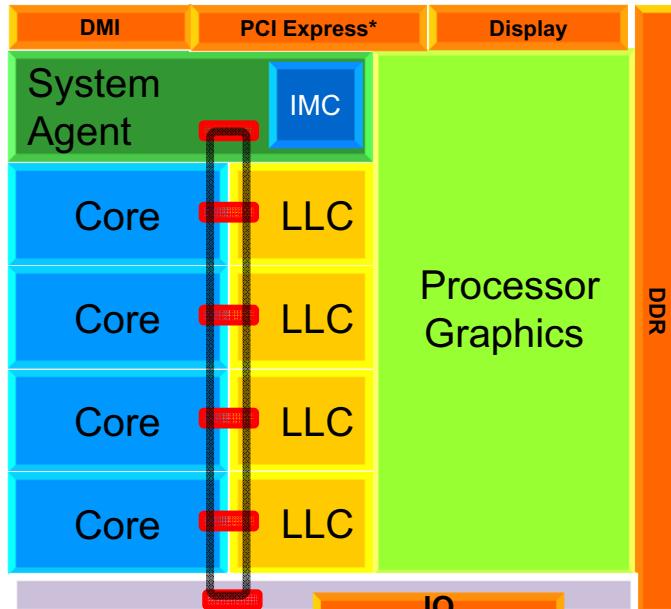
- Each die has different Subthreshold/Junction leakage trade-off
 - Hence different optimal VSS_WL per die
- Die-by-die VSS_WL setting achieves significantly improved bit-fail distribution

1Gbit Array Shmoo

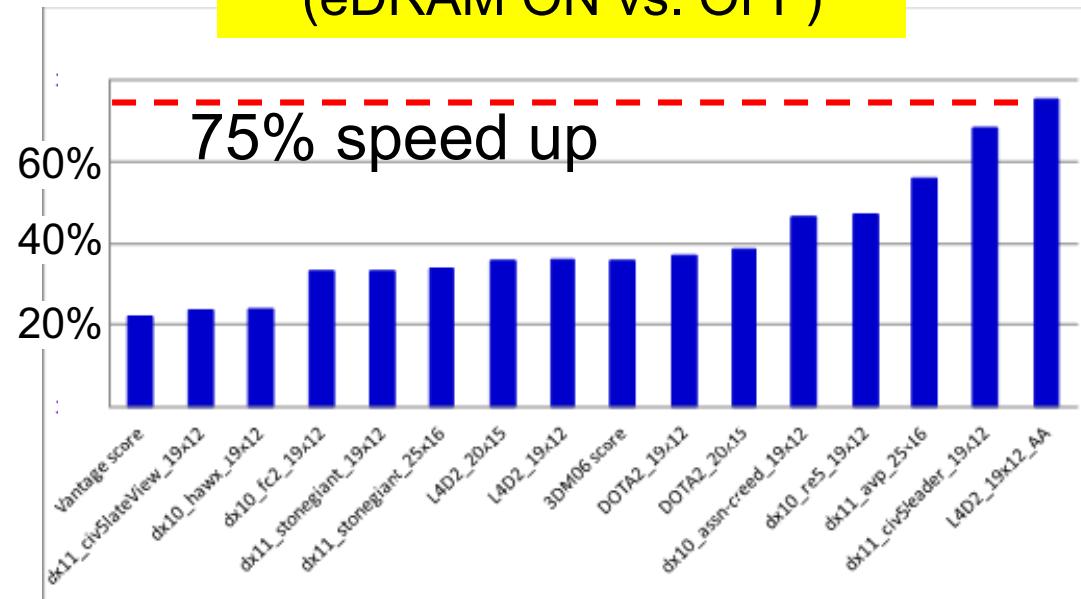


- 0.7V/1GHz – 1.05V/2GHz operation @95C, 100 μ s RT
 - 3ns Random Cycle Time @1.05V
- 64GB/s Read + 64GB/s Write Bandwidth @1.05V

Iris Pro™ with 1Gbit MCP eDRAM



Performance Improvement
(eDRAM ON vs. OFF)



N. Kurd, ISSCC'14, "Haswell: A Family of IA 22nm Processors"

- Iris Pro™ with 1Gb eDRAM MCP is commercially available
- up to 75% Graphics Performance improvement with 1Gbit eDRAM

Outline

- Motivation
- eDRAM Process in 22nm Tri-Gate
- Sub-Array Design
- Integrated Charge Pumps
- Die Architecture and Protocol
- Silicon Data
- Conclusions

Conclusions

- ❑ A record density eDRAM Technology is developed in 22nm Tri-Gate Technology to provide high-BW and low Power Memory access
- ❑ 1Gb eDRAM is packaged with high-end graphics processor to improve the performance up to 75% for wide spectrum of applications.
- ❑ The product is commercially available as Iris Pro™

Q&A

A 14nm FinFET 128Mb 6T SRAM with V_{MIN} Enhancement Techniques for Low-Power Applications

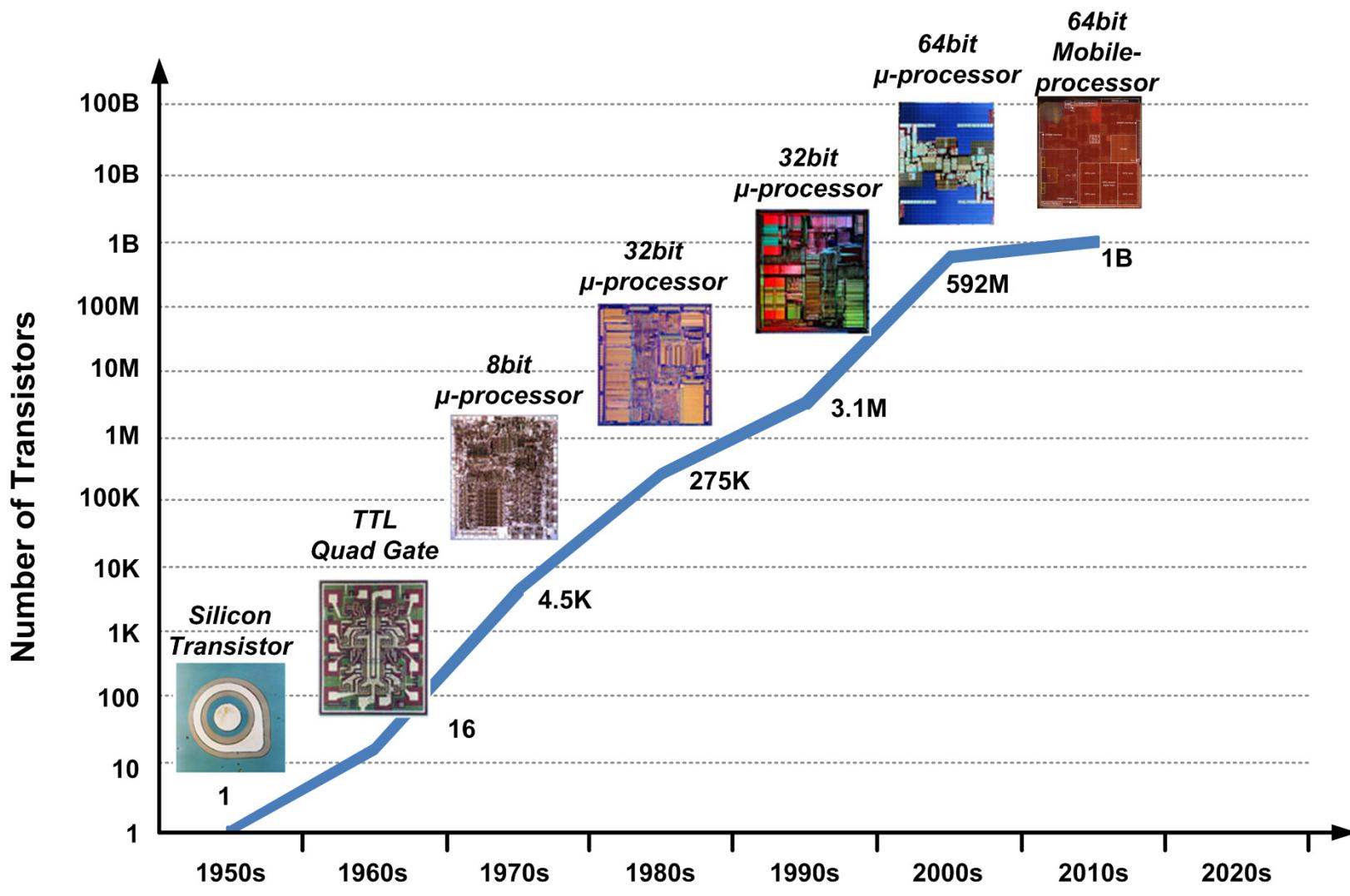
Taejoong Song, Woojin Rim, Jonghoon Jung,
Giyong Yang, Jaeho Park, Sunghyun Park, Kang-Hyun Baek,
Sanghoon Baek, Sang-Kyu Oh, Jinsuk Jung, Sungbong Kim,
Gyuhong Kim, Jintae Kim, Youngkeun Lee, Kee Sup Kim,
Sang-Pil Sim, Jong Shik Yoon, Kyu-Myung Choi

Samsung Electronics, Yongin, Korea

Outline

- **Motivation**
- **FinFET : Opportunity and Challenges to SRAM**
- **Conventional SRAM Assist Techniques**
- **Proposed SRAM Assist Scheme**
- **Implementation and Measurement Results**
- **Conclusions**

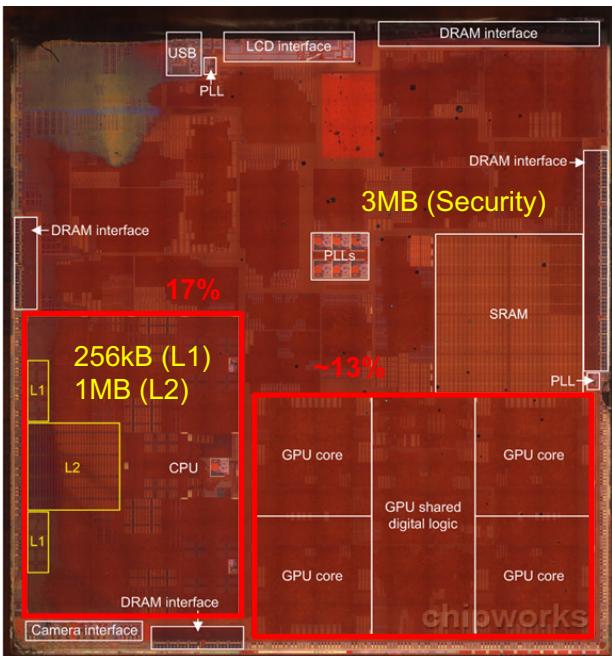
Silicon Scaling and SoC



Source : Computer History Museum (www.computerhistory.org)

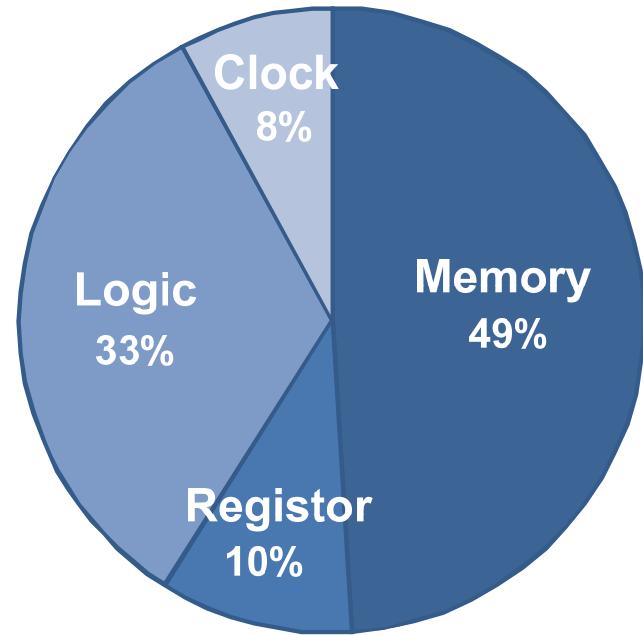
Cache (SRAM) Occupies Huge Area, Power

- SRAM takes up to 20%~30% area of SoC
- SRAM consumes 40%~50% of total chip power



An SoC (A7 Processor, 28nm) Die Photo

Source : Chipworks (www.chipwoks.com)

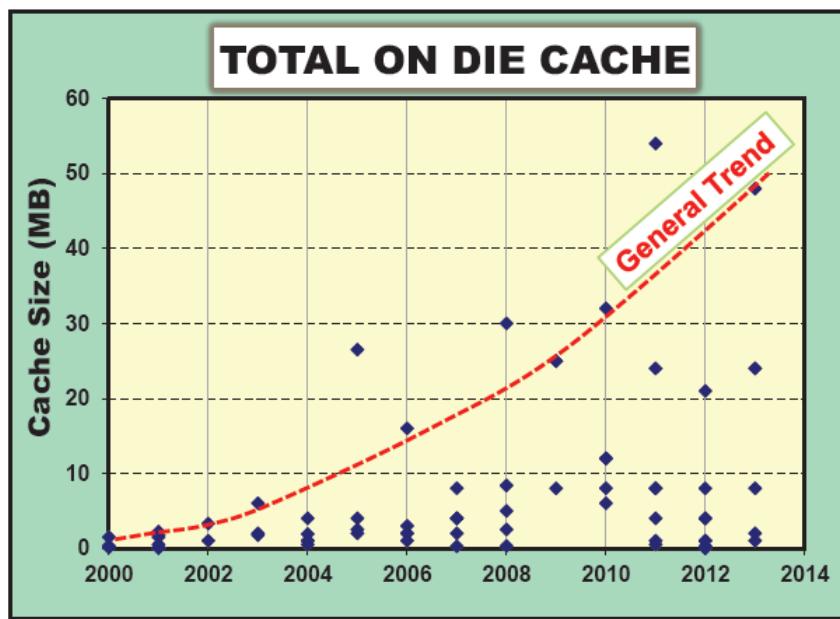


Power Consumption Distribution in an SoC

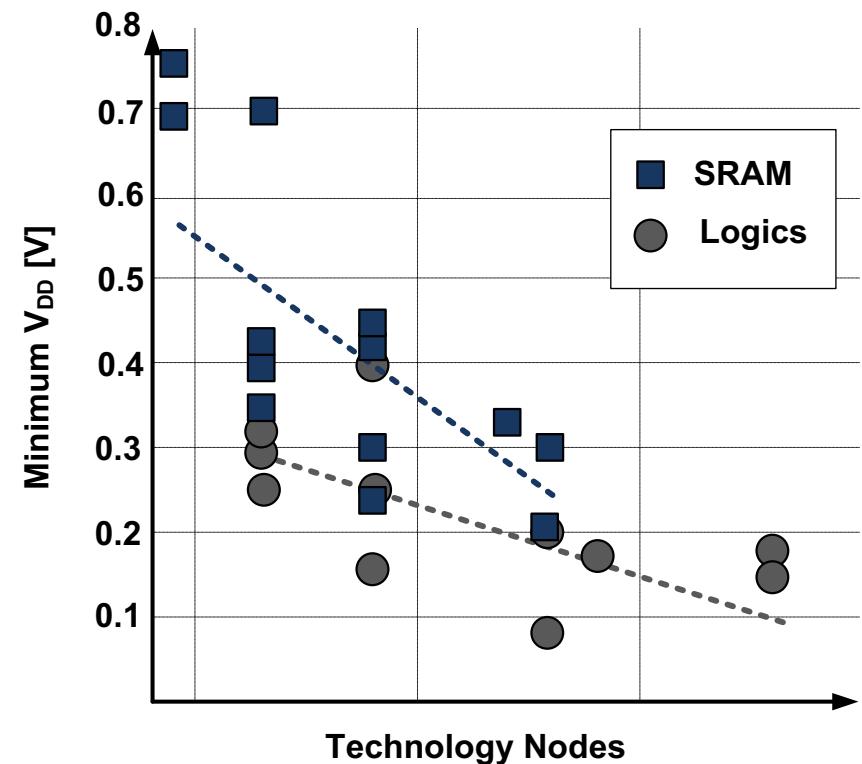
Source : Calypto Design System & Virage Logic

SRAM Capacity and V_{MIN} Trend

- Cache size increases linearly
- SRAM V_{MIN} does not scale-down as much as Logic



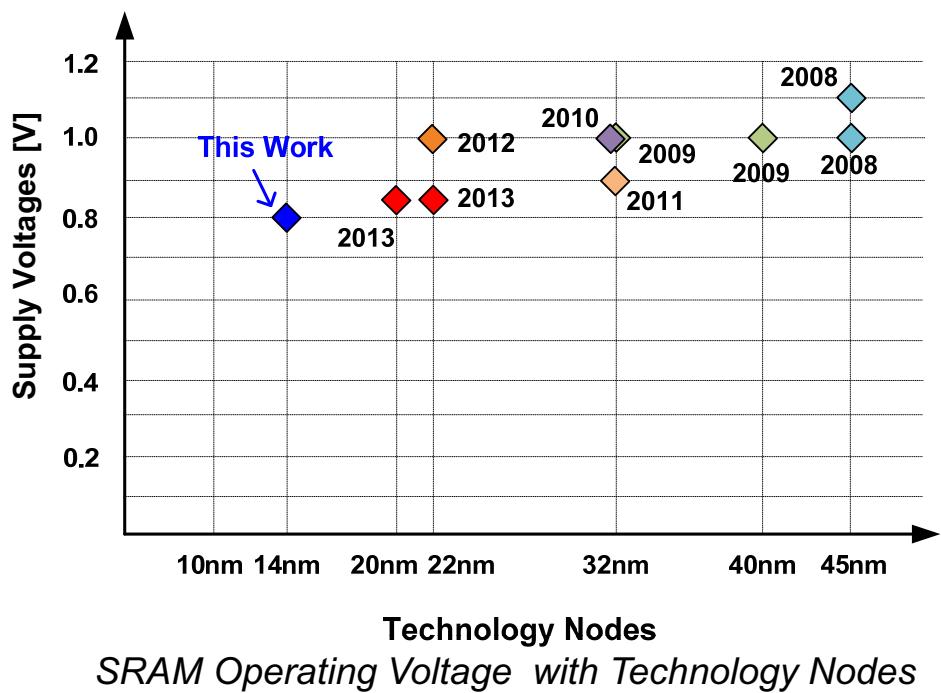
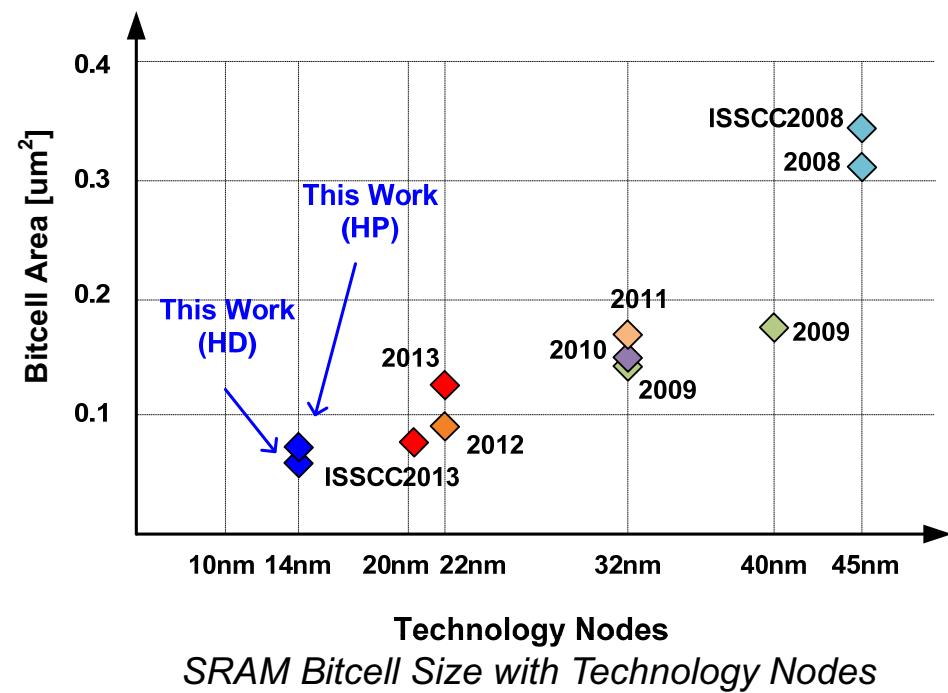
Source : ISSCC 2013 Trends



V_{MIN} trends of SRAM and Logic

SRAM Bitcell Size and V_{DD} Trends

- Bitcell size has been scaled-down
- V_{DD} scaling has been stuck
- SRAM power is the major portion of SoC power
- V_{MIN} should be improved for low-power SRAM



For Improving SRAM V_{MIN}

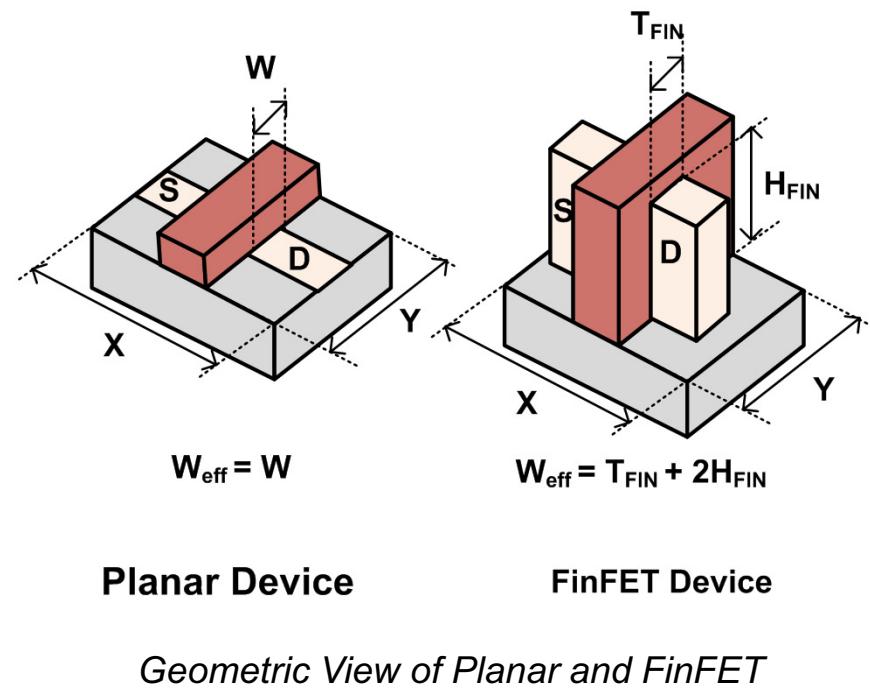
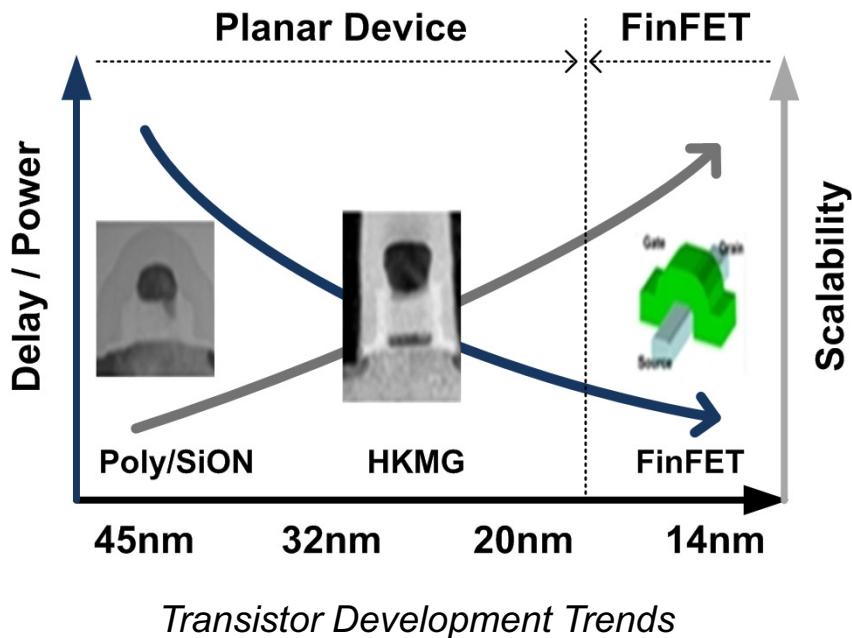
- Bitcell and Process
 - Minimize device fluctuation
 - Optimize bitcell design (PU,PG,PD)
 - Improve performance vs leakage
 - Design
 - Hierarchical WL, BL structure
 - Increased redundancy
 - ECC (Error Collection Code)
 - Write/read assist techniques
- need new transistor and technology-level effort
- Area / structure overhead
- Novel circuit design techniques

Outline

- Motivation
- FinFET: Opportunity and Challenges to SRAM
- Conventional SRAM Assist Techniques
- Proposed SRAM Assist Scheme
- Implementation and Measurement Results
- Conclusions

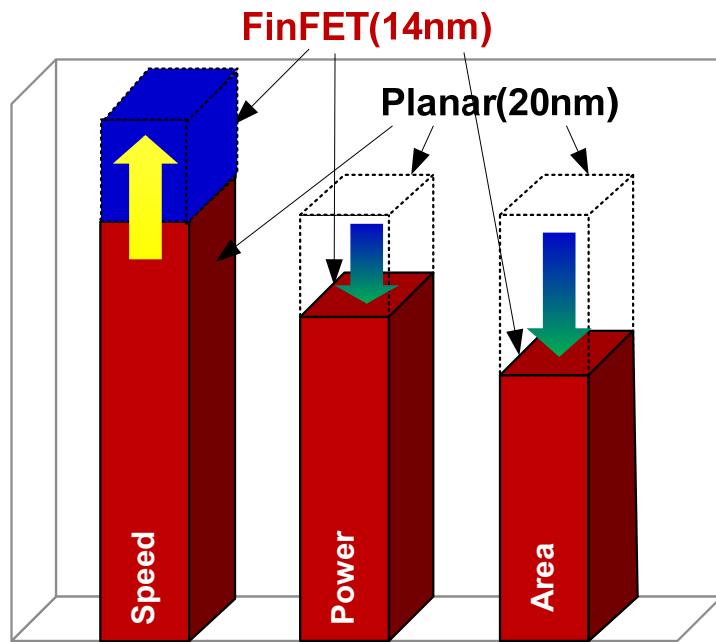
New Transistor: FinFET

- 3D Structure: excellent gate controllability
- Superior scaling of L_g and V_{TH}
- Higher mobility, lower ΔV_{TH}

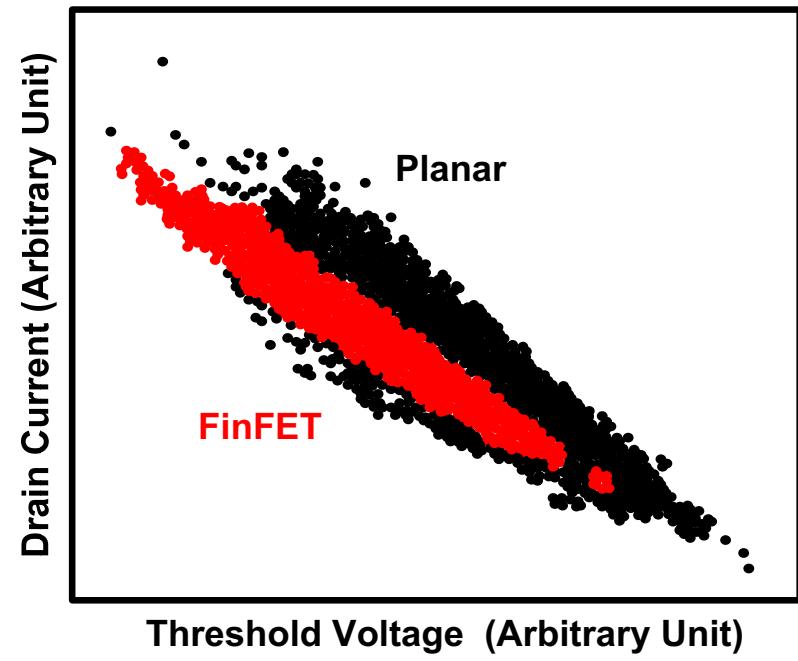


FinFET Benefits: PPA

- Improved short channel effect and subthreshold swing
- Improved speed, power, and area versus Planar



PPA (14nm FinFET vs 20nm Planar)

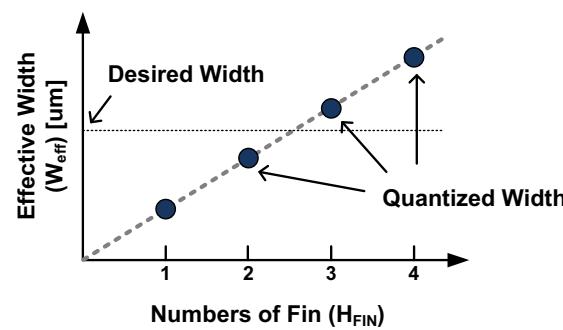
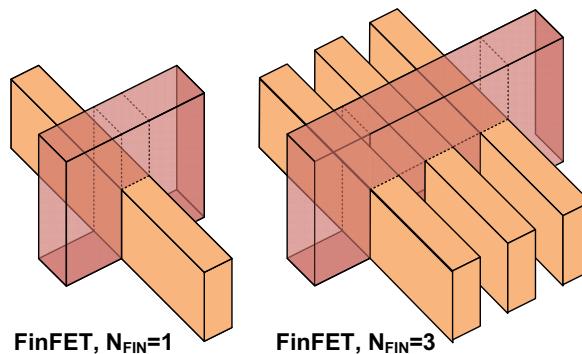


Voltage and Current Mismatch

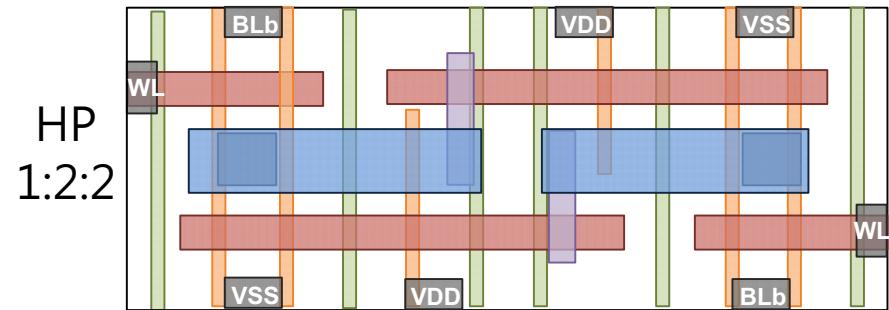
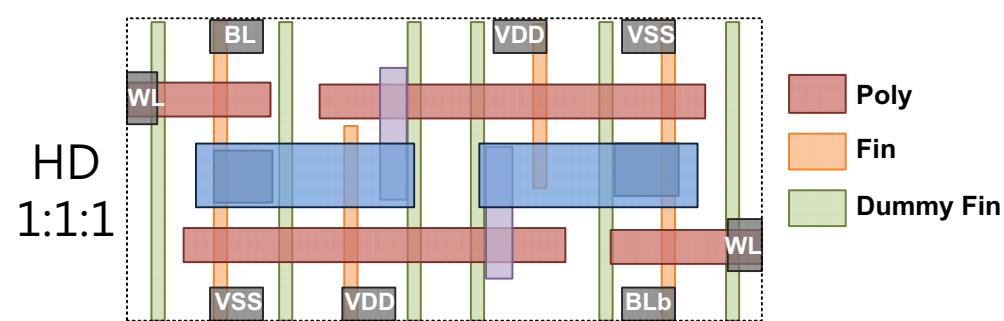
Bitcell Optimization? Too Difficult

- (1) Quantized Width

- In Planar, width of the device is a continuous parameter
- In FinFET, width is given by $W_{eff} = (T_{FIN} + 2H_{FIN}) \times N_{FIN}$
- This property restricts the design optimization



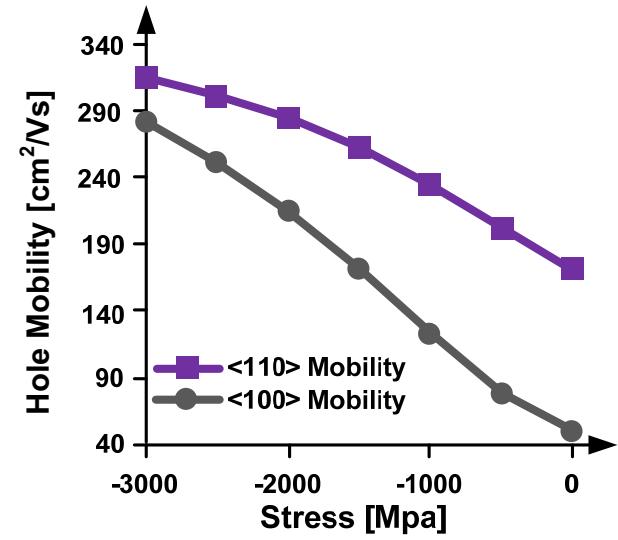
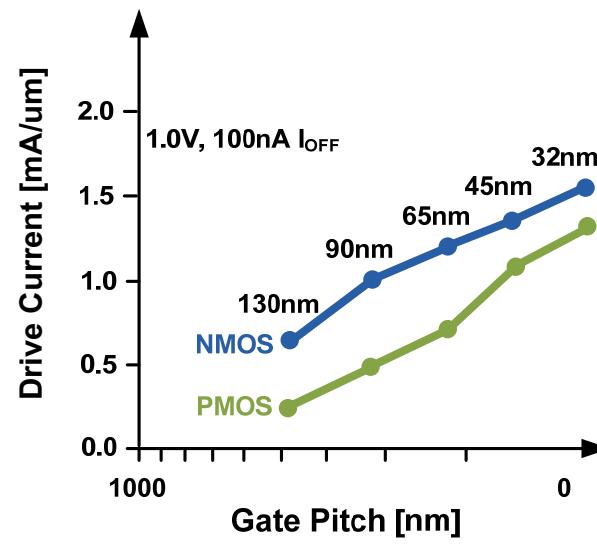
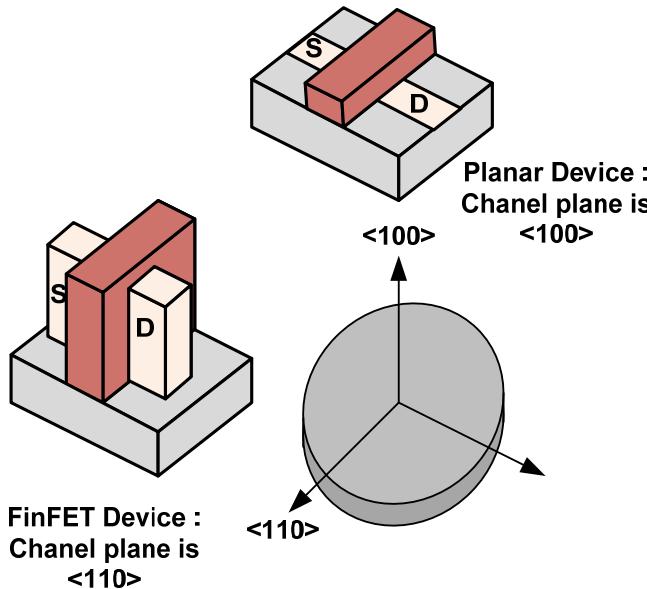
Width quantization property



Fin-based SRAM bitcell design

Bitcell Optimization? Too Difficult

- (2) Strong PMOS
 - Higher hole mobility in $<110>$ plane and electron mobility in $<100>$ plane
 - Faster PMOS in FinFET
 - Worse write-ability of the high-density bitcell
($N_{FIN,PU}:N_{FIN,PD}=1:1$)



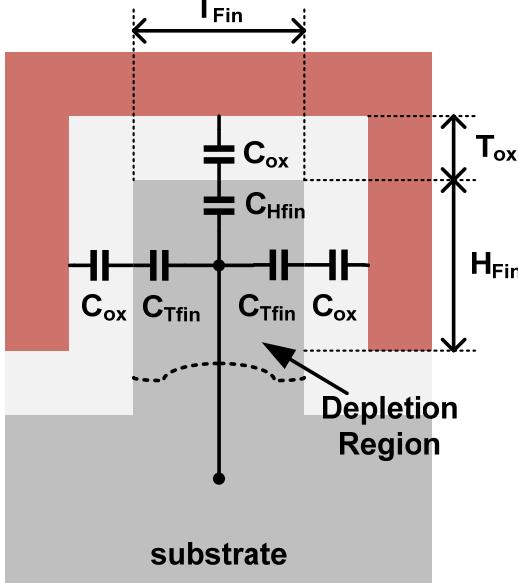
Source : N. Xu et al., IEEE Electron Device Letter, Vol.33. Mar. 2012
M. Bohr, ISSCC 2009

Bitcell Optimization? Too Difficult

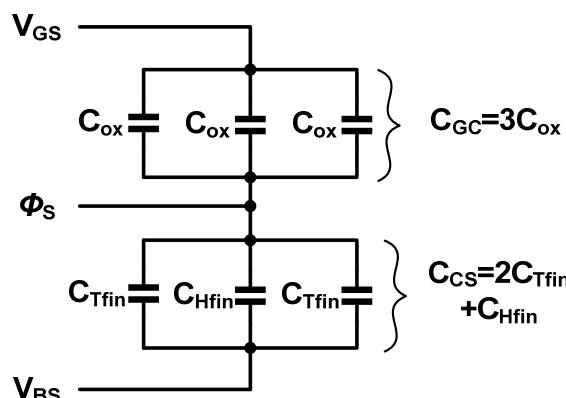
- (3) Less Body Biasing Effect
 - Fins are almost fully depleted : very little body effect
 - Designer loses one of knobs in FinFET
 - Body-Effect

$$\gamma = \frac{dV_{TH}}{dV_{BS}} \approx \frac{C_{CS}}{C_{GS}}$$

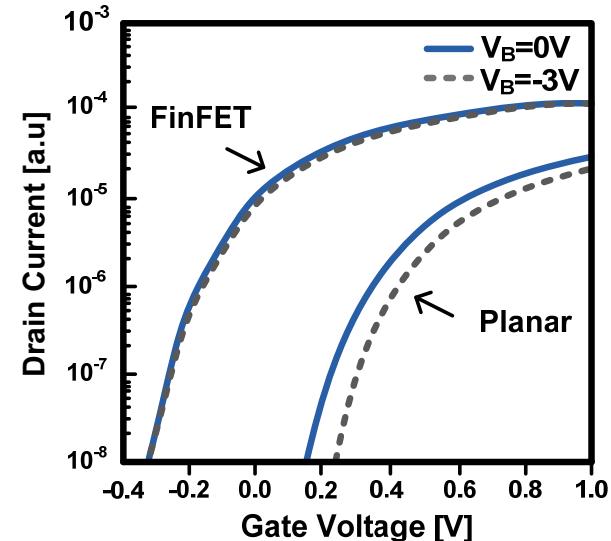
C_{CS} is smaller and C_{GS} is larger than planar



Cross-section view of FinFET



Capacitive equivalent circuit of Fin



Body biasing dependency

Bitcell Optimization? Too Difficult

- **SRAM bitcell optimization is too difficult**
 - Quantized width
 - Strong PMOS
 - Little body-biasing effect
- **Designer needs to add assist circuit to ensure reliable SRAM operation**

Outline

- Motivation
- FinFET: Opportunity and Challenges to SRAM
- **Conventional SRAM Assist Techniques**
- Proposed SRAM Assist Scheme
- Implementation and Measurement Results
- Conclusions

Conventional SRAM Assist Schemes

- SRAM assist helps write or read operation to overcome the weakness of bitcell itself

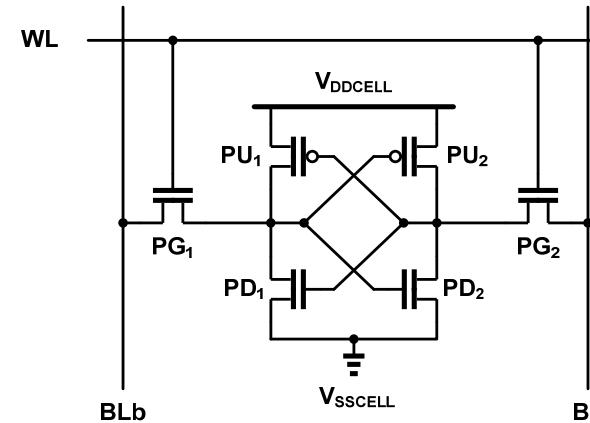
- Write assist

- WL_{OD}** (*WL Overdrive*)
- VDCL** (V_{DDCELL} Lowering)
- NBL** (*Negative BL*)

- Stability assist

- WLUD** (*WL Underdrive*)

- Trade-off between PPA (power, performance, area) and write/stability margin

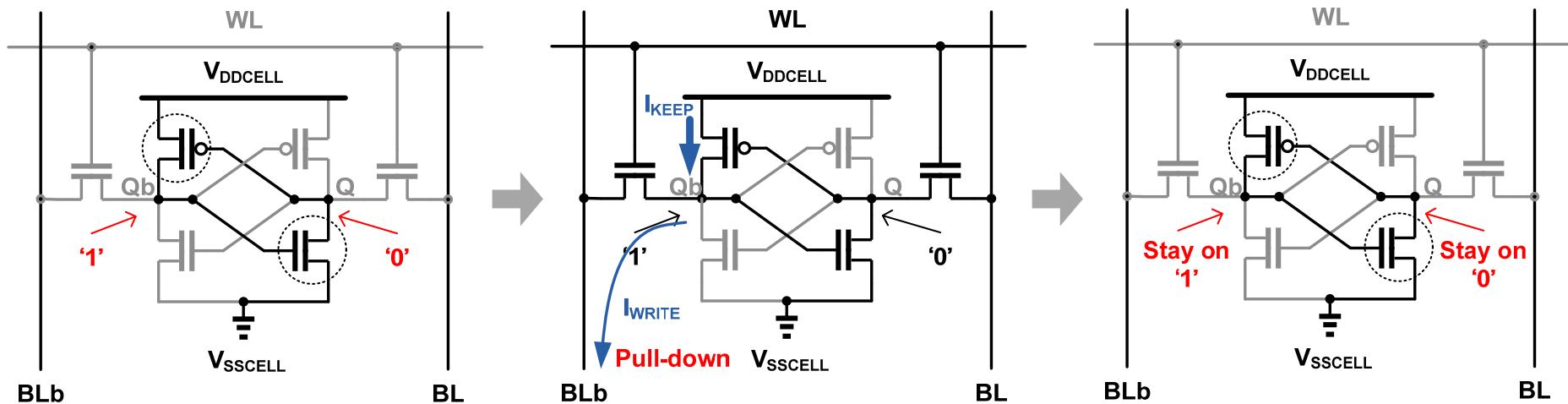


WL Overdrive (WLOD)		WL Underdrive (WLUD)	
overhead		overhead	
Timing	O	Timing	O
Area	O	Area	Δ
Instability	O	Instability	X

V _{DDCELL} Lowering (VDCL)		Negative BL (NBL)	
overhead		overhead	
Timing	X	Timing	Δ
Area	Δ	Area	O
Instability	O	Instability	X

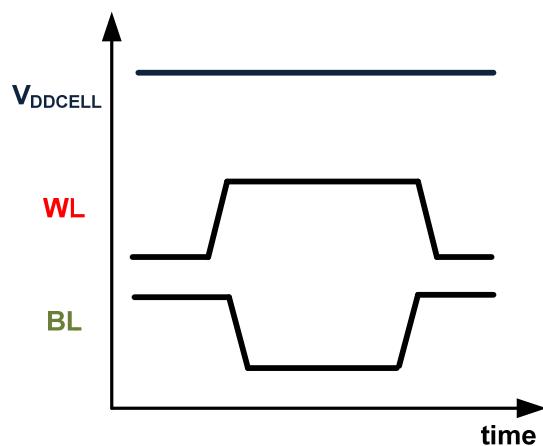
Write Assist? Write Failure Mechanism

- With process variations, PG is not strong enough to overpower PU and pull the internal node to ground
- Process variation also reduced the trip point of inverter holding the state '1', resulting in write failure

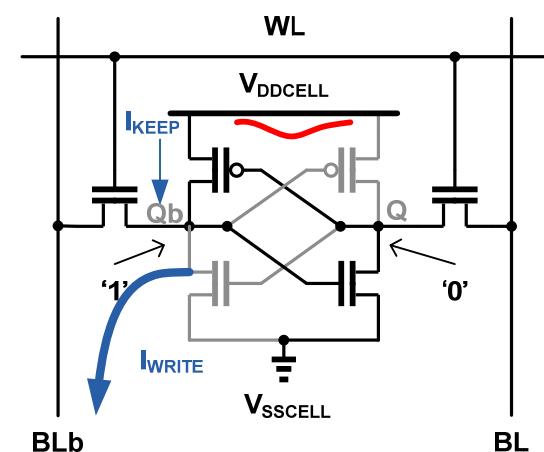
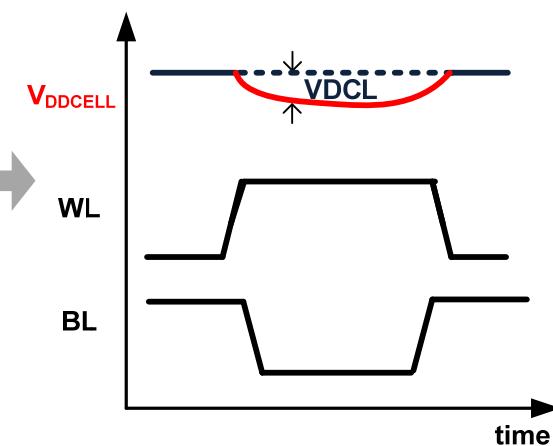


Write Assist? Principles of VDCL

- To allow safe write-operation, I_{WRITE} needs to be increased and/or I_{KEEP} decreased
- VDCL decreases I_{KEEP} with lower V_{GS} of PU
- VDCL can be lowered dynamically or statically



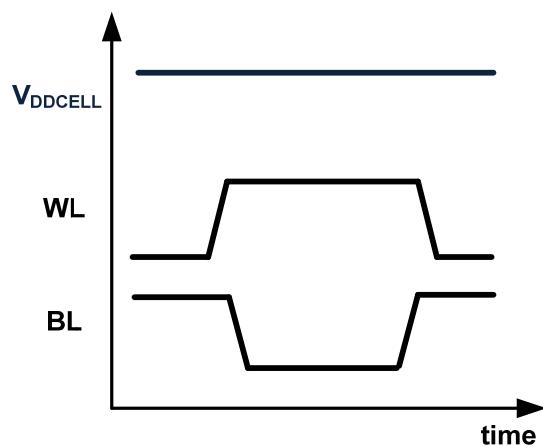
Timing diagram of Write Assists (VDCL)



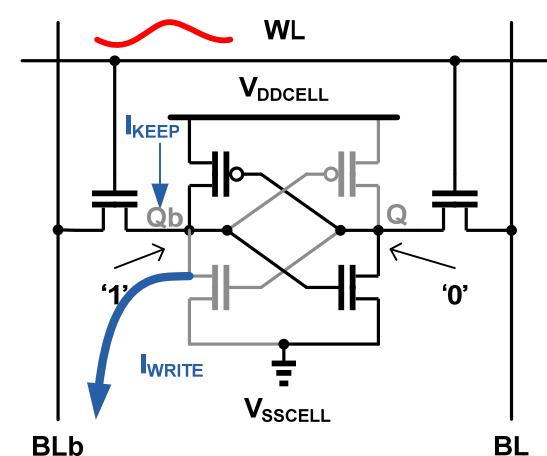
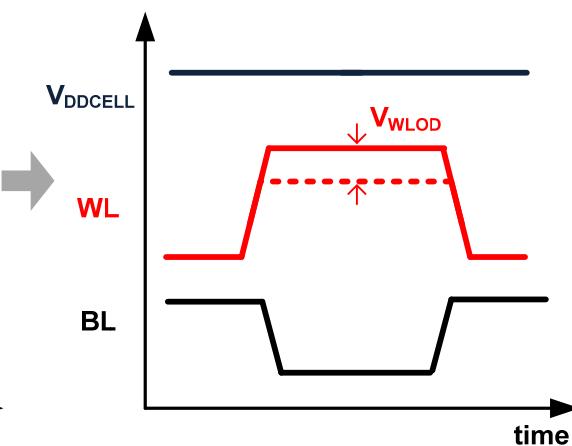
Current flow in Bitcell during Write Assist

Write Assist? Principles of WLOD

- To allow safe write-operation, I_{WRITE} needs to be increased and/or I_{KEEP} decreased
- WLOD increases I_{WRITE} with higher V_{GS} of PG
- WLOD can be raised dynamically or statically



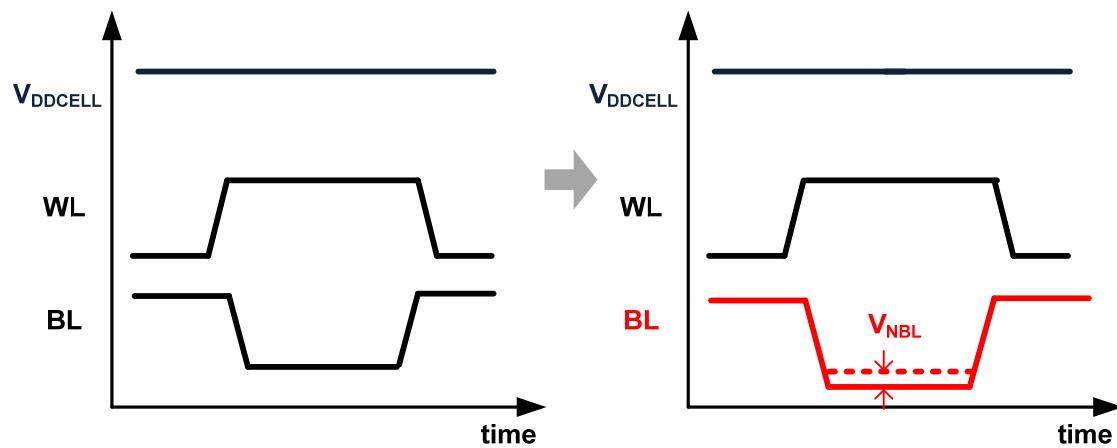
Timing diagram of Write Assists (WLOD)



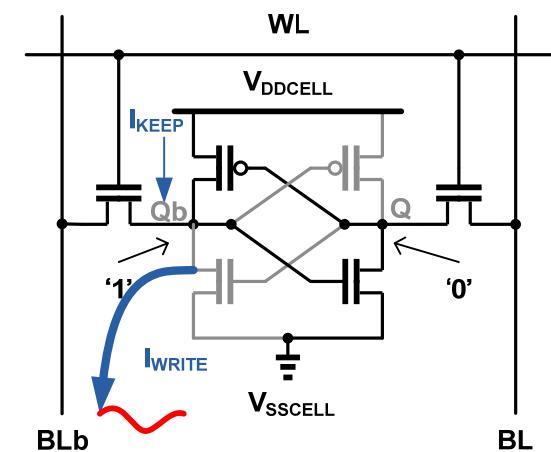
Current flow in Bitcell during Write Assist

Write Assist? Principles of NBL

- To allow safe write-operation, I_{WRITE} needs to be increased and/or I_{KEEP} decreased
- *NBL* increases I_{WRITE} with higher V_{GS} of PG
- *NBL* needs to be negatively boosted dynamically



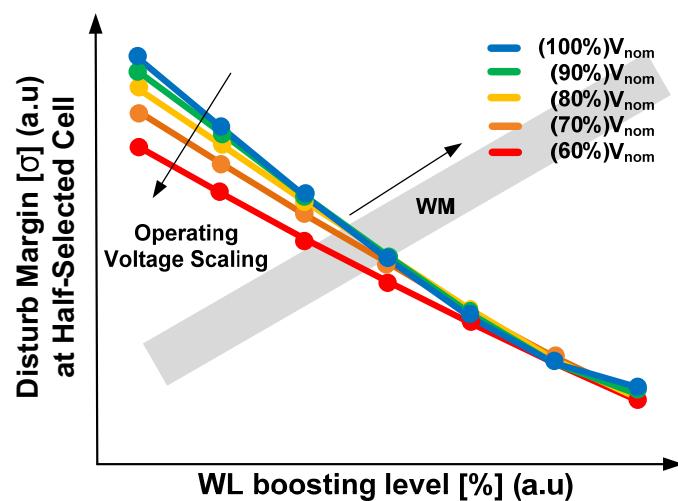
Timing diagram of Write Assists (NBL)



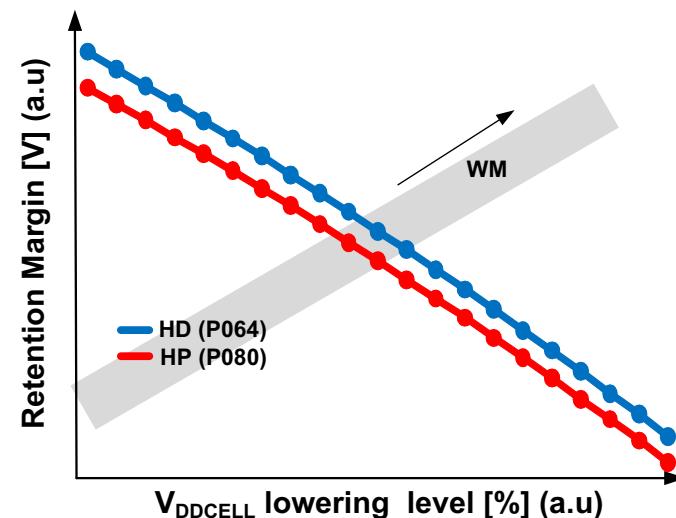
Current flow in Bitcell during Write Assist

Drawback of Conventional WOLD and VDCL

- *WLOD* impacts half-selected bitcell in a row with higher WL with disturbance noise
- *VDCL* impacts de-selected bitcells in a column with retention noise
- *NBL* has less drawback than *WLOD* or *VDCL* regarding disturbance margin and retention margin



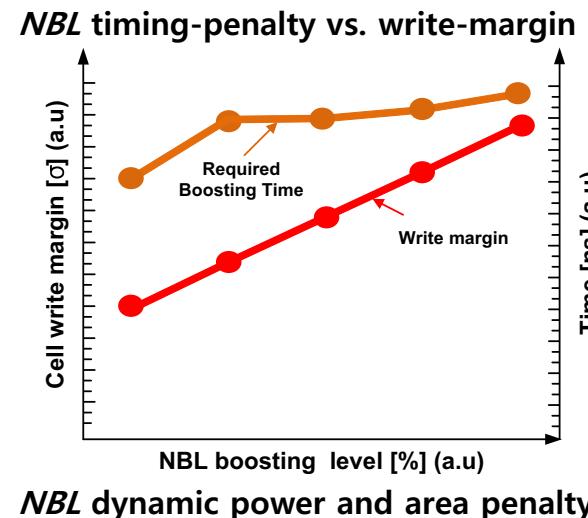
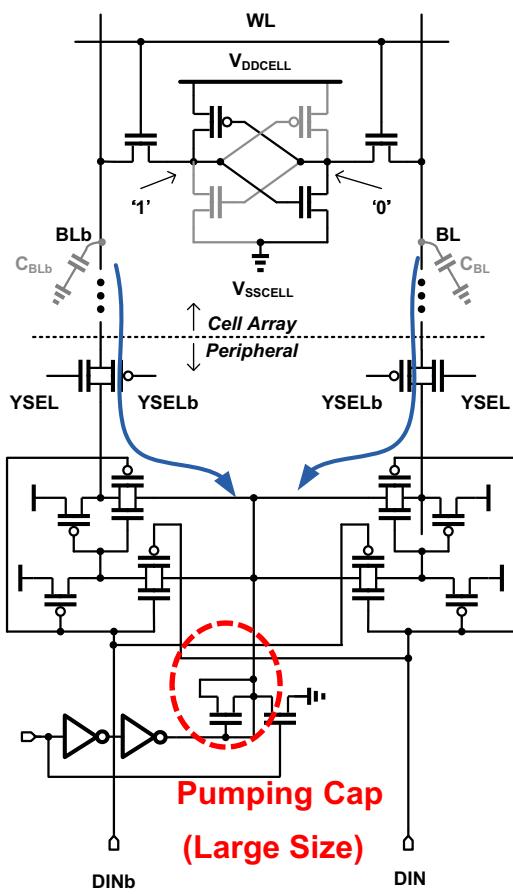
WLOD level vs. Cell Disturb Margin



VDCL vs. retention margin

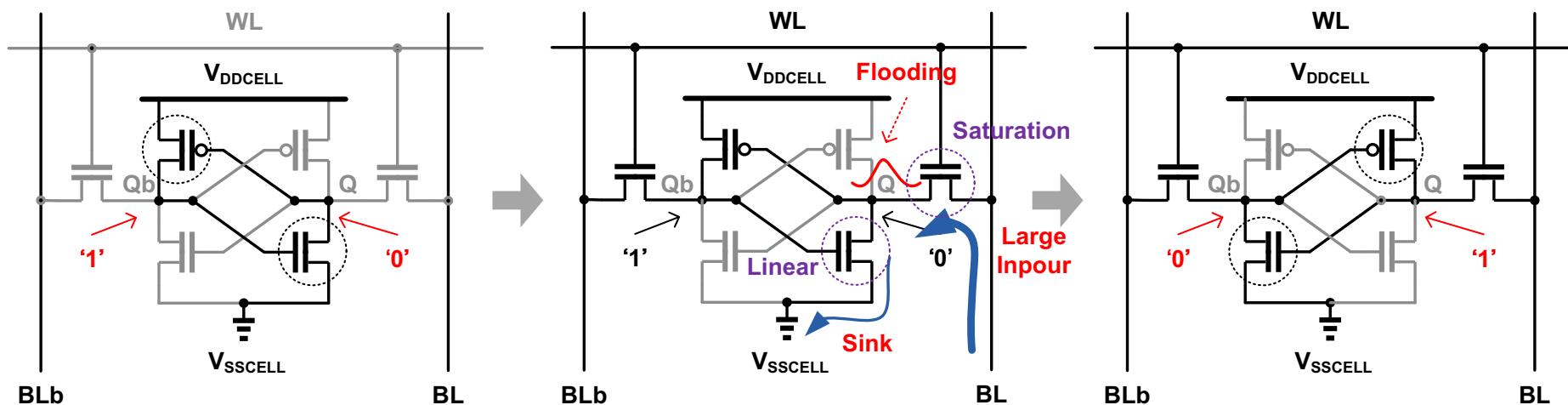
Drawback of NBL

- NBL needs to be optimized by considering timing, power, and area penalty versus necessary WM



Stability Assist? Disturb Failure Mechanism

- After fast turn-on of PG, charge in BLs flow into the latch-nodes of bitcell
- If the data-'0'-node voltage reaches the trip voltage of the cell inverter, bitcell flips by destroying data



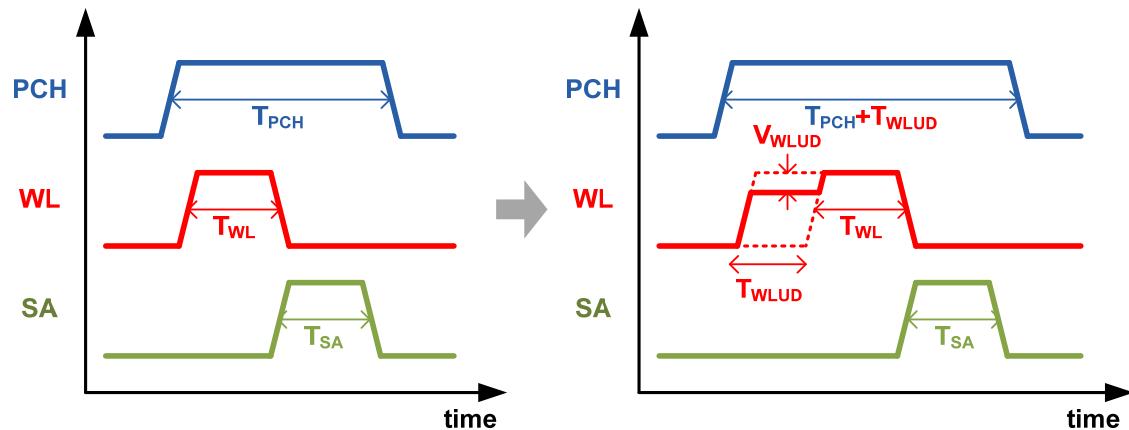
Principles of Stability Assists (Disturb)

- Parameters of Disturbance :

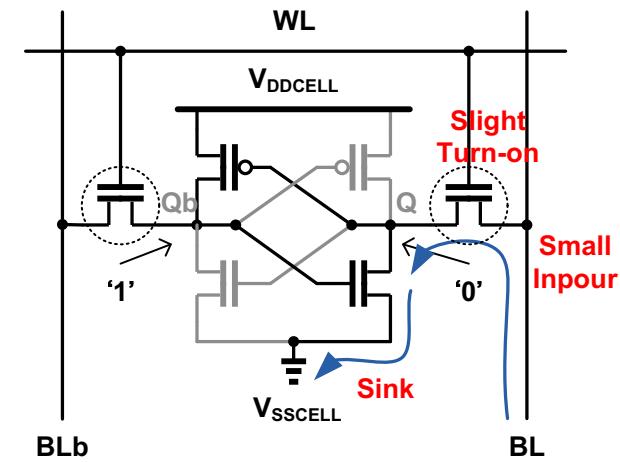
- (1) Device mismatch → Designer can't handle it
- (2) BL capacitance (Rows per BL) → Area & speed penalty
- (3) Precharge voltage level → Need additional analog circuits
- (4) Charge-injection speed

- WLUD Technique

- Turn WL on slowly before full-access



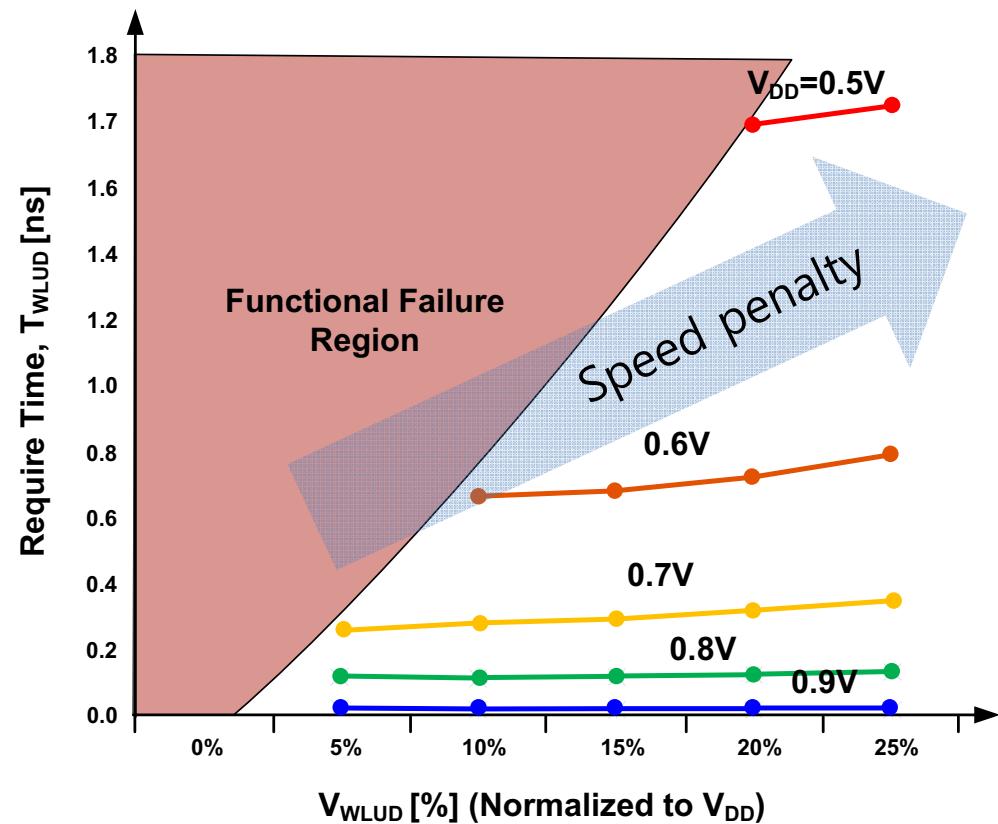
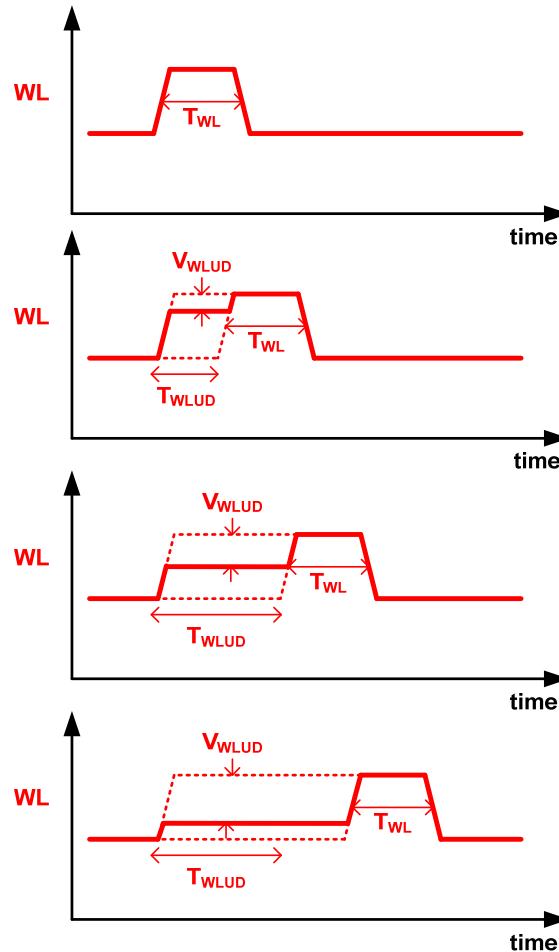
Typical WLUD timing diagram



Current flow in Bitcell during WLUD

Drawback of Conventional Stability Assist

- WLUD needs additional timing to synchronize BL with cell data slowly → speed penalty



Required time of WLUD for enough ADM
(Worst bitcell Corner,)

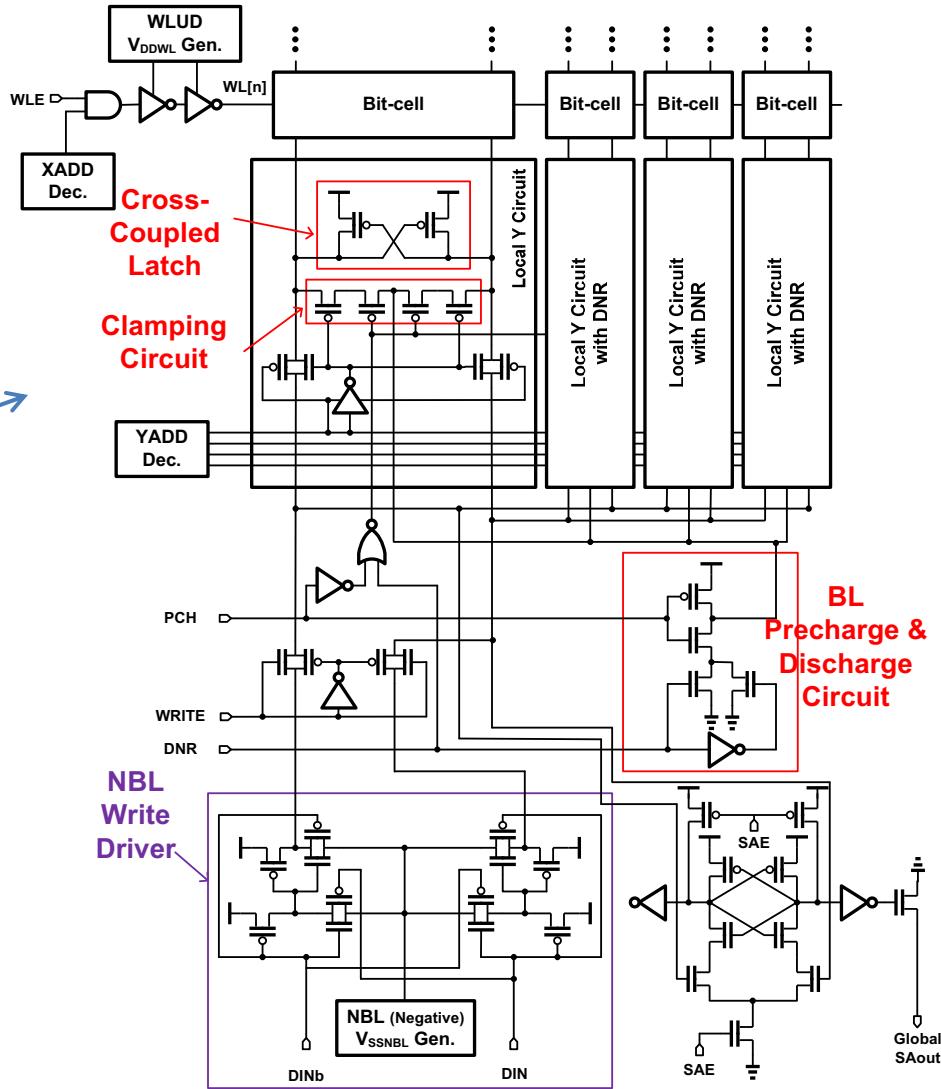
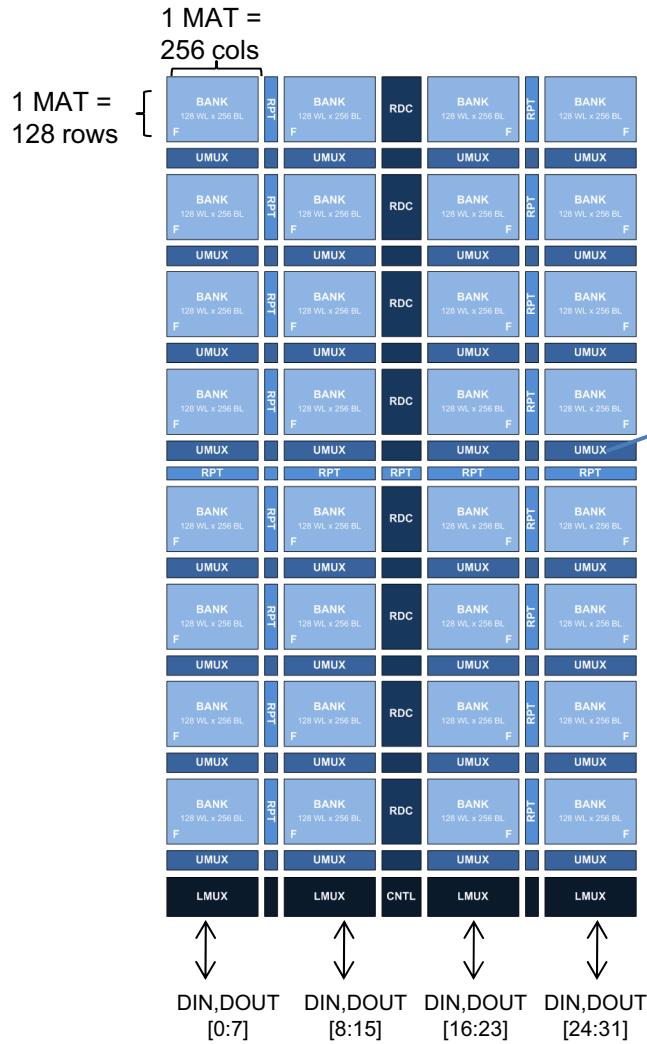
Outline

- Motivation
- FinFET: Opportunity and Challenges to SRAM
- Conventional SRAM Assist Techniques
- Proposed SRAM Assist Scheme
- Implementation and Measurement Results
- Conclusions

Assist Schemes in this Work

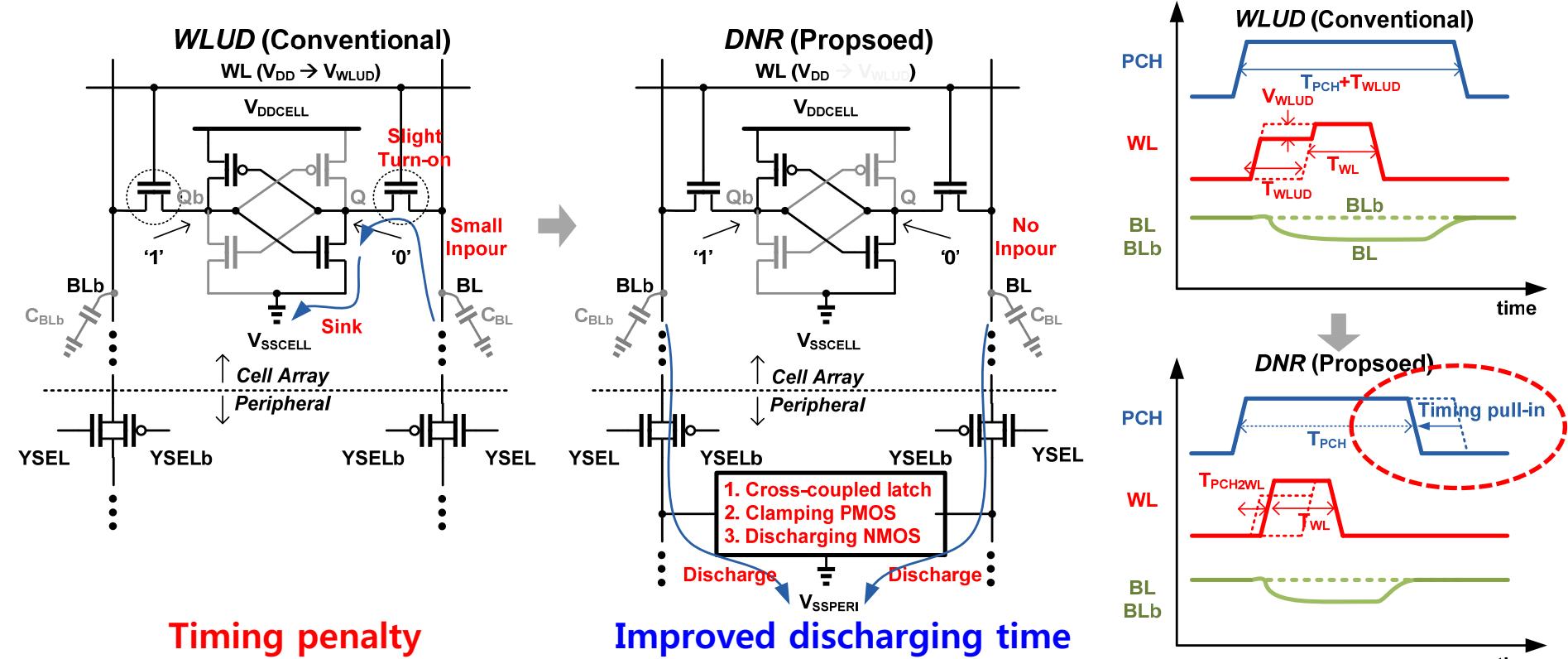
- **Disturbance Noise Reduction (*DNR*) (Proposed)**
 - Reduce BL noise **dynamically** at WL-enabling timing
 - Discharge BL noise with **strong** driver in Ypath
 - Keep BL level **safe** with cross-coupled latch and clamping circuit
- **Negative BL (*NBL*) Technique (Conv.)**
 - Drive negative BL voltage to allow higher I_{WRITE}
 - Free of half-selected problem or retention problem

Schematic (*This Work*)



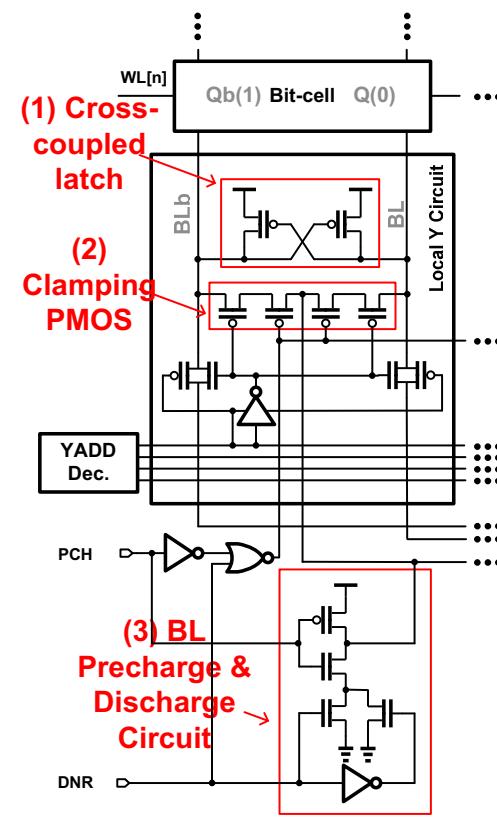
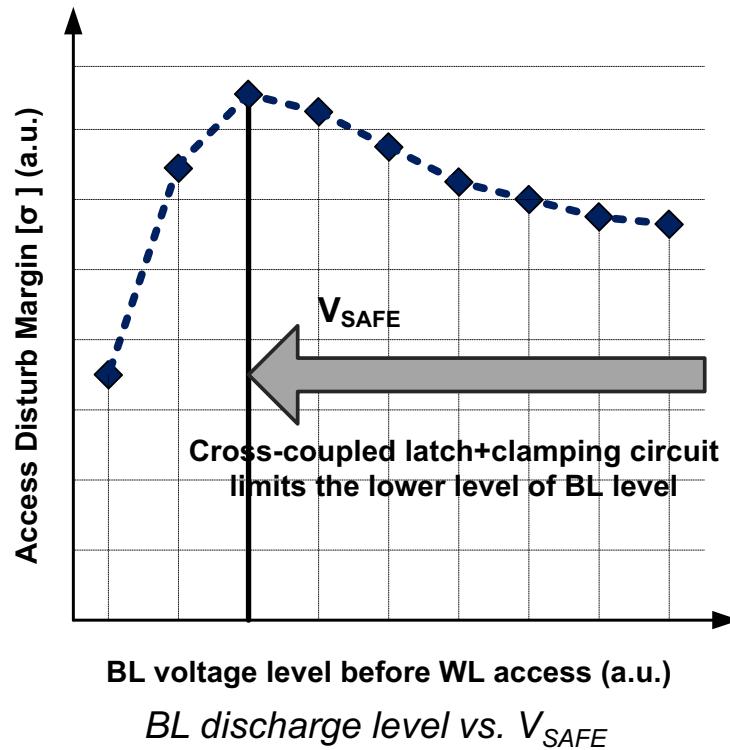
Principle of DNR (Proposed)

- DNR discharges BLs to reduce charge injection using **strong NMOS** in peripheral circuit
- DNR doesn't need T_{WLUD}



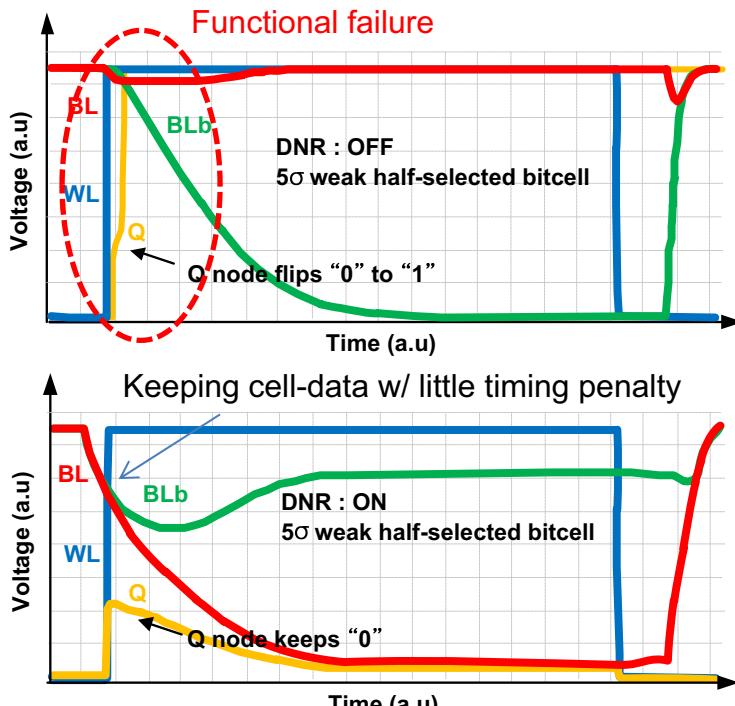
Limitation of DNR: V_{SAFE}

- V_{SAFE} is the minimum voltage level not to induce ‘0’ noise to $Q_b(1)$ through BL_b
- Cross-coupled latch and clamping circuit guarantees $BL_b=‘1’$ with the self-biasing of BL and $Q(0)$

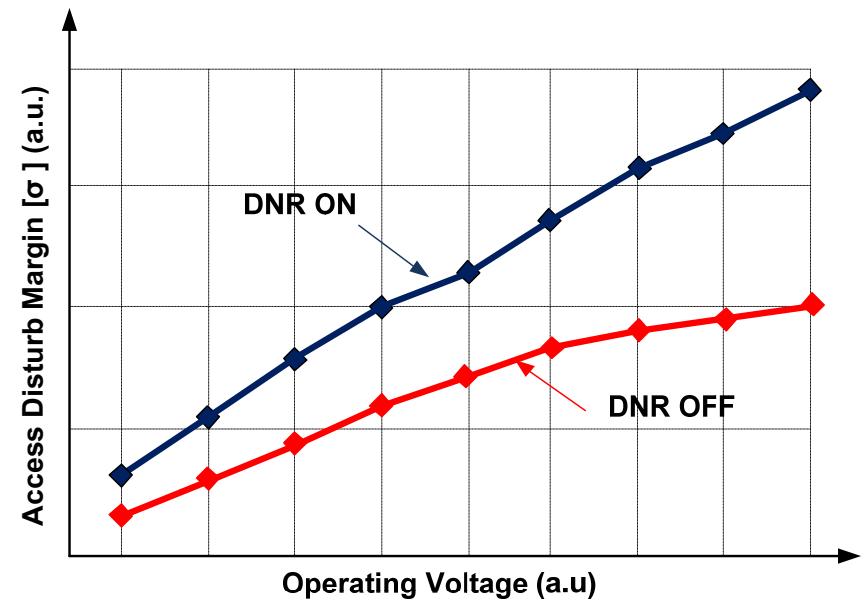


DNR : Simulation Results

- DNR keeps cell-data by reducing disturbance-noise to bitcell
- DNR shows little timing penalty, since BL is forced to lower level with strong NMOS in peripheral



Waveform (DNR OFF vs. ON) of Weak Cell



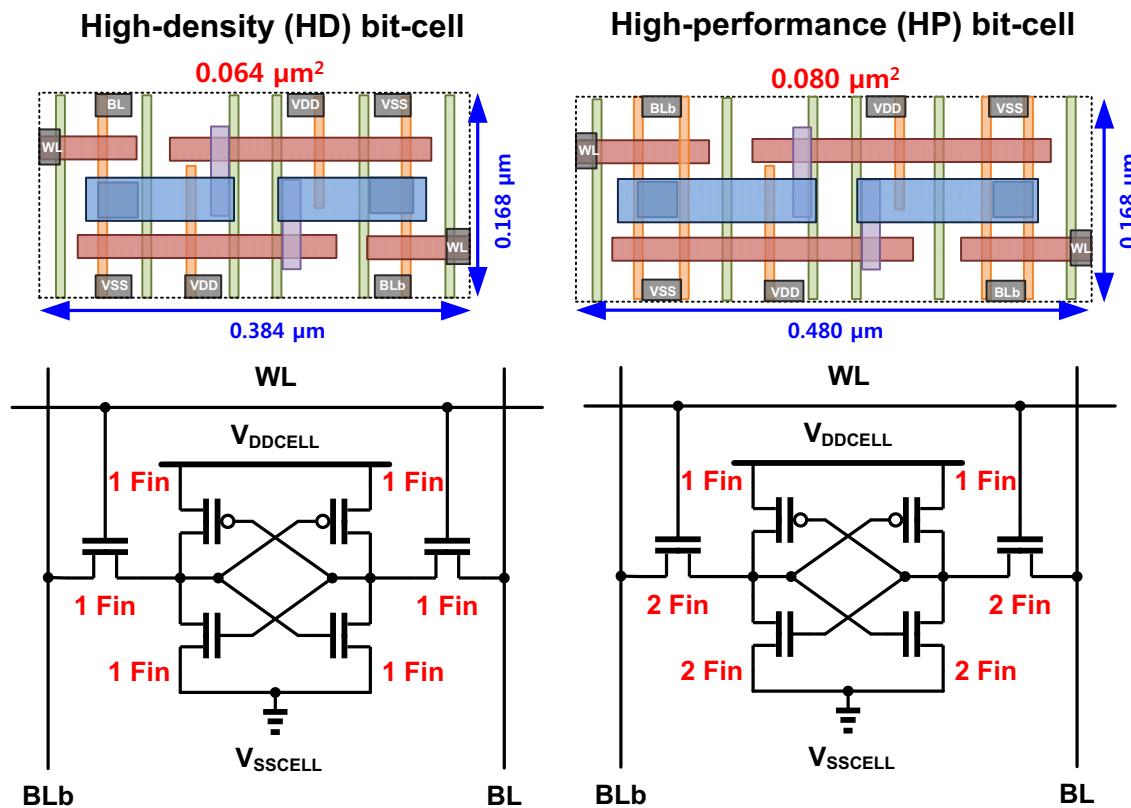
ADM (DNR OFF vs. ON) of Weak Cell

Outline

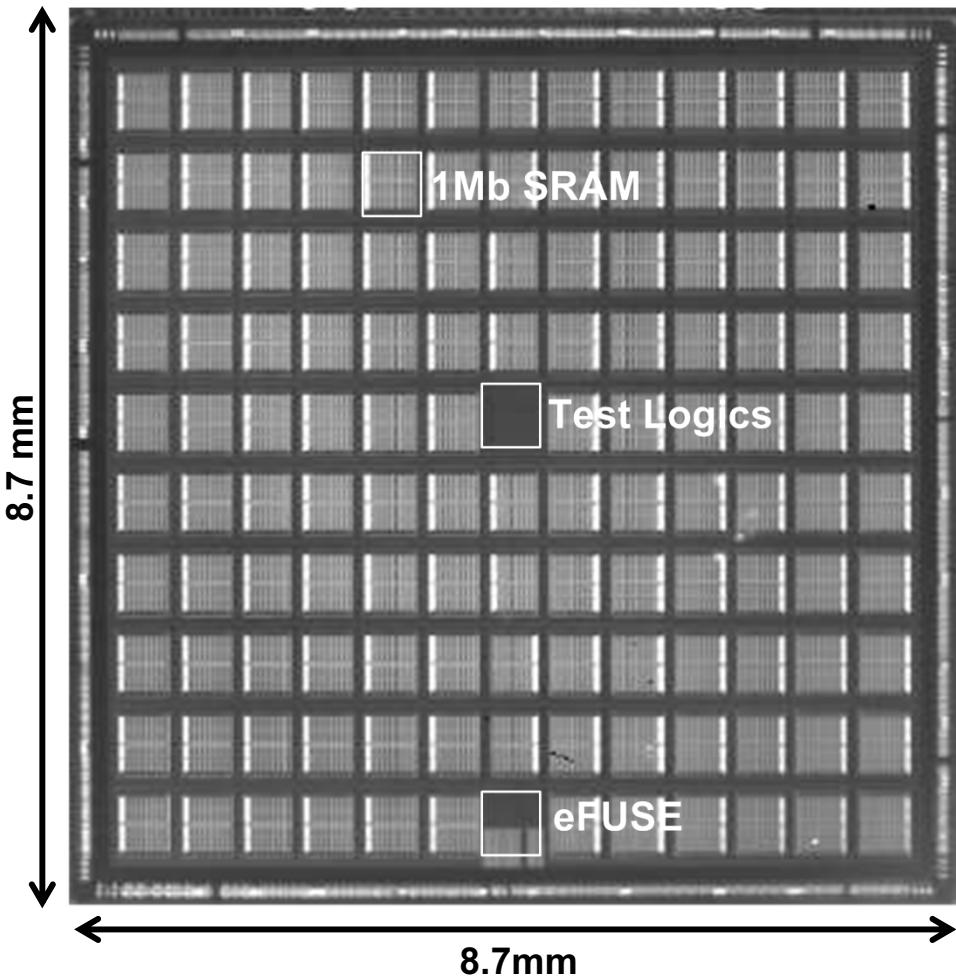
- Motivation
- FinFET: Opportunity and Challenges to SRAM
- SRAM Assist Techniques
- Proposed SRAM Assist Scheme
- Implementation and Measurement Results
- Conclusions

14nm FinFET 6T SRAM Bitcell

- High-density (**smallest**) and high-performance bitcell
- HD $0.064 \mu\text{m}^2$ with PU:PG:PD=1:1:1
- HP $0.080 \mu\text{m}^2$ with PU:PG:PD=1:2:2



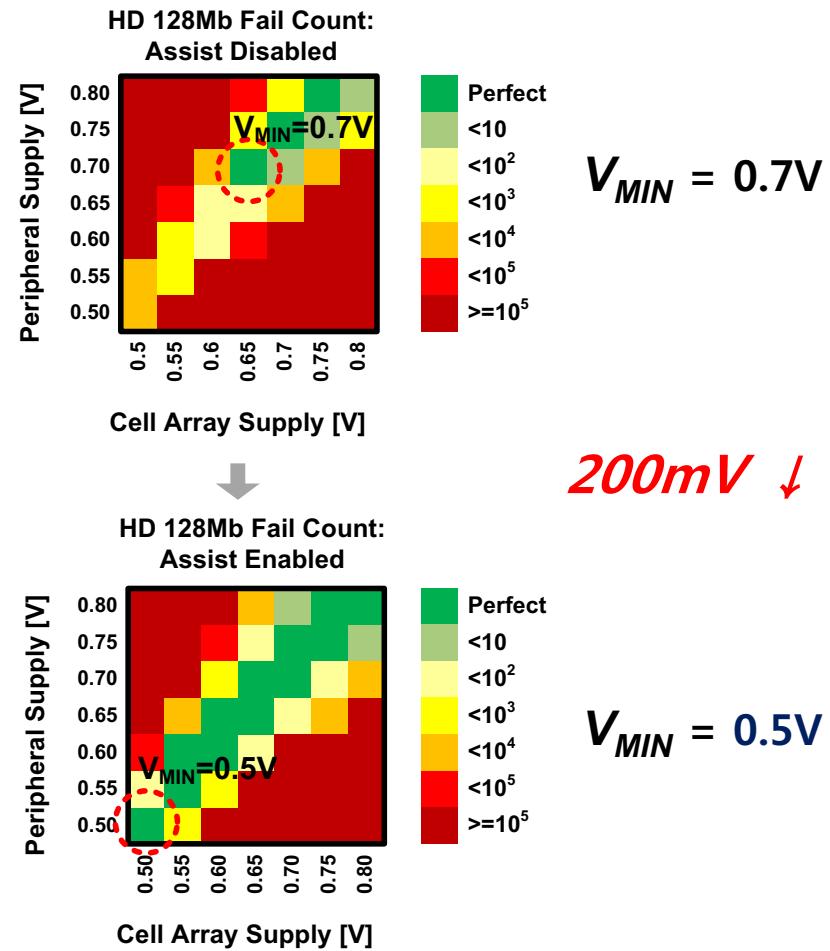
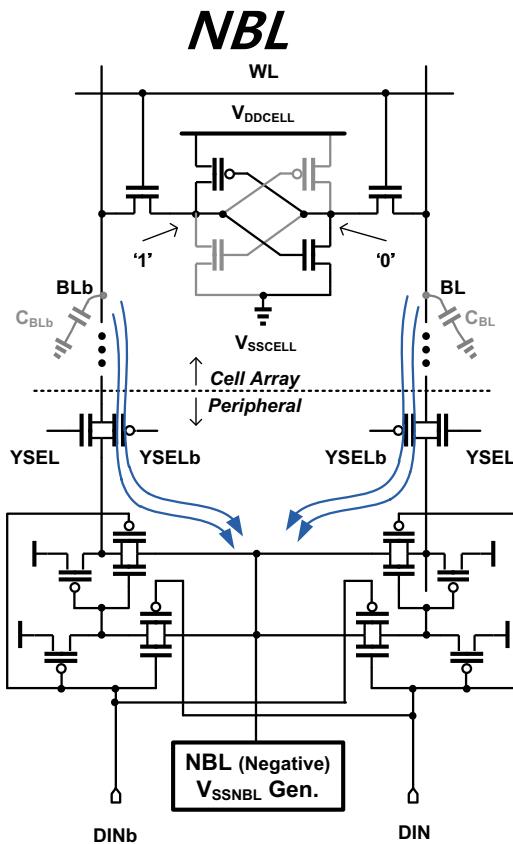
Die Micrograph of 128Mb SRAM



Technology	14nm Bulk FinFET
Chip Size	75.6mm ²
Density	128Mb (8 Muxed IO in Chip, 32 IO per Macro)
Metal	9 Metal for Chip (5 Metal Probe Available for Macro)
IPs	1Mb SRAM x 128, eFUSE, 1.8V General Purpose IO, Standard Cells
Power Supply	0.8V (nominal)

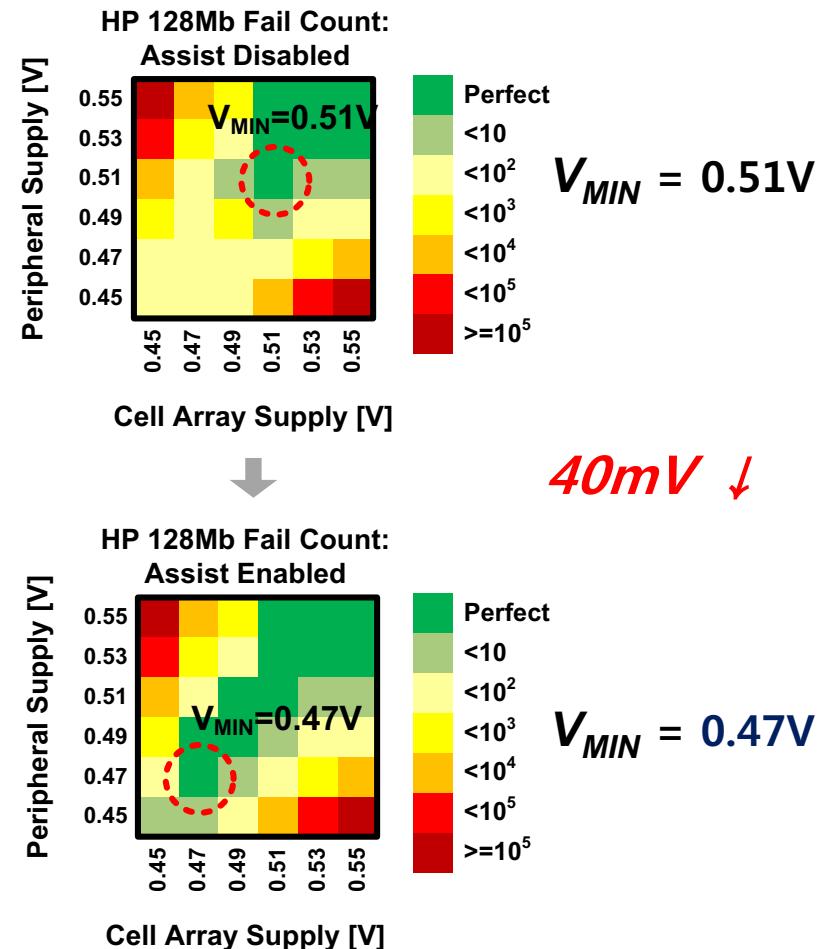
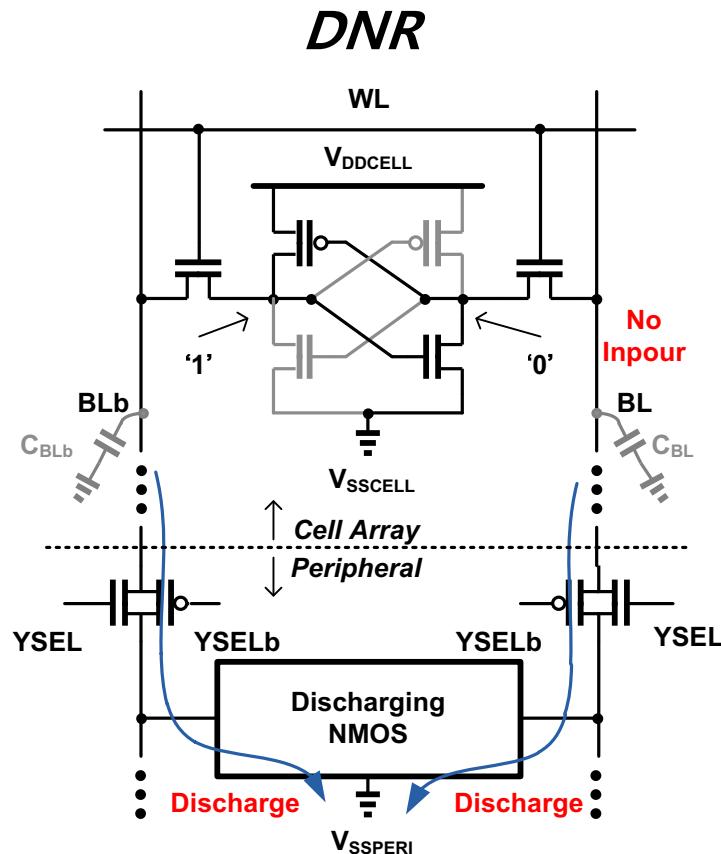
Measurement Results (HD 128Mb SRAM)

- **NBL** helps to lower V_{MIN} of 6T-HD SRAM by 200mV
- **NBL** does not make disturbance noise nor retention penalty



Measurement Results (HP 128Mb SRAM)

- DNR helps to lower V_{MIN} of 6T-HP SRAM by 40mV
- DNR has little timing penalty and 0.87% area penalty



Outline

- Motivation
- FinFET: Opportunity and Challenges to SRAM
- SRAM Assist Techniques
- Proposed SRAM Assist Scheme
- Implementation and Measurement Results
- Conclusions

Conclusion

- 14nm FinFET 128Mb 6T SRAM is fully demonstrated
- 6T-HD $0.064 \mu\text{m}^2$ shows 200mV V_{MIN} improvement with *NBL* write-assist
 - $0.064 \mu\text{m}^2$ is the smallest bitcell published
- 6T-HP $0.080 \mu\text{m}^2$ shows 40mV V_{MIN} improvement with *DNR* disturbance-assist
- Assist circuits enable low-voltage SRAM for low-power SoC

20nm High-Density Single-Port and Dual-Port SRAMs with Wordline-Voltage-Adjustment System for Read/Write Assist

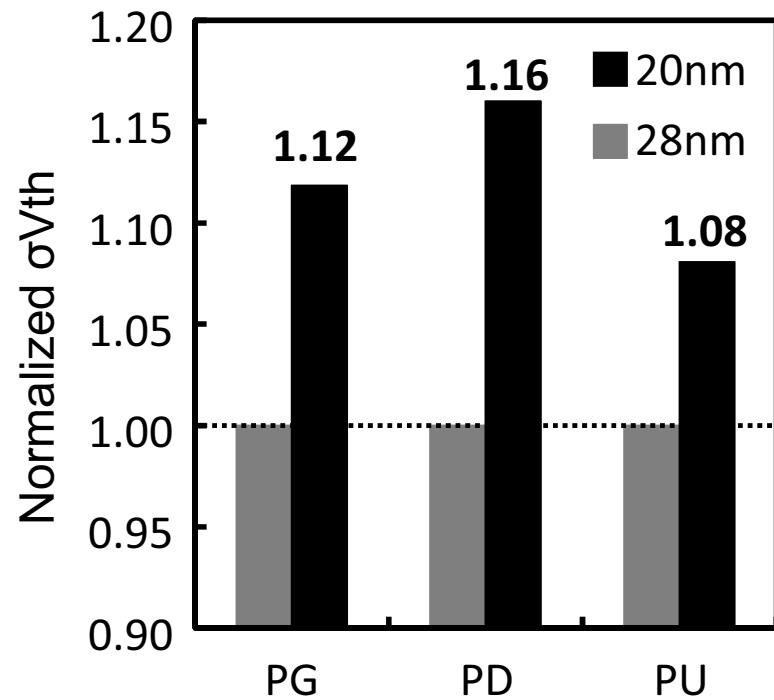
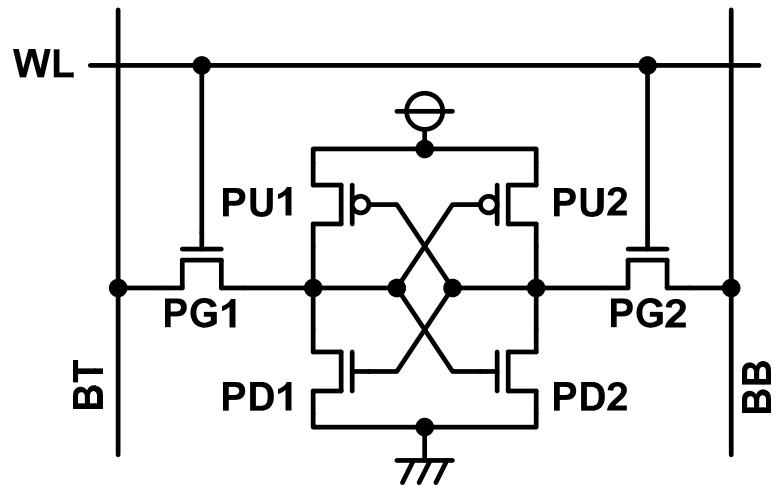
Makoto Yabuuchi, Yasumasa Tsukamoto,
Masao Morimoto, Miki Tanaka, Koji Nii



Overview

- Background
- Wordline-Voltage-Adjustment system
- 20nm design and evaluation of test chips
- Conclusion

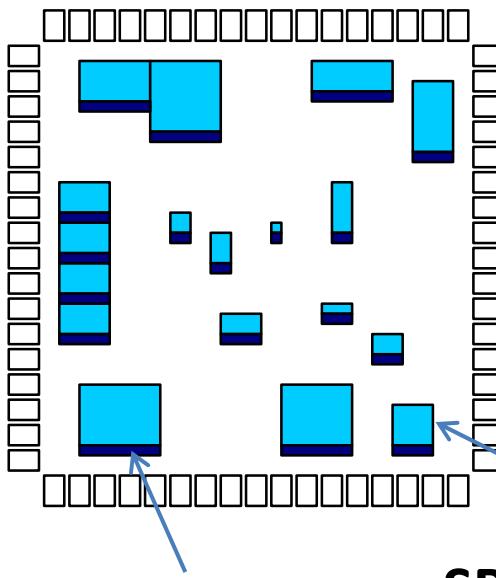
Background



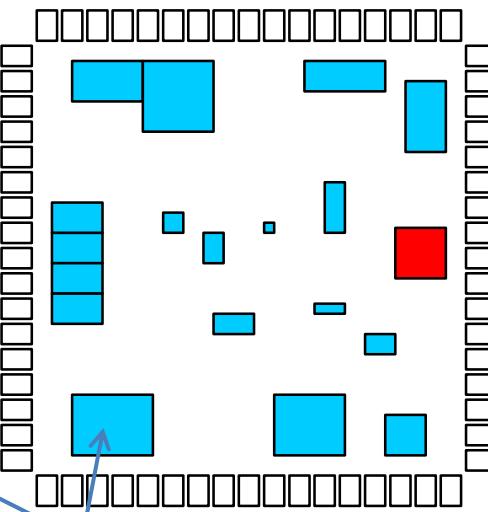
**Simple scaling induces local variation, which degrades SRAM characteristics
→ Need assist circuitries for read/write operation**

Background

w/ conv. Assist

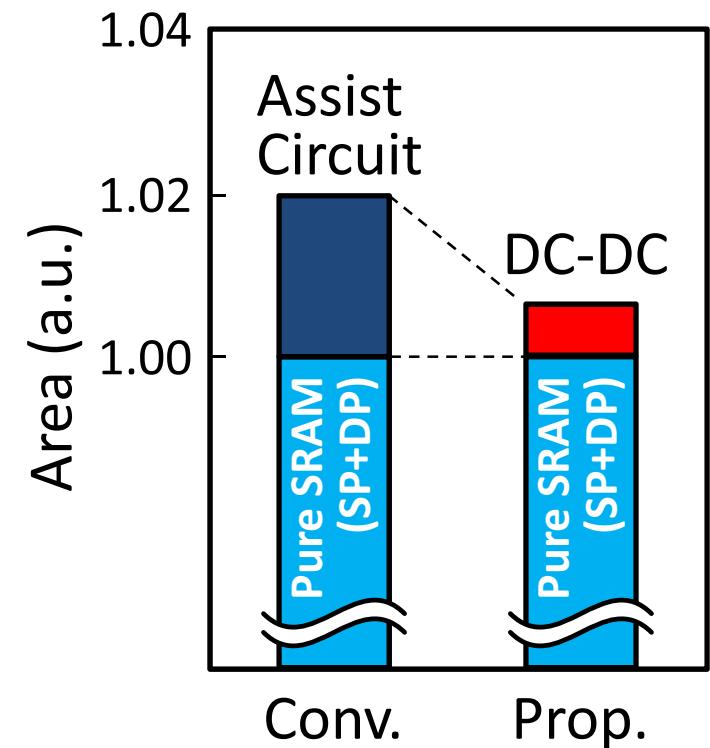


Proposed



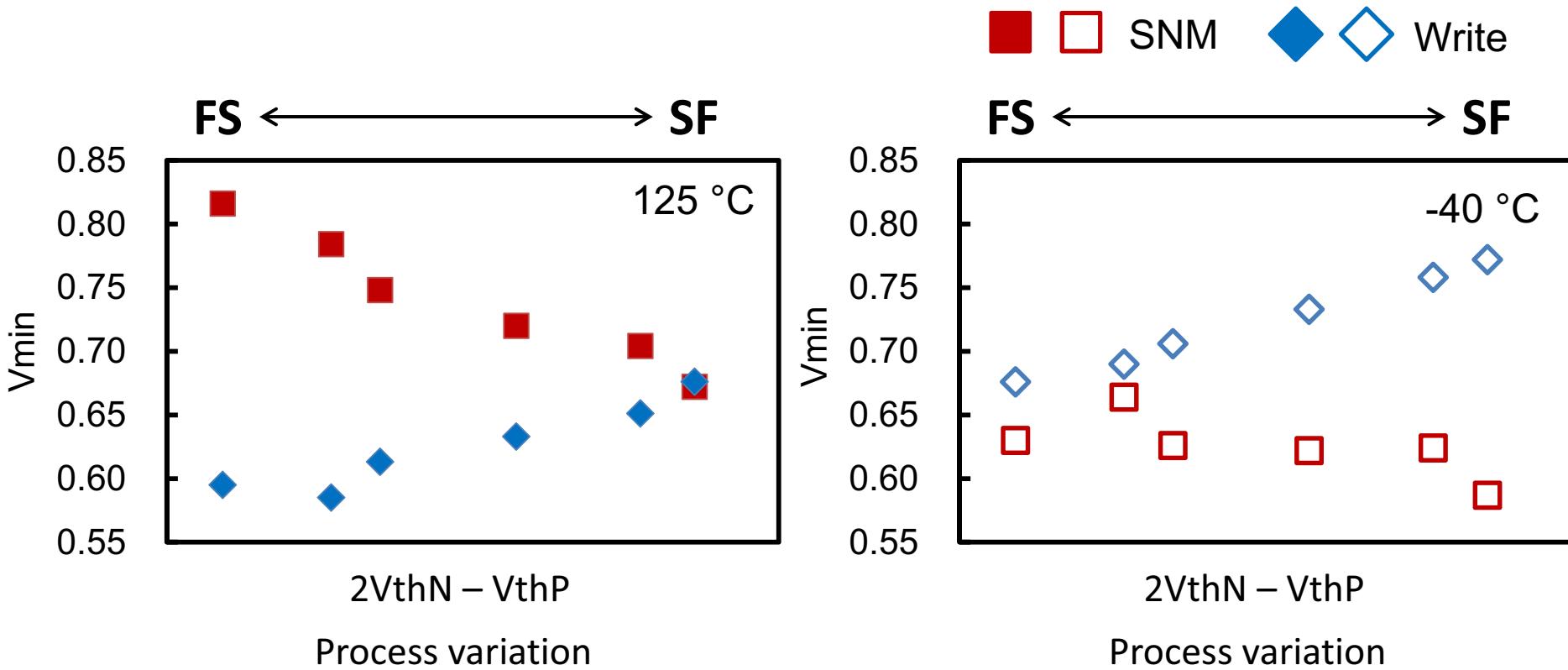
Assist circuit

SP/DP SRAM HM



SRAM hard macros(HM) with various bit/word config.
Chip level dual supply + DC-DC > assist circuit w/ each HM

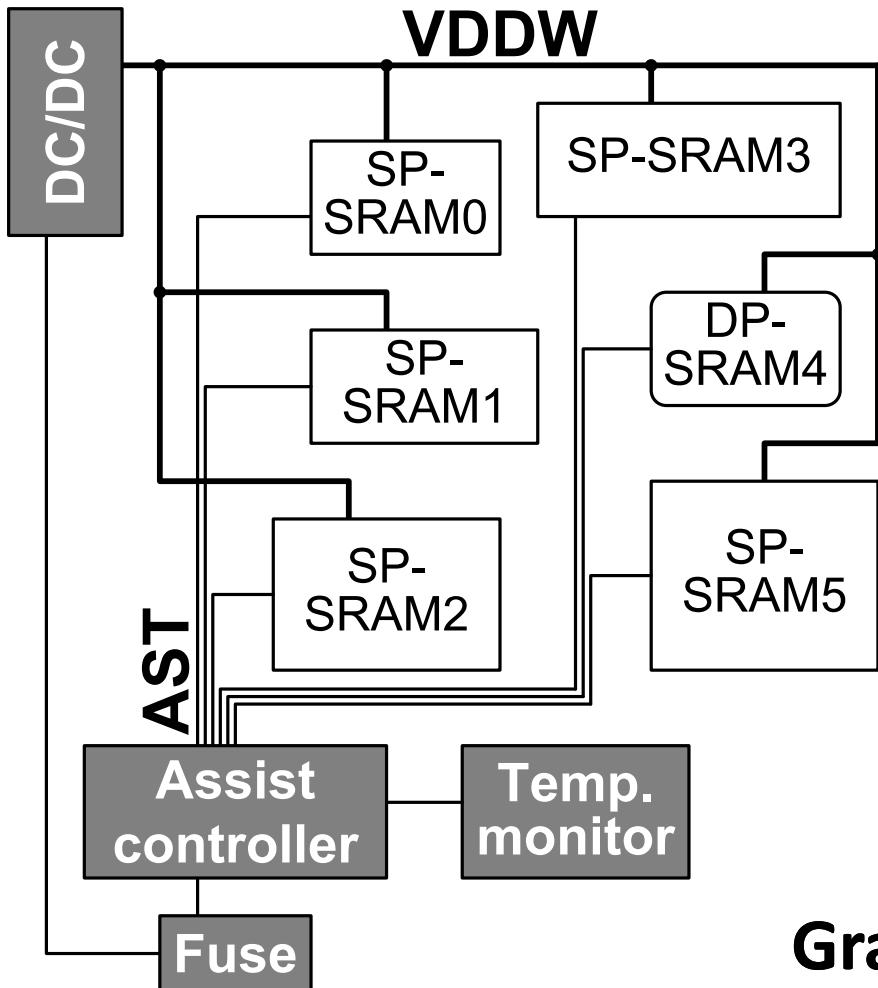
SRAM Vmin correlation



Two worst cases to be considered for Vmin improvement

- **High Temp & FS Corner for SNM → Lowered WL**
- **Low Temp & SF Corner for Write → Ramped up WL**

Wordline Voltage Adjustment system

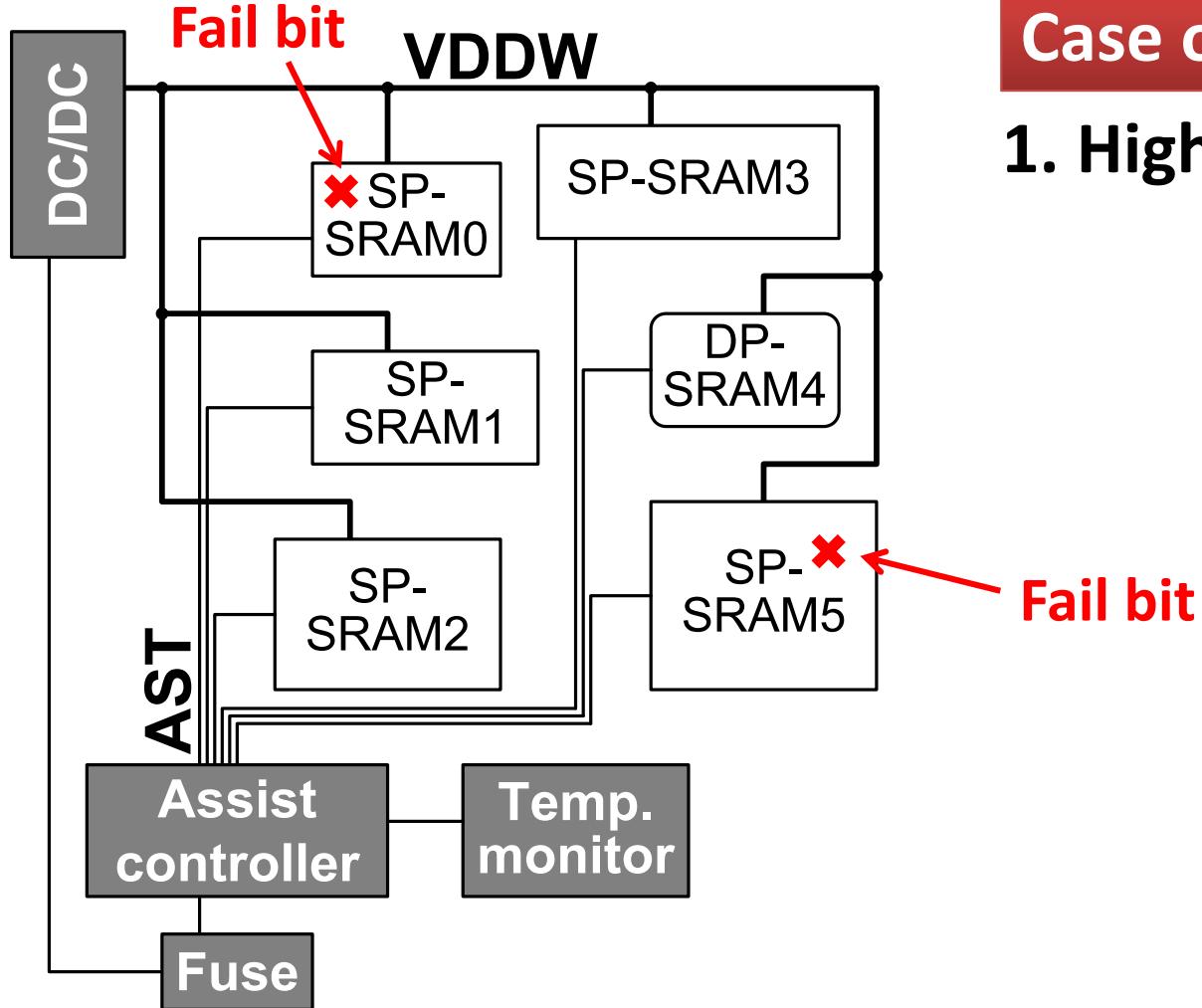


Block function

- DC/DC block
VDD for wordline(VDDW) supply
- AST signal
Wordline level selector to VDD or VDDW
- Assist controller block
Control AST signals for each macro
- Fuse block
Store block location, WL Voltage
- Temp. monitor block

**Grain wordline voltage control
improve Vmin**

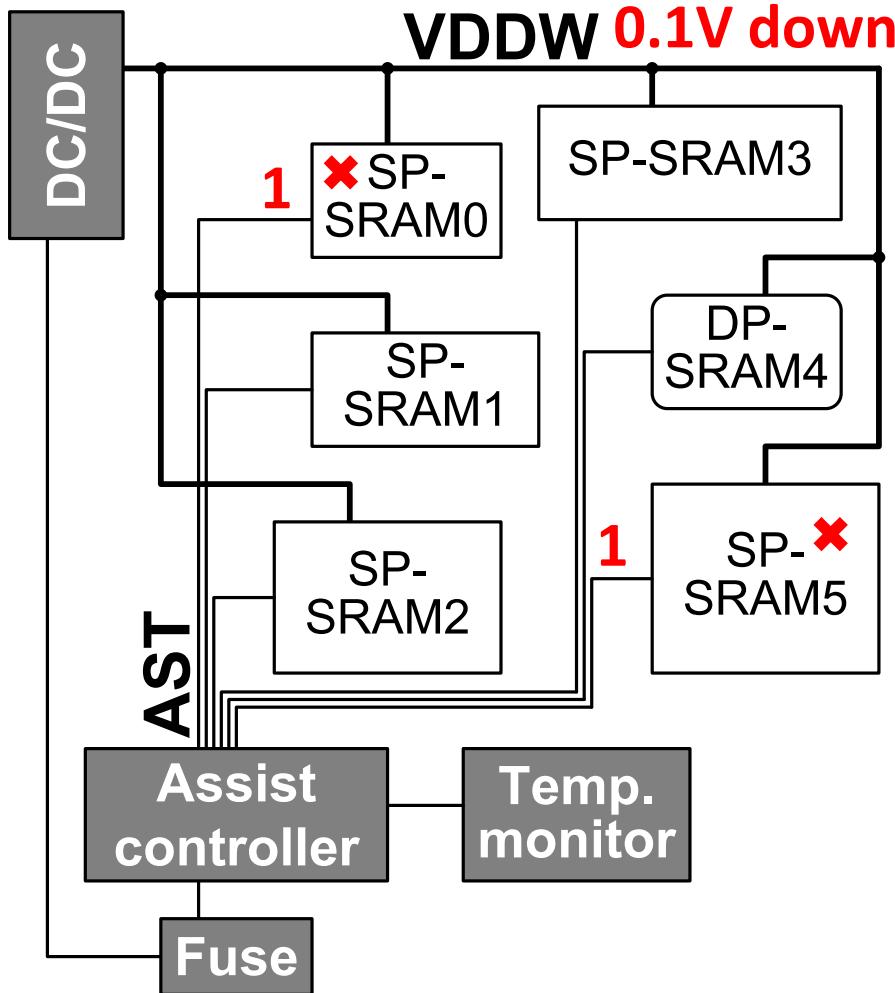
WDVA system sequence



Case of SNM fail chip

1. High temp. test

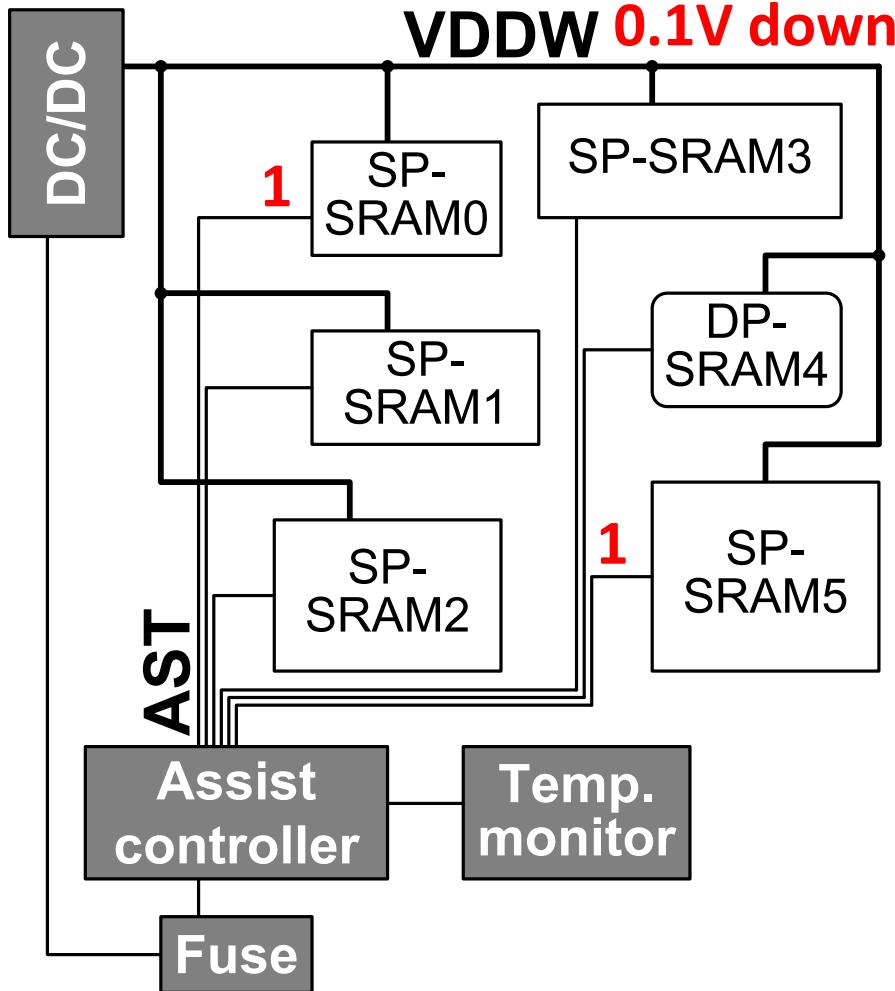
WDVA system sequence



Case of SNM fail chip

1. High temp. test
2. VDDW level down
AST signal enable

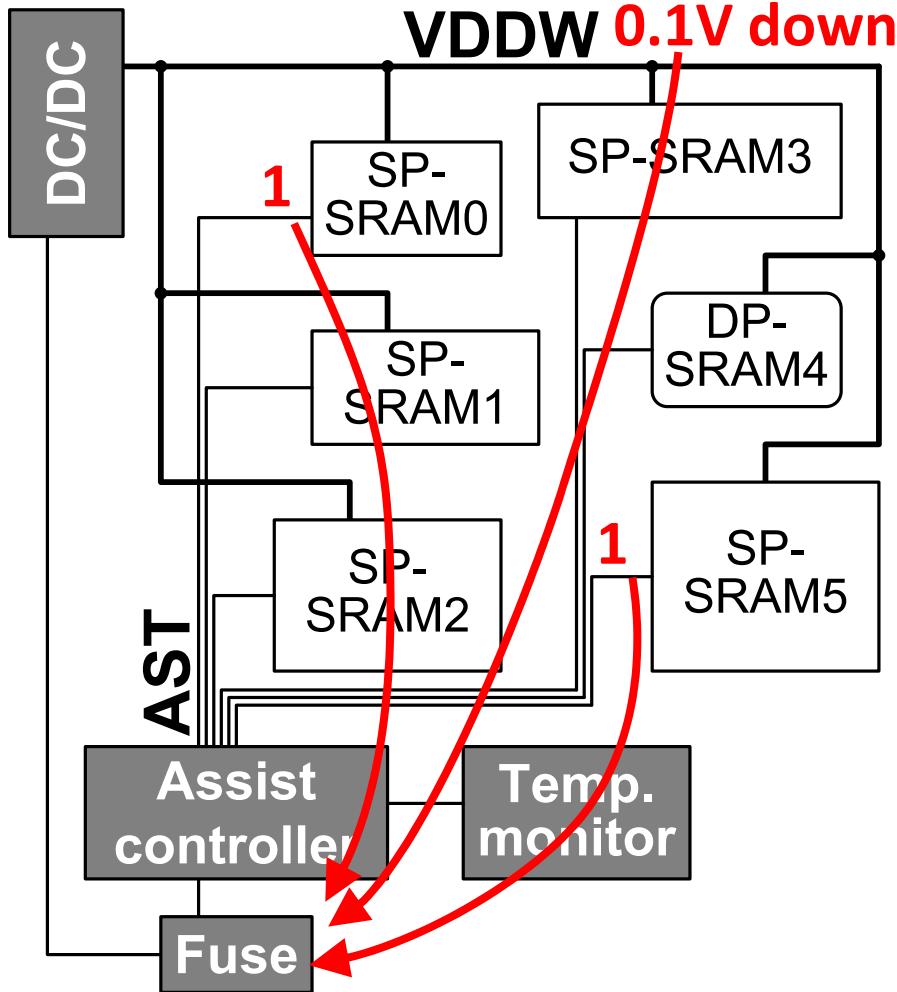
WDVA system sequence



Case of SNM fail chip

1. High temp. test
2. VDDW level down
AST signal enable
3. Re-test

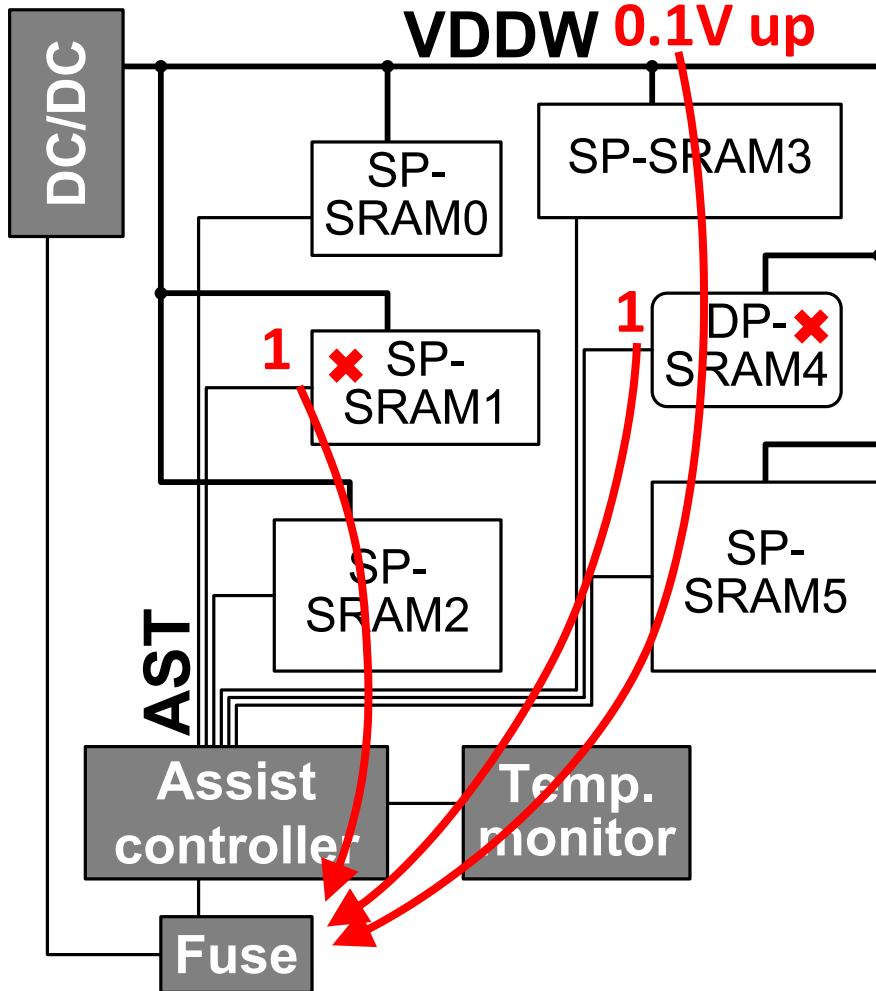
WDVA system sequence



Case of SNM fail chip

1. High temp. test
2. VDDW level down
AST signal enable
3. Re-test
4. Programming fuse
 - H.T.
 - $\text{VDDW} = \text{VDD} - 0.1\text{mV}$
 - **SRAM0 and SRAM5**

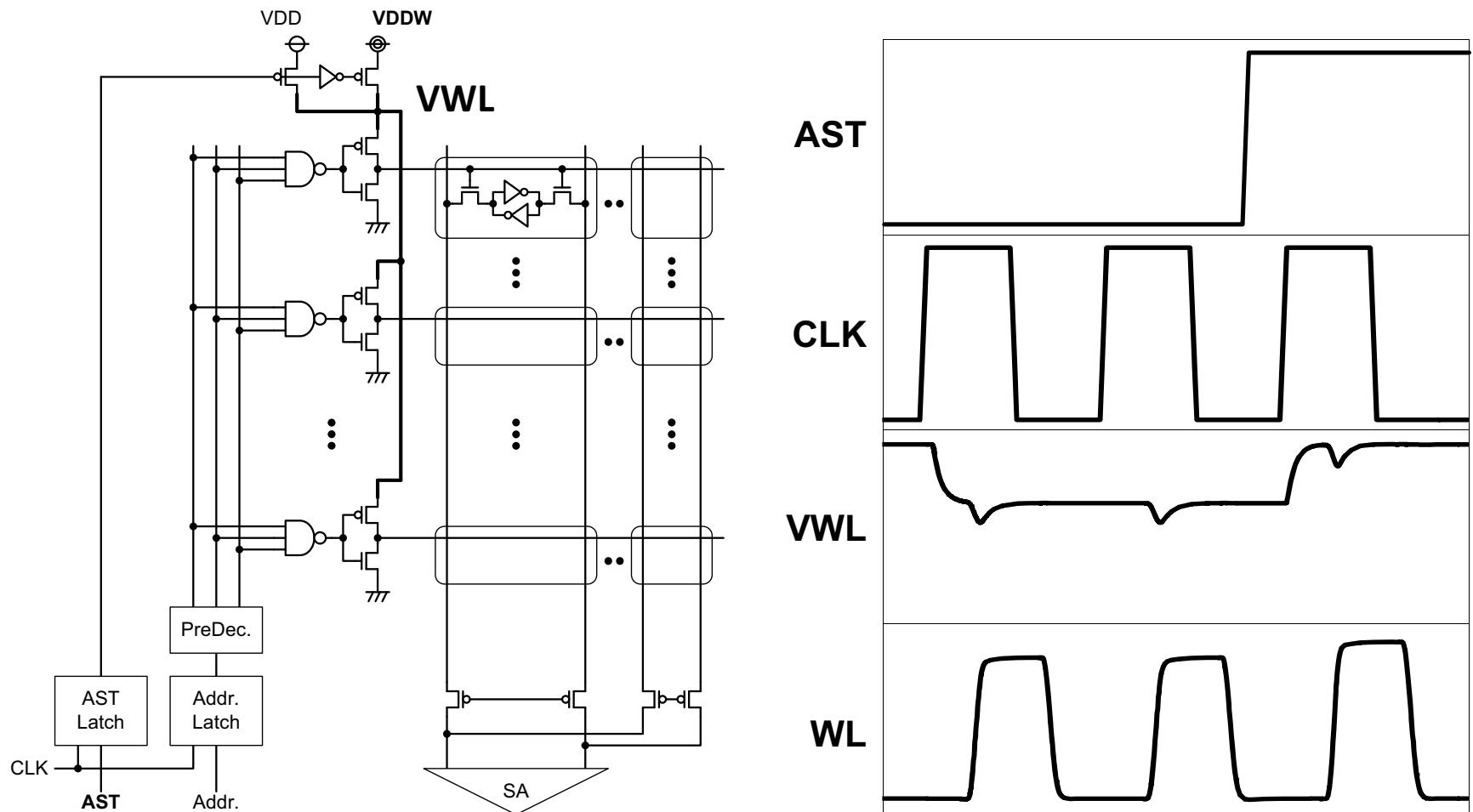
WDVA system sequence



Case of write fail chip

1. High temp. test
→ Pass
2. Low temp. test
3. VDDW level up
AST signal enable
4. Re-test
4. Programing fuse
 - L.T.
 - $VDDW = VDD + 0.1V$
 - SRAM1 and SRAM4

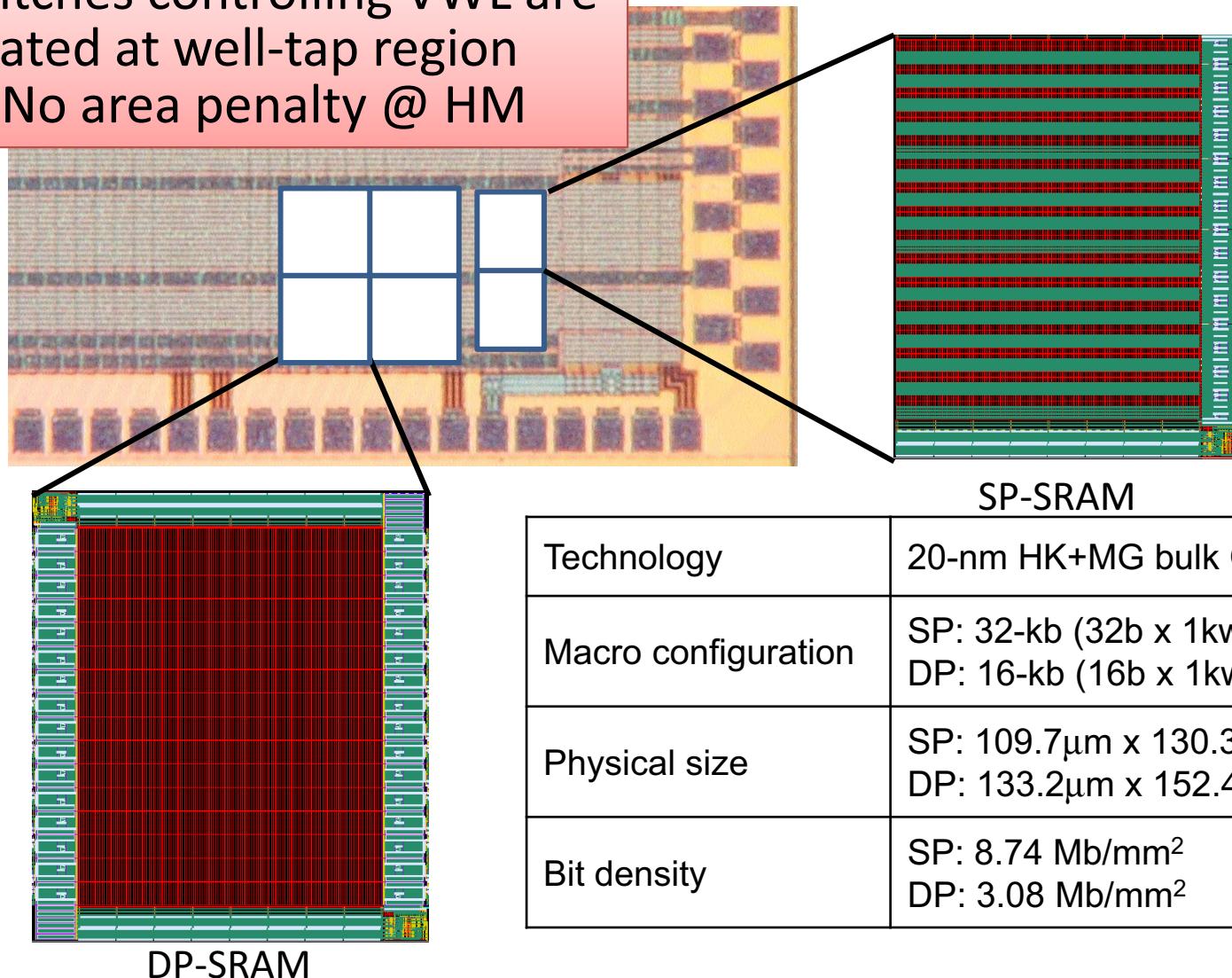
Circuit of WDVA SRAM macro



Transition of VWL is completed before each WL is activated, so no impact to speed degradation

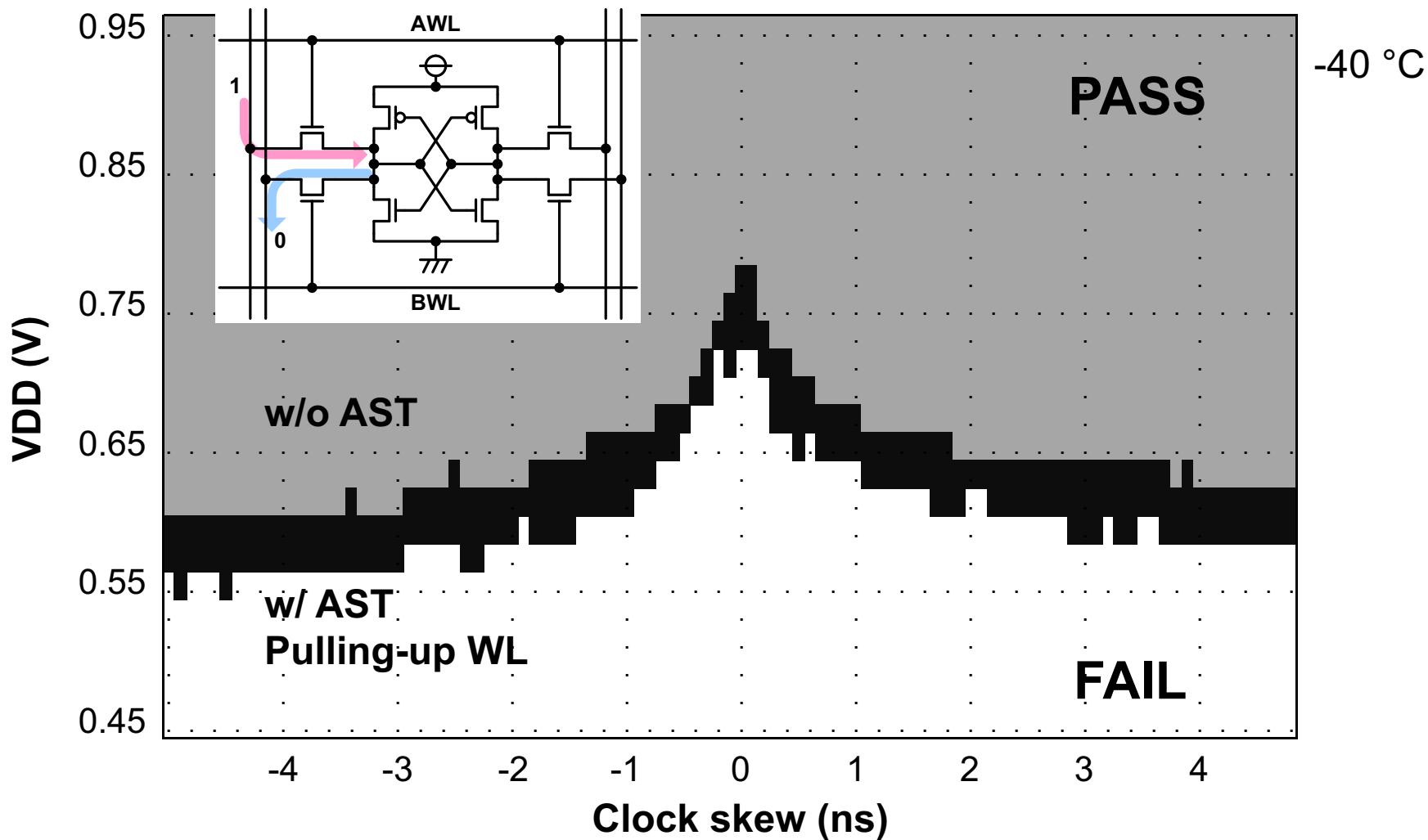
Die micrograph

Switches controlling VWL are located at well-tap region
→ No area penalty @ HM



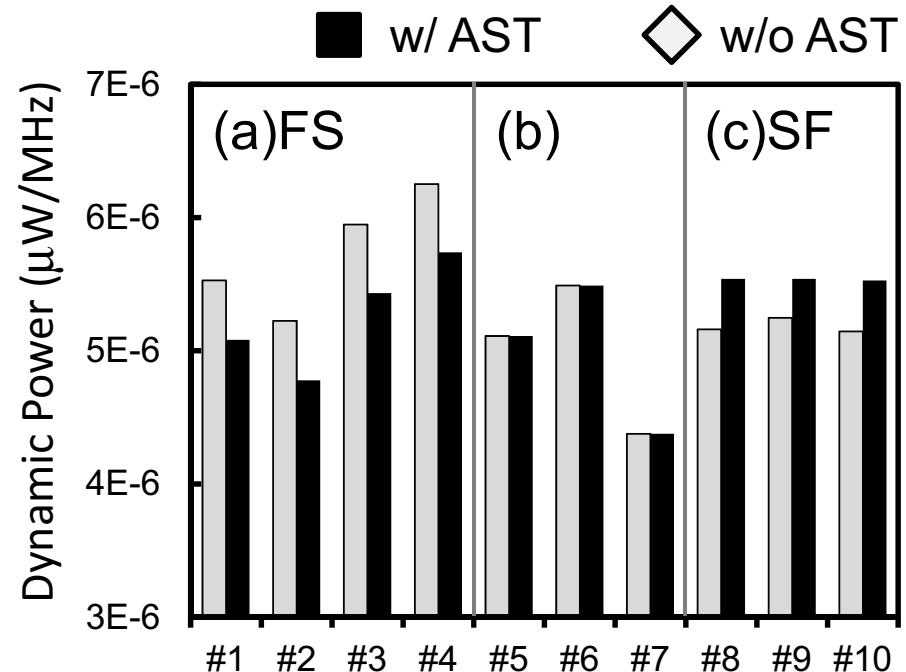
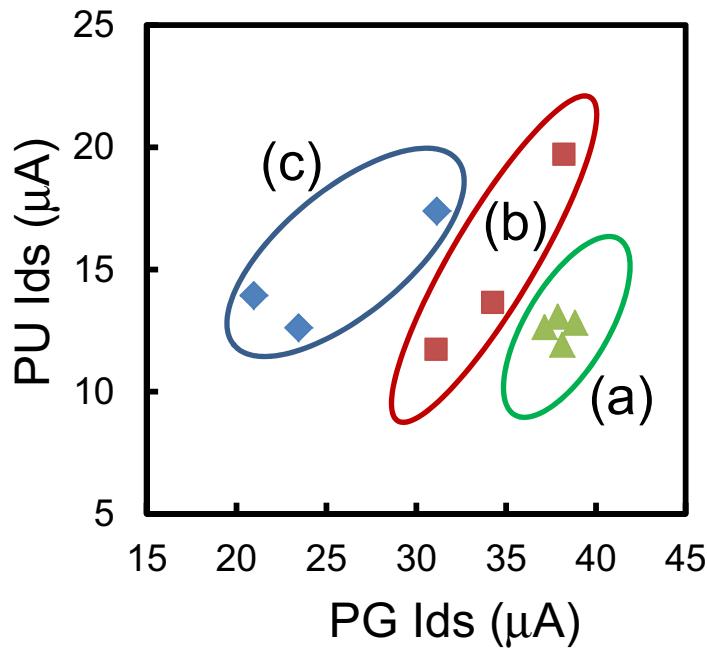
Technology	20-nm HK+MG bulk CMOS (planar)
Macro configuration	SP: 32-kb (32b x 1kw) x 2; Total 64-kb DP: 16-kb (16b x 1kw) x 4; Total 64-kb
Physical size	SP: 109.7µm x 130.3 µm @ 32-kb DP: 133.2µm x 152.4 µm @ 16-kb
Bit density	SP: 8.74 Mb/mm ² DP: 3.08 Mb/mm ²

Measurement of dual-port clock skew



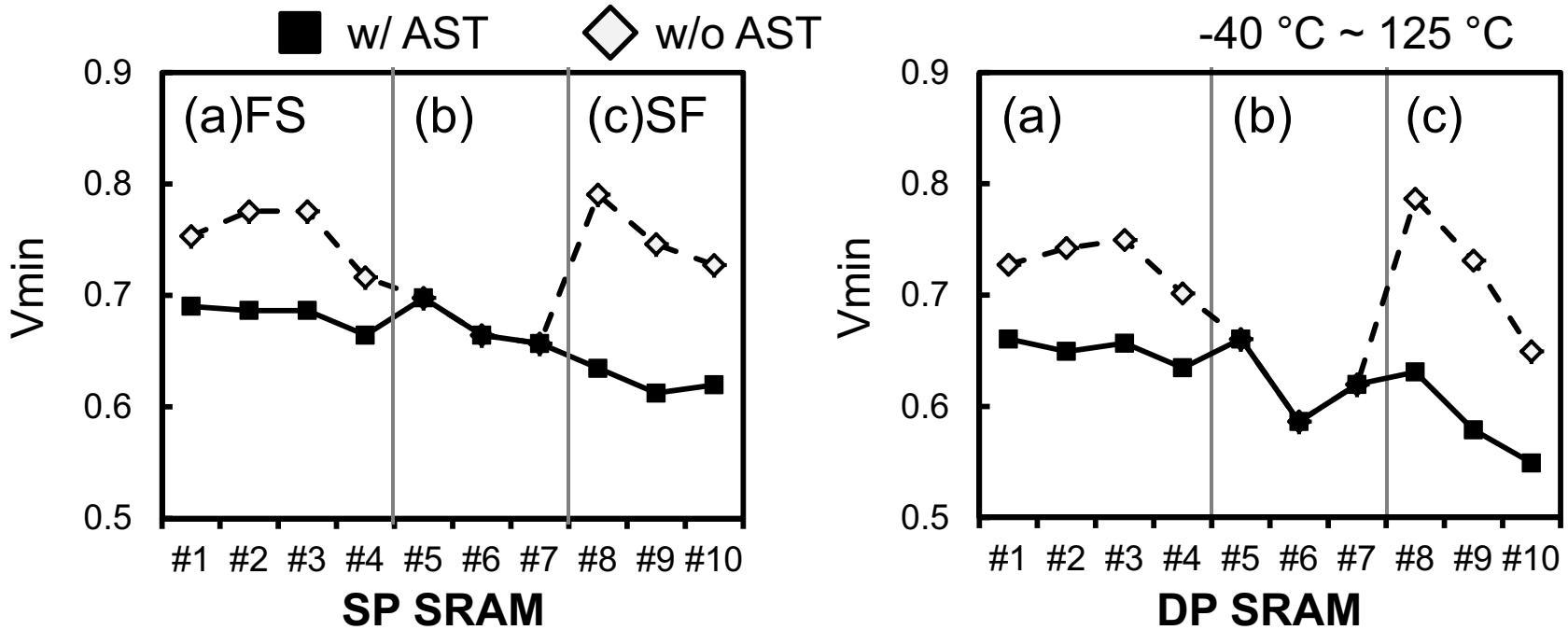
V_{min}@ Disturb is not degraded though VWL is raised.

Measurement of Dynamic power



Assist adjustment for each chip
For (a)FS, power can be reduced by lowering VWL

Measurement of SRAM Vmin



By adjusting assist level for chip by chip, achieves effective improvement of Vmin

Conclusion

- We proposed a new V_{min} improvement method called WDVA System.
- WDVA achieved 0.1V V_{min} reduction and power reduction for the worst condition.
- Designed and fabricated in 20nm planar bulk process.
- The bit density of SP-SRAM is 8.74Mb/mm² and that of DP-SRAM is 3.08Mb/mm².

A 7ns-Access-Time 25μW/MHz 128kb SRAM for Low-Power Fast Wake-Up MCU in 65nm CMOS with 27fA/b Retention Current

T. Fukuda¹, K. Kohara¹, T. Dozaka¹, Y. Takeyama¹,
T. Midorikawa¹, K. Hashimoto², I. Wakiyama², S. Miyano¹,
T. Hojo¹

¹Toshiba, Kawasaki, Japan

²Toshiba Microelectronics, Kawasaki, Japan

Outline

- Motivation
- Reduction in Leakage Current
- Reduction in Active Power
- Conclusion

Motivation

- Low-power MCUs are used in battery driven systems such as wearable devices, healthcare tools, smart meters, etc.
- **STRONG DEMANDS** for long battery lifetime of 10 years

Comparison of back-up memory type

memory	perform- ance	active power	leakage current	additional process
FRAM	slow	high	low	necessary
SRAM	fast	low	high	not necessary

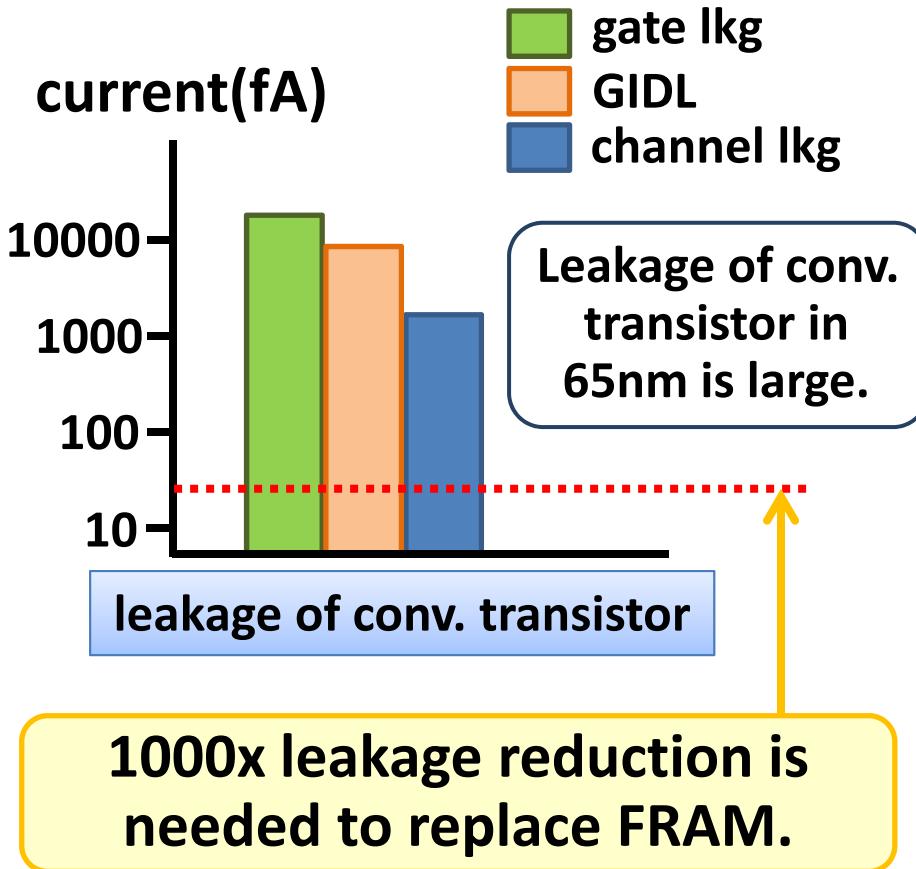
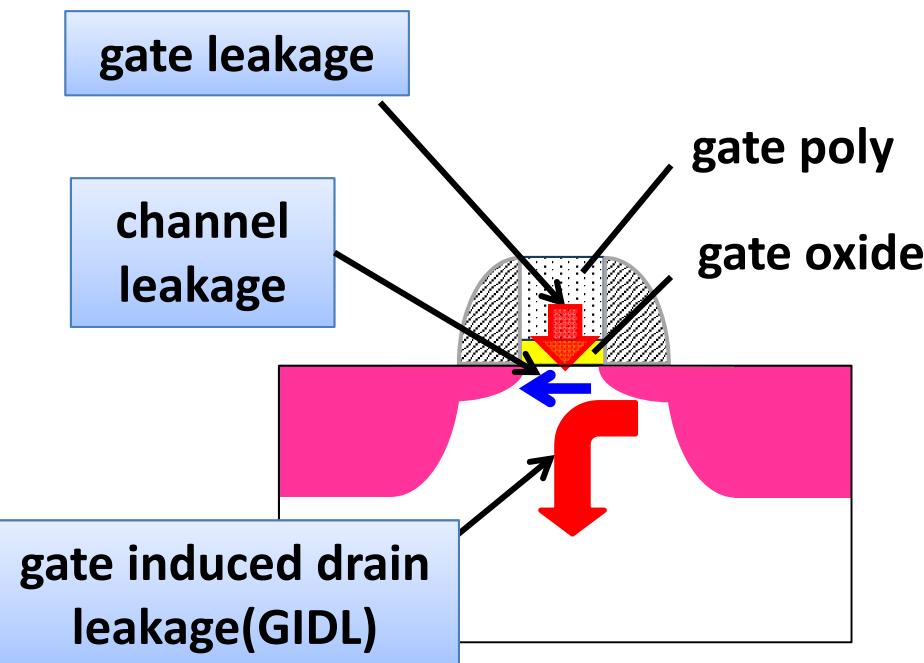
→ SRAM is suitable, except LEAKAGE.

We propose **eXtremely Low Leakage SRAM (XLL SRAM)** whose leakage is low enough to replace FRAM used as back-up RAM of low-power MCU.

Outline

- Motivation
- Reduction in Leakage Current
- Reduction in Active Power
- Conclusion

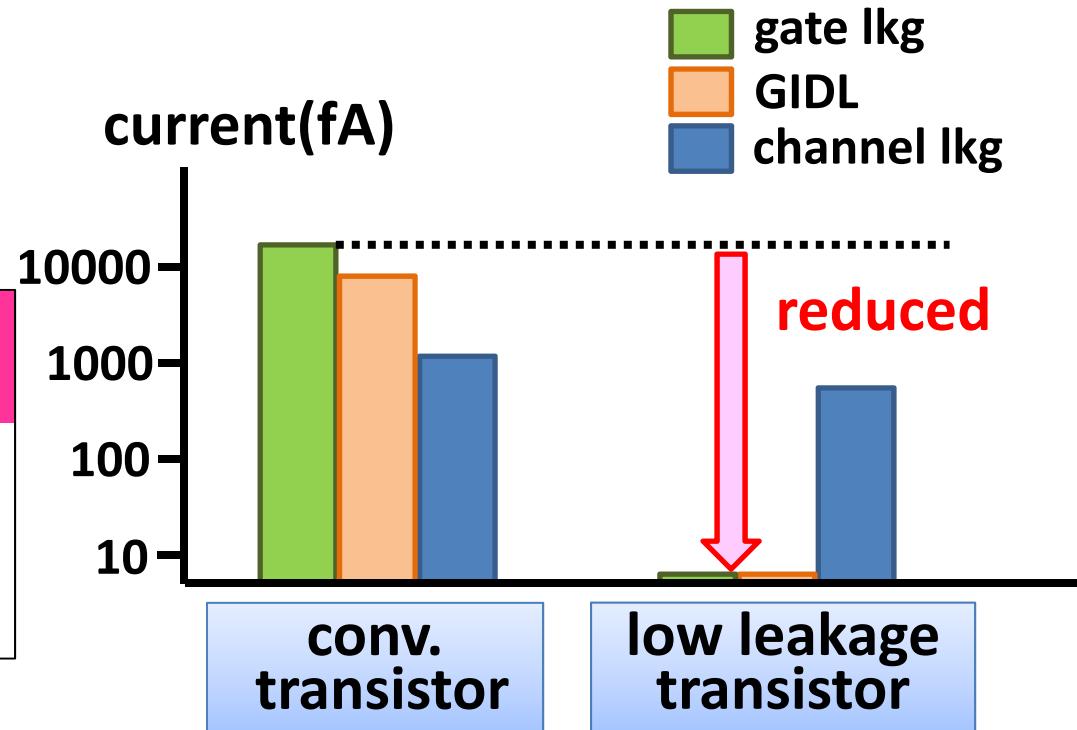
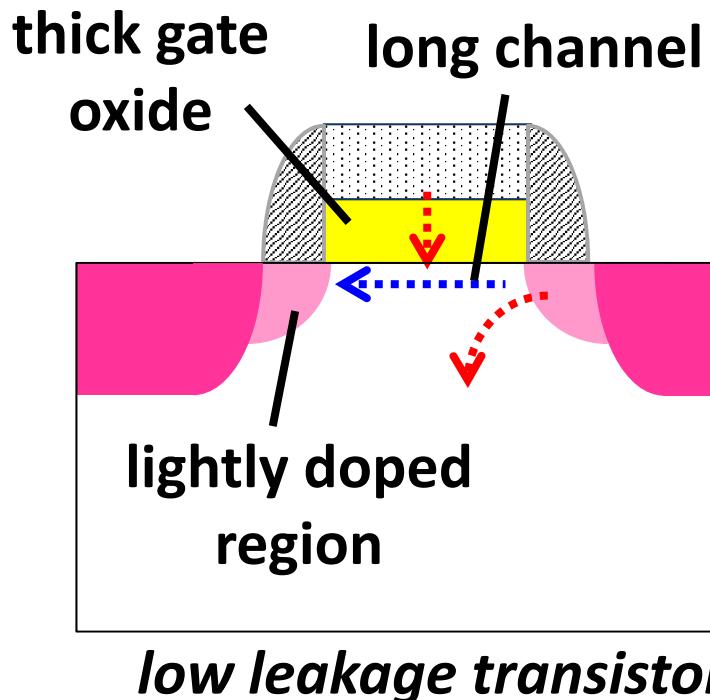
Direction of Leakage Reduction



All factors of leakage need to be reduced.

- Gate leakage and GIDL are reduced by device approach.
- Channel leakage is reduced by circuit approach.

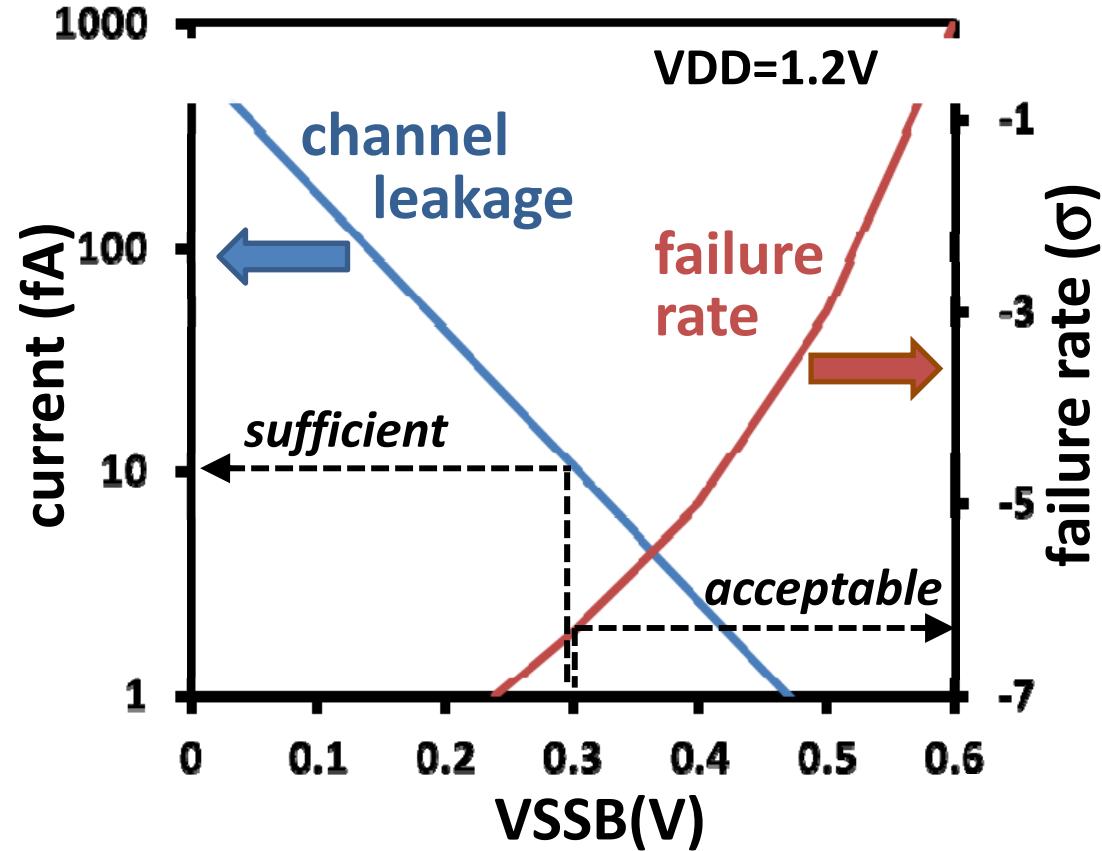
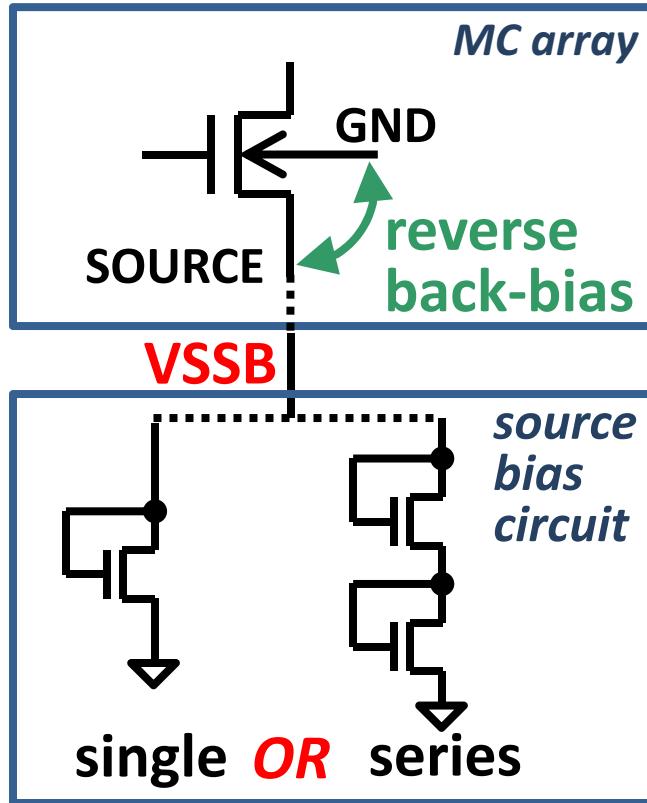
Gate Leakage & GIDL Reduction



- Thick gate oxide reduces gate leakage
- Long channel and lightly doped region reduce GIDL

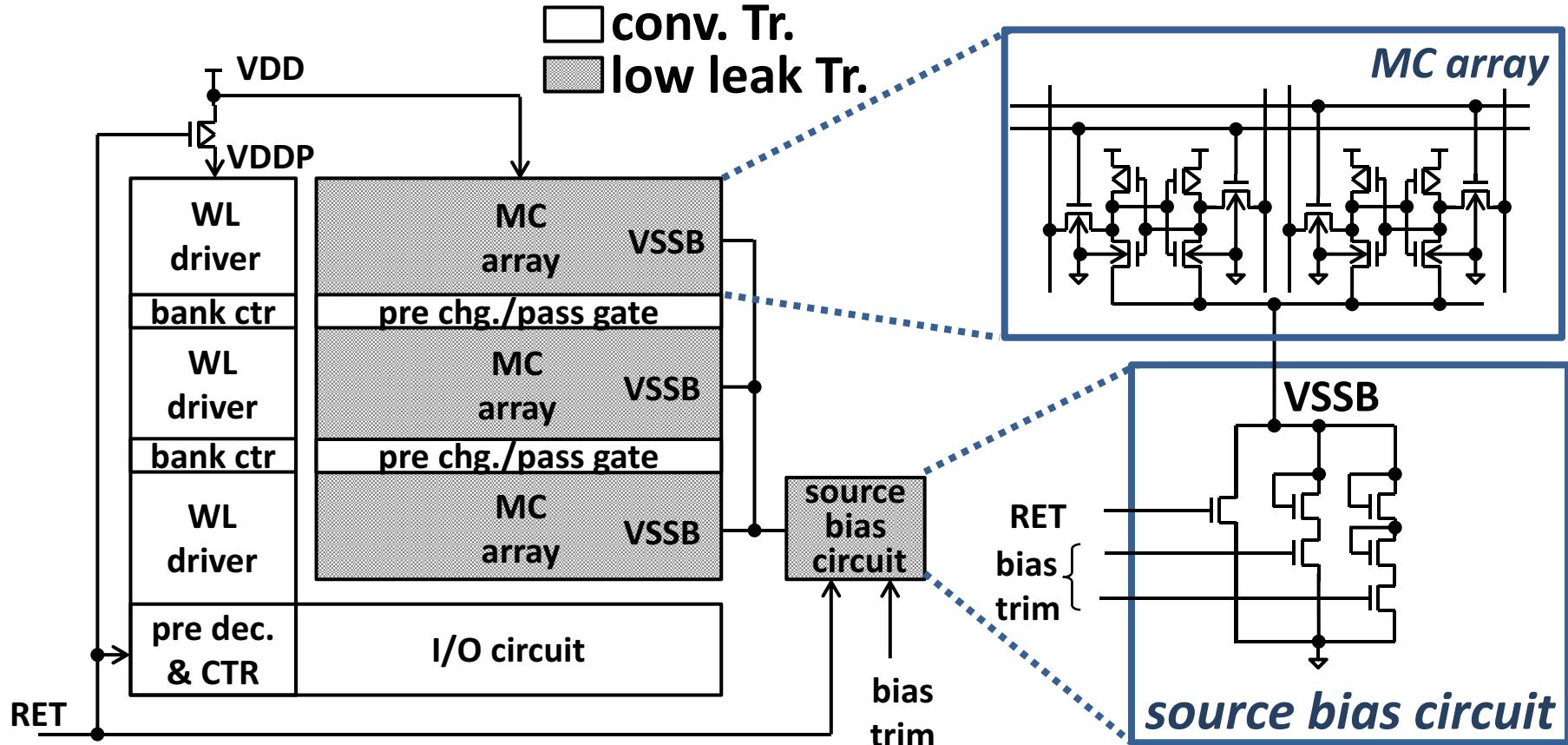
To reduce leakage current **extremely** in 65nm,
we adopt **larger transistor** to SRAM memory cell.

Channel Leakage Reduction



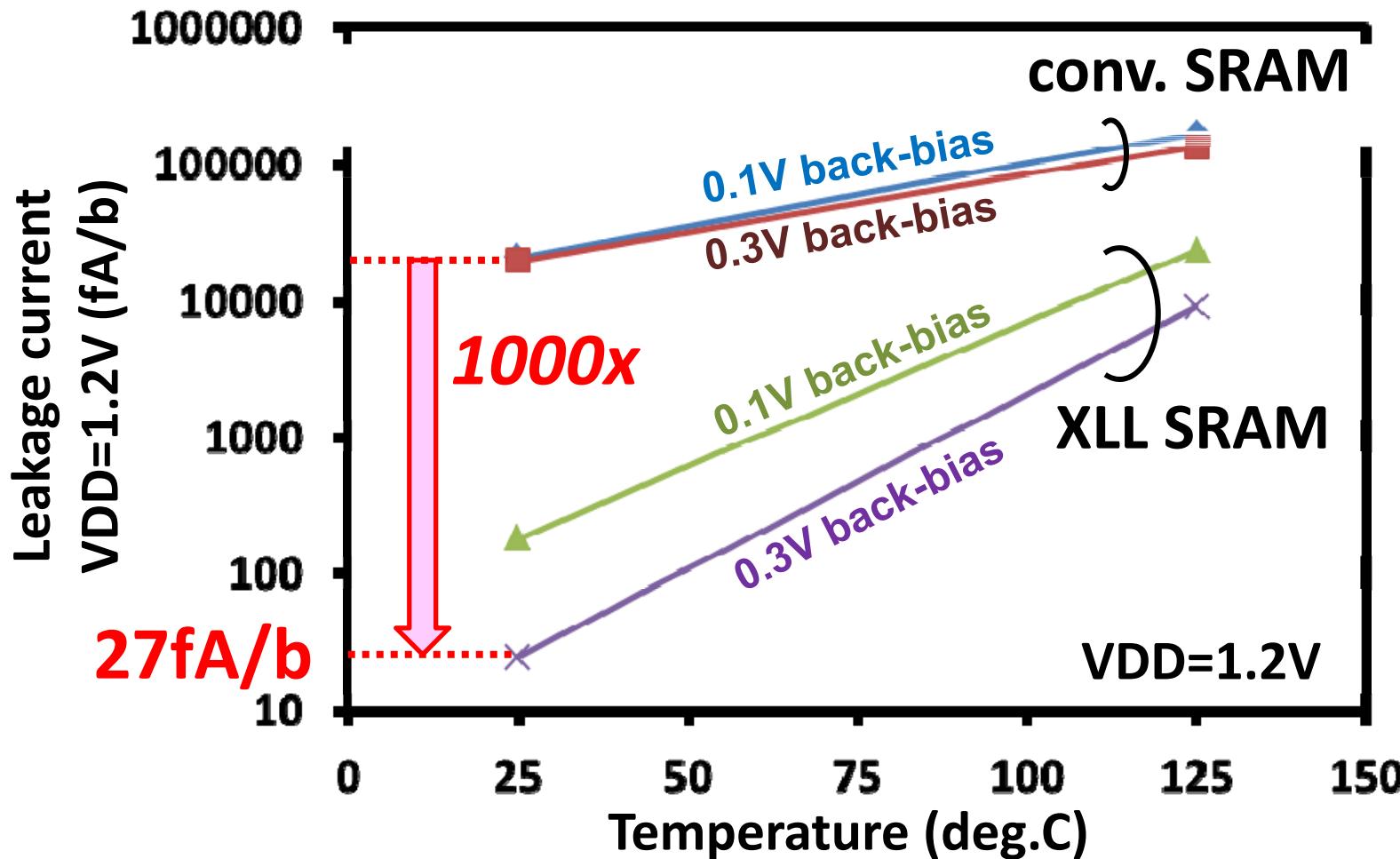
- Employ source bias circuit with MOS diode to apply reverse back-bias to memory cell transistors
- Apply 0.3V of VSSB to reduce leakage and to keep acceptable memory failure rate
- Adjust the VSSB by varying number of diode and channel width

Block Diagram of 128kb XLL SRAM



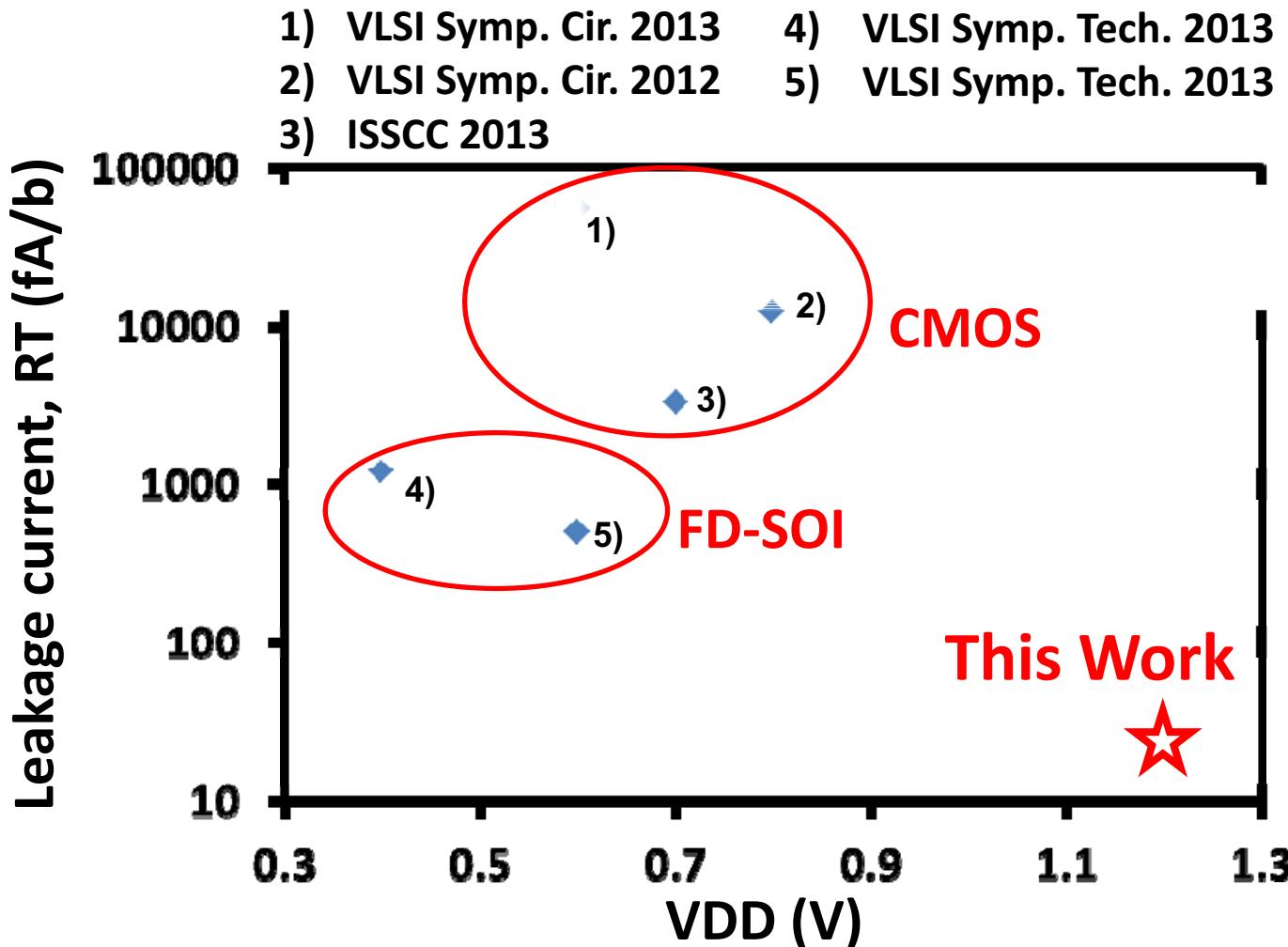
- Use low leakage transistor for memory cell and source bias circuit
- NMOS source node of memory cell is reverse biased via source bias circuit.
- Supply voltage of peripheral circuit is cut off in back-up mode.

Measured Leakage Current



1000x lower leakage compared to conventional SRAM
by low leakage transistor and back bias effect.

Comparison to Previous Work



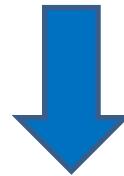
Lower than published data beyond 65nm tech.

Outline

- Motivation
- Reduction in Leakage Current
- Reduction in Active Power
- Conclusion

Reduction in Active Power

Because low leakage transistor is larger than conventional transistor, **active power dissipation increases.**

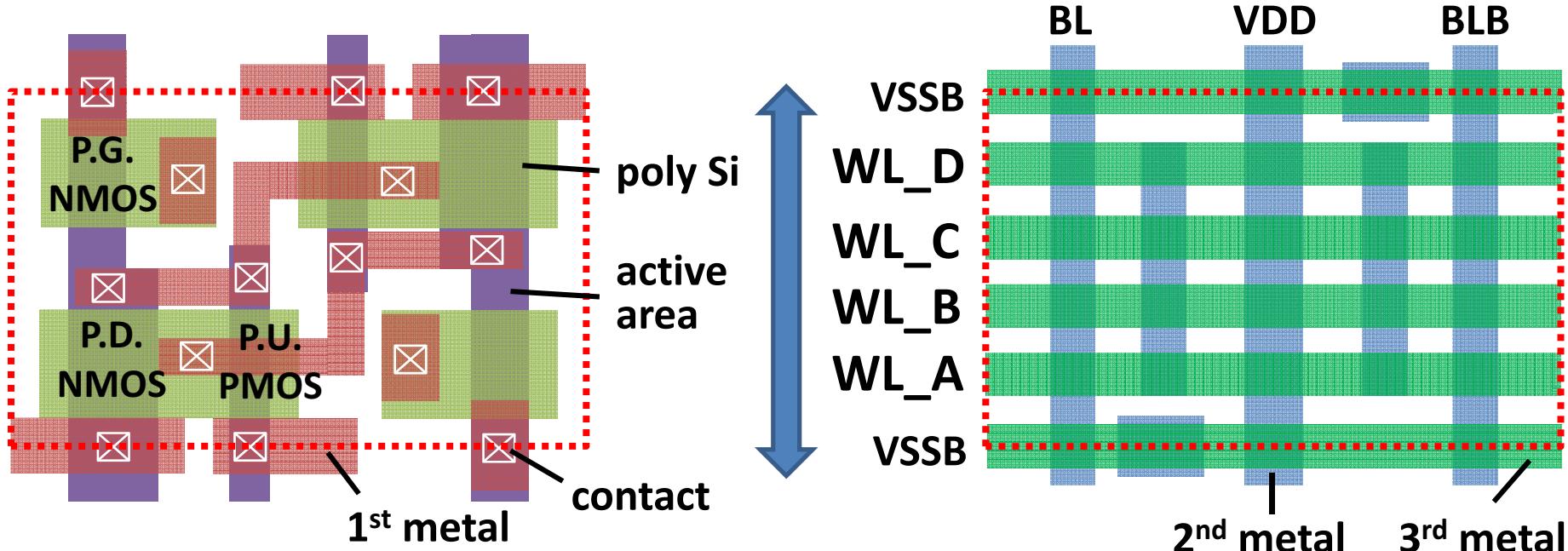


avoid this increasing

Reduction in Active Power Dissipation

- Quarter array activation scheme (QAAS)
- Charge-shared hierarchical bitline (CSHBL)

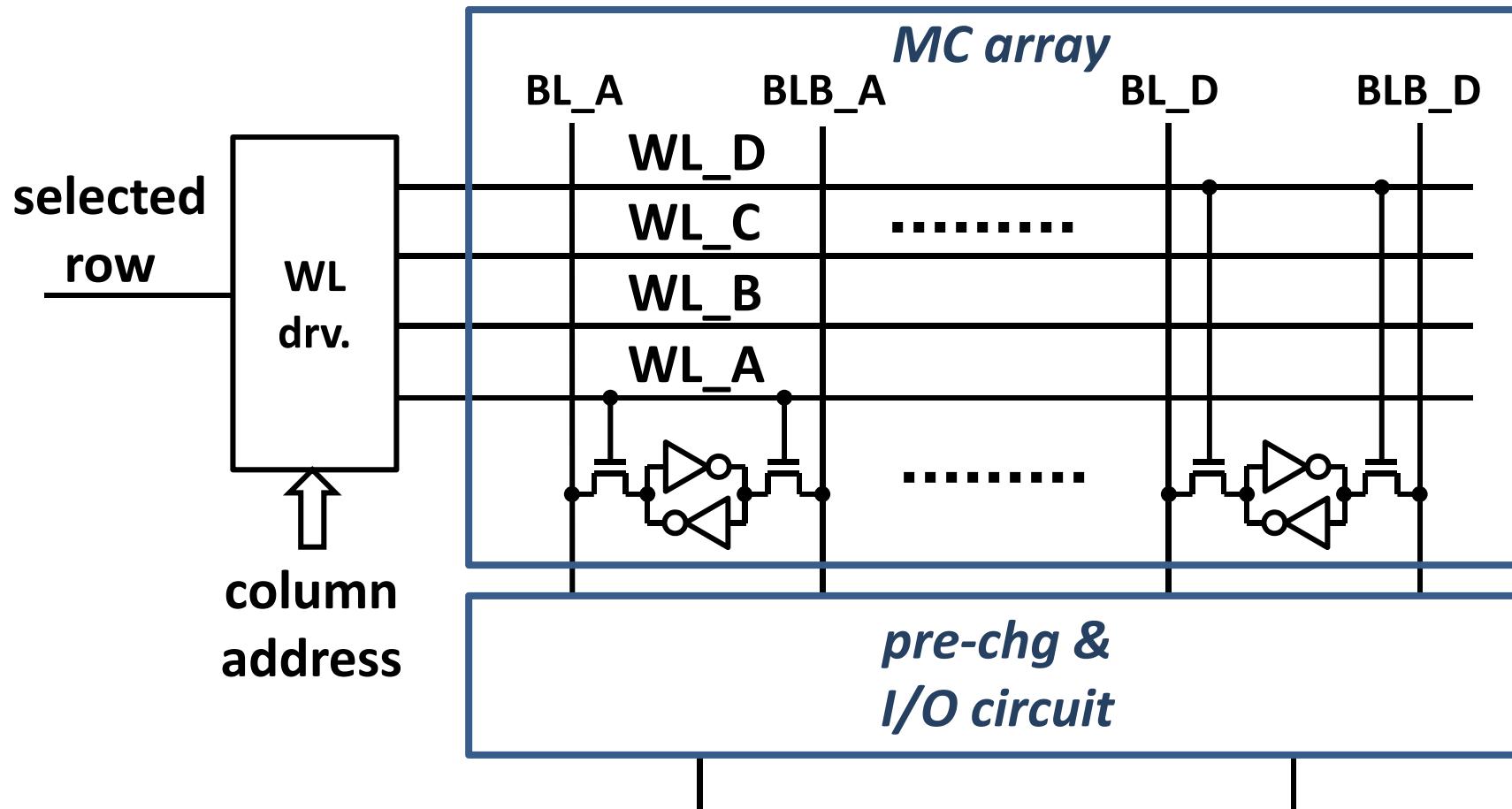
Memory Cell Layout



Extended by long channel length transistors

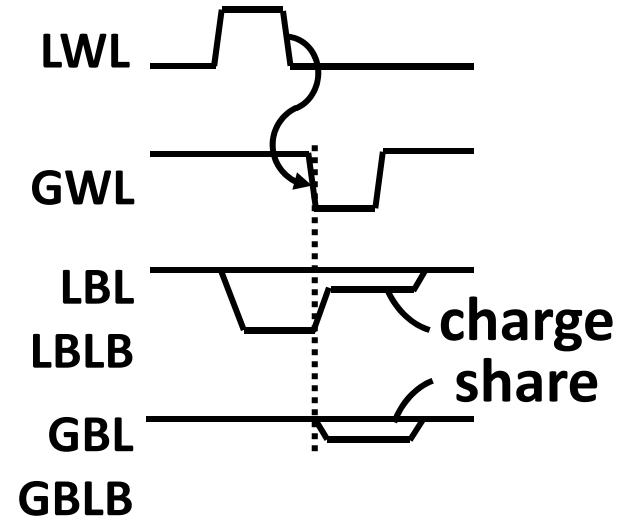
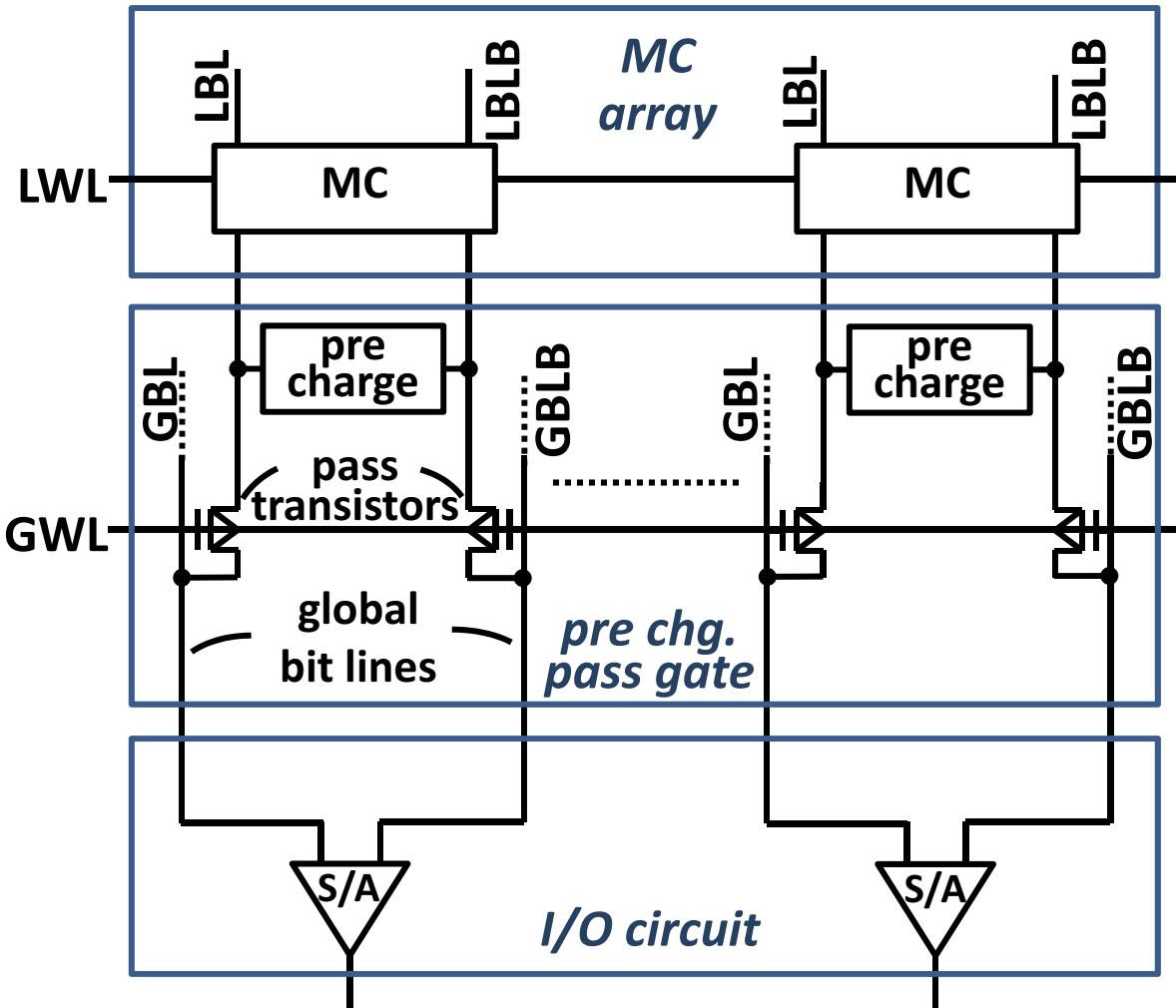
Four wordlines routed over a memory cell are used for active power reduction.

Reduction in Active Power : QAAS



Active duration of bitlines decreases, so active power is reduced.

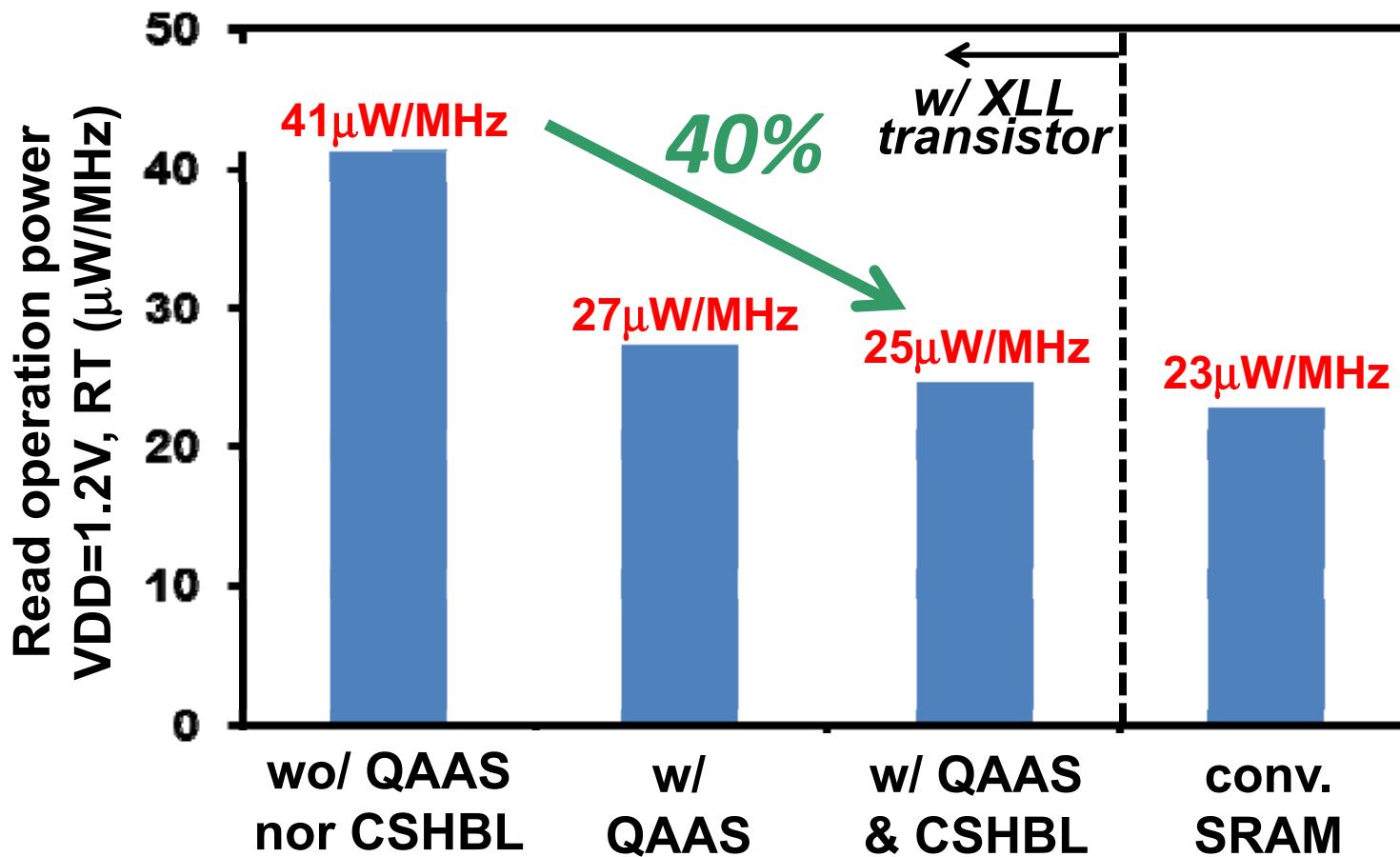
Reduction in Active Power : CSHBL



cf.) S. Miyano et al., JSSC 2013.

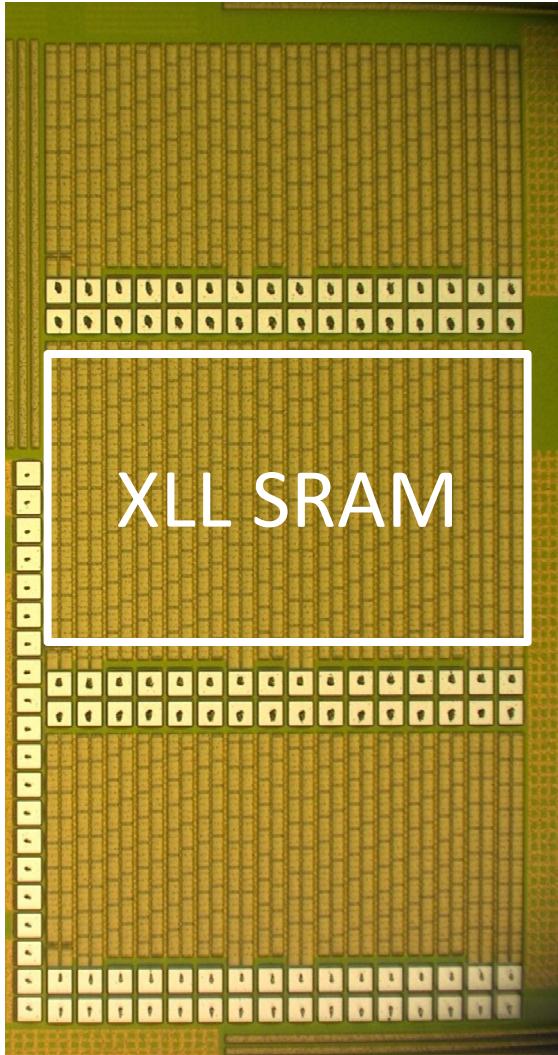
Active duration of bitlines decreases, so active power is reduced.

Measured Active Power Dissipation



- 40% reduction by adopting QAAS & CSHBL
- Equivalent to conventional SRAM

Chip Micrograph & Key Features



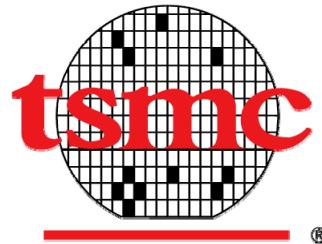
Technology	65nm CMOS
Power supply	1.2V
Cell structure	6T
Cell size	2.159μm²
Capacity	128kbit
Macro area	0.443mm ²
Read access time	7ns
Power dissipation	Operation
	Standby (27fA/b)

Conclusion

- We have developed **eXtremely Low Leakage SRAM (XLL SRAM)** in 65nm. Its leakage is low enough to replace FRAM used as back-up RAM of low-power MCU.
- **1000x leakage current reduction** compared to conventional SRAM by low leakage transistor for memory cell and back-bias circuit.
- Active power is equivalent to conventional SRAM by adopting QAAS & CSHBL.

A 16nm 128Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low- V_{MIN} Applications

[Yen-Huei Chen](#), Wei-Min Chan, Wei-Cheng Wu, Hung-Jen Liao, Kuo-Hua Pan, Jhon-Jhy Liaw, Tang-Hsuan Chung, Quincy Li, George H. Chang, Chih-Yung Lin, Mu-Chi Chiang, Shien-Yang Wu, Sreedhar Natarajan, Jonathan Chang



Outline

- Motivation
 - How FinFET impacts on SRAM cell design?
 - Design considerations of write assist techniques
- Proposed low V_{MIN} design techniques
 - Suppressed Coupling Signal Negative Bit-Line (SCS-NBL) scheme
 - Write Recover Enhancement Lower Cell VDD (WRE-LCV) scheme
- Silicon results
- Summary

Outline

- Motivation
 - How FinFET impacts on SRAM cell design?
 - Design considerations of write assist techniques
- Proposed low V_{MIN} design techniques
 - Suppressed Coupling Signal Negative Bit-Line (SCS-NBL) scheme
 - Write Recovery Enhancement Lower Cell VDD (WRE-LCV) scheme
- Silicon results
- Summary

Motivation

- FinFET has become the main-stream solution
 - Better short channel effect
 - Lower device mismatch
- Quantized sizing (W , L) limits the SRAM cell Design

6T FinFET HD SRAM bit cell layout

$$\text{SRAM Area} = (\text{X-pitch}) \times (\text{Y-pitch})$$

Geometric ratio

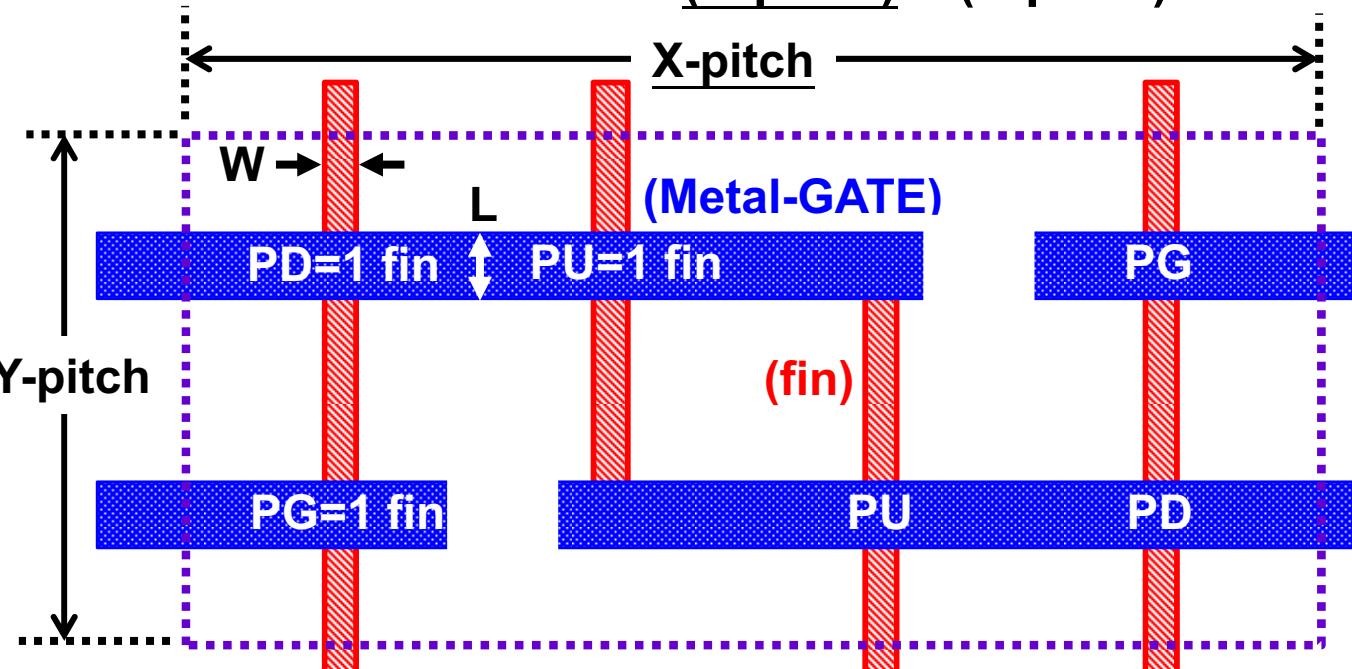
(W/L)

PU:PD:PG=1:1:1



SRAM operation

issue!

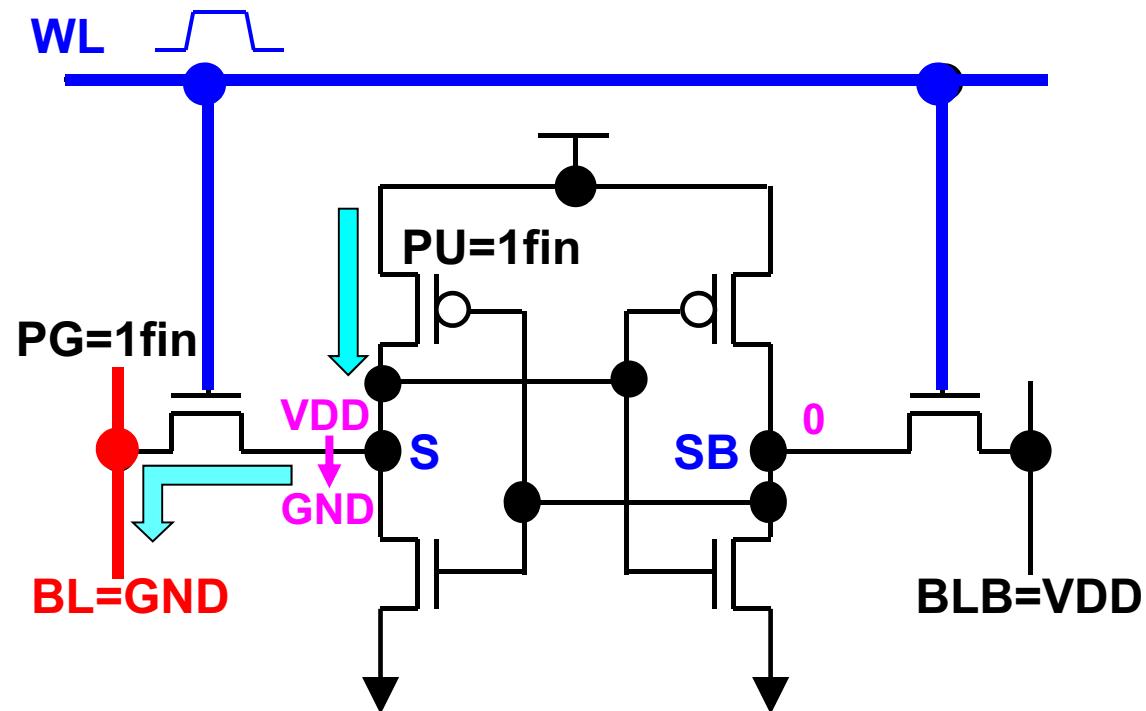


13.5: A 16nm 128Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low- V_{MIN} Applications

SRAM Write Operation Issue

- With device variation, strength of PU > PG
- PG is unable to pull “S” node to GND
 - Contention write failure
- Write assist is needed for FinFET HD SRAM cell

FinFET HD SRAM cell
Geometric ratio (W/L)
PU:PG=1:1

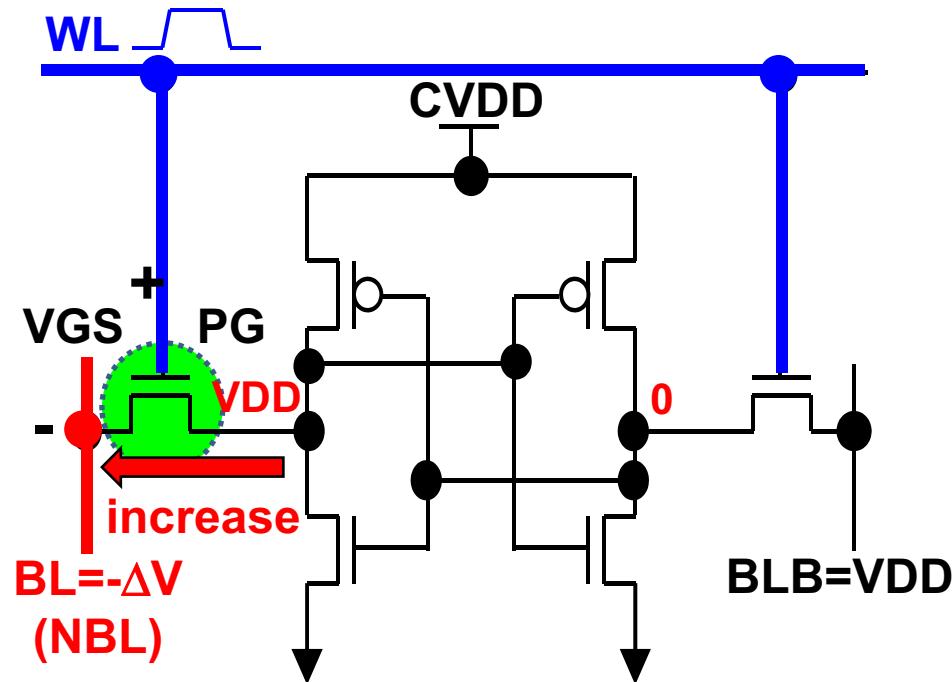


13.5: A 16nm 128Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low-V_{MIN} Applications

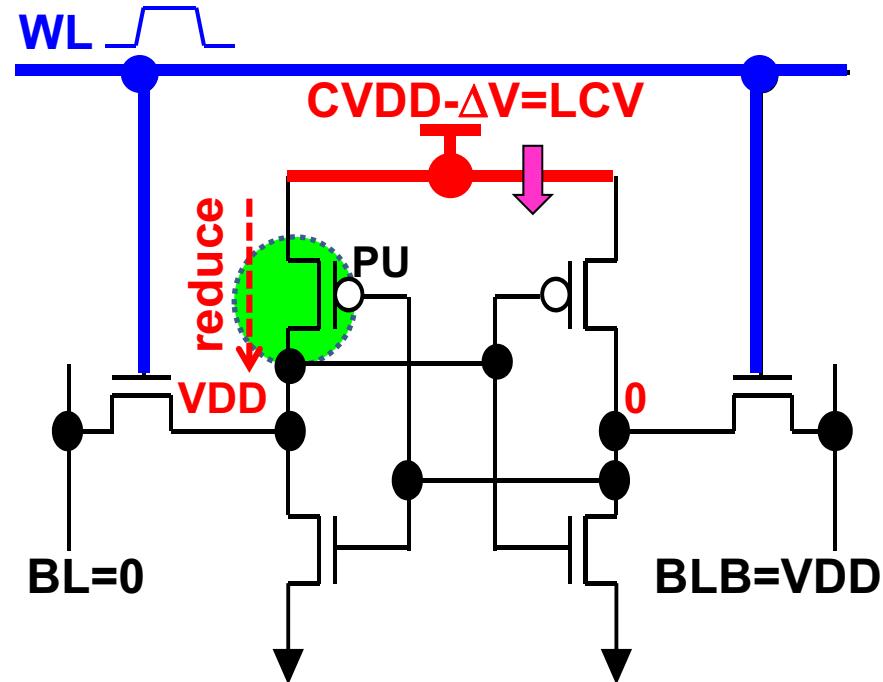
Write Assist Techniques

- Negative Bit-Line (NBL): increase PG strength
- Lower Cell-VDD (LCV): reduce PU strength
- Improve SRAM cell write ability

NBL: increase PG strength

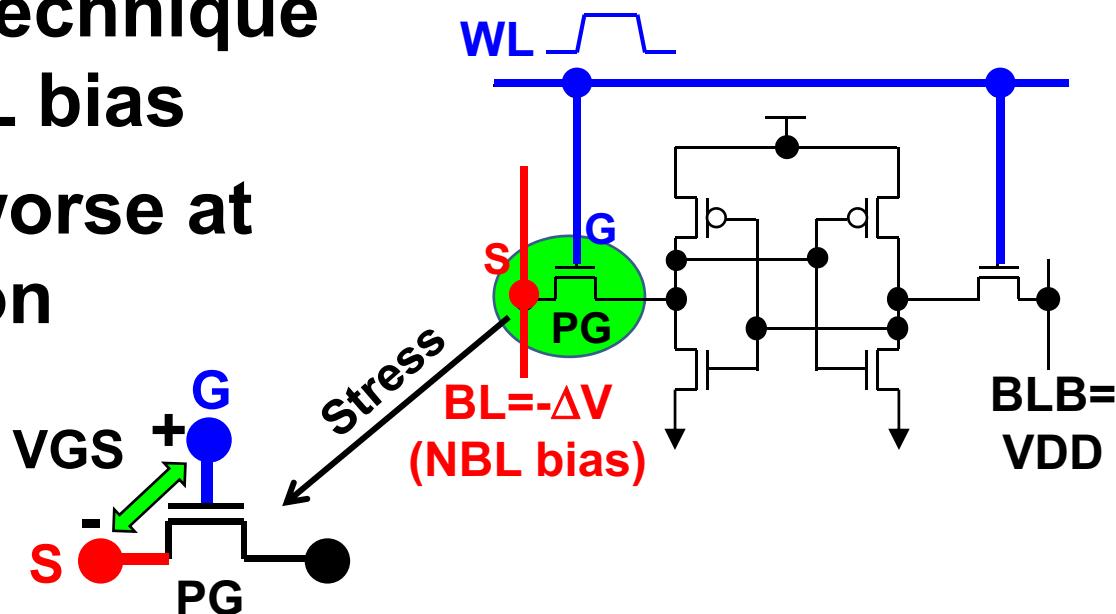


LCV: reduce PU strength



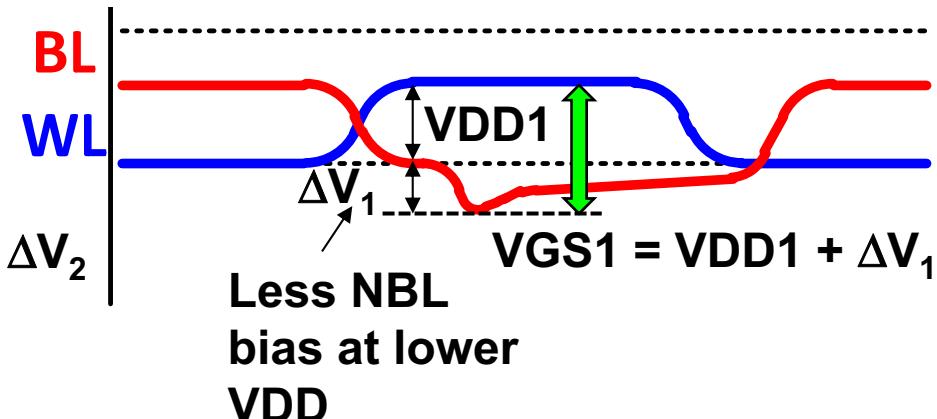
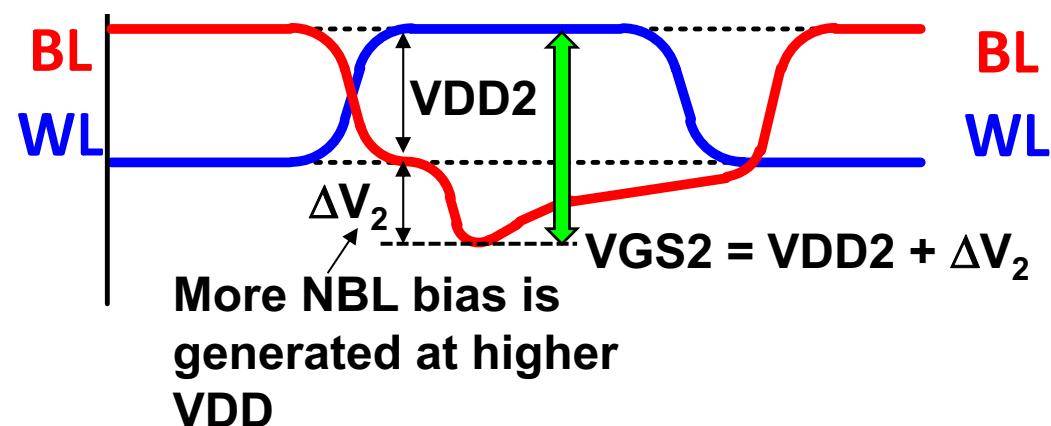
Design Consideration of NBL

- Adopt coupling technique to generates NBL bias
- Stress of PG is worse at higher VDD region



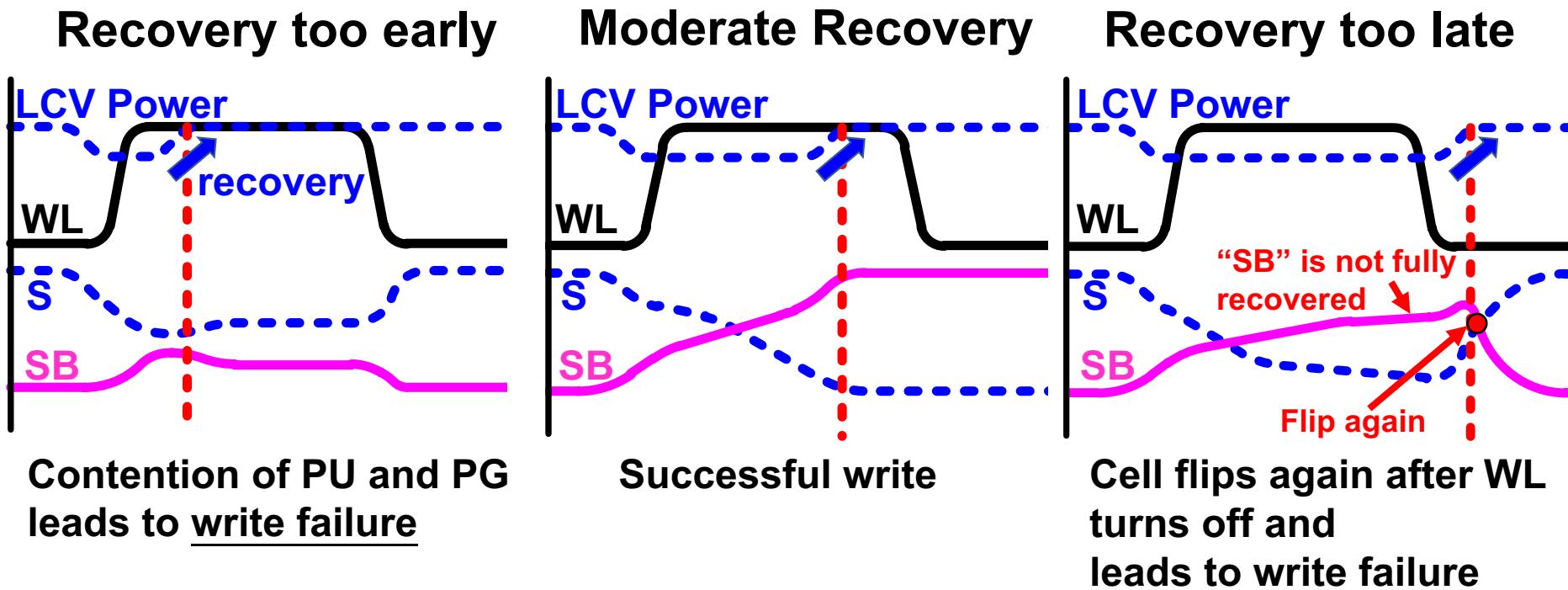
Higher VDD operation

Lower VDD operation



Design Consideration of LCV

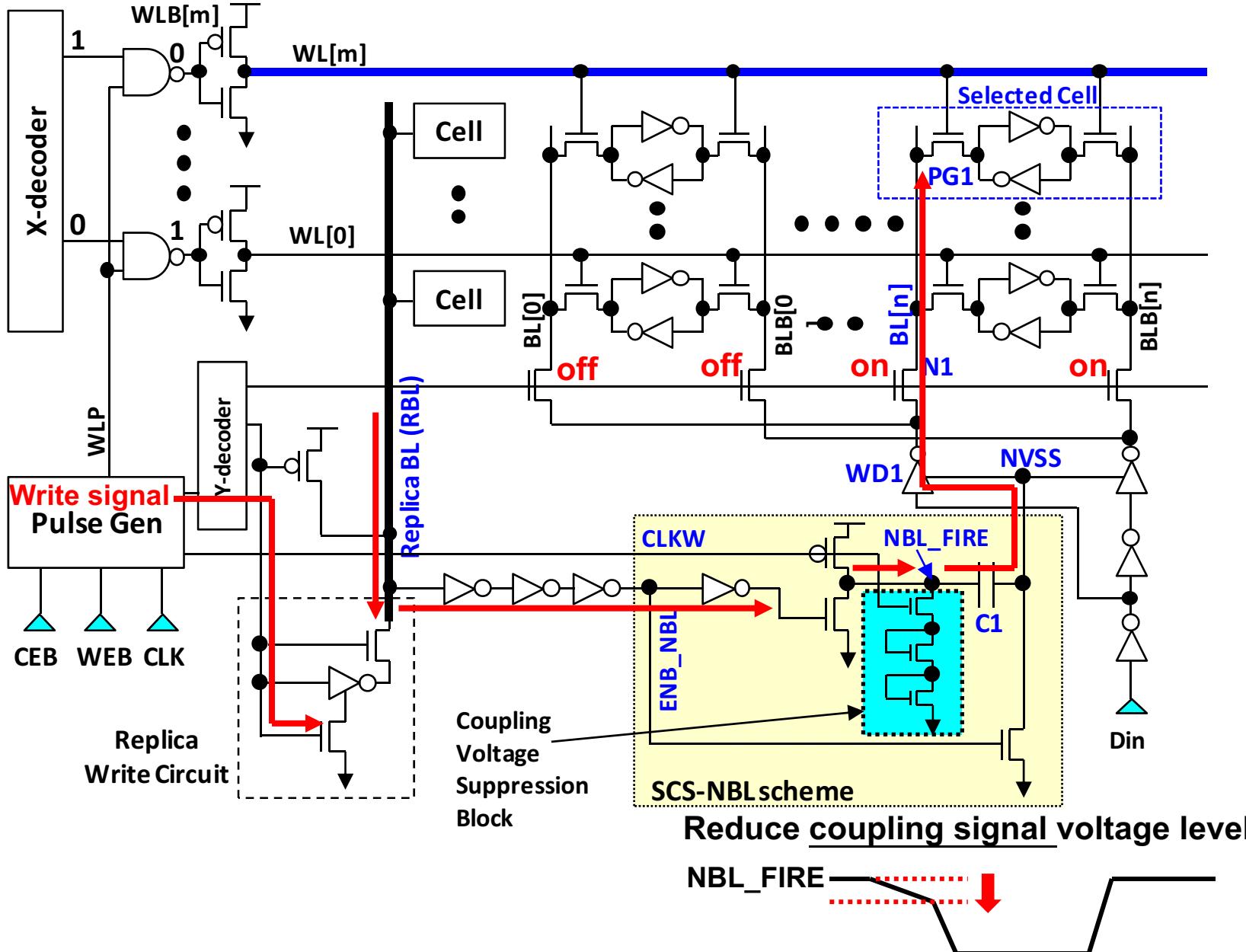
- LCV power recovery timing is important
- With moderate LCV power recovery, solve
 - Contention of PU and PG issue
 - Storage node recovery issue



Outline

- Motivation
 - How FinFET impacts on SRAM cell design?
 - Design considerations of write assist techniques
- Proposed low V_{MIN} design techniques
 - Suppressed Coupling Signal Negative Bit-Line (**SCS-NBL**) scheme
 - Write Recover Enhancement Lower Cell VDD (**WRE-LCV**) scheme
- Silicon results
- Summary

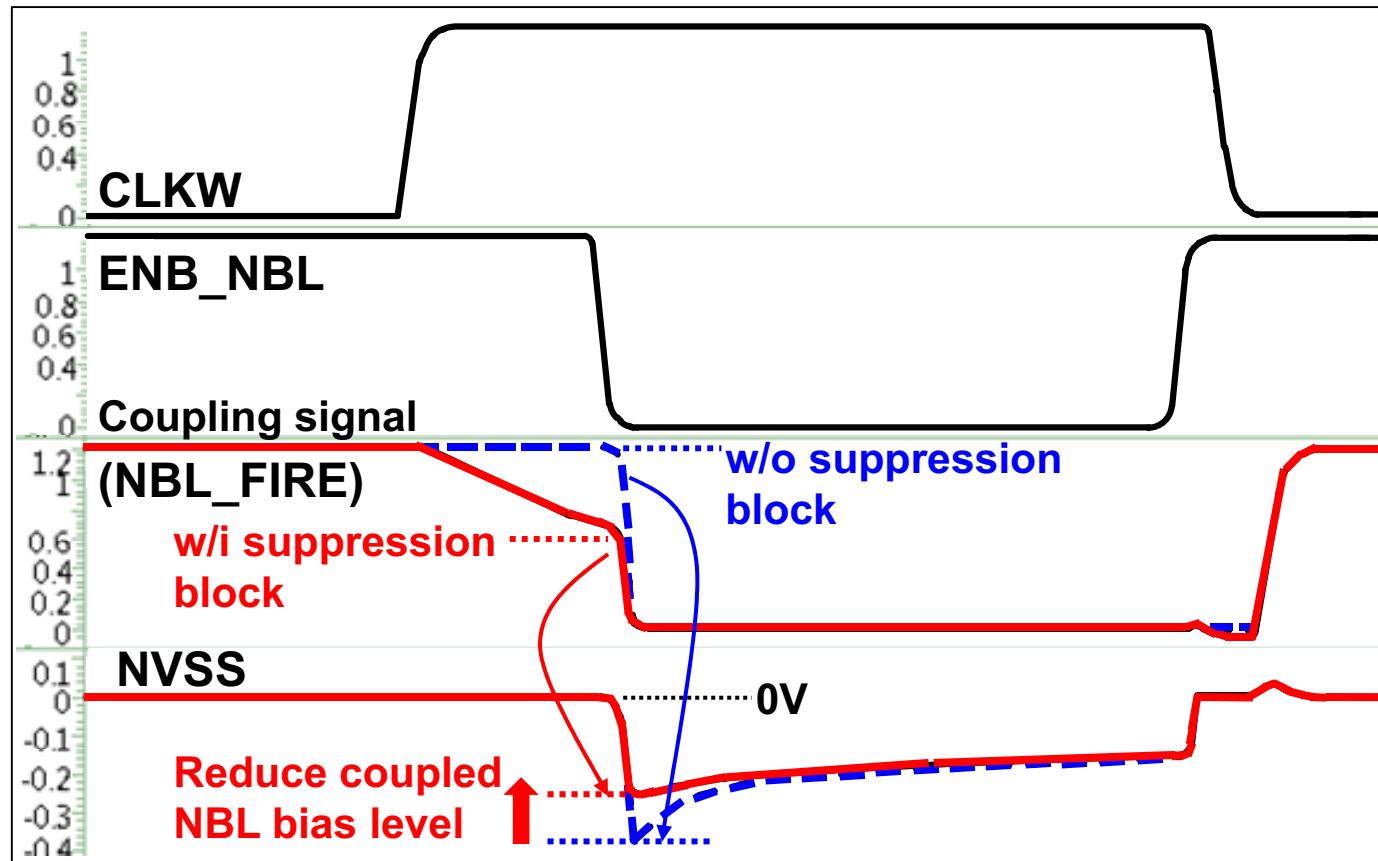
SCS-NBL Scheme



13.5: A 16nm 128Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuity for Low-V_{MIN} Applications

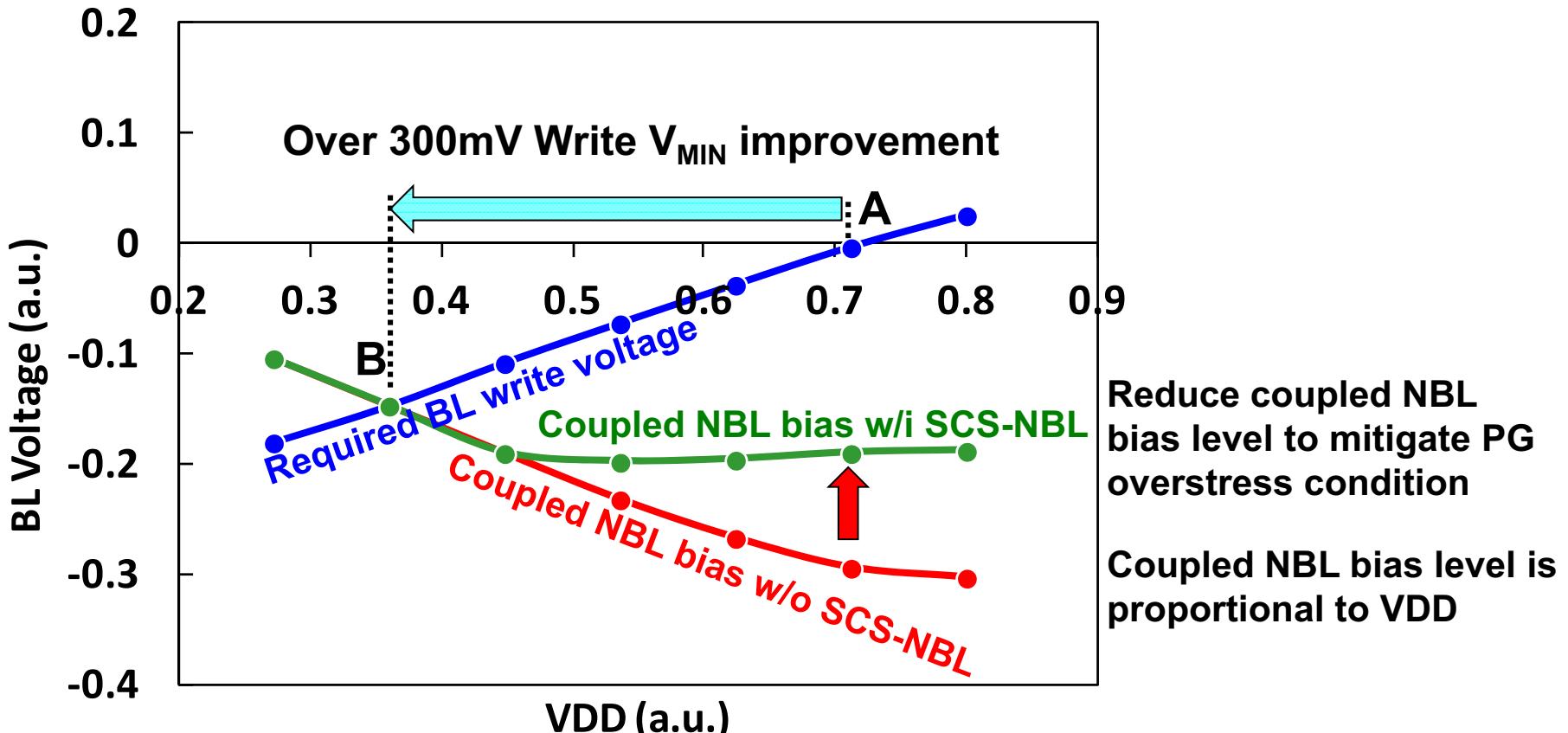
Simulated Waveforms of SCS-NBL

- With SCS-NBL scheme, coupled NBL bias level can be reduced to mitigate PG overstress



Simulated Write V_{MIN} of SCS-NBL

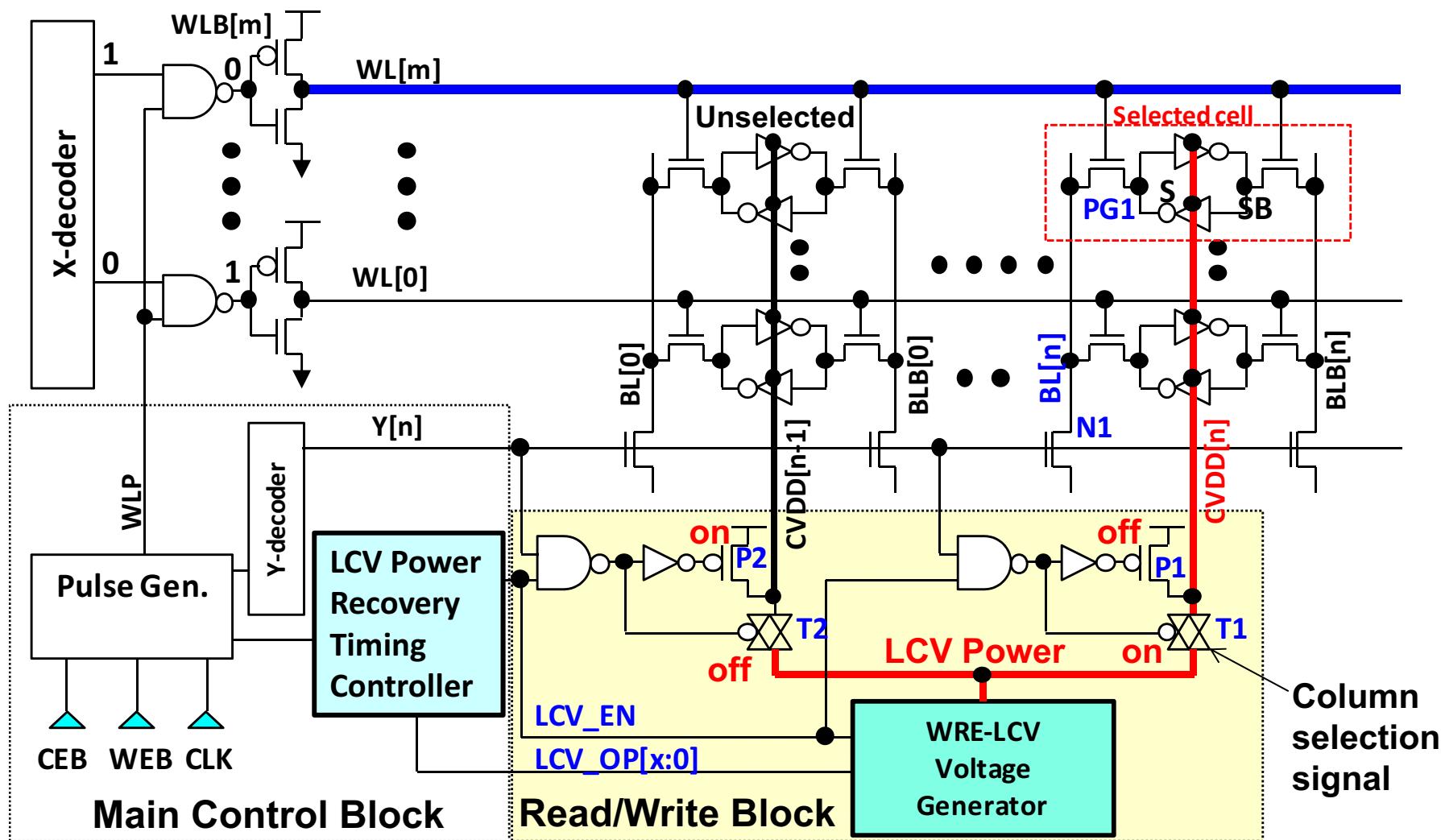
- Point A: intrinsic SRAM write V_{MIN}
- Point B: improved SRAM write V_{MIN} with NBL-WAS
- SCS-NBL scheme reduces the coupled NBL bias level



Outline

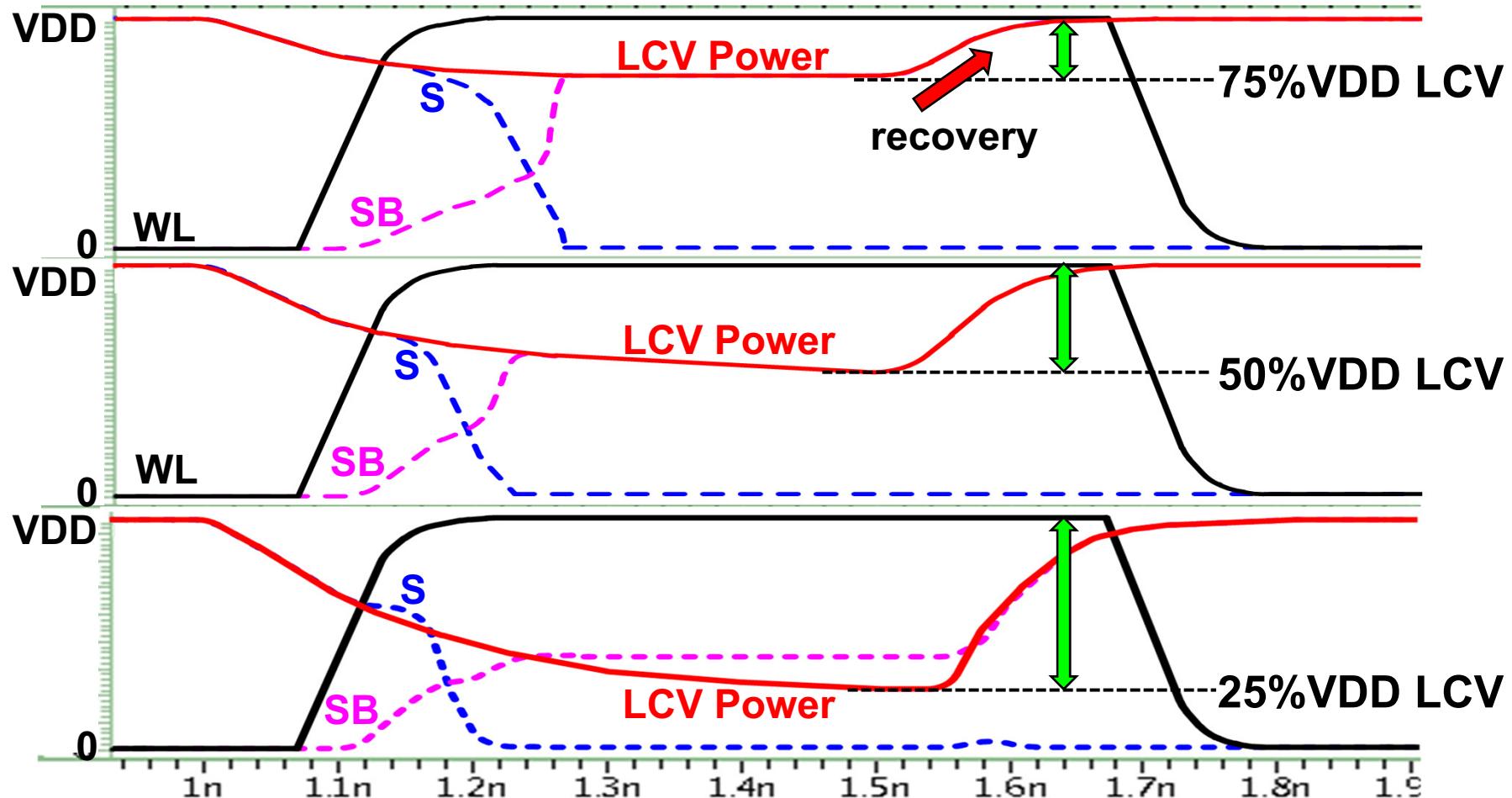
- Motivation
 - How FinFET impacts on SRAM cell design?
 - Design considerations of write assist techniques
- Proposed low V_{MIN} design techniques
 - Suppressed Coupling Signal Negative Bit-Line (SCS-NBL) scheme
 - Write Recover Enhancement Lower Cell VDD (WRE-LCV) scheme
- Silicon results
- Summary

WRE-LCV Scheme



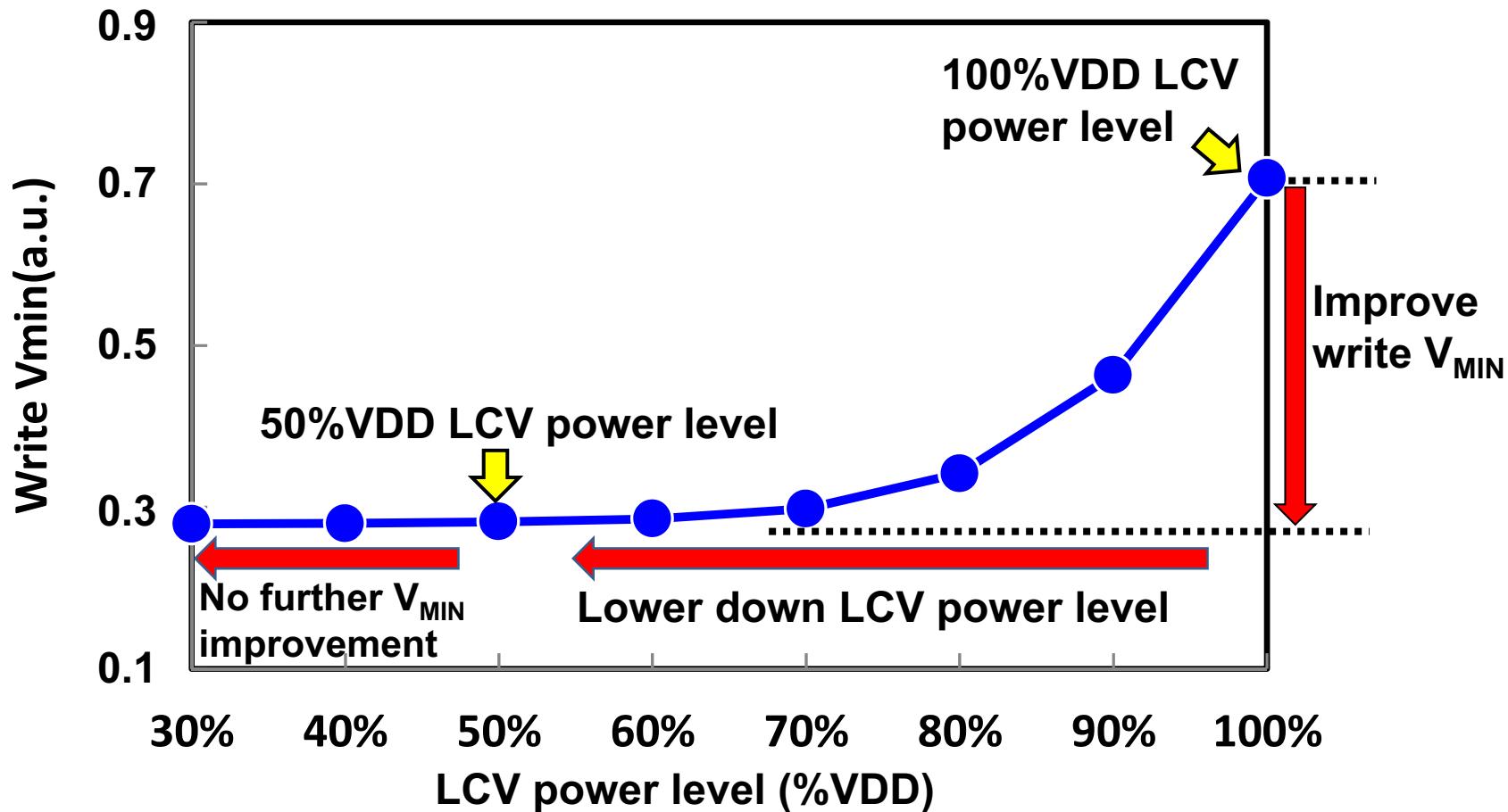
Simulated Waveforms of WRE-LCV

- LCV power recovery control as moderate recovery
- Offer 3 LCV power options to mitigate variation



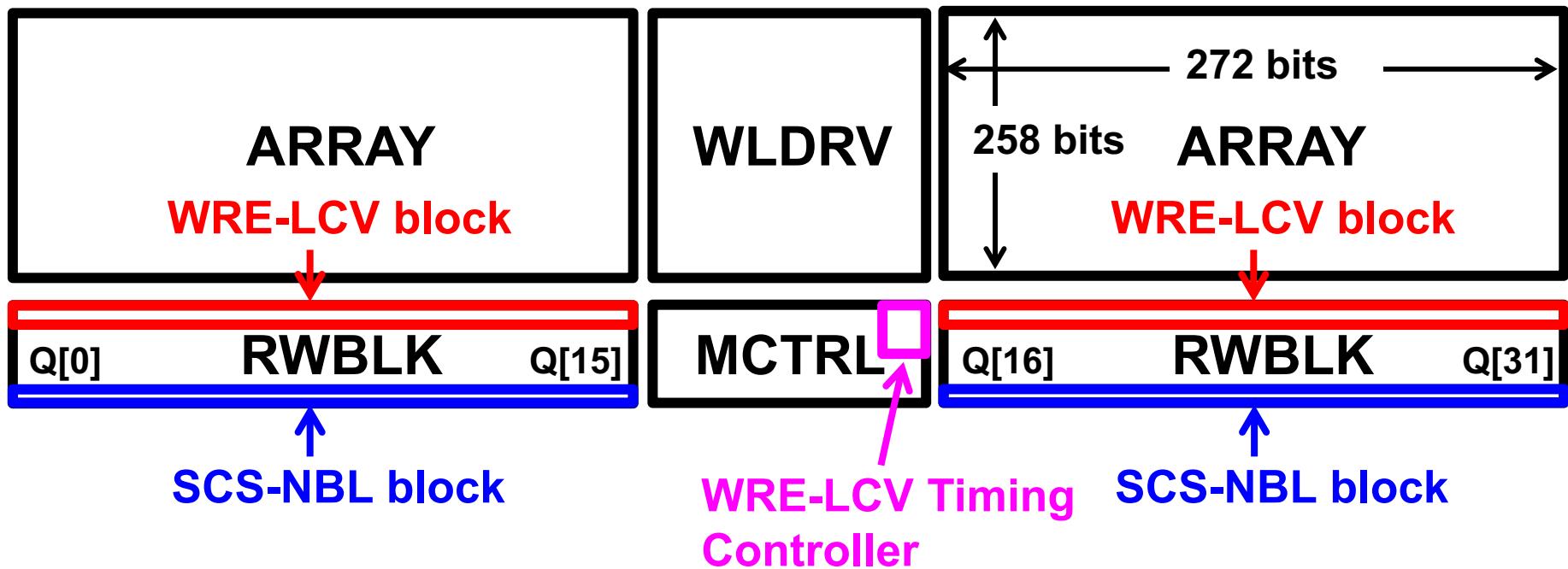
Simulated Write V_{MIN} of WRE-LCV

- Write V_{MIN} improvement saturates at 50%VDD LCV level
- No further V_{MIN} improvement for LCV power < 50%VDD



SRAM Macro Floor Plan

- 2% area overhead for SCS-NBL scheme
- 3% area overhead for WRE-LCV scheme

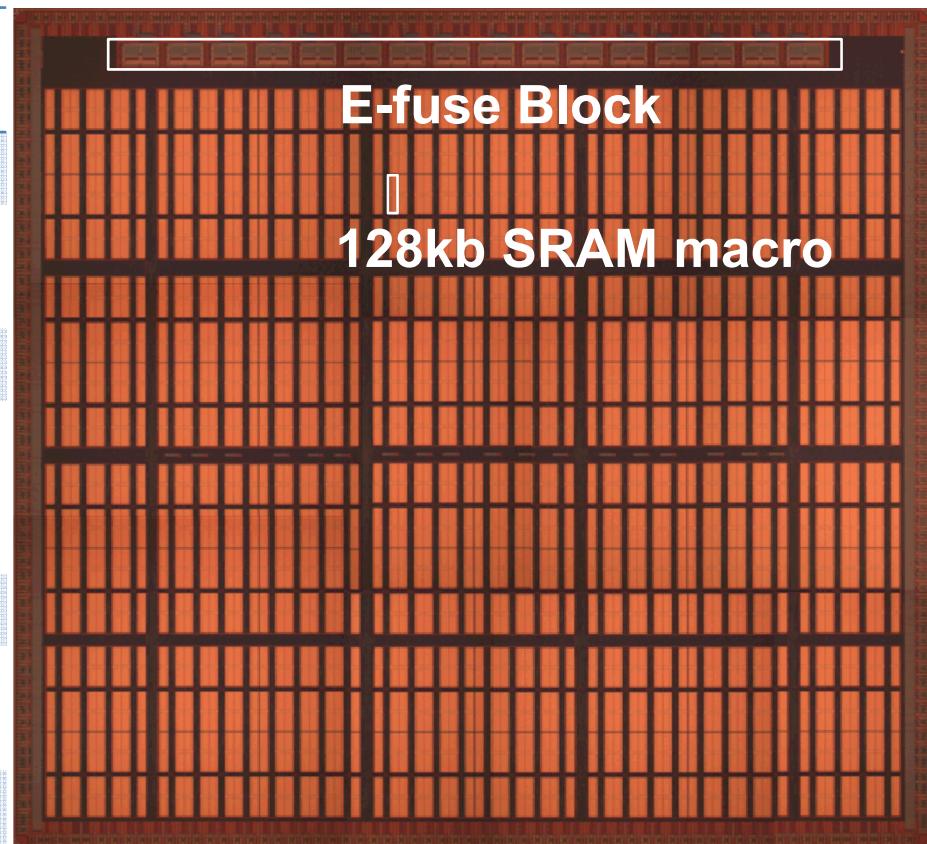


Outline

- Motivation
 - How FinFET impacts on SRAM cell design?
 - Design considerations of write assist techniques
- Proposed low V_{MIN} design techniques
 - Suppressed Coupling Signal Negative Bit-Line (SCS-NBL) scheme
 - Write Recover Enhancement Lower Cell VDD (WRE-LCV) scheme
- Silicon results
- Summary

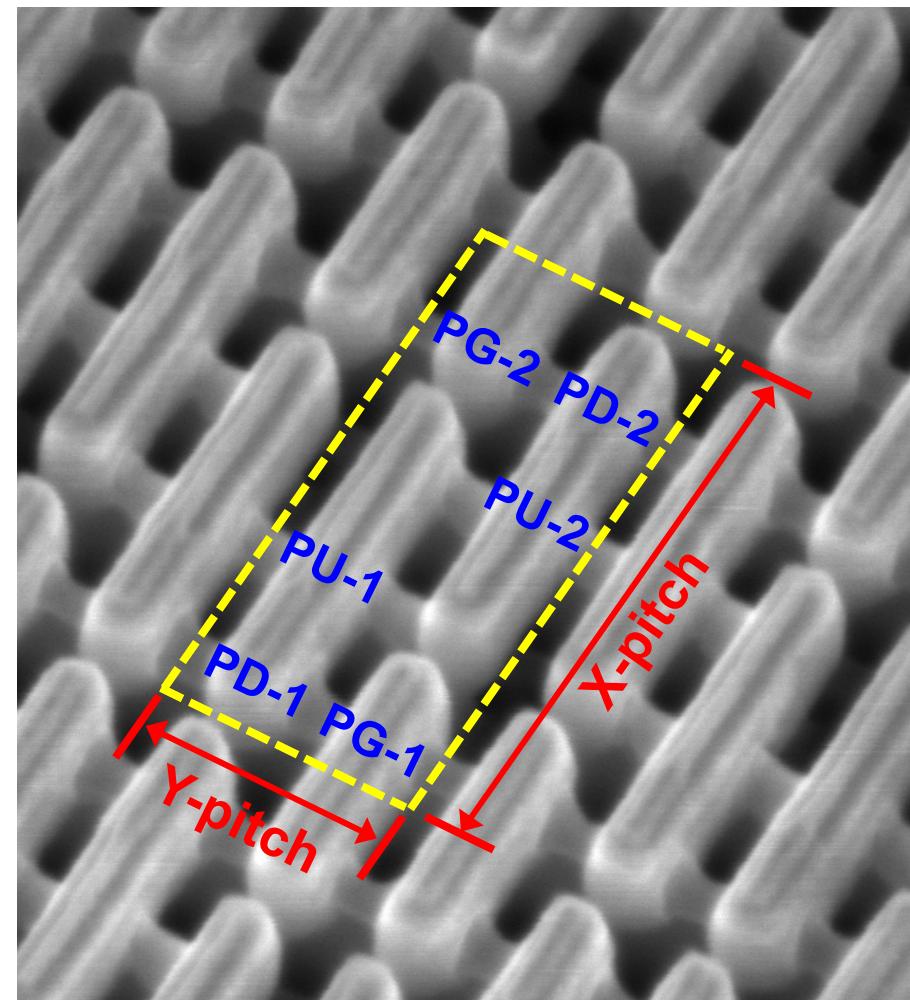
Test Chip Information

Technology	16nm HK-MG FinFET
Metal scheme	1P7M
Supply voltage	Core: 0.85V IO: 1.8V
Bit cell size	0.07 μm^2
SRAM macro configuration	4096x32 MUX=16 258 bits/BL, 272 bits/WL
SRAM capacity	128Mb
Test Features	Row/Column Redundancy Programmable E-fuse
Chip size	6740 μm x 6240 μm = 42mm ²



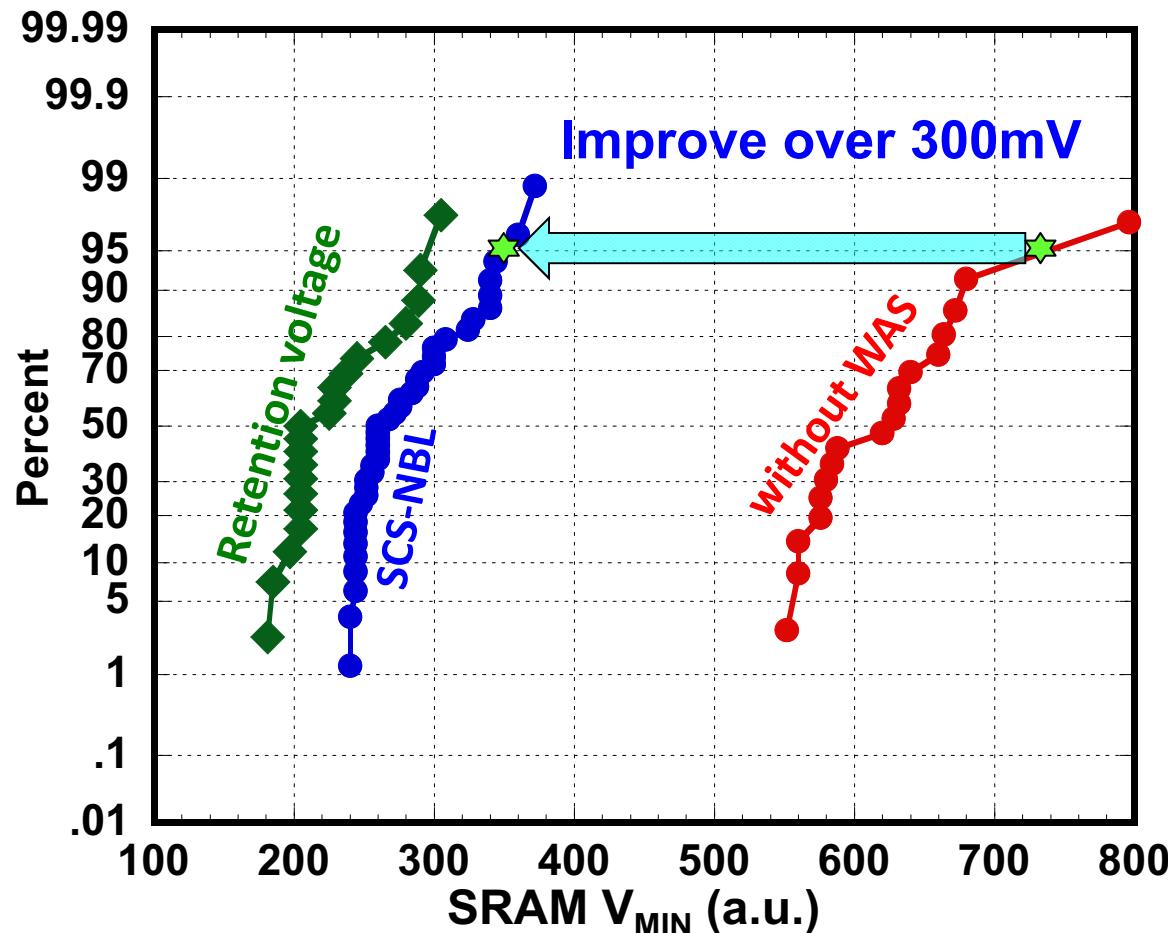
SEM of FinFET HD SRAM Bit Cell

- PU, PG and PD are sized as 1 fin to minimize the SRAM cell area
- The SRAM bit cell area is $0.07\mu\text{m}^2$



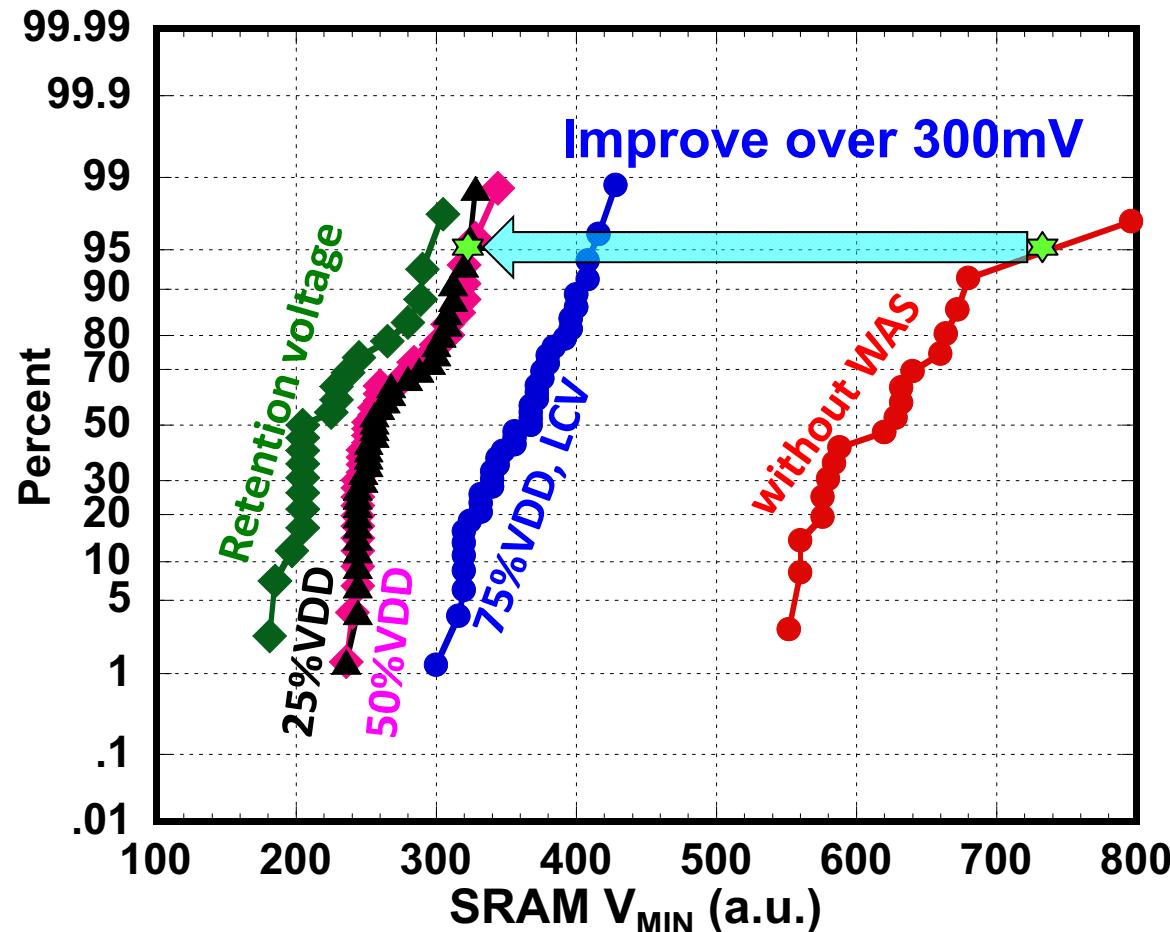
SRAM V_{MIN} Cumulative Plot SCS-NBL

- SCS-NBL improve SRAM V_{MIN} over 300mV at 95 percentile



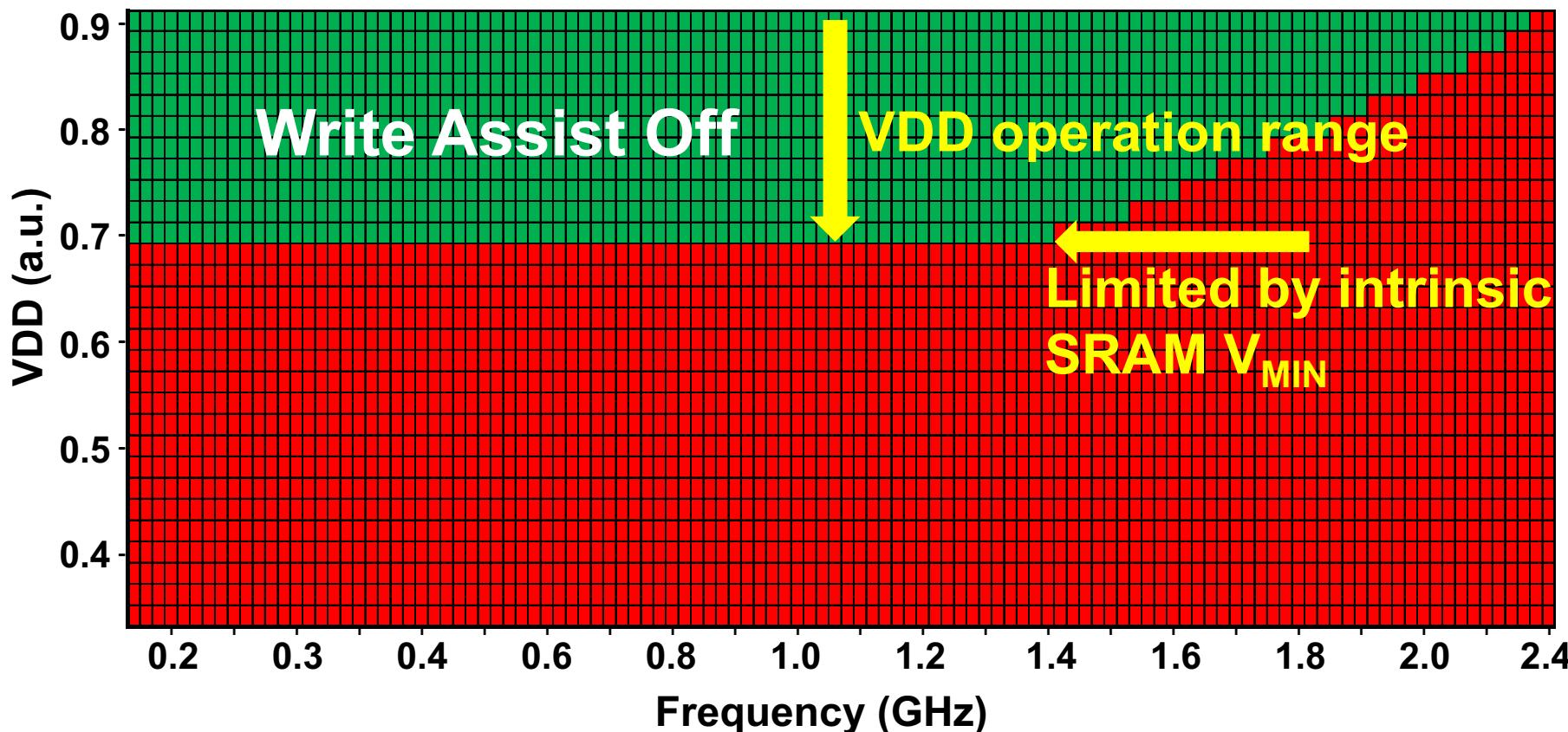
SRAM V_{MIN} Cumulative Plot WRE-LCV

- SRAM V_{MIN} improvement saturates at 50%VDD LCV
- No further V_{MIN} improvement for 25%VDD LCV



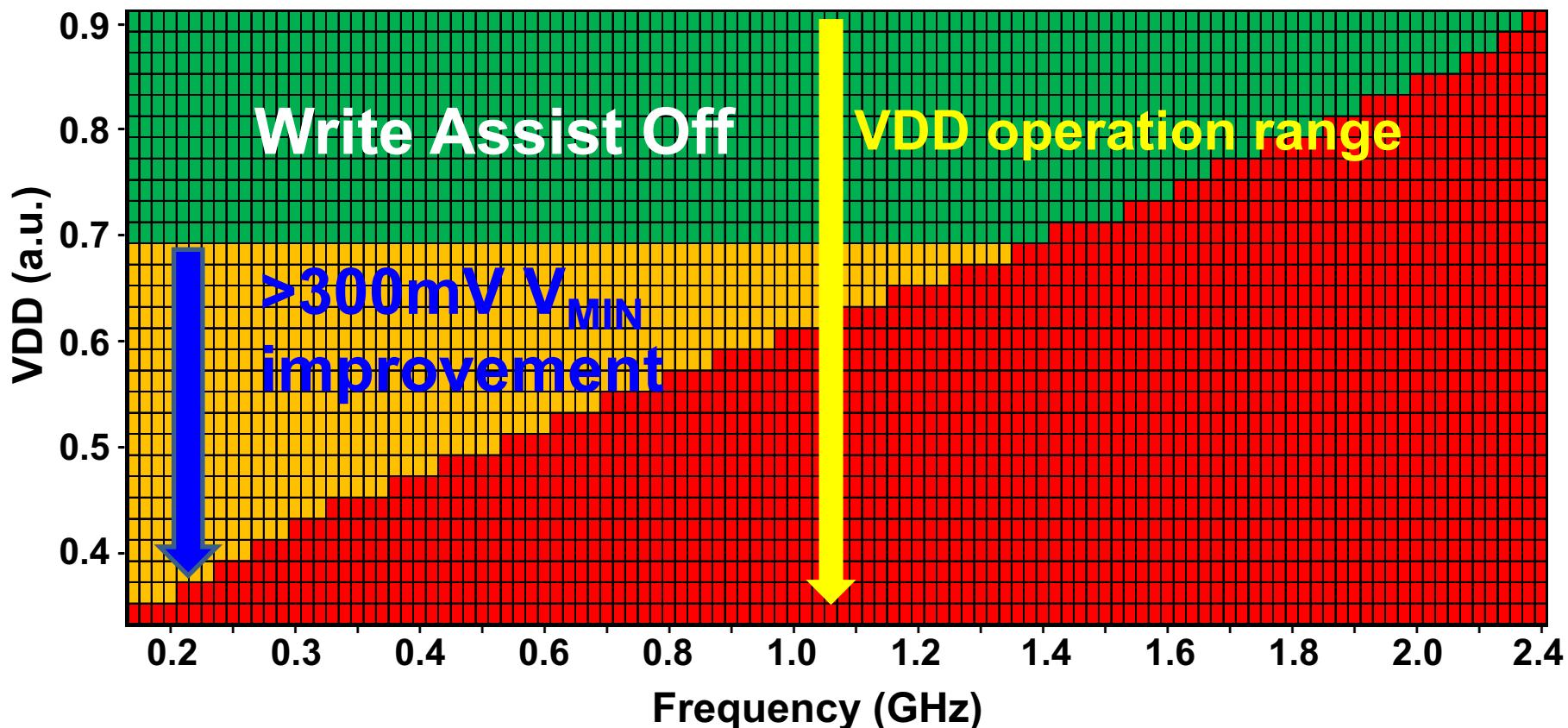
Speed Shmoo Plot w/o WAS

- Without WAS, VDD operation range is limited by intrinsic SRAM V_{MIN}



Speed Shmoo Plot w/i WAS

- With WAS, effectively enlarge the VDD operation range
- Over 300mV V_{MIN} improvement by proposed WAS



Outline

- Motivation
 - How FinFET impacts on SRAM cell design?
 - Design considerations of write assist techniques
- Proposed low V_{MIN} design techniques
 - Suppressed Coupling Signal Negative Bit-Line (SCS-NBL) scheme
 - Write Recover Enhancement Lower Cell VDD (WRE-LCV) scheme
- Silicon results
- Summary

Summary

- Successfully demonstrated a fully functional $0.07\mu\text{m}^2$ SRAM bit cell with proposed write assist techniques in a 16nm FinFET 128Mb SRAM test chip
 - Developed SCS-NBL and WRE-LCV schemes for write assist design solutions
 - The area overhead is 2% for SCS-NBL and 3% for WRE-LCV schemes
 - Overall SRAM V_{MIN} improvement is over 300mV
- The proposed write assist design solutions can enable the lower V_{MIN} applications

Thank you for your attention!

A 28nm 400MHz 4-Parallel 1.6Gsearch/s 80Mb Ternary CAM

Koji Nii¹, Teruhiko Amano², Naoya Watanabe²,
Minoru Yamawaki³, Kenji Yoshinaga³,
Mihoko Wada³ Isamu Hayashi²,

¹ Renesas Electronics, Kodaira, Japan

² Renesas Electronics Itami, Japan

³ Renesas System Design, Itami, Japan

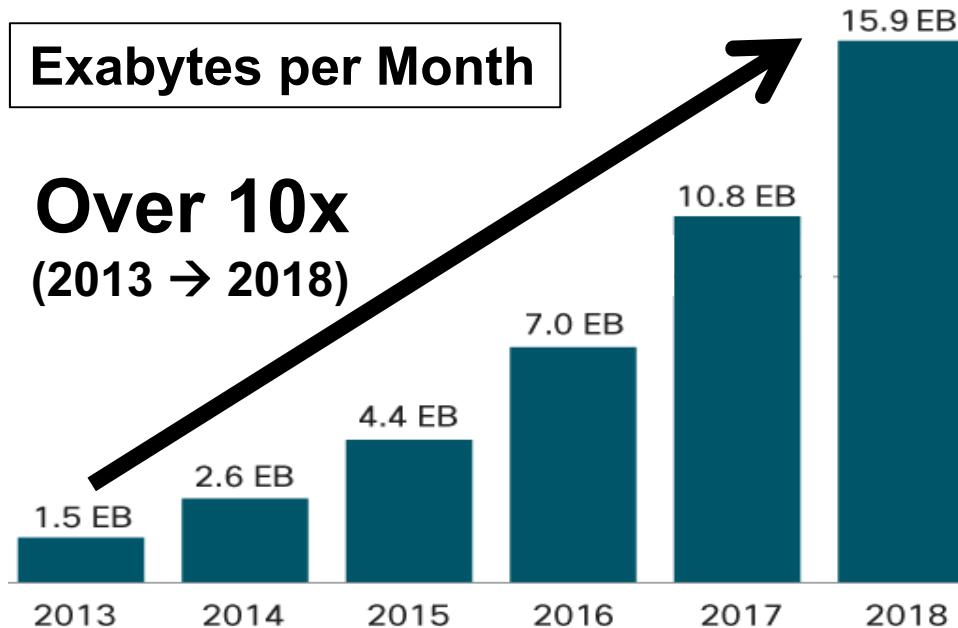


Outline

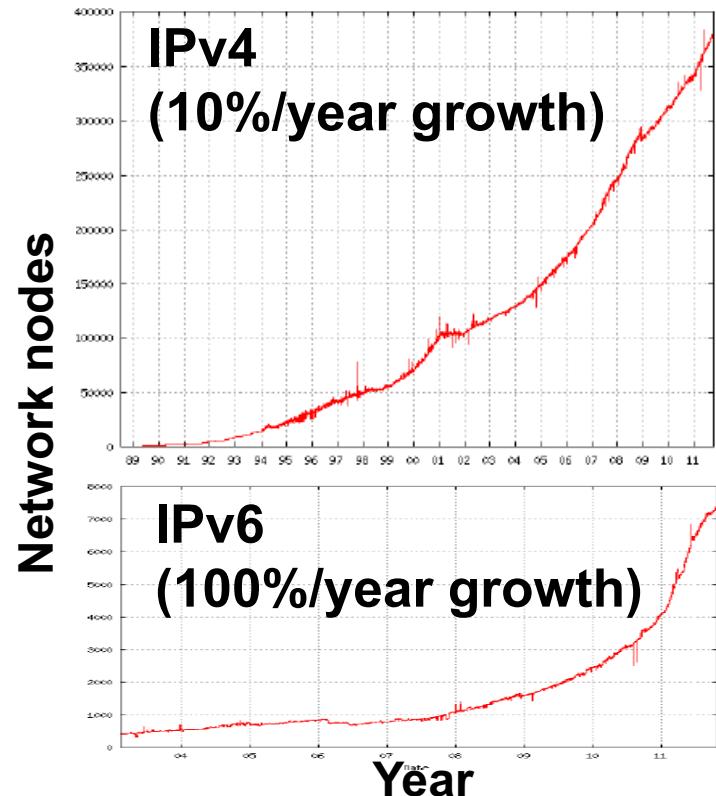
- ✓ Background, Motivation
- ✓ Proposed 80Mb TCAM
 - Dual and Quad search mode w/ flexible search key
 - Shift redundancy
 - Power saving by Valid-bit cell
 - High-speed search w/ Multi-Vt Match sense amp.
- ✓ Measurement results
- ✓ Conclusion

Background

- ✓ Mobile data traffic is rapidly increasing year by year due to growing Smartphone/Tablet and Internet-of-Things (IOT).
- ✓ # of IPv4 routing table entries is growing at a rate of 10% per year. → Risks depletion of IPv4 addresses, need IPv6.

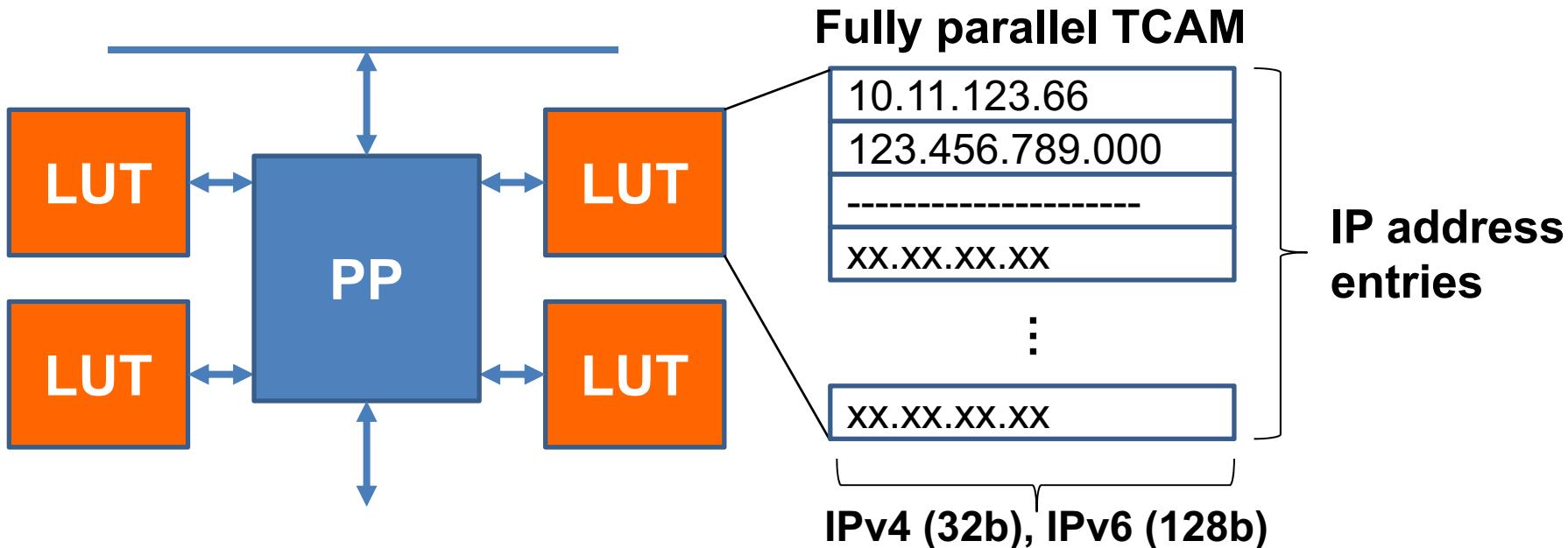


Source: CISCO VNI Mobile, 2014



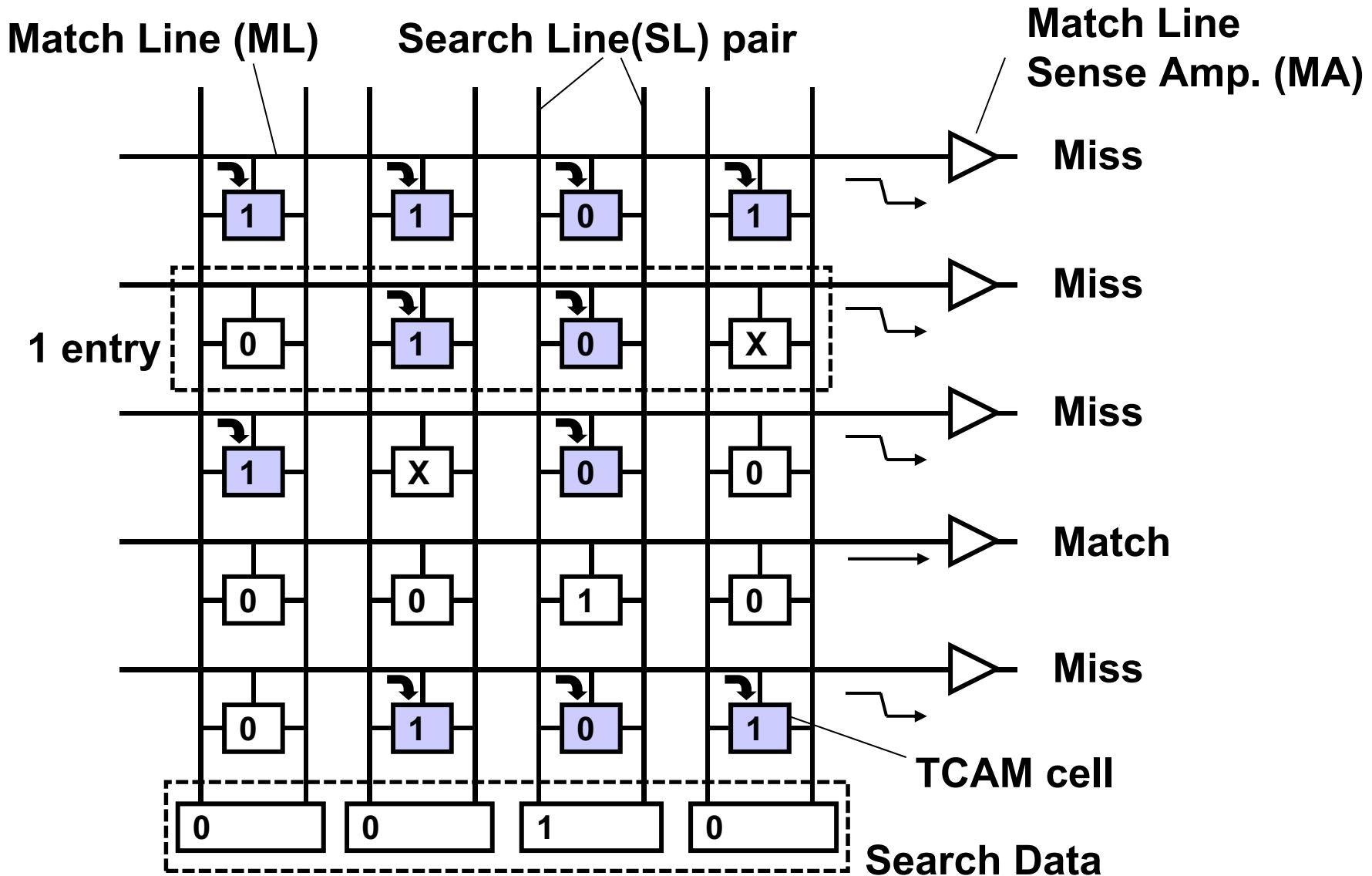
Packet Processing (PP) w/ LUT

- ✓ Demands for routers and switches capable of packet processing with high-throughput and low-latency.
- ✓ Fully parallel TCAMs are the key devices to handle the large number of routes in the look-up table (LUT).

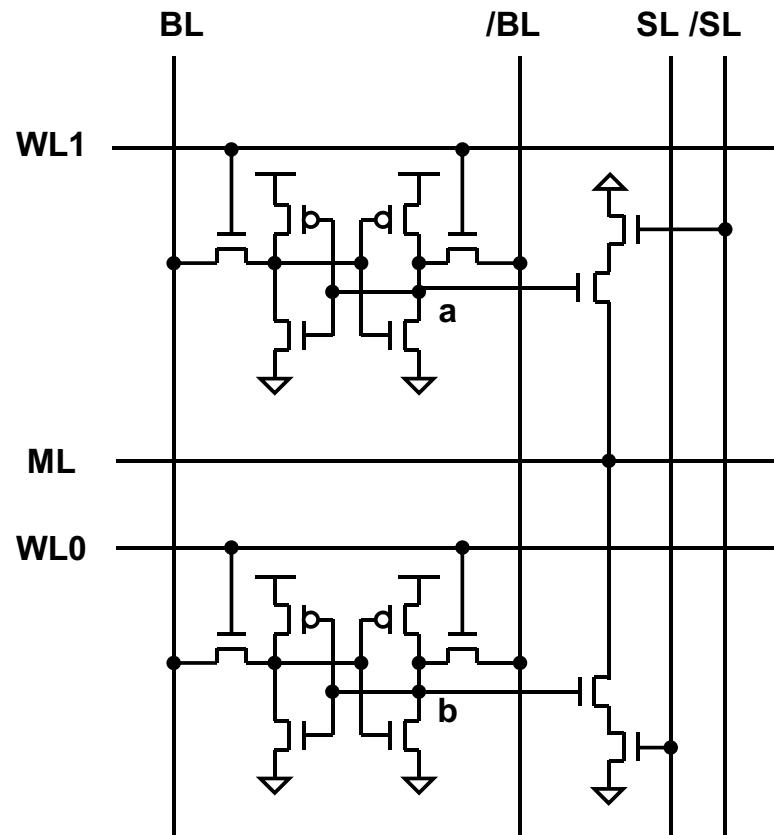


Realizing both high-density (high-capacity) and high-speed search operation for TCAM is a big challenge.

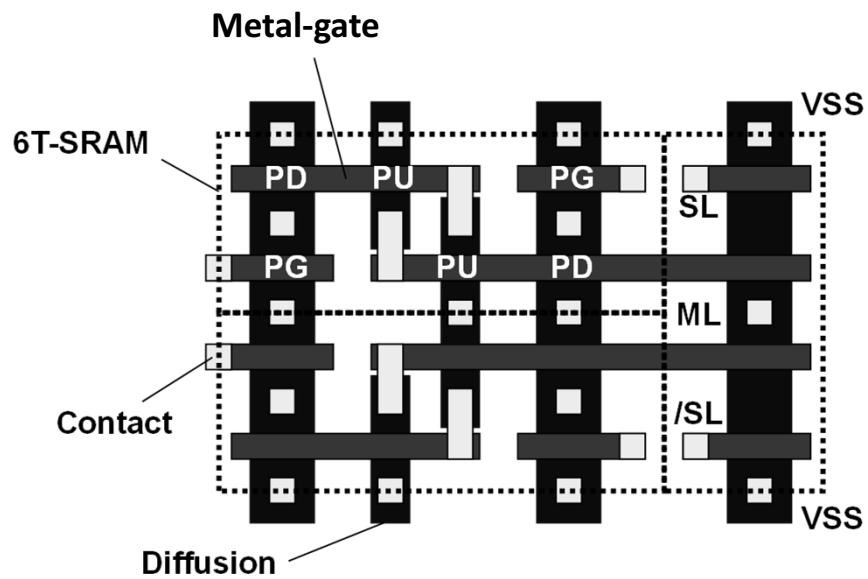
Ternary CAM



TCAM Bitcell



a	b	Stored data
0	1	0
1	0	1
0	0	X (don't care)
1	1	Inhibit

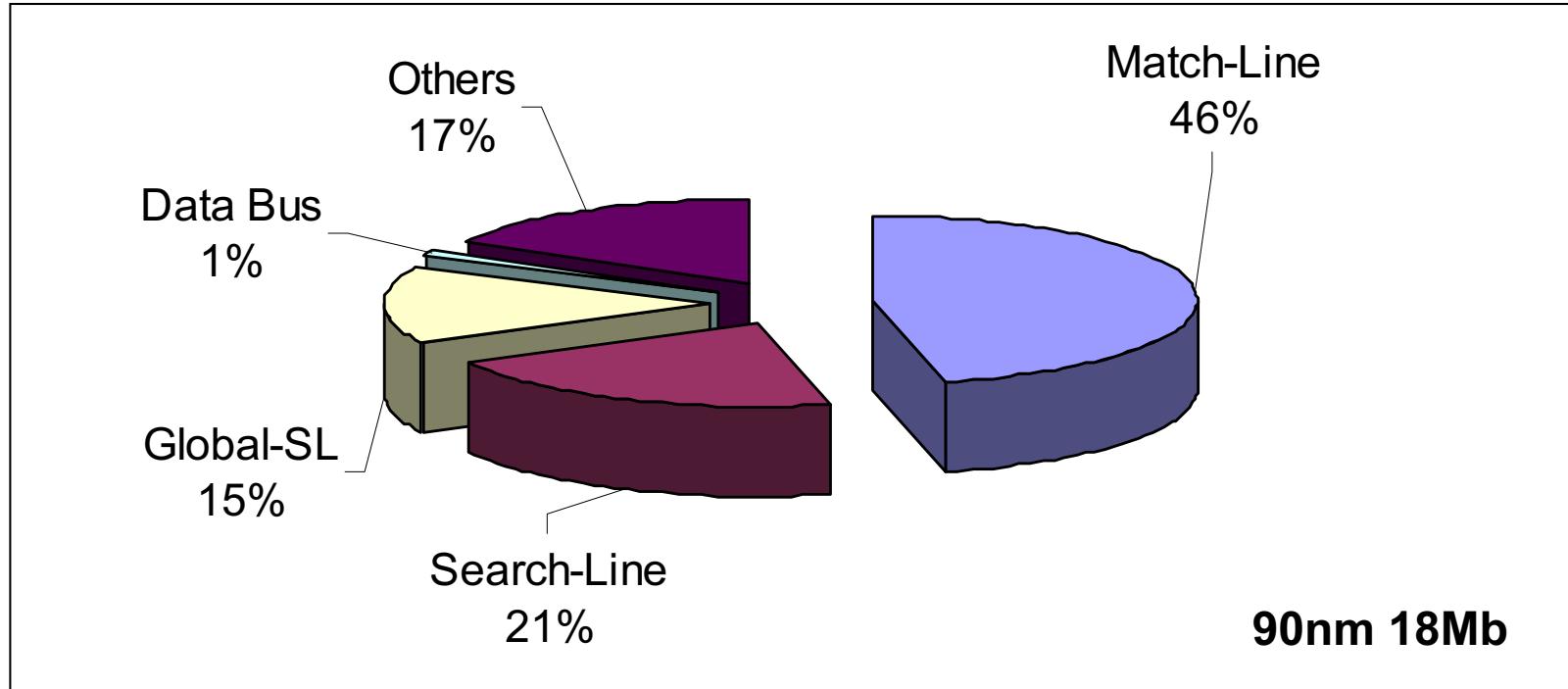


16T TCAM bitcell
including two 6T SRAMs

→ **High-density (high-capacity)
but much power consumption.**

Motivation

Power distribution of TCAM search operation



Source: I. Hayashi et al., ASSCC 2012

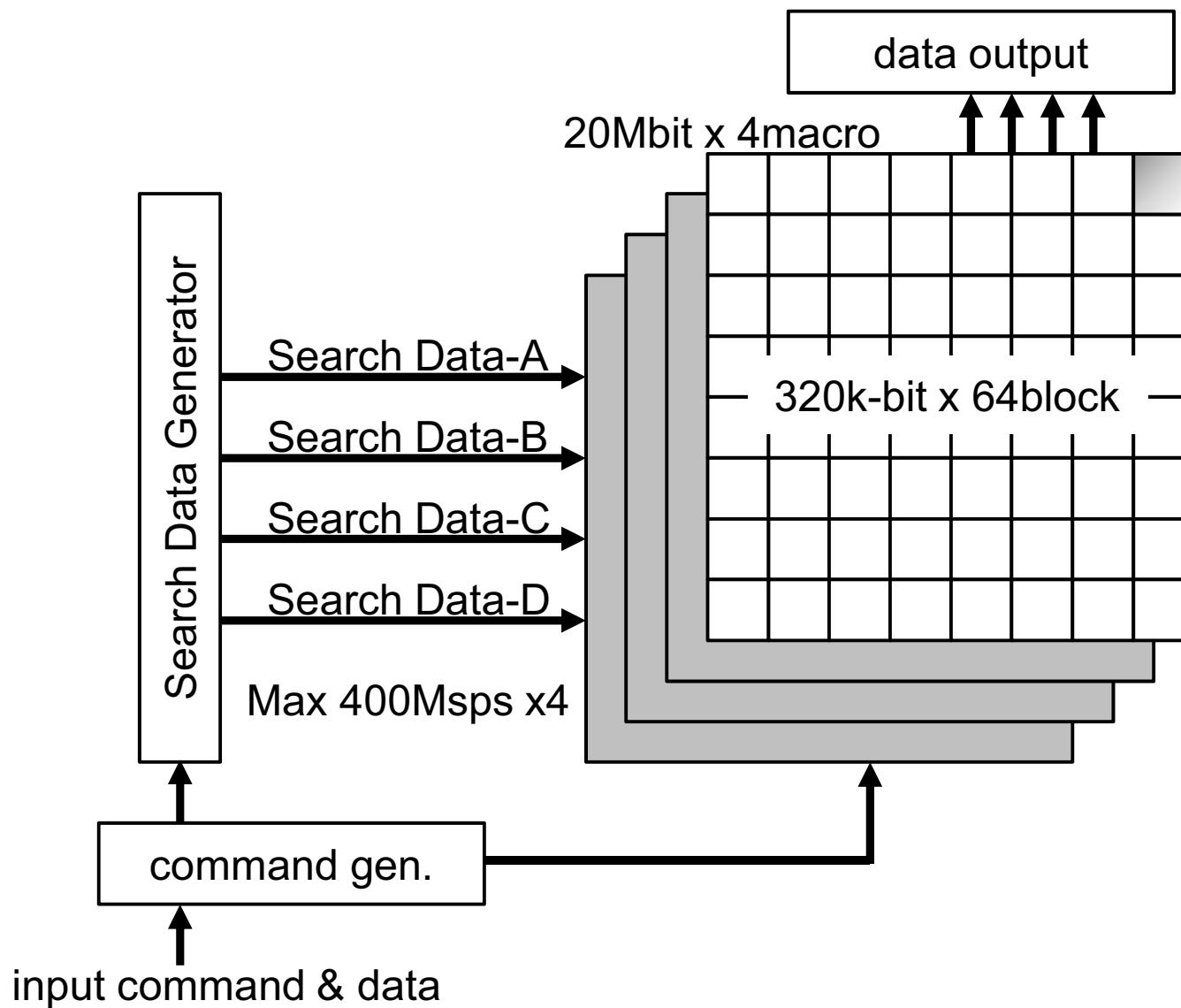
Match-line charge/dis-charge power is dominant.

→ Need to reduce the dynamic power of match-line.

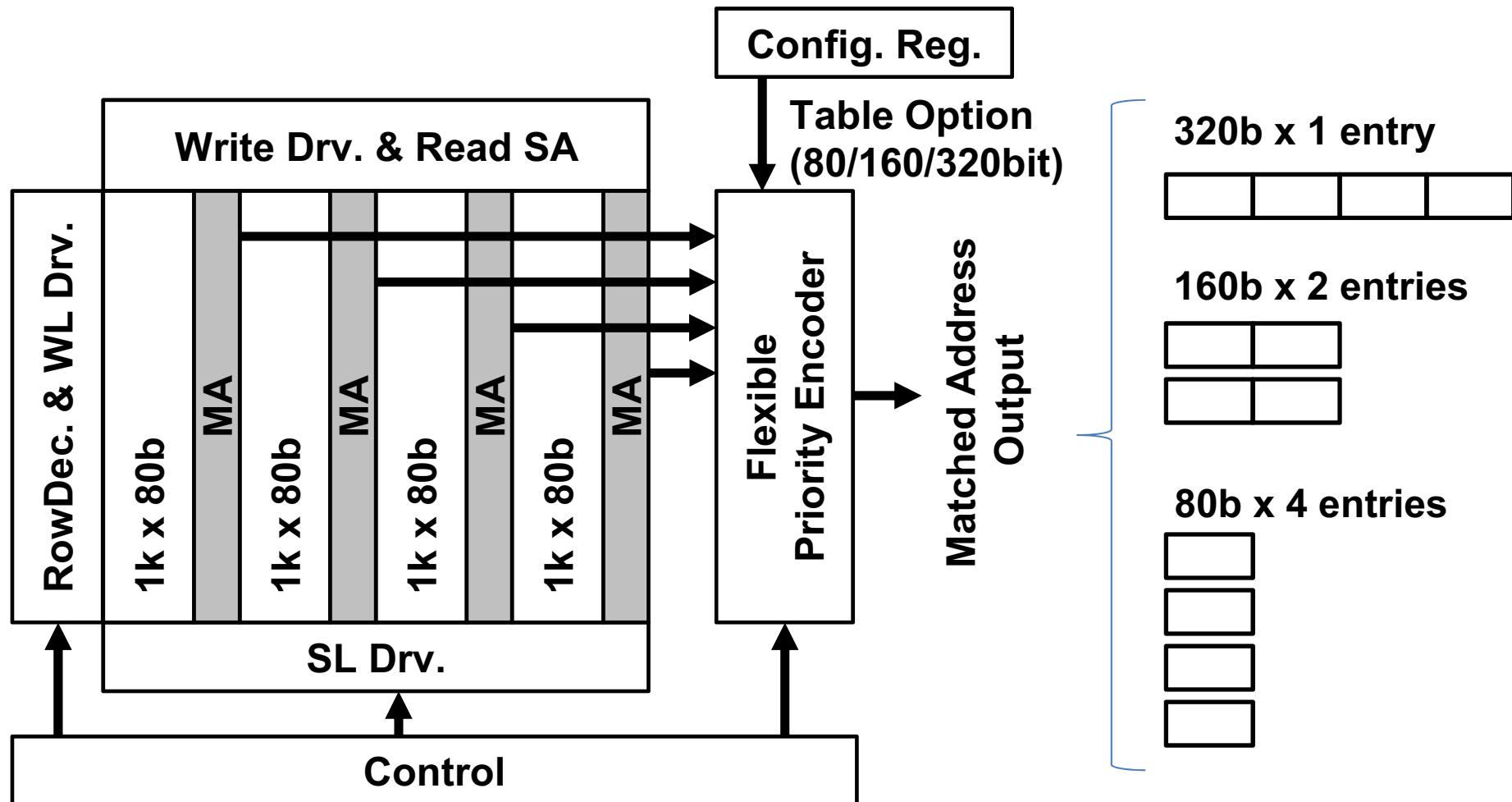
Outline

- ✓ Background, Motivation
- ✓ Proposed 80Mb TCAM
 - Dual and Quad search mode w/ flexible search key
 - Shift redundancy
 - Power saving by Valid-bit cell
 - High-speed search w/ Multi-Vt Match sense amp.
- ✓ Measurement results
- ✓ Conclusion

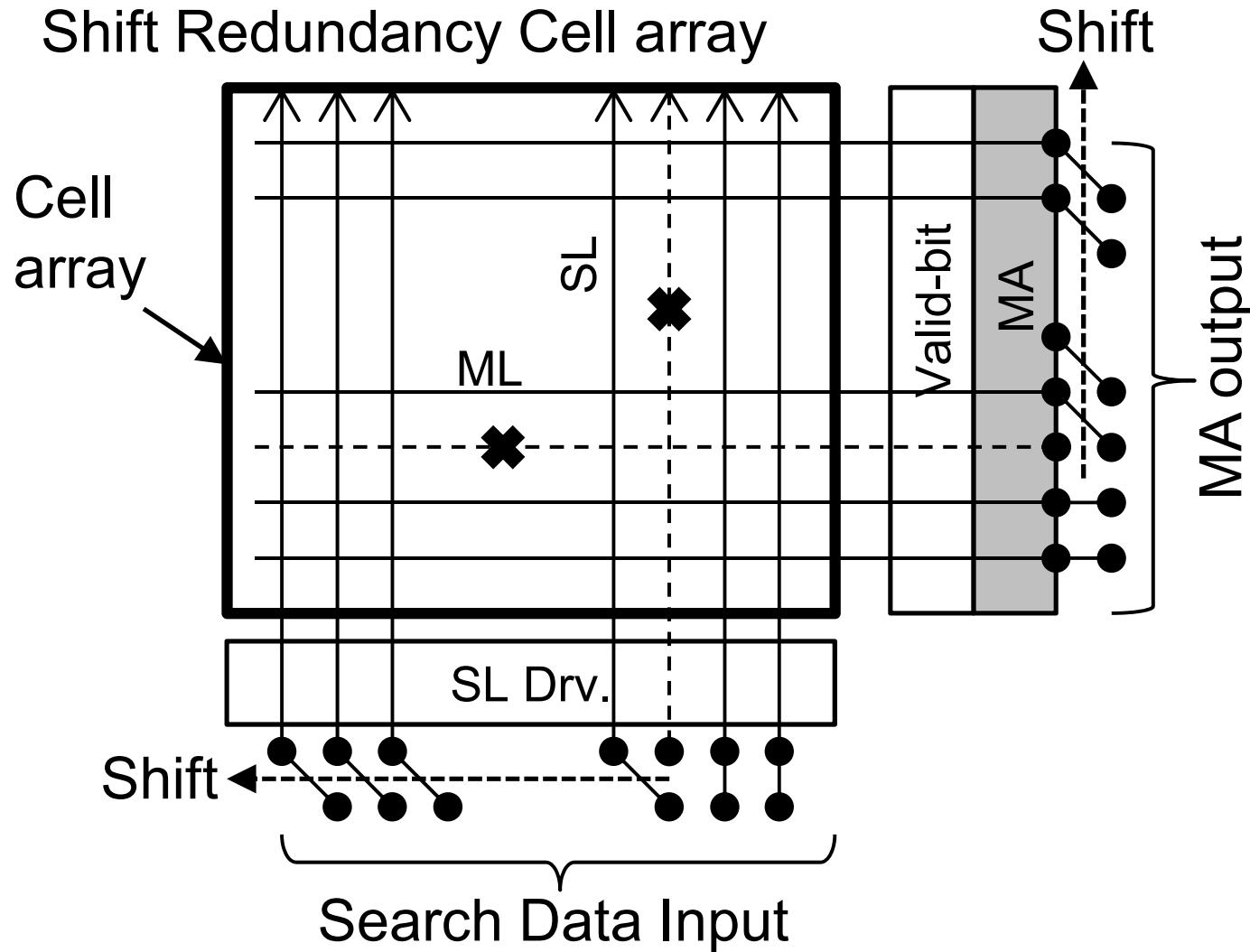
80Mb TCAM Block Diagram



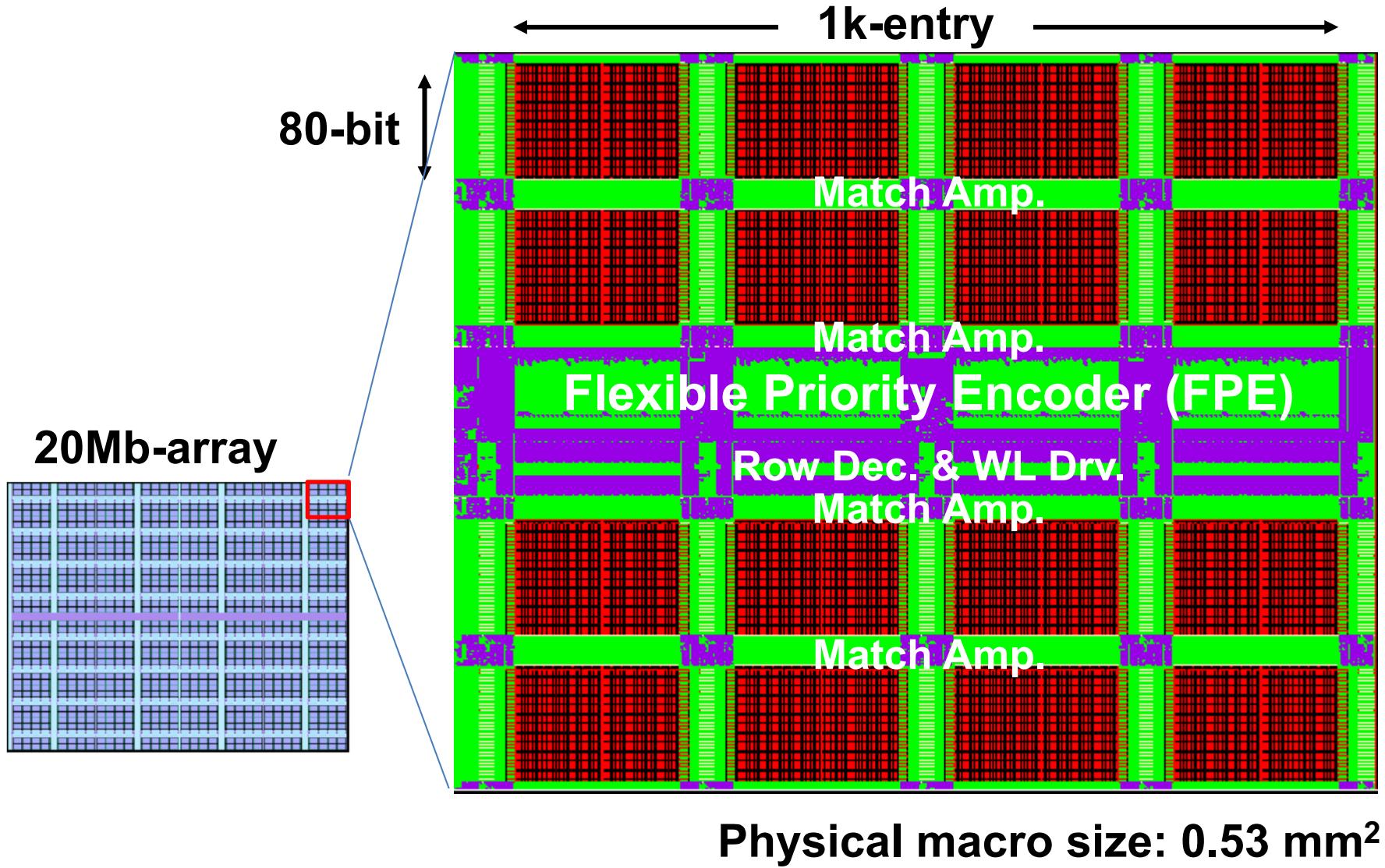
320kb Sub-array Block Diagram



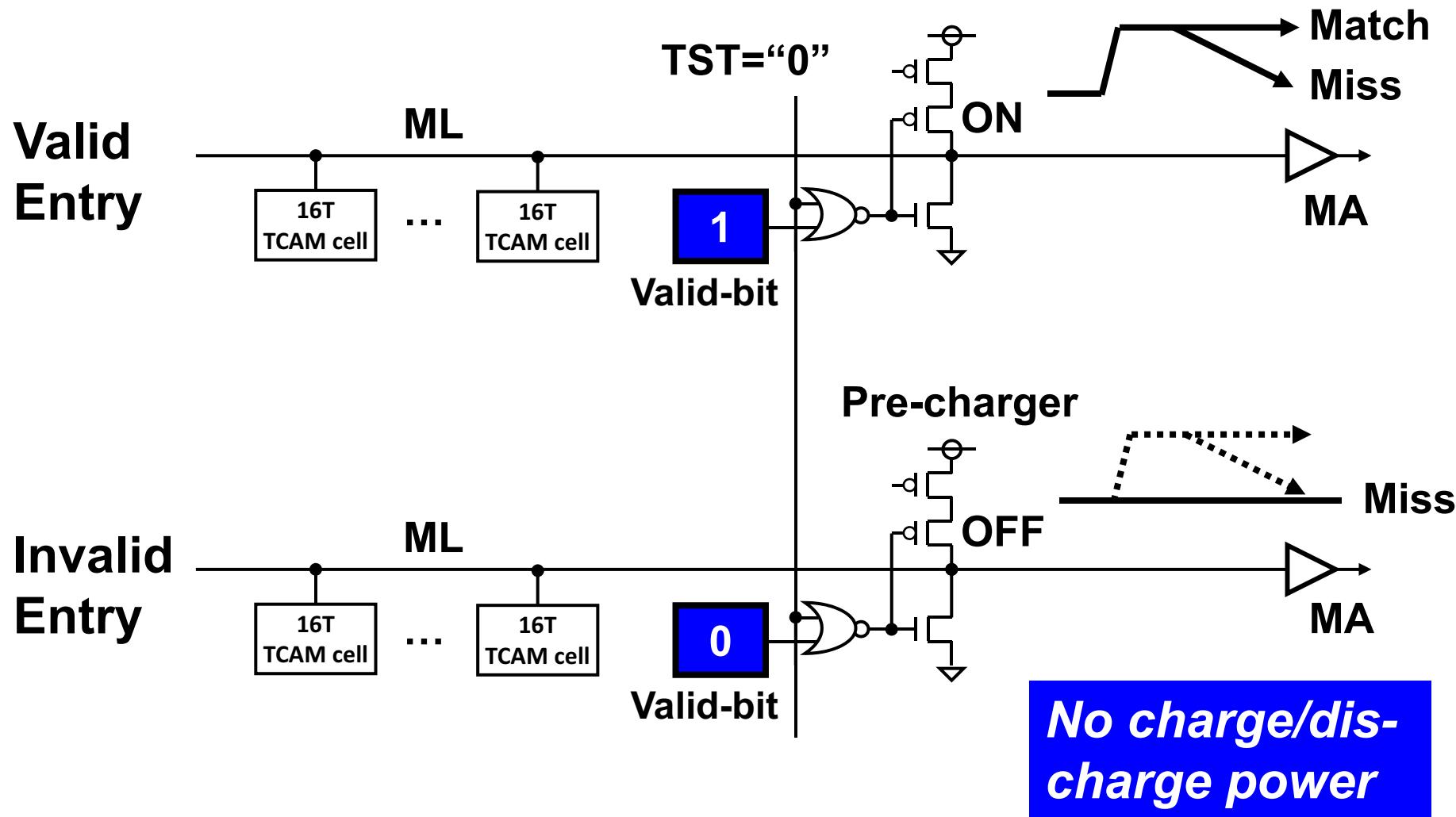
80kb Sub-array w/ Shift Redundancy



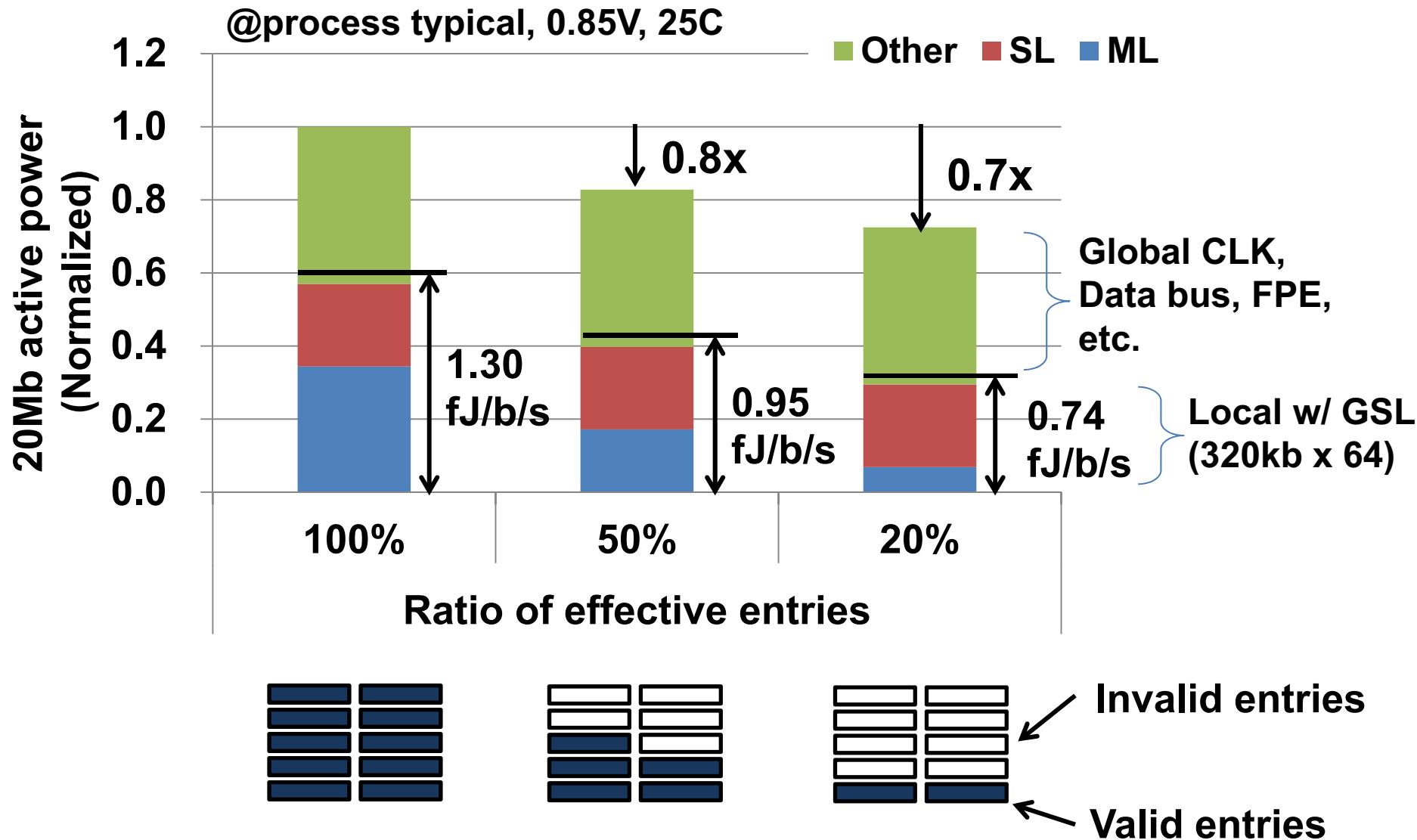
Layout Plots of 320kb Sub-array



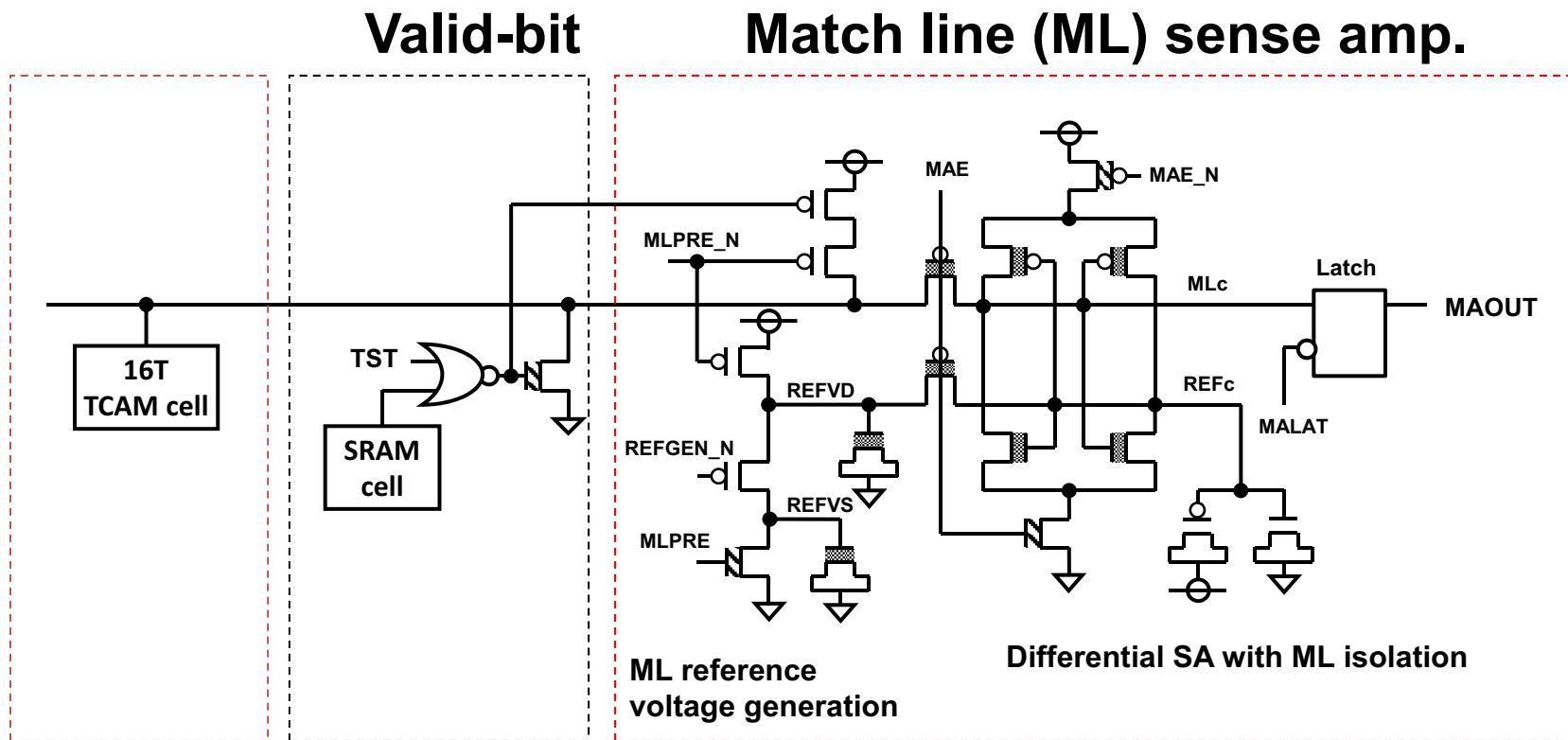
Power Saving by Valid-bit Cell



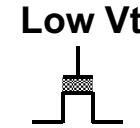
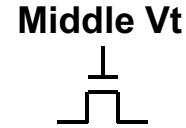
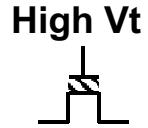
Estimated Power Reduction



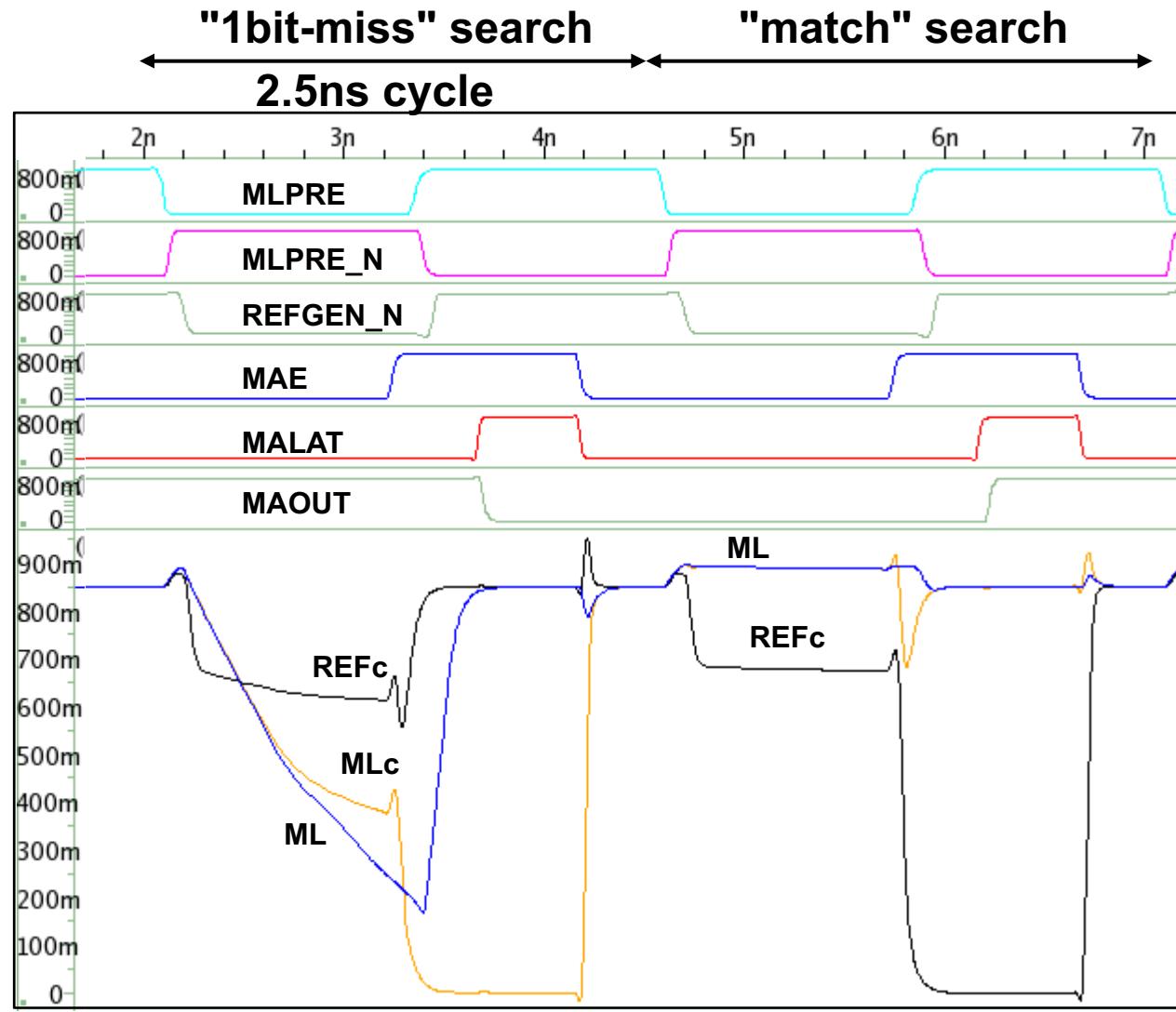
Multi-Vt Match Sense Amplifier



**80-bit
TCAM-array**



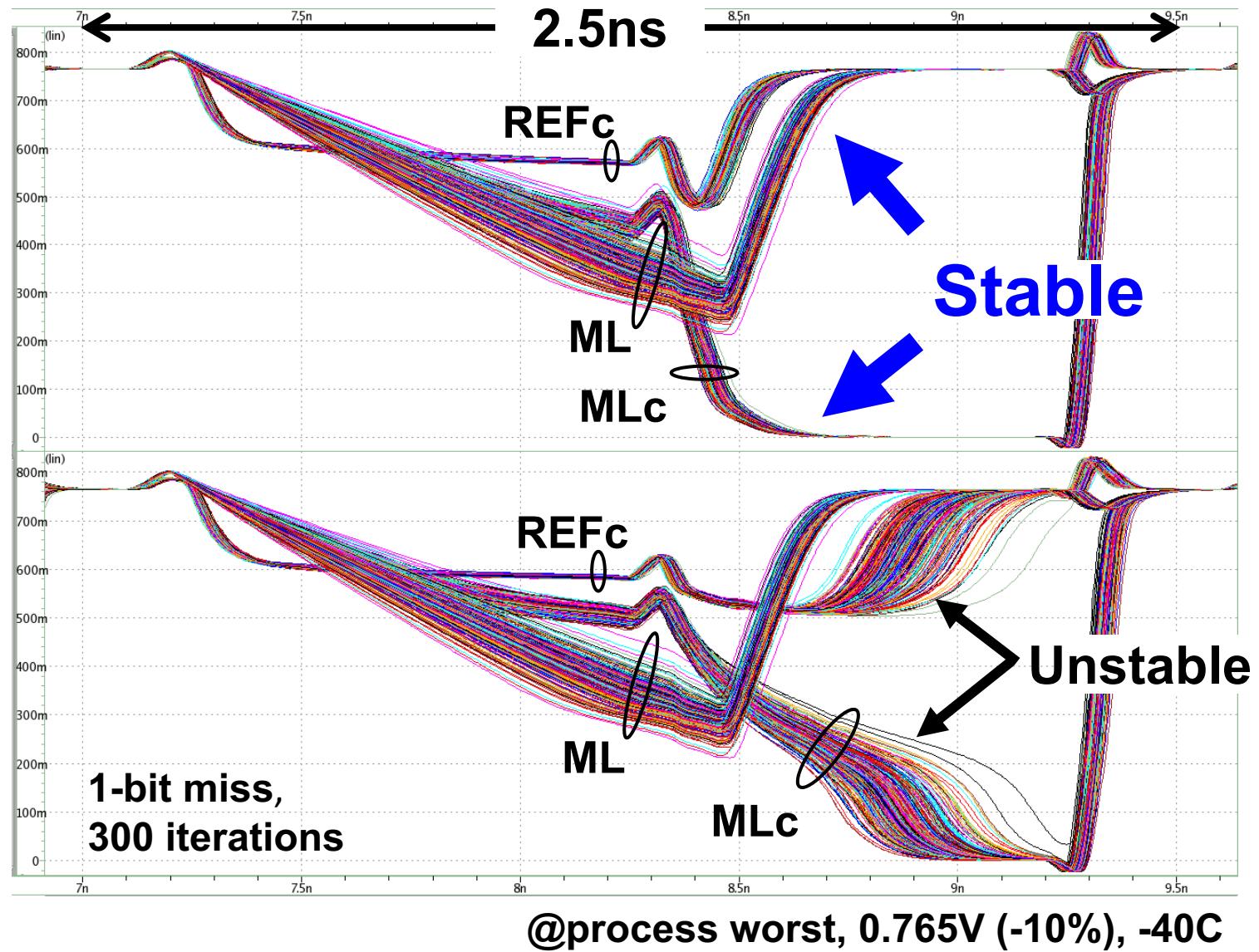
Simulation Waveform



@process worst, 0.85V, 25C

Monte Carlo Sim.

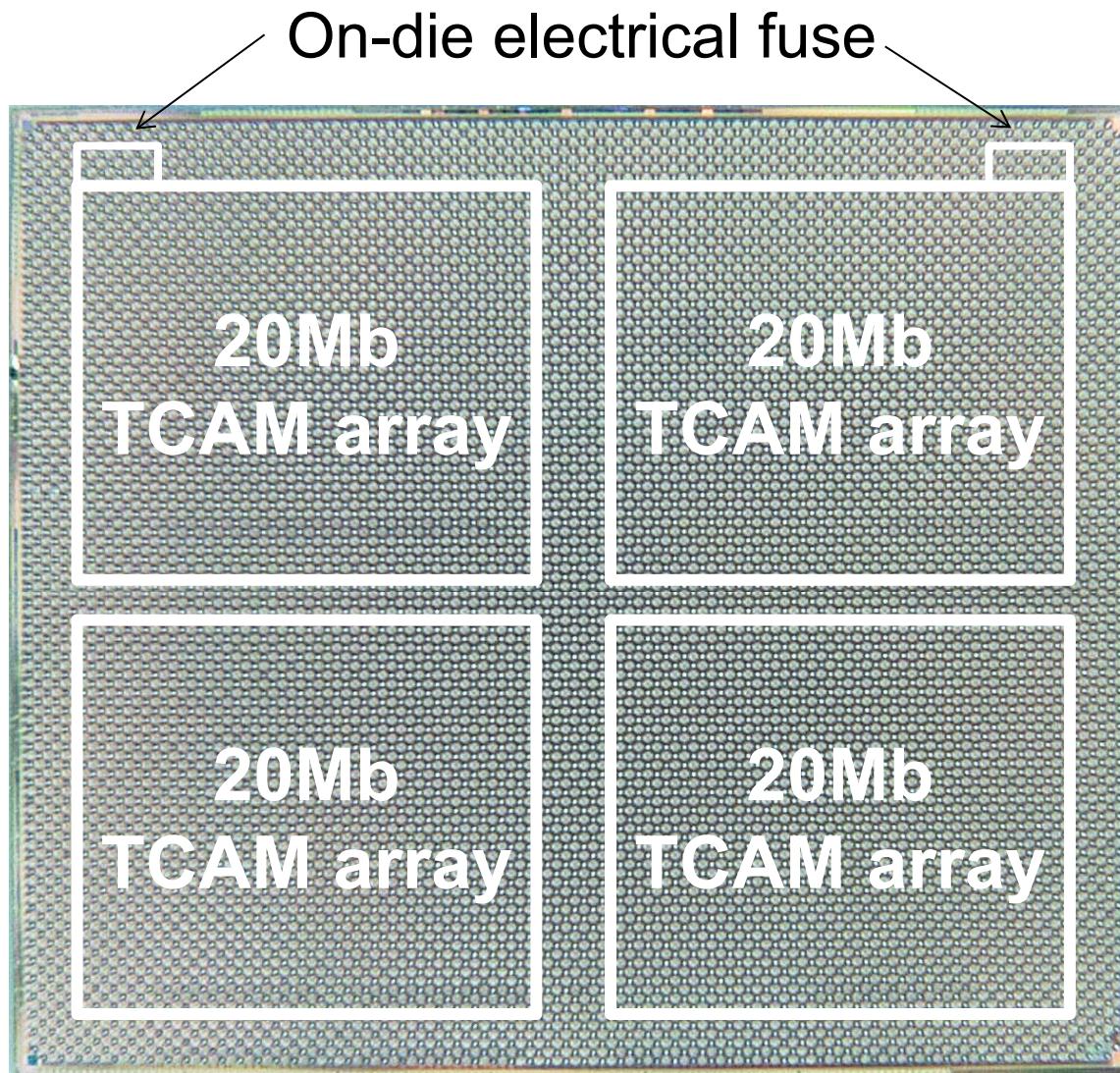
Multi-Vt
(Prop.)



Outline

- ✓ Background, Motivation
- ✓ Proposed 80Mb TCAM
 - Dual and Quad search mode w/ flexible search key
 - Shift redundancy
 - Power saving by Valid-bit cell
 - High-speed search w/ Multi-Vt Match sense amp.
- ✓ Measurement results
- ✓ Conclusion

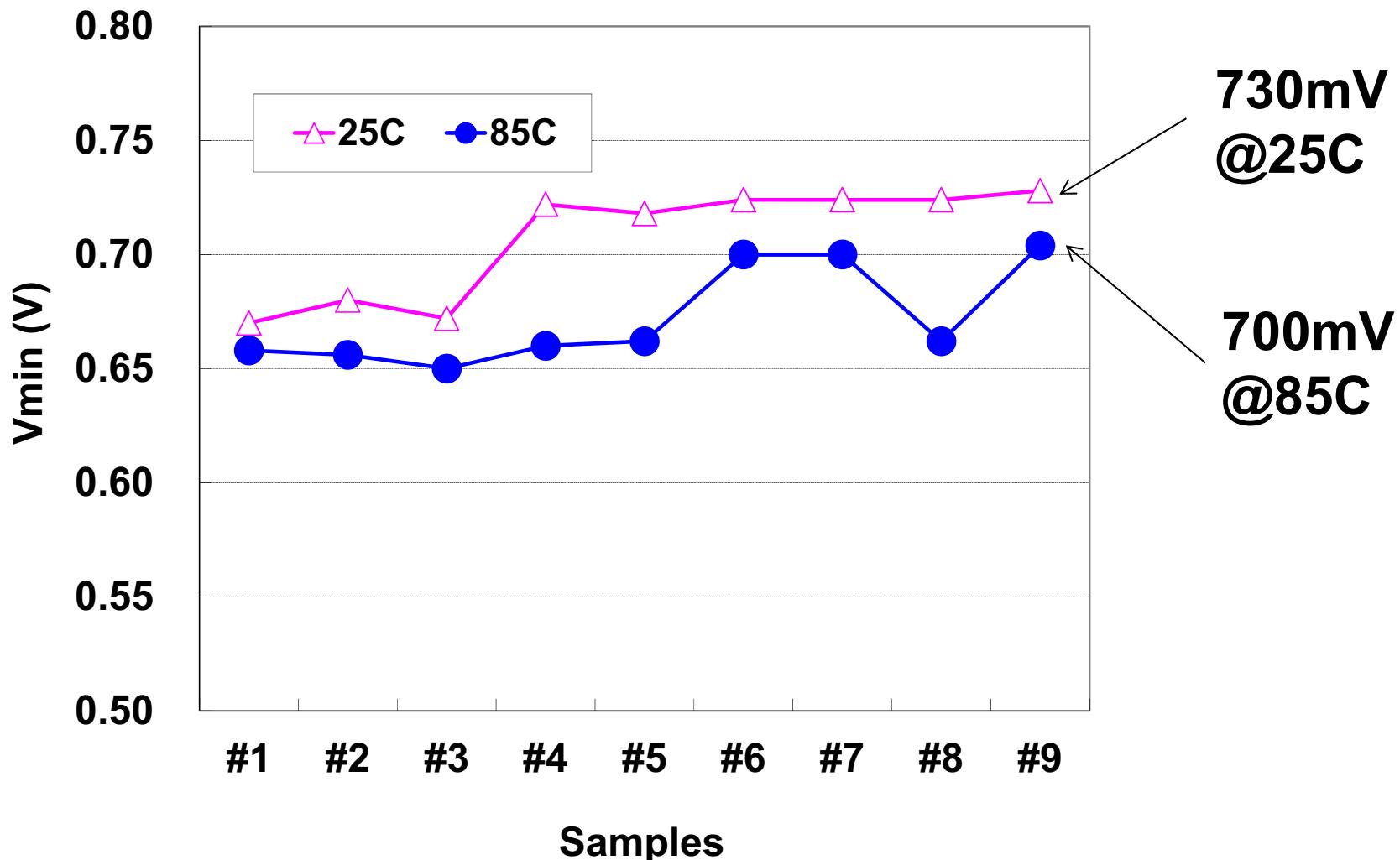
Die Micrograph



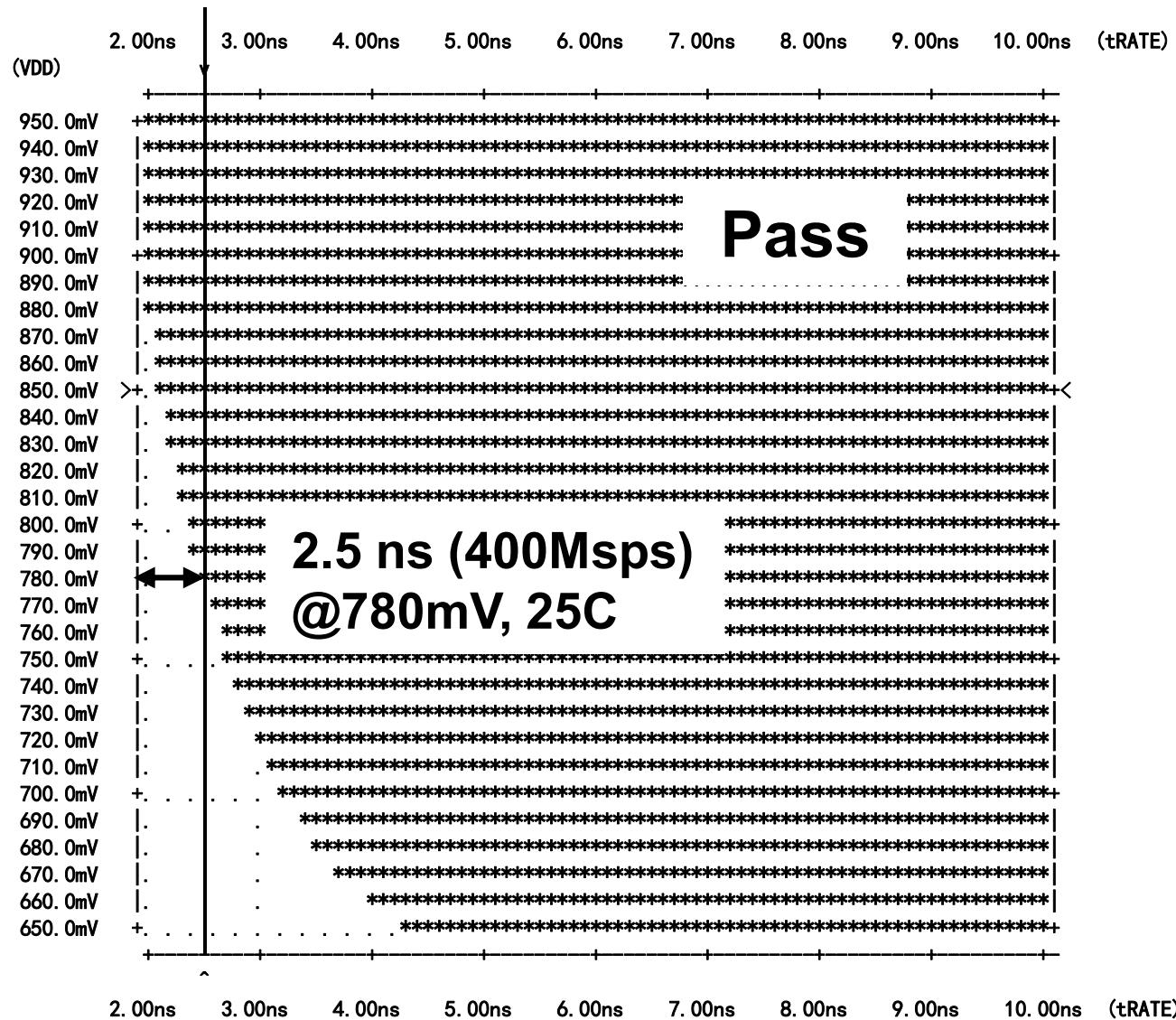
Chip Features

Technology:	28nm HKMG high-performance bulk CMOS with 10-Cu metals and AL-top-metal
Supply voltage:	0.85 V
Clock frequency:	Max 400MHz
Total capacity:	80-Mbit
Configuration:	320k (4k x 80b) x 64 blocks x 4 macros
Physical macro size:	0.53 mm ² @ 320-kbit sub-array
TCAM bitcell:	16T bitcell type (including two 6T SRAMs)
Max search speed:	400M-sps (single mode), 800M-sps (dual mode), 1.6G-sps (quad mode) @ 0.85 V, 25C
Power consumption:	8.4 W (Meas.) @ 20Mb, 250M-sps, 50% entries, 0.85 V, 25C

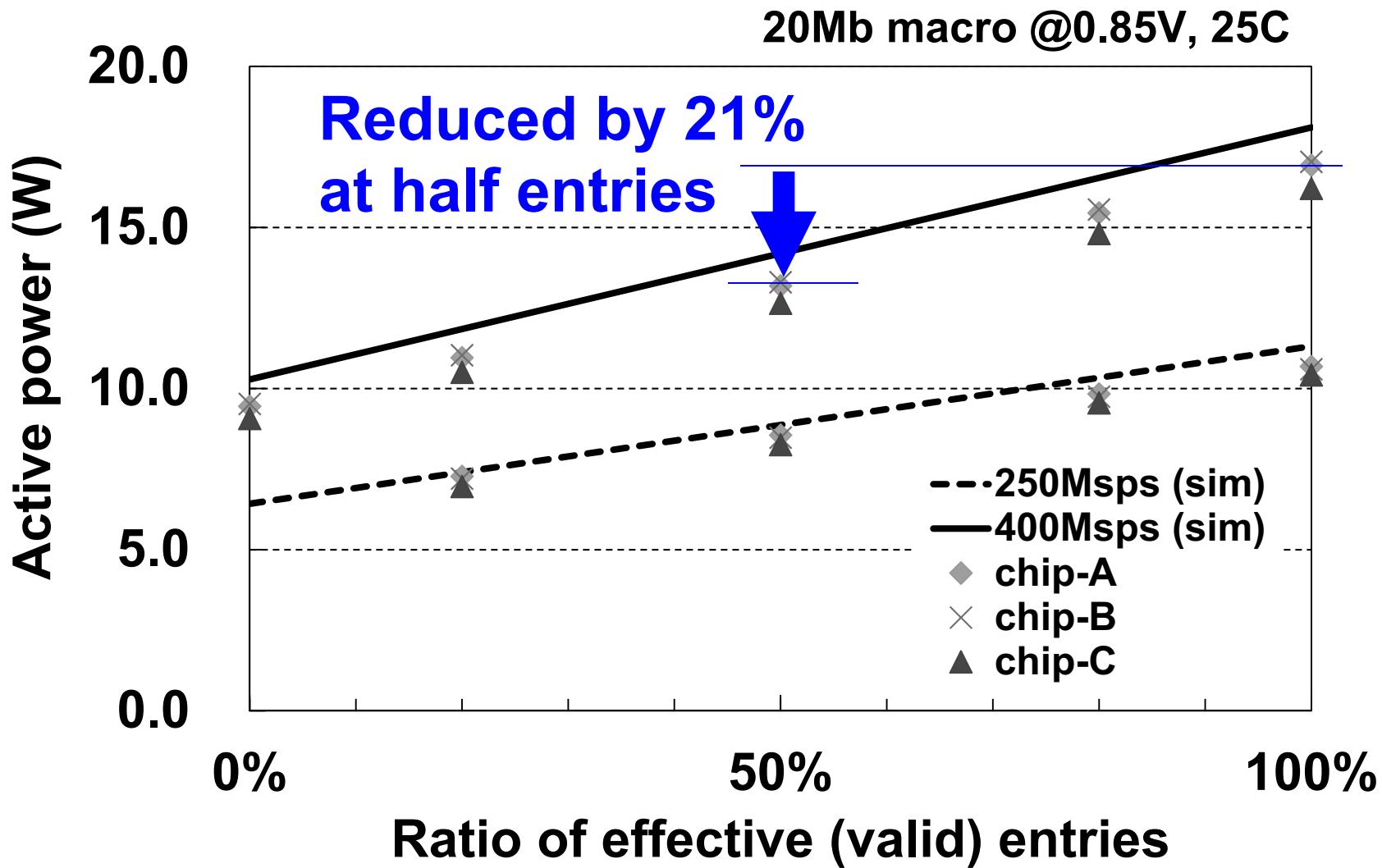
Measured Vmin



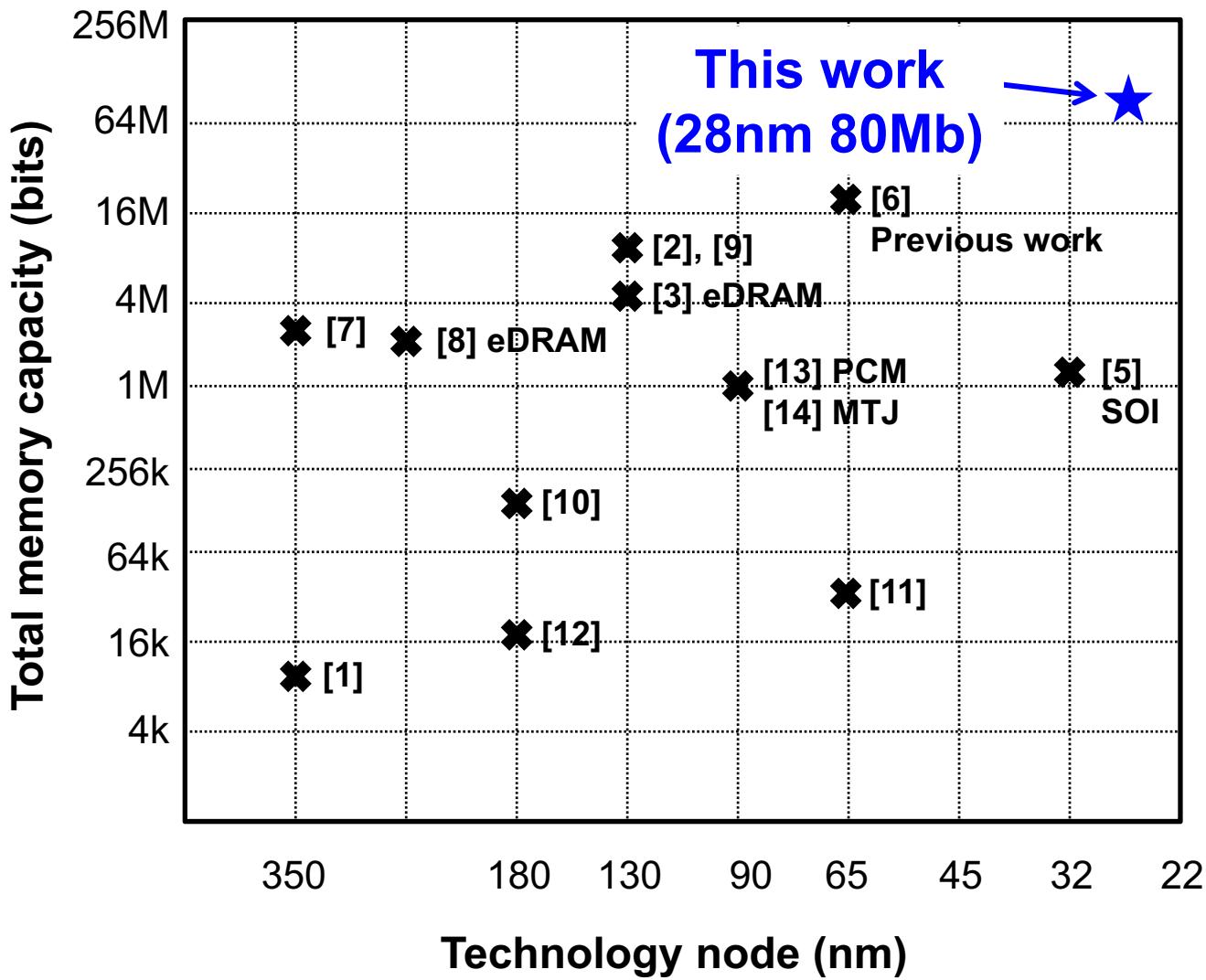
SHMOO Plot



Measured Power Consumption



Comparison with Previous Works



- [1] H. Miyatake, JSSC, 2001.
- [2] G. Kasai, CICC, 2003.
- [3] H. Noda, JSSC, 2005.
- [4] K. Pagiamtzis, JSSC, 2006.
- [5] I. Arsovski, JSSC, 2013
- [6] Isamu Hayashi, JSSC, 2013.
- [7] F. Shafai, JSSC, 1998.
- [8] V. Linesi, IWM, 2000.
- [9] A. Rothi , CICC, 2004.
- [10] T. Kusumoto, ASSCC, 2008.
- [11] P.-T. Huanget, JSSC, 2011.
- [12] B.-D. Yang , TCAS-I, 2011.
- [13] J. Li, VLC, 2013.
- [14] S. Matsunaga , VLC, 2013.

Comparison with Previous Works

	JSSC 2013 [5]	Previous work [6]	This work
Technology	32nm HKMG SOI	65nm CMOS	28nm HKMG CMOS
Supply Voltage	0.95 V	1.0 V	0.85 V
Total Capacity	1.25 Mb	18 Mb (4.5M x 4macro)	80 Mb (20M x 4macro)
Density	0.84 Mb/mm ²	0.24 Mb/mm ²	0.61 Mb/mm ²
Max. Search Rate	1Gsearch/s	250Msearch/s	400Msearch/s (single) 800Msearch/s (dual) 1.6Gsearch/s (quad)
Energy Efficiency @1.28Mb (w/o FPE)	1.04 fJ/bit/search (worst case) 0.58 fJ/bit/search (best case)	1.38 fJ/bit/search	1.30 fJ/bit/search (100% full valid entries) 0.74 fJ/bit/search (20% valid entries)

→ Total capacity and Max. search rate are current industry best.

Conclusion

- ✓ We successfully designed and fabricated 80Mb TCAM using 28nm HKMG bulk CMOS.
- ✓ Dual and Quad search modes are available with flexible search key-width (80/160/320/640-bit)
- ✓ Proposed valid-bit cell reduces 21% search power at half effective entries and multi-V_t match sense amplifier enables 2.5 ns match access time, achieving 1.6Gsearch/s at quad mode.
- ✓ Measured V_{min} are below 0.75 V and observed good distribution by effective redundancy.



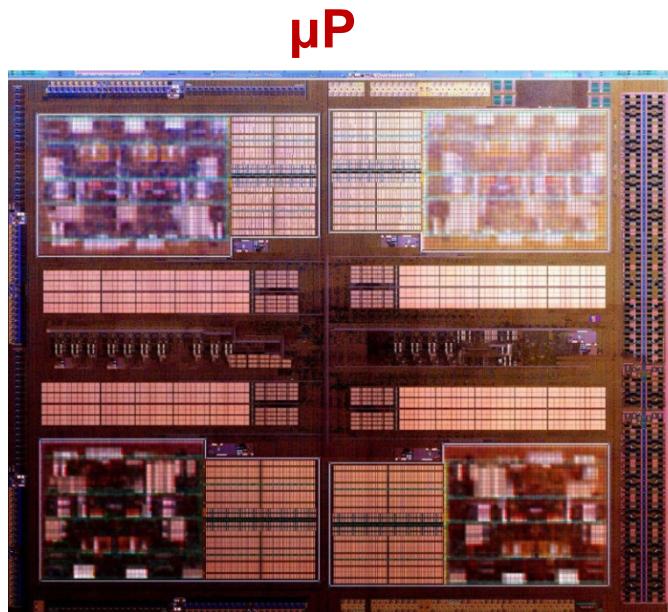
A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

Bharan Giridhar, Nathaniel Pinckney,
Dennis Sylvester, David Blaauw

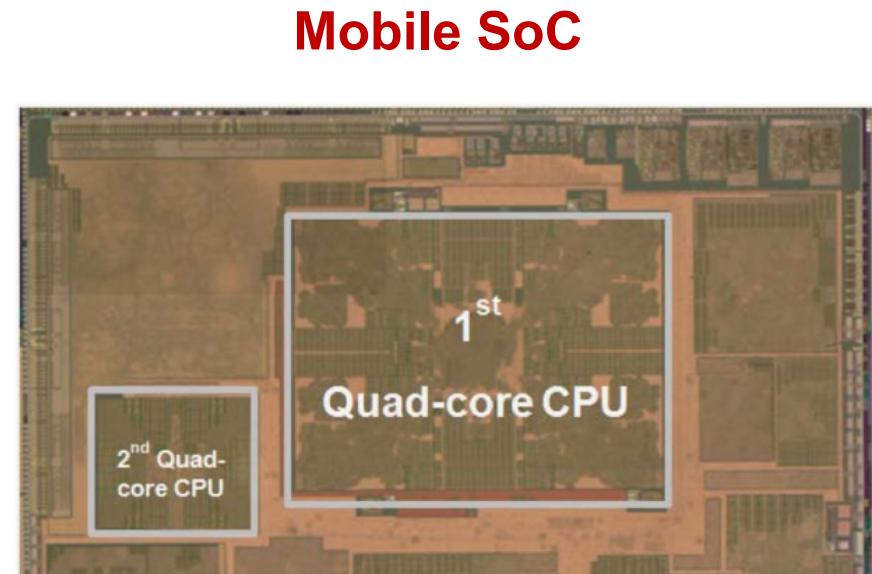
University of Michigan, Ann Arbor, MI

Motivation

- High speed SRAMs critical in μ Ps, SoCs



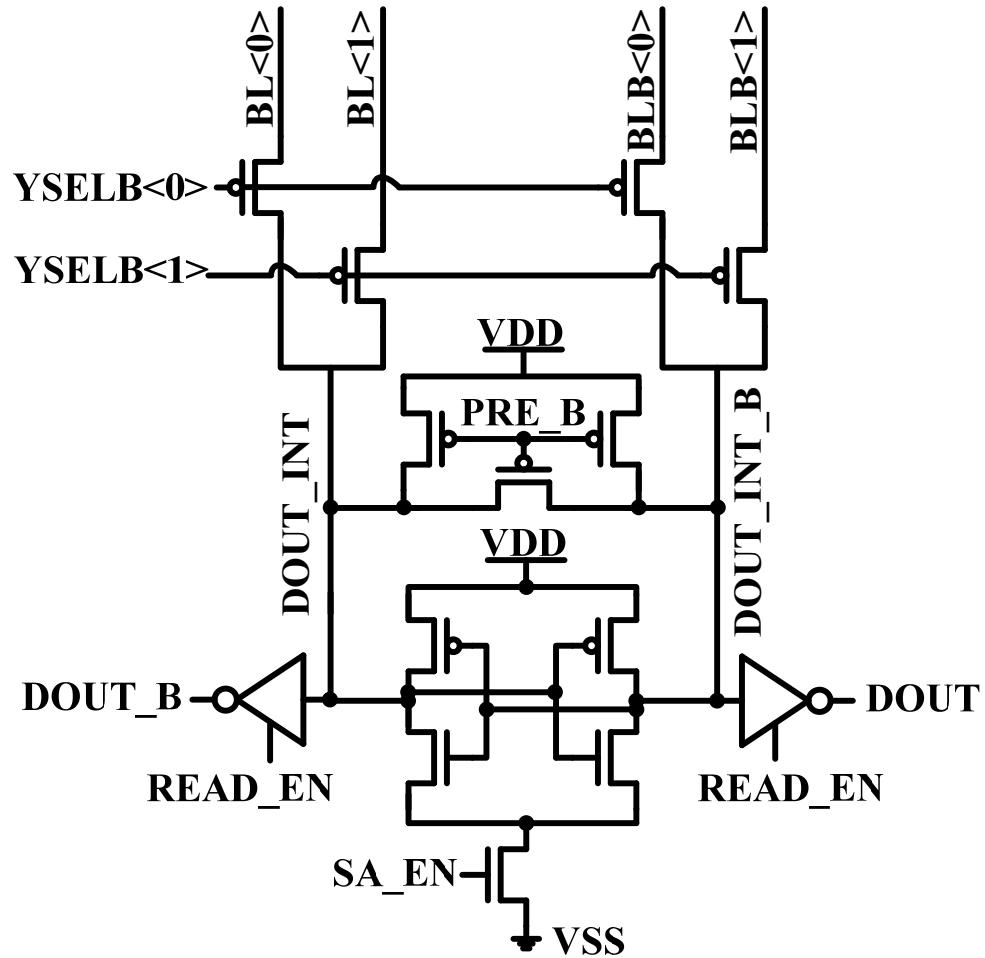
2.3GHz, 64Mb L3 SRAM in 32nm
D. Weiss et al., ISSCC'11



1.8GHz, 2Mb L2 SRAM in 28nm
Y. Shin et al., ISSCC'13

Motivation

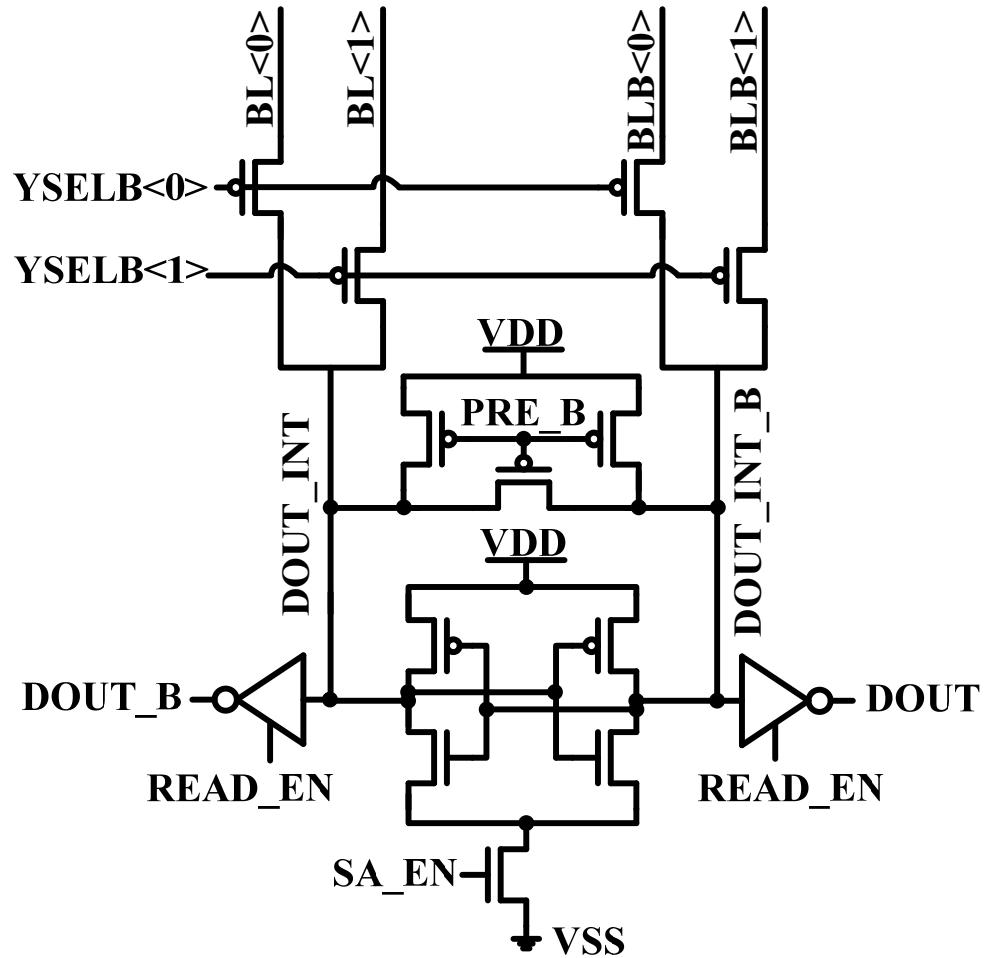
- High speed SRAMs critical in μPs, SoCs
 - Bitline sensing speed key component



Conventional SA

Motivation

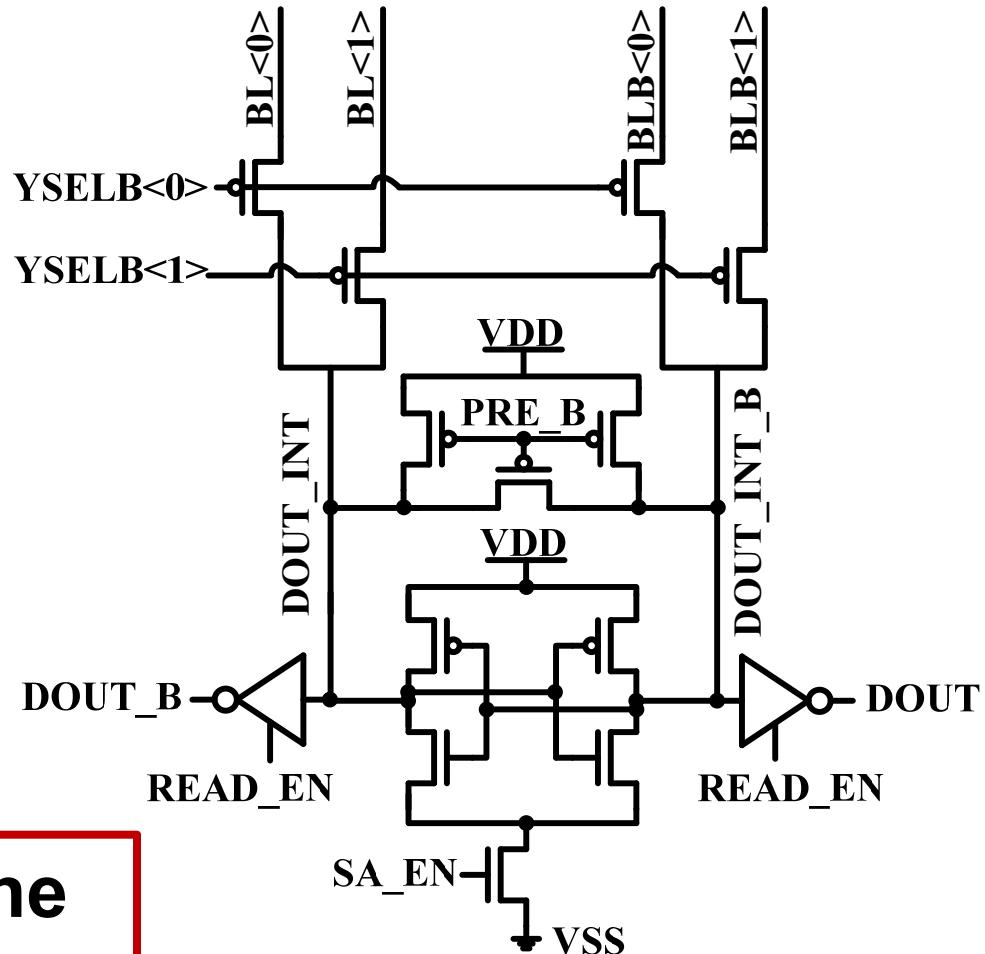
- High speed SRAMs critical in μPs, SoCs
 - Bitline sensing speed key component
 - Additional variation with scaling
 - Lower array efficiency, or
 - Read performance



Conventional SA

Motivation

- High speed SRAMs critical in μ Ps, SoCs
 - Bitline sensing speed key component
 - Additional variation with scaling
 - Lower array efficiency, or
 - Read performance

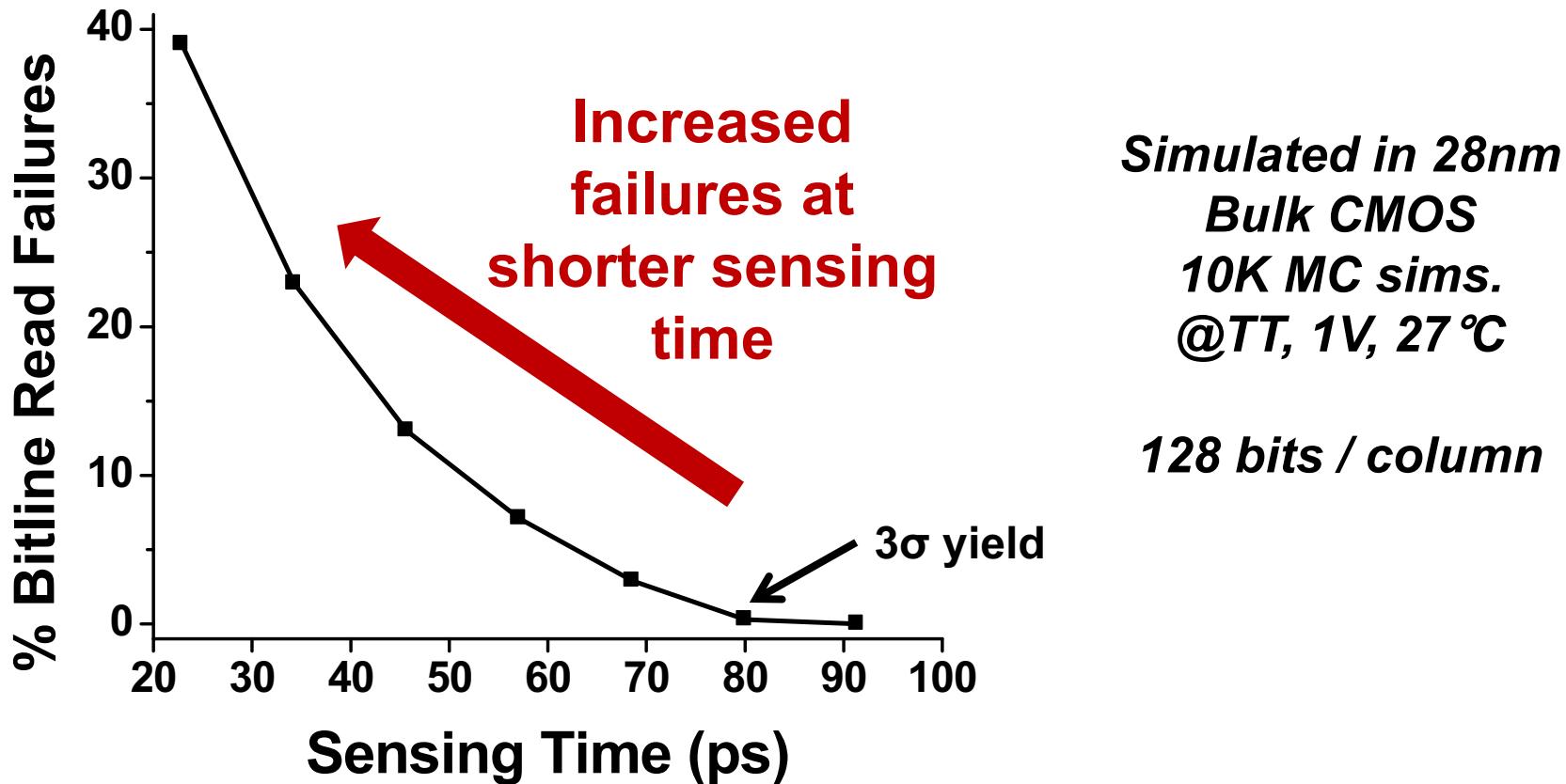


Conventional SA

Fast, area-efficient bitline
sensing is desired

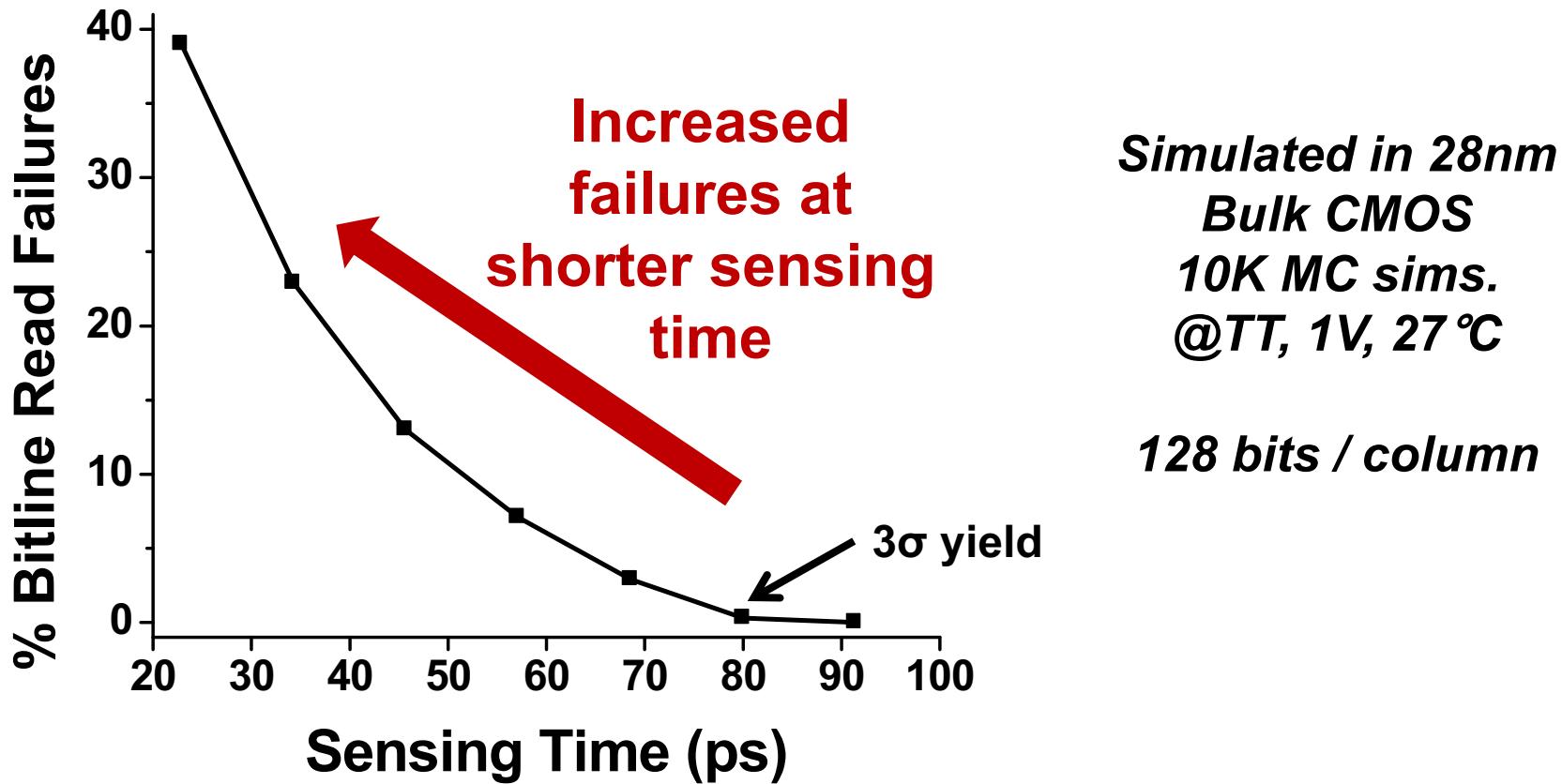
Motivation

- Variation increases bitline read failures
 - Requires margining by increasing sensing time



Motivation

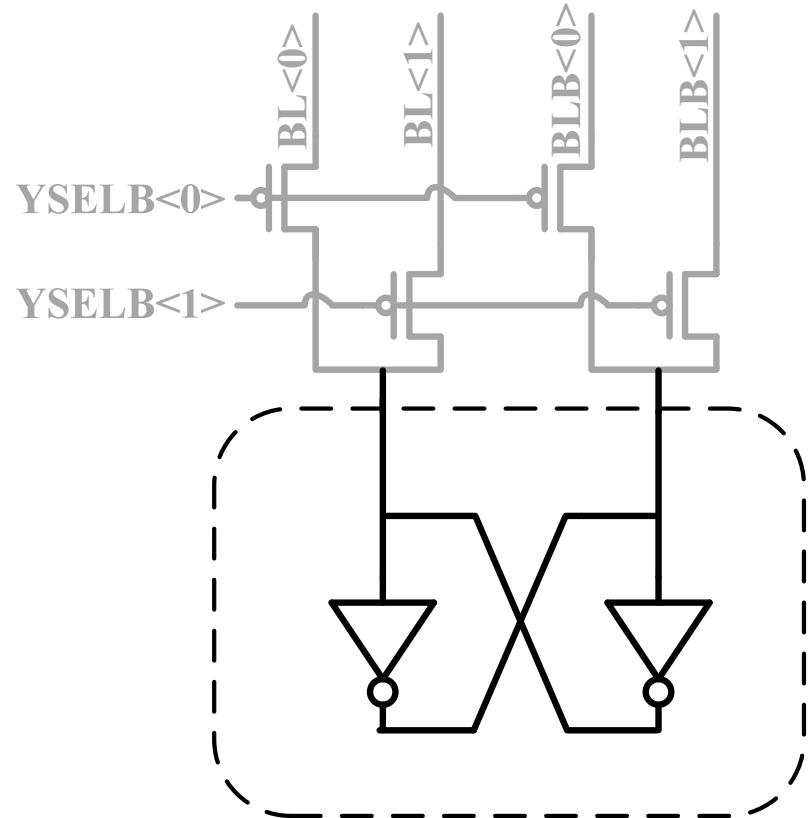
- Variation increases bitline read failures
 - Requires margining by increasing sensing time



Recover robustness using SA reconfiguration

Proposed Concept - VTS

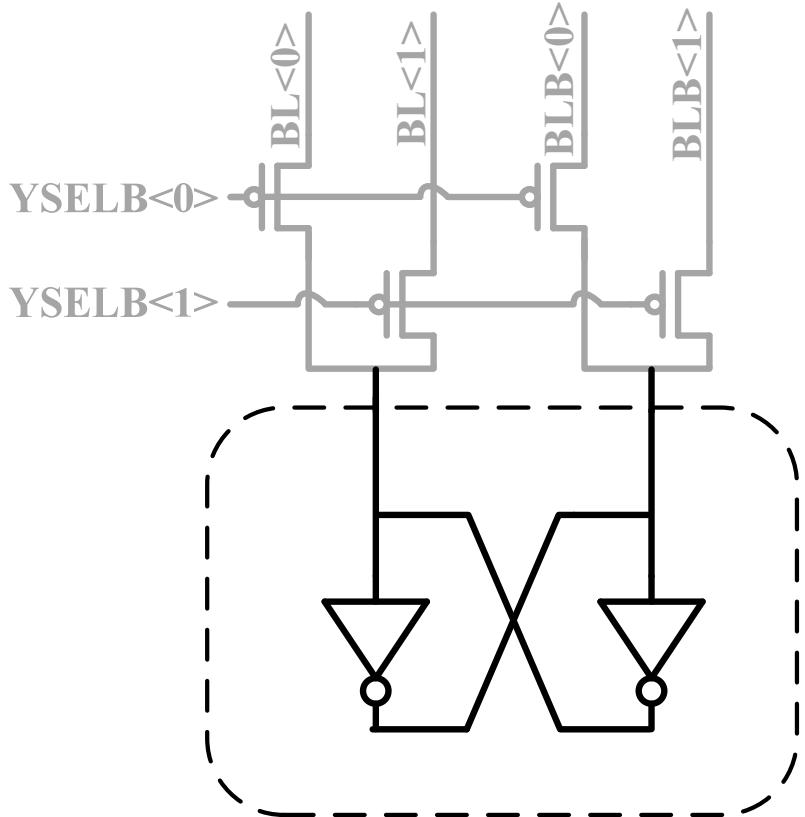
- Additional connections to switch SA between amp and latch configurations



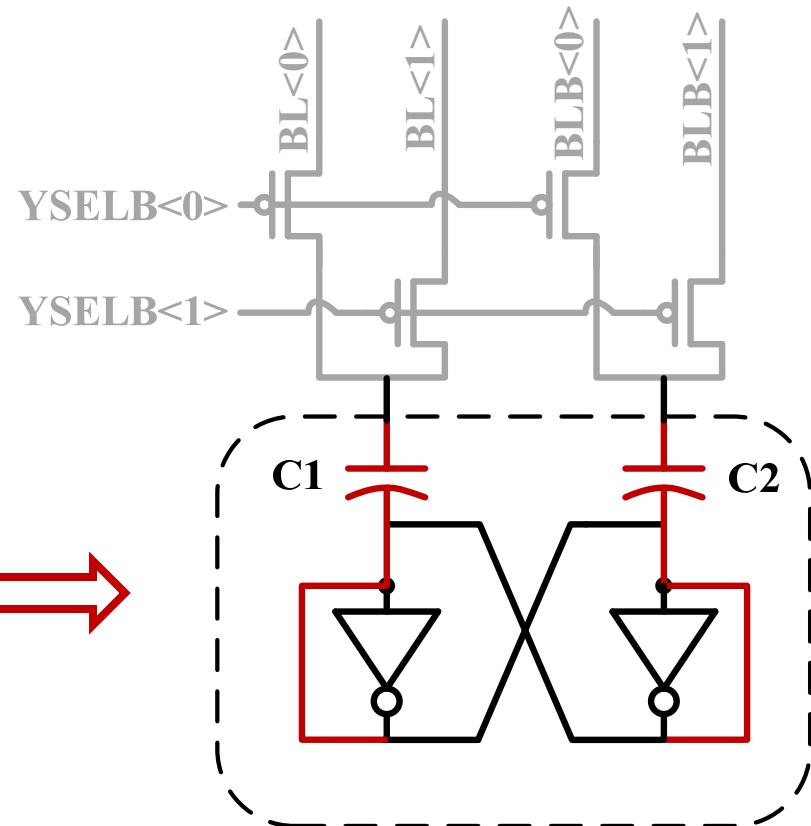
Conventional SA topology

Proposed Concept - VTS

- Additional connections to switch SA between amp and latch configurations



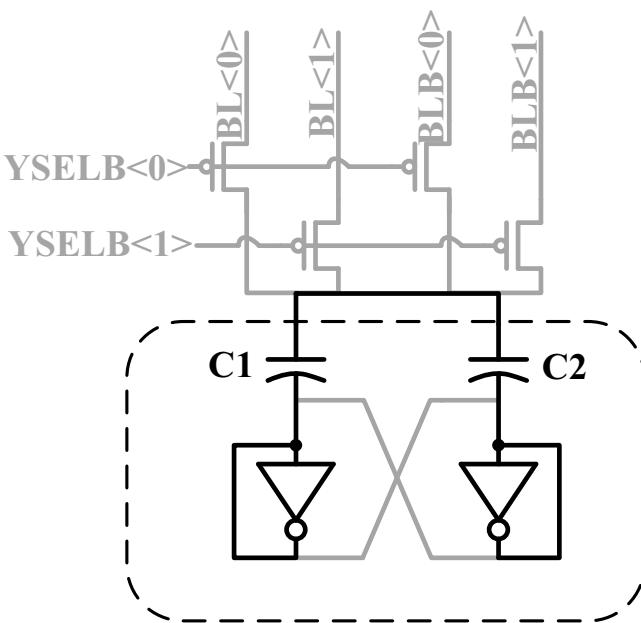
Conventional SA topology



VTS-SA topology

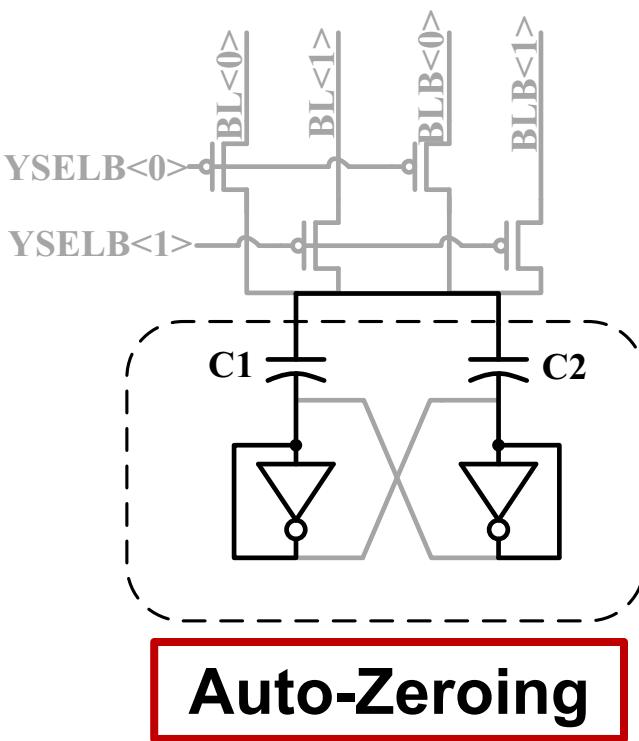
Proposed Concept - VTS

- Reconfigure inverter pair for VTS



Proposed Concept - VTS

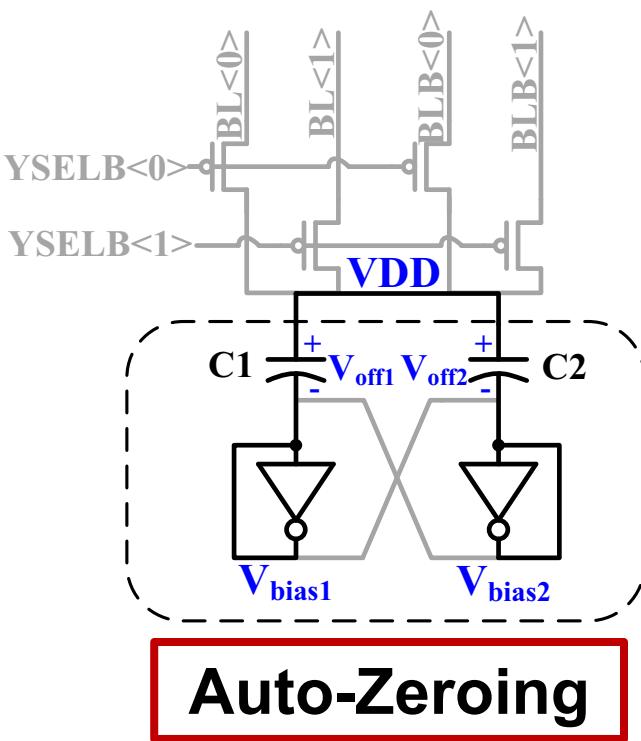
- Reconfigure inverter pair for VTS
 - Auto-zeroing



Auto-Zeroing

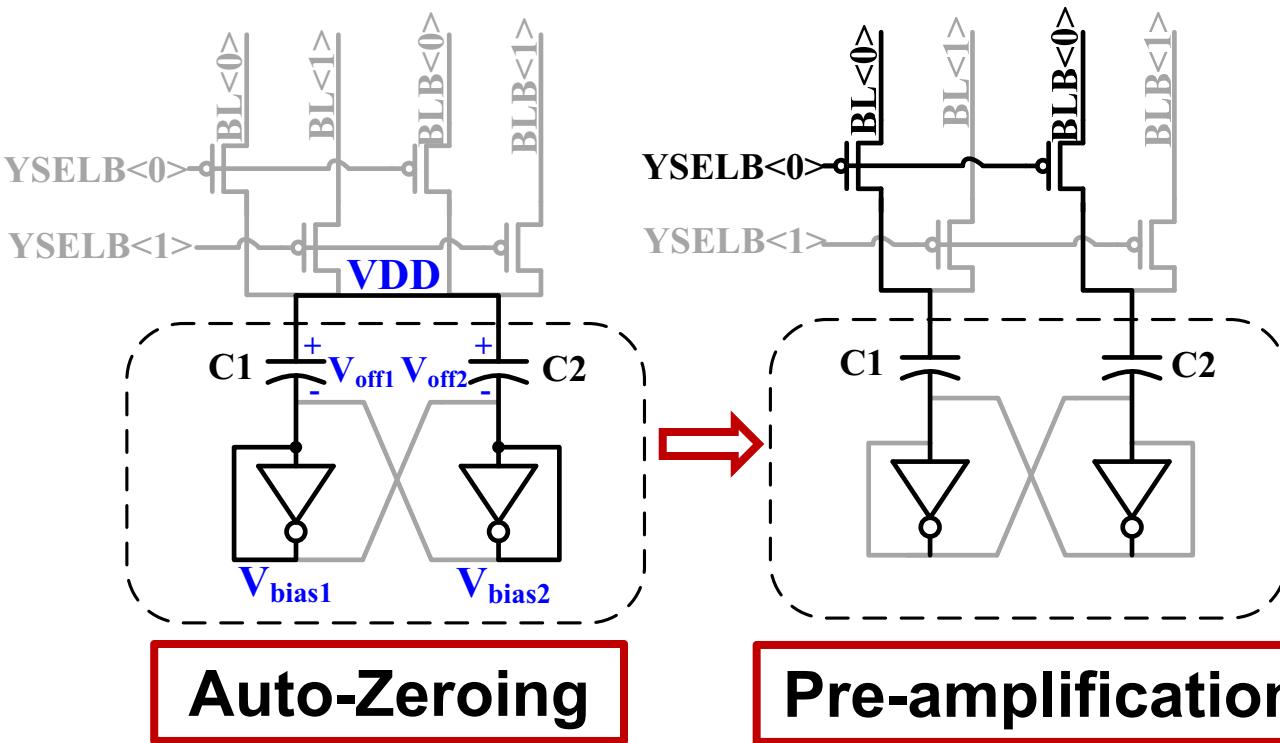
Proposed Concept - VTS

- Reconfigure inverter pair for VTS
 - Auto-zeroing



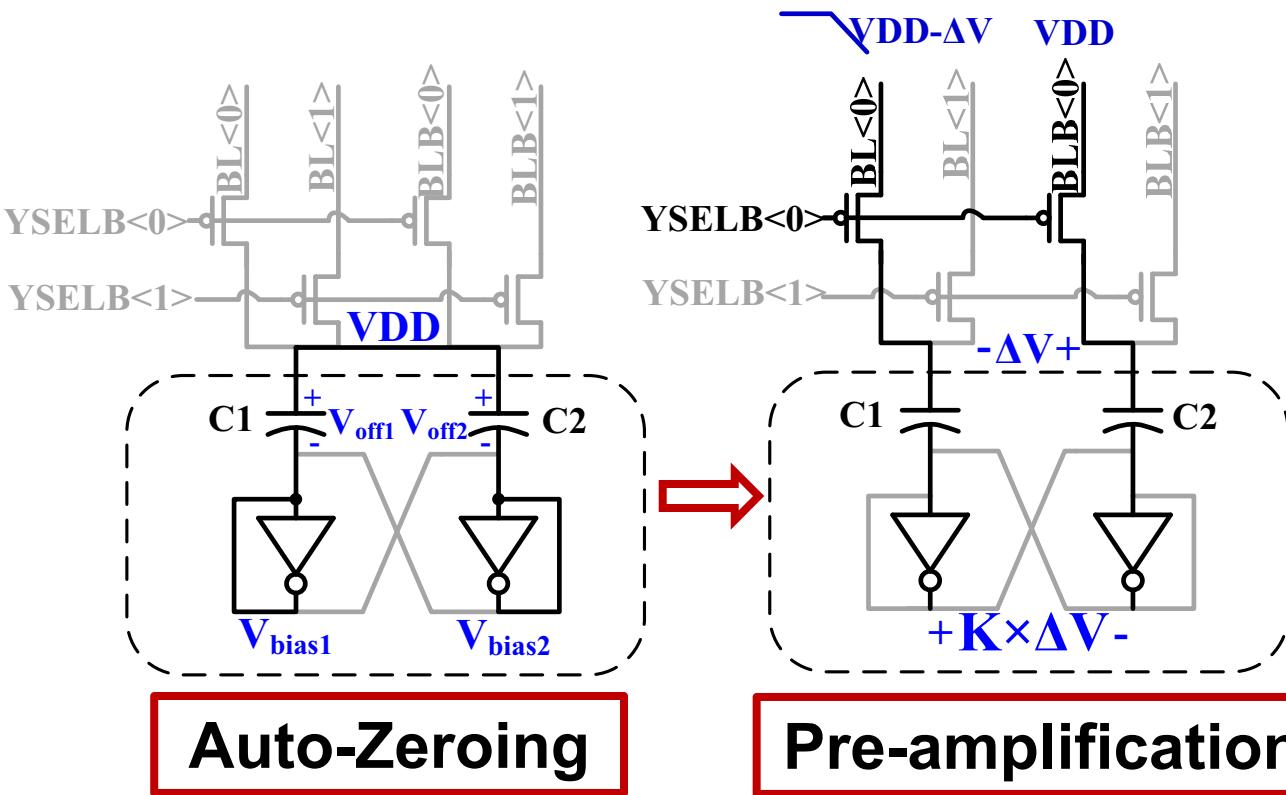
Proposed Concept - VTS

- Reconfigure inverter pair for VTS
 - Auto-zeroing
 - Pre-amplification of bitline differential



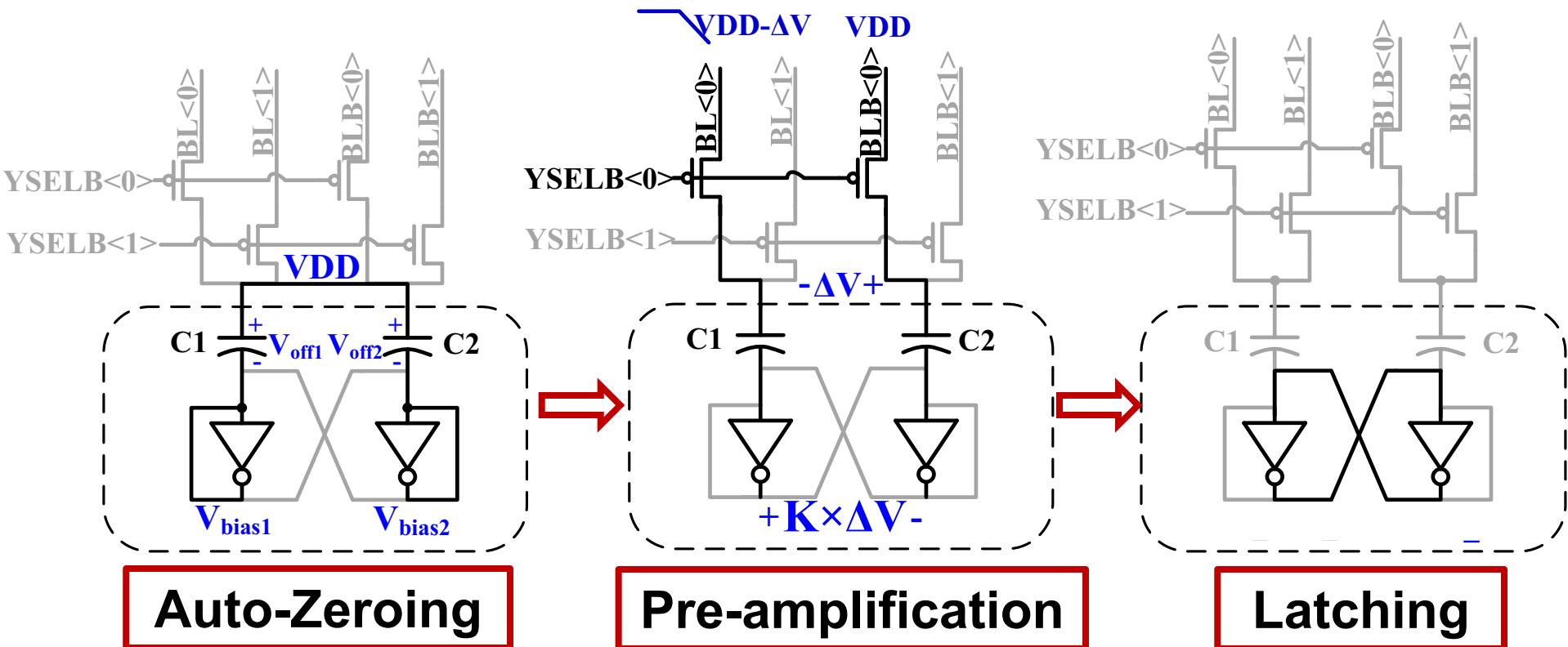
Proposed Concept - VTS

- Reconfigure inverter pair for VTS
 - Auto-zeroing
 - Pre-amplification of bitline differential



Proposed Concept - VTS

- Reconfigure inverter pair for VTS
 - Auto-zeroing
 - Pre-amplification of bitline differential
 - Latching



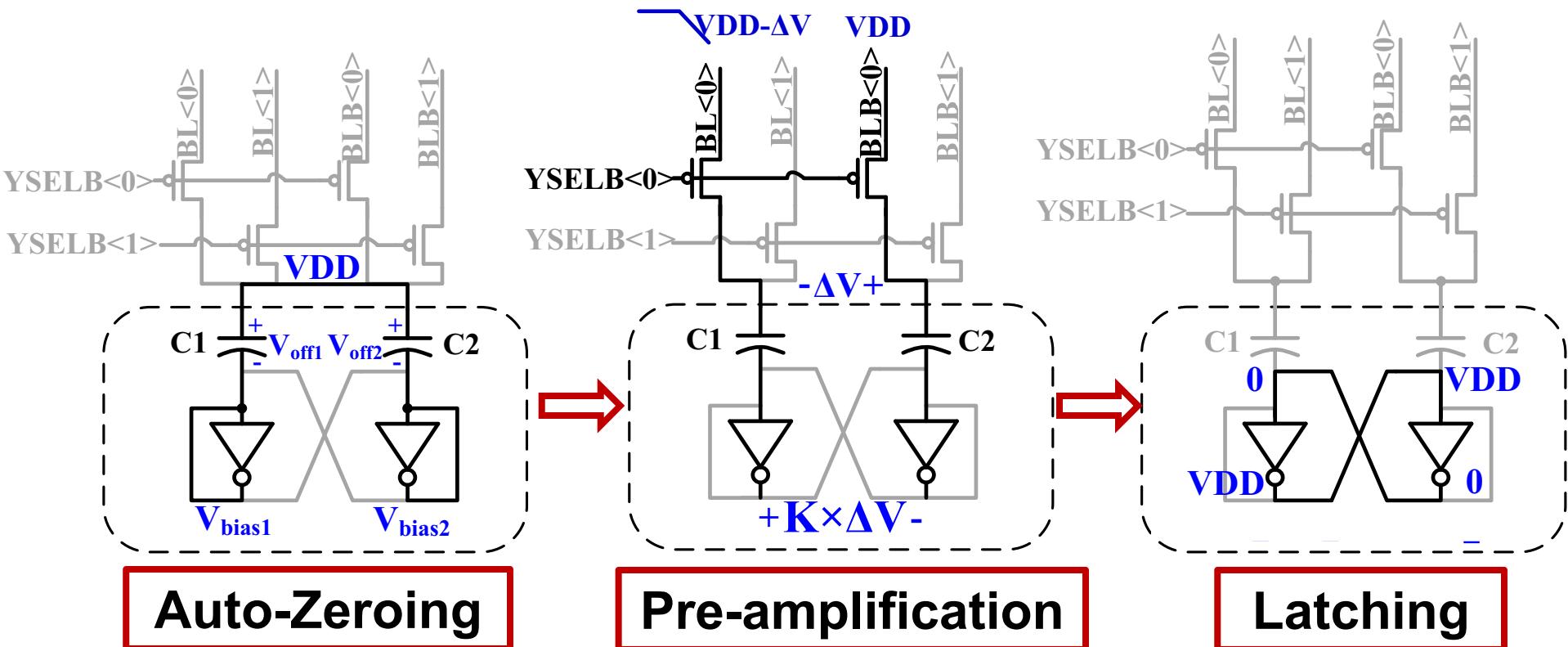
Auto-Zeroing

Pre-amplification

Latching

Proposed Concept - VTS

- Reconfigure inverter pair for VTS
 - Auto-zeroing
 - Pre-amplification of bitline differential
 - Latching

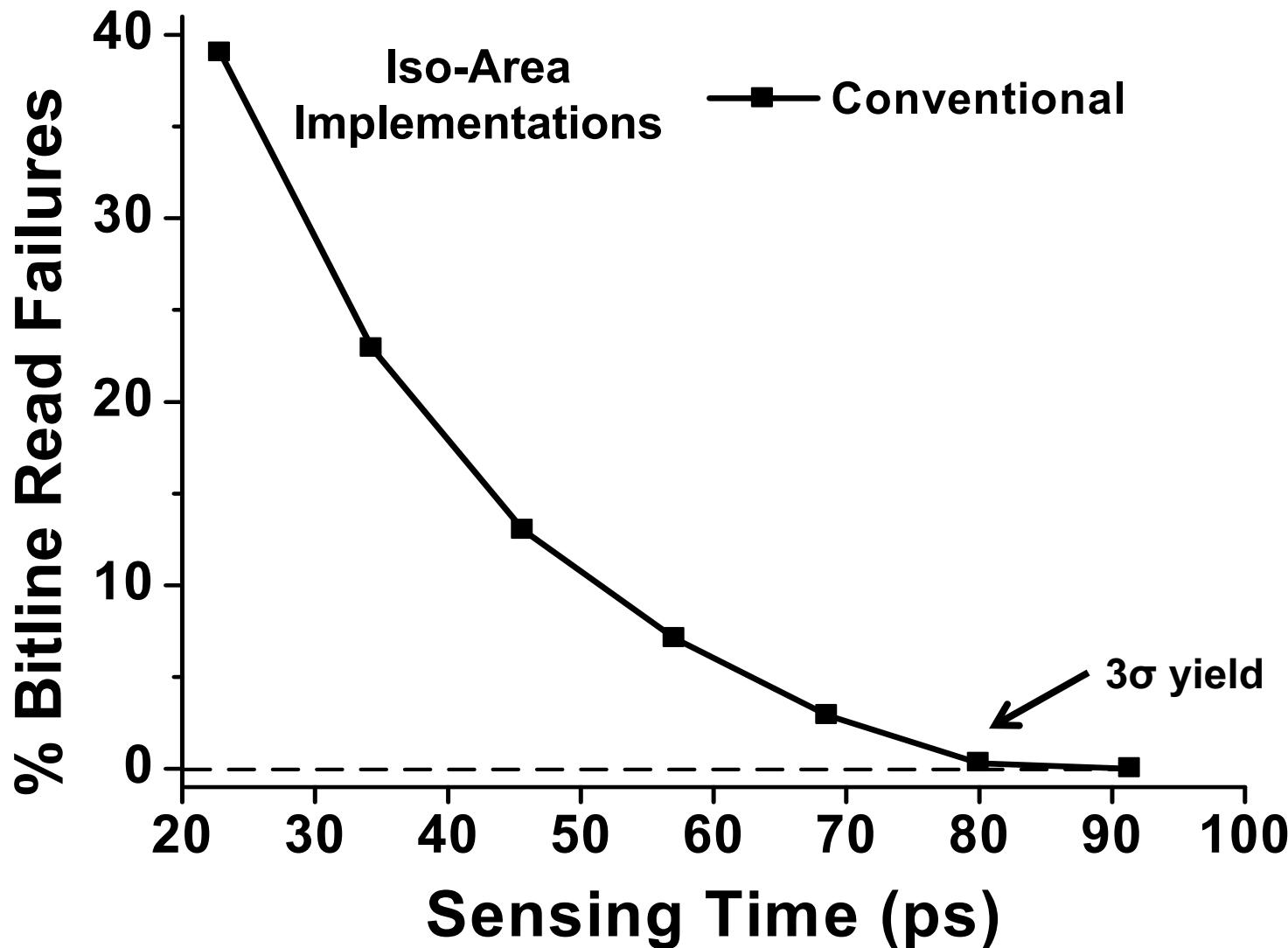


Auto-Zeroing

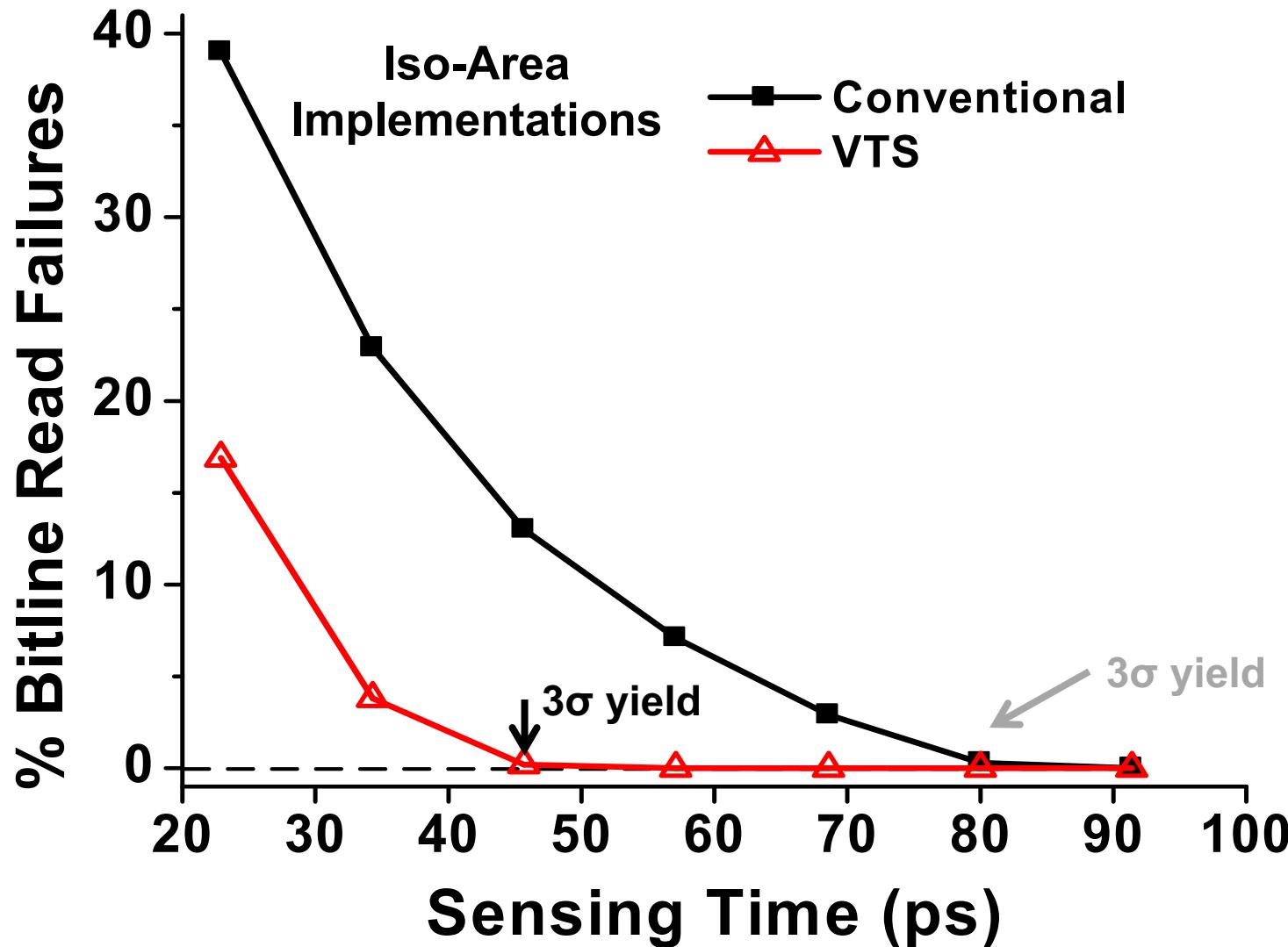
Pre-amplification

Latching

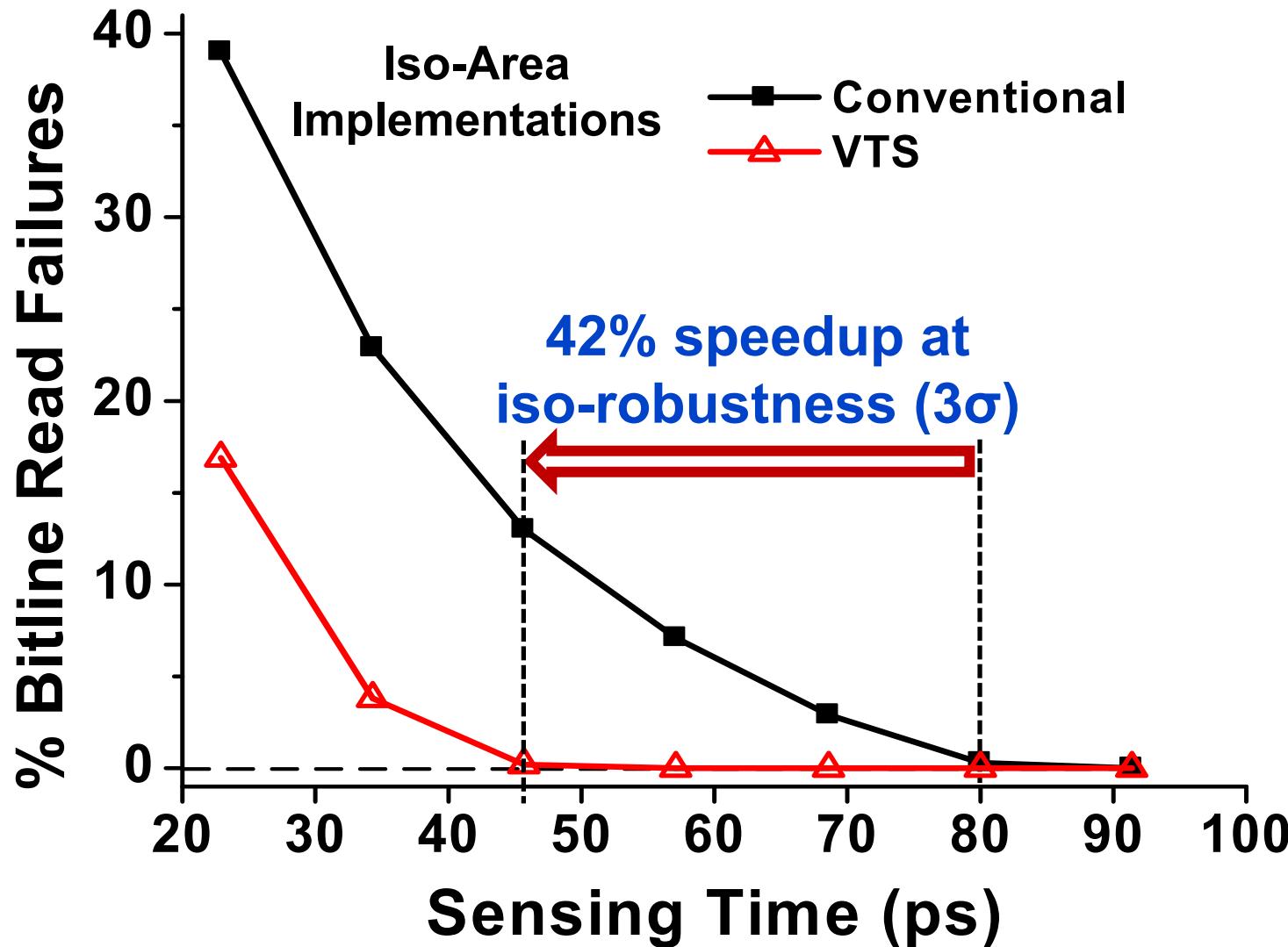
Simulated Improvement in 28nm CMOS



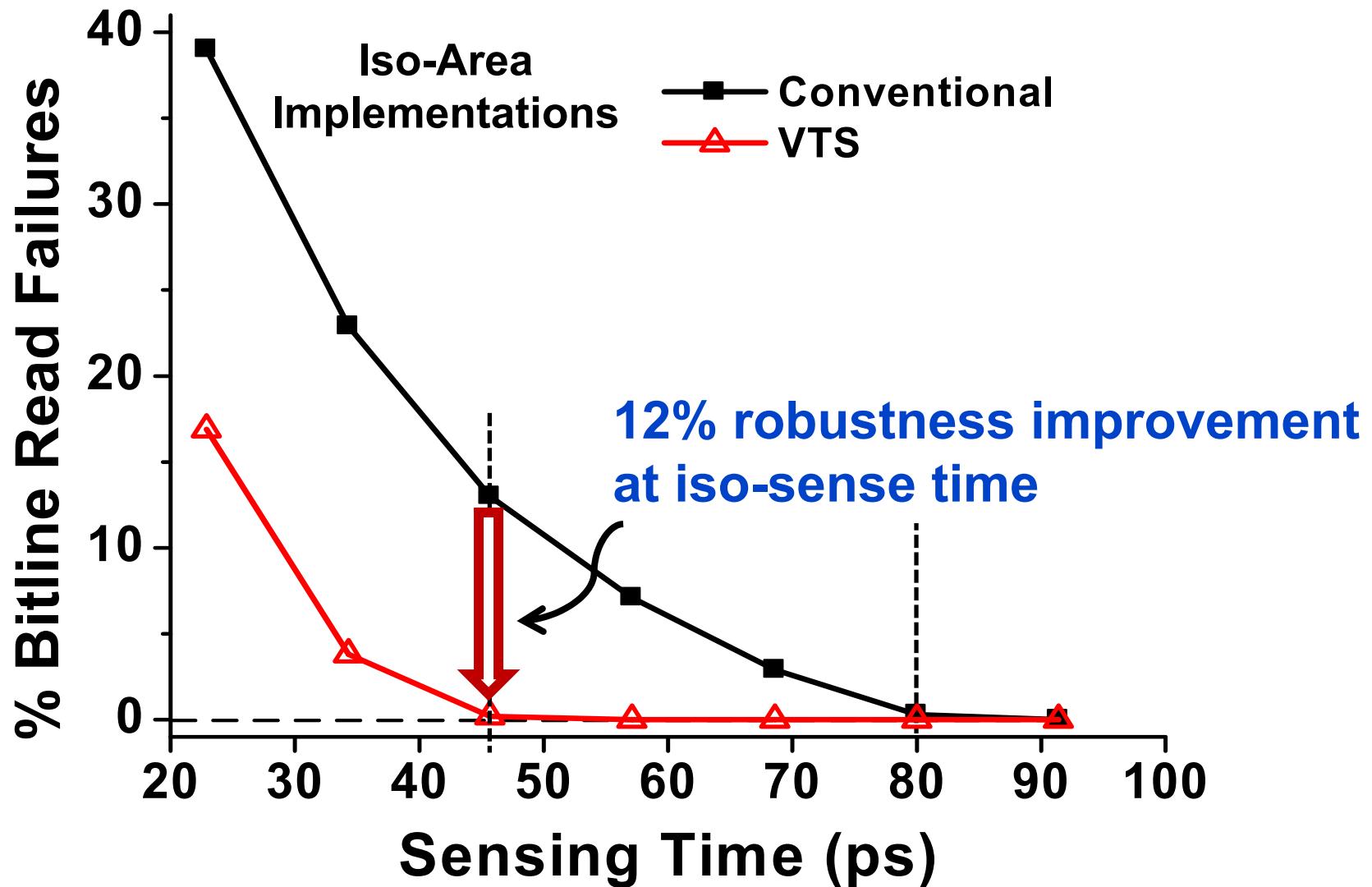
Simulated Improvement in 28nm CMOS



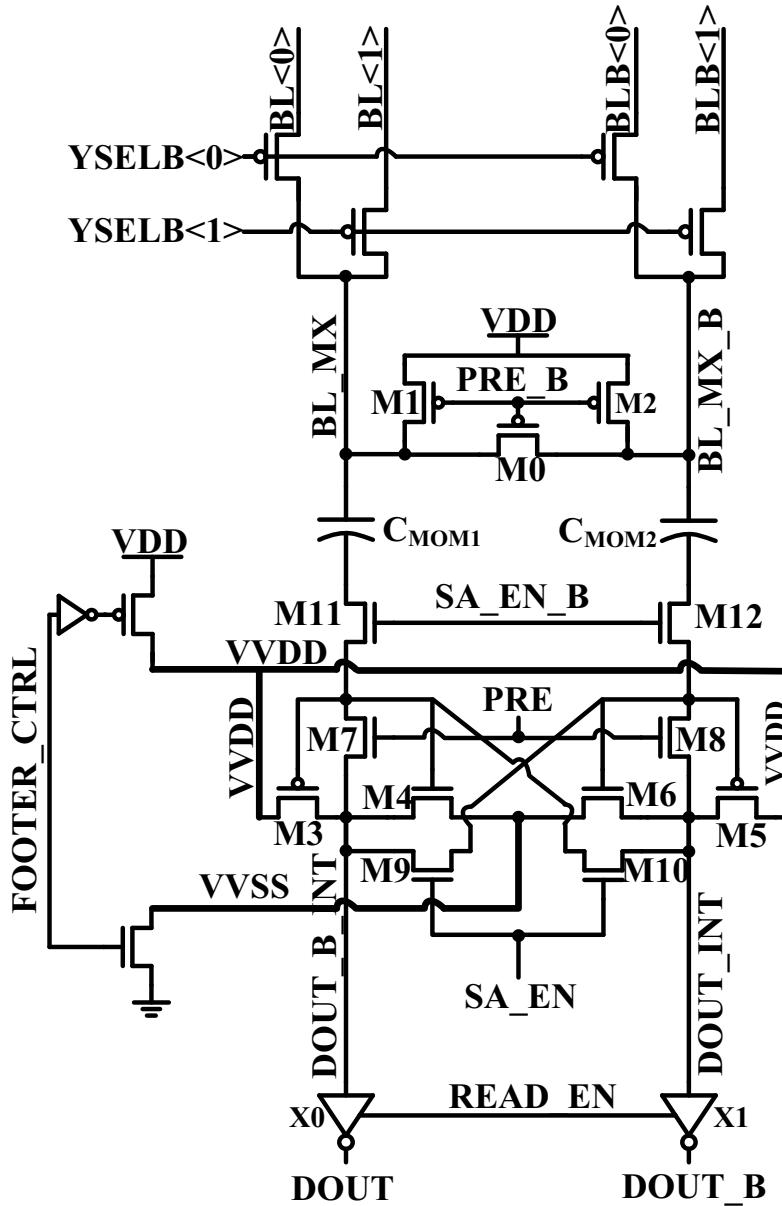
Simulated Improvement in 28nm CMOS



Simulated Improvement in 28nm CMOS



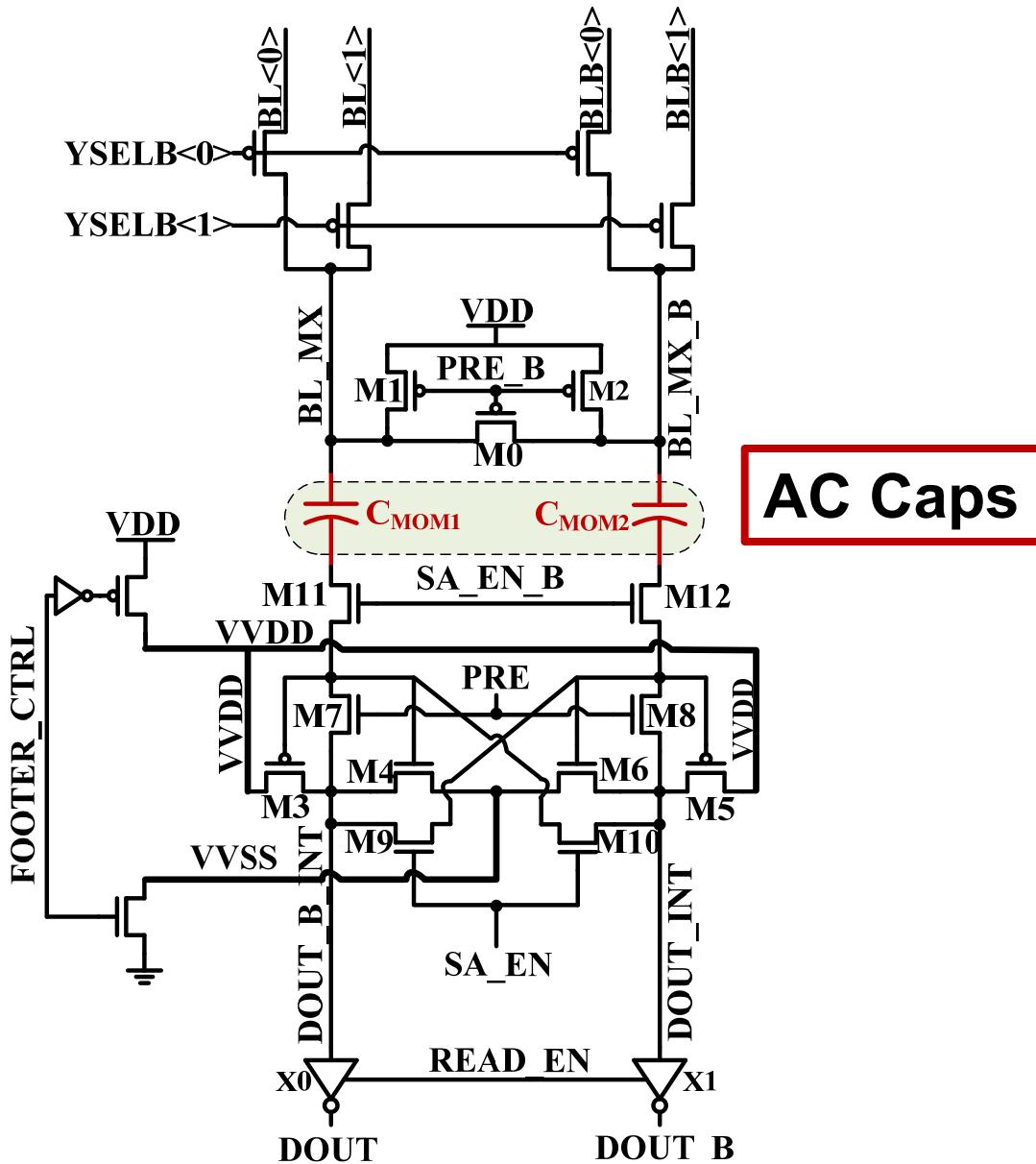
VTS Circuit Schematic



13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

VTS Circuit Schematic

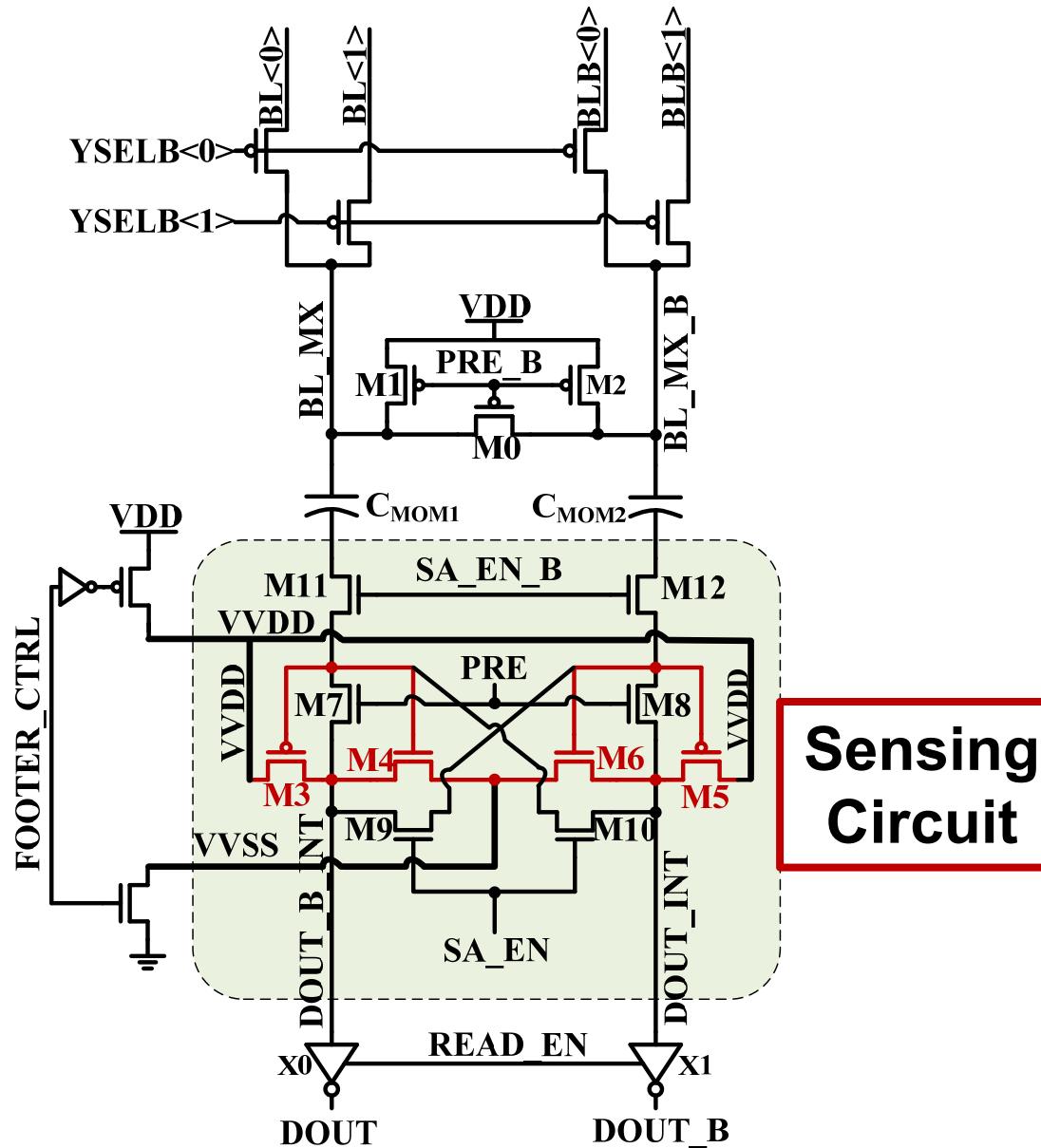
- C_{MOM1}, C_{MOM2}
 - MOM caps



13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

VTS Circuit Schematic

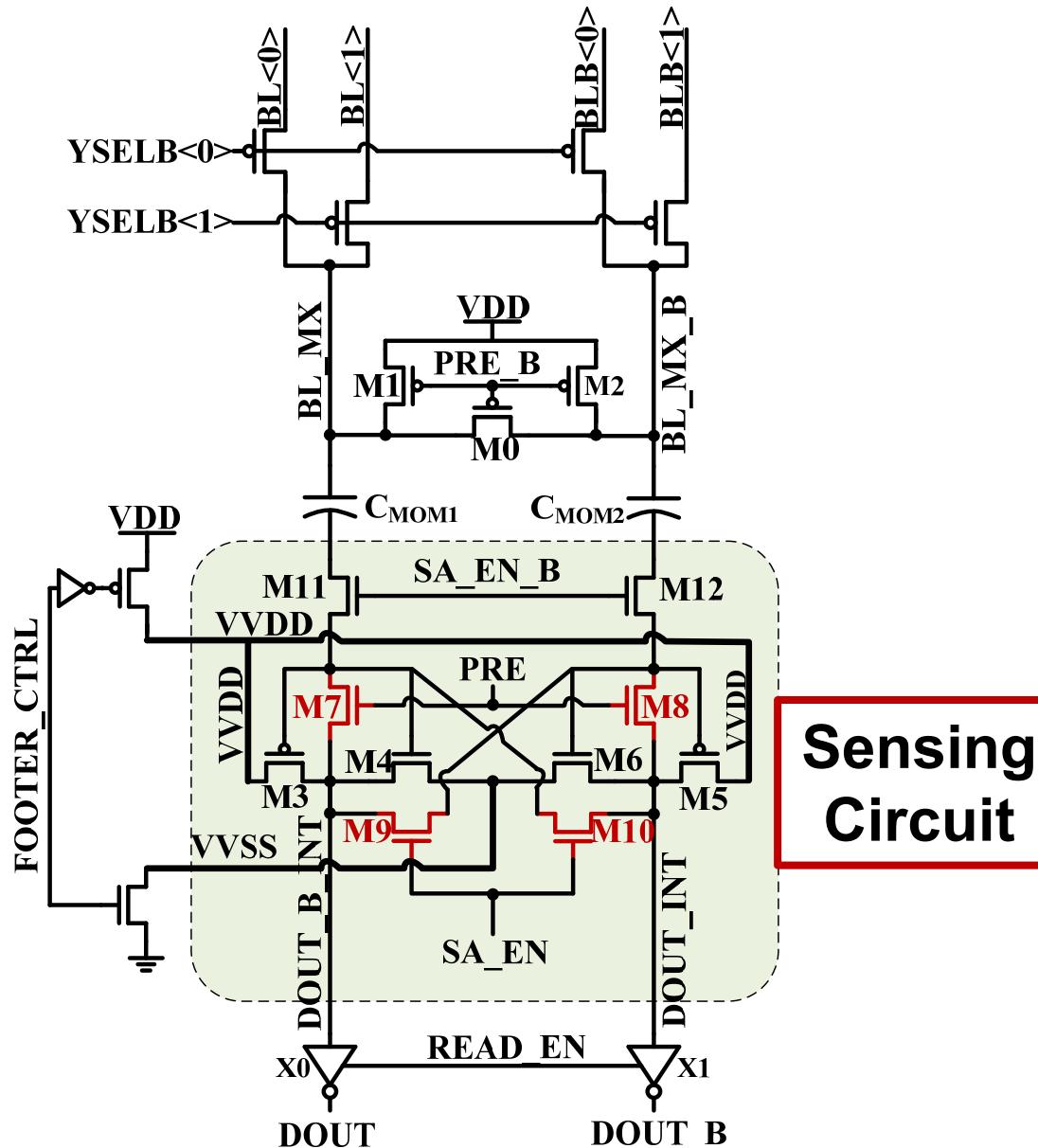
- C_{MOM1}, C_{MOM2}
 - MOM caps
- M3-M4, M5-M6
 - SA inverters



**Sensing
Circuit**

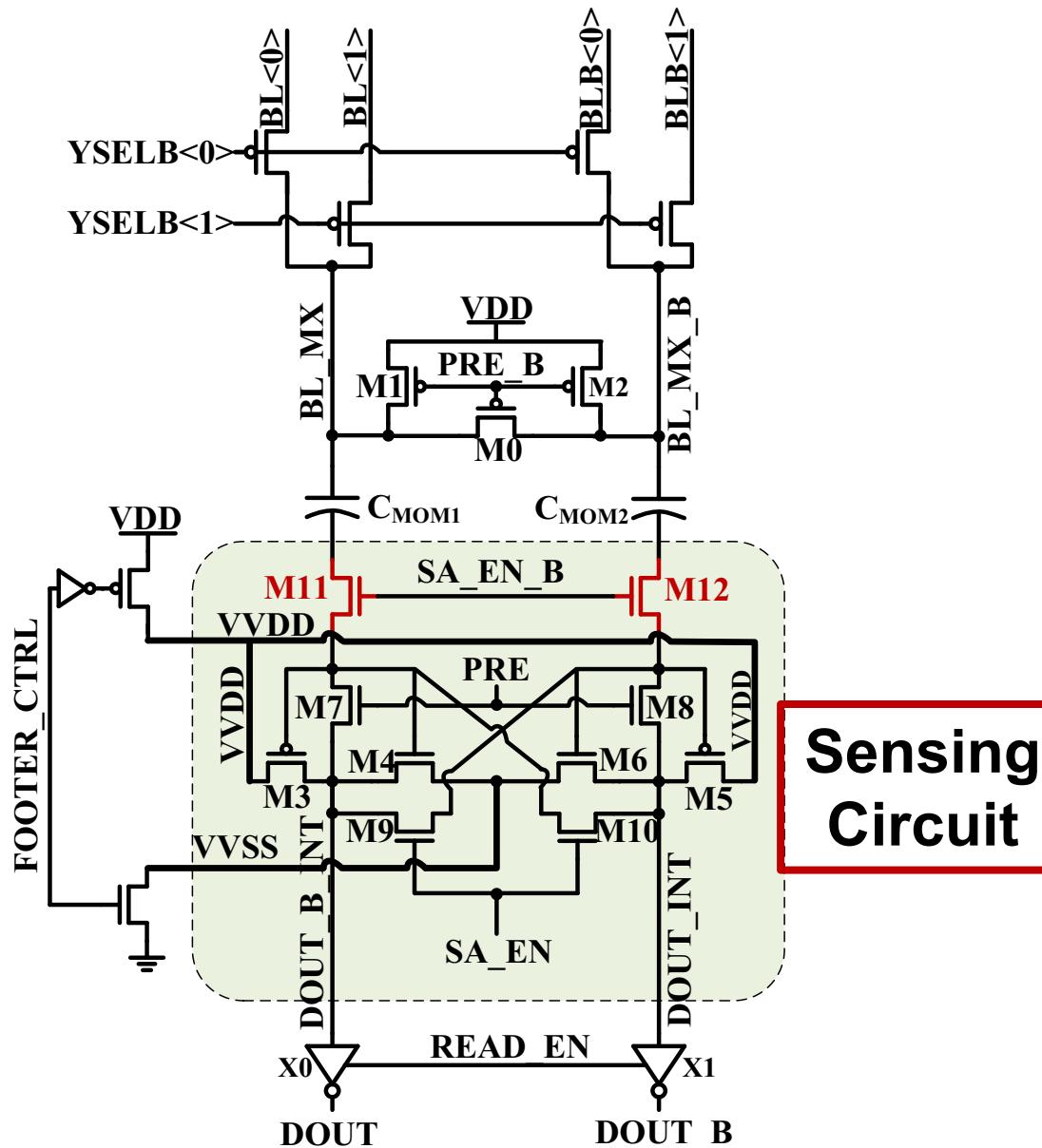
VTS Circuit Schematic

- C_{MOM1}, C_{MOM2}
 - MOM caps
- M3-M4, M5-M6
 - SA inverters
- M7-10
 - Reconfig. switches



VTS Circuit Schematic

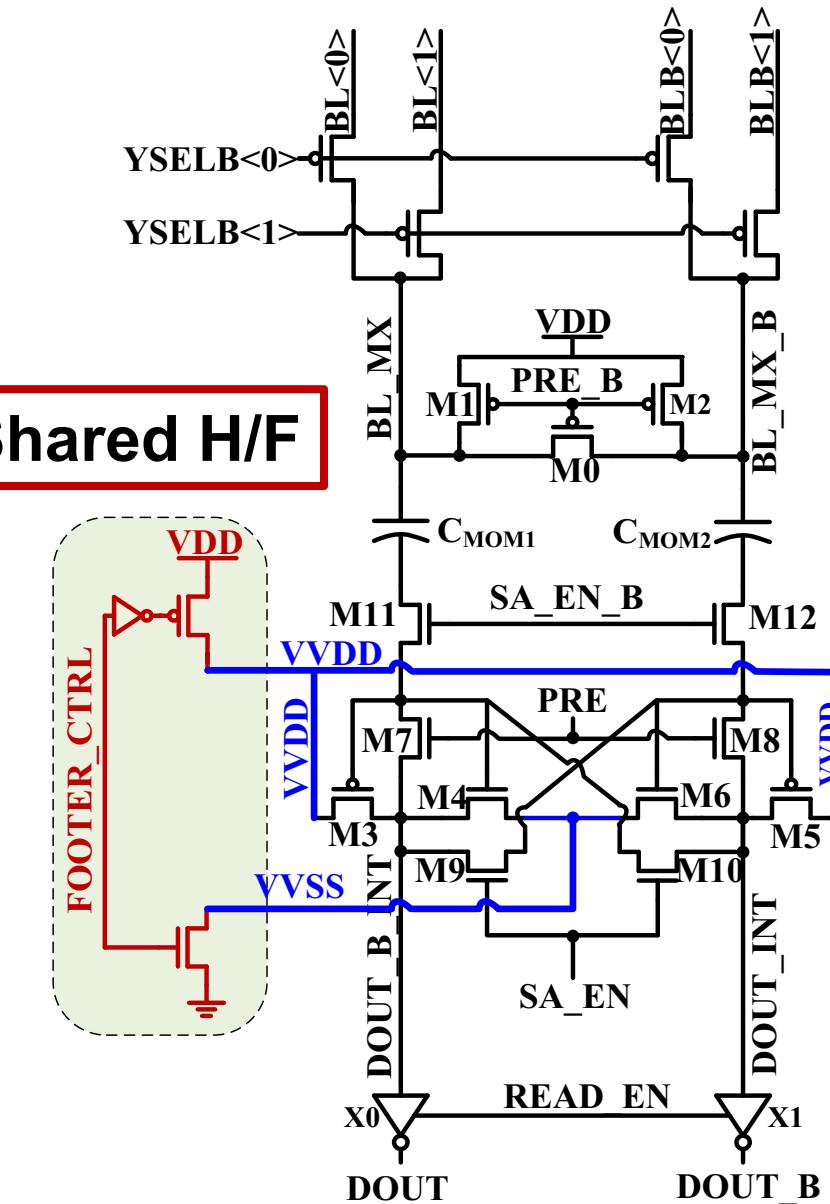
- C_{MOM1}, C_{MOM2}
 - MOM caps
- M3-M4, M5-M6
 - SA inverters
- M7-10
 - Reconfig. switches
- M11, M12
 - Cap isolation



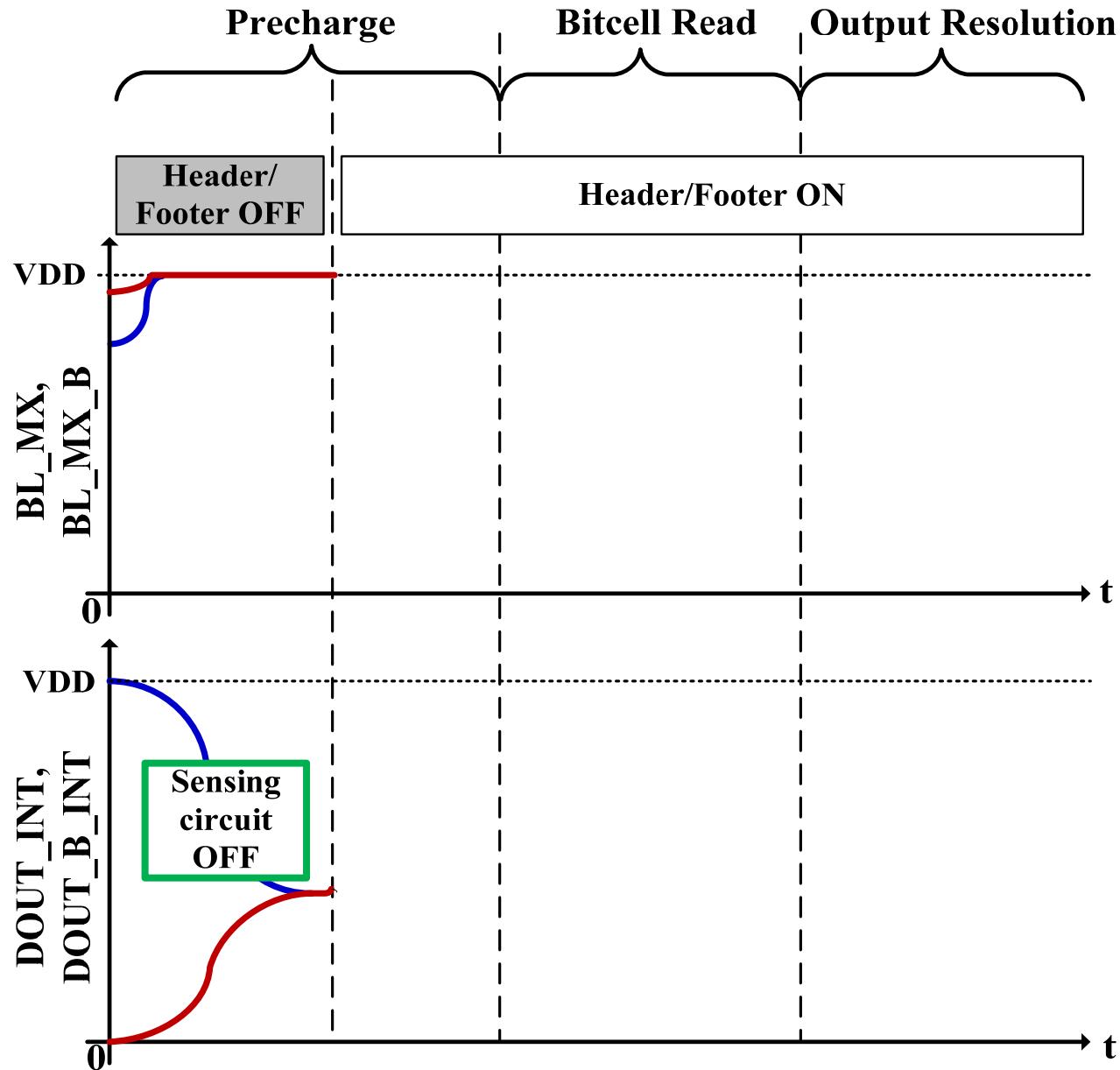
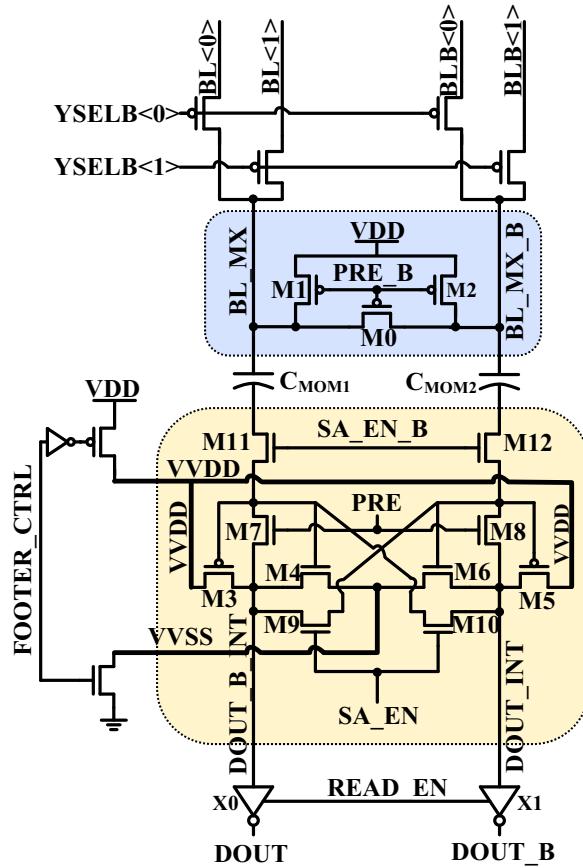
Sensing
Circuit

VTS Circuit Schematic

- C_{MOM1}, C_{MOM2}
 - MOM caps
- M3-M4, M5-M6
 - SA inverters
- M7-10
 - Reconfig. switches
- M11, M12
 - Cap isolation
- Header/footer shared across 16 SAs
 - Reduces power by 26% (*measured*)

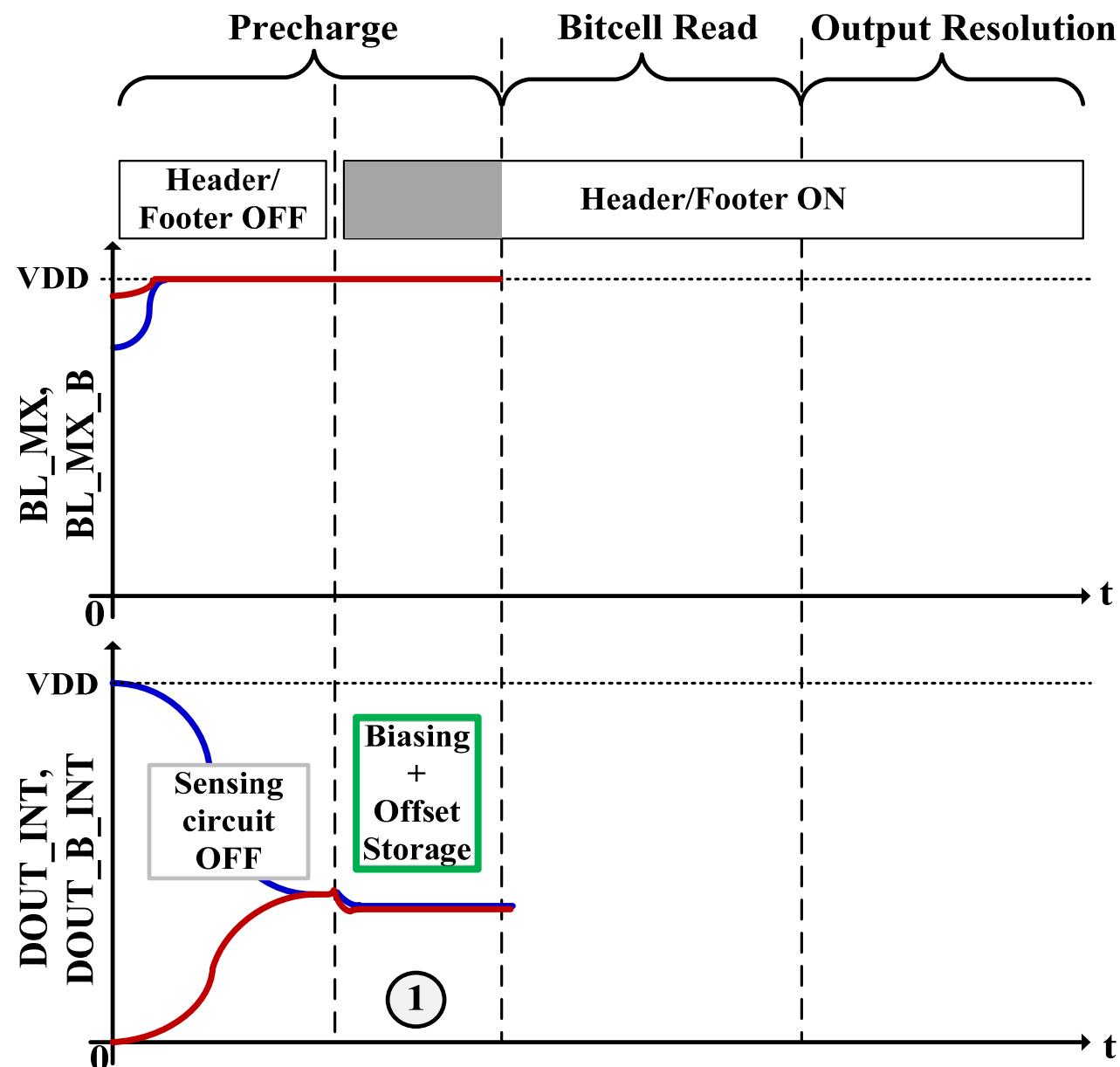
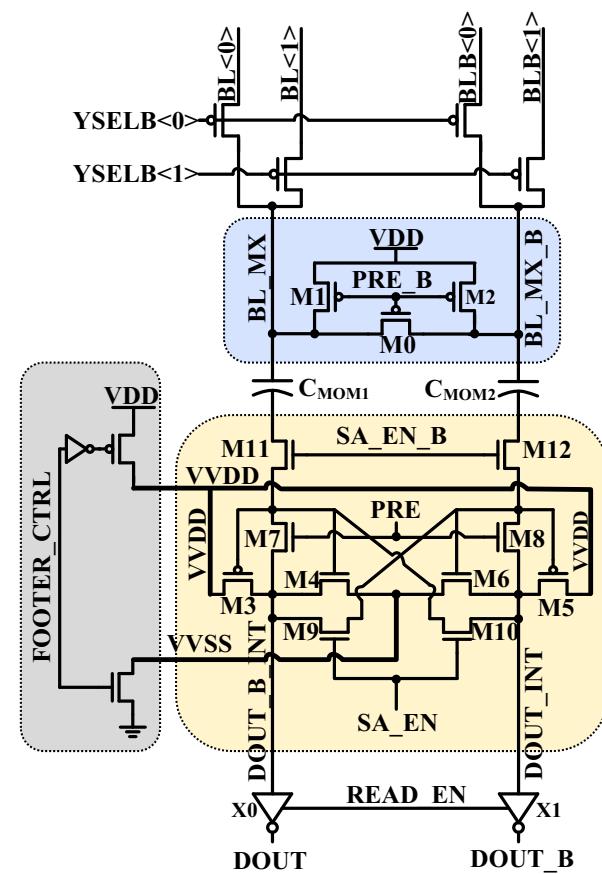


VTS Operation Phases



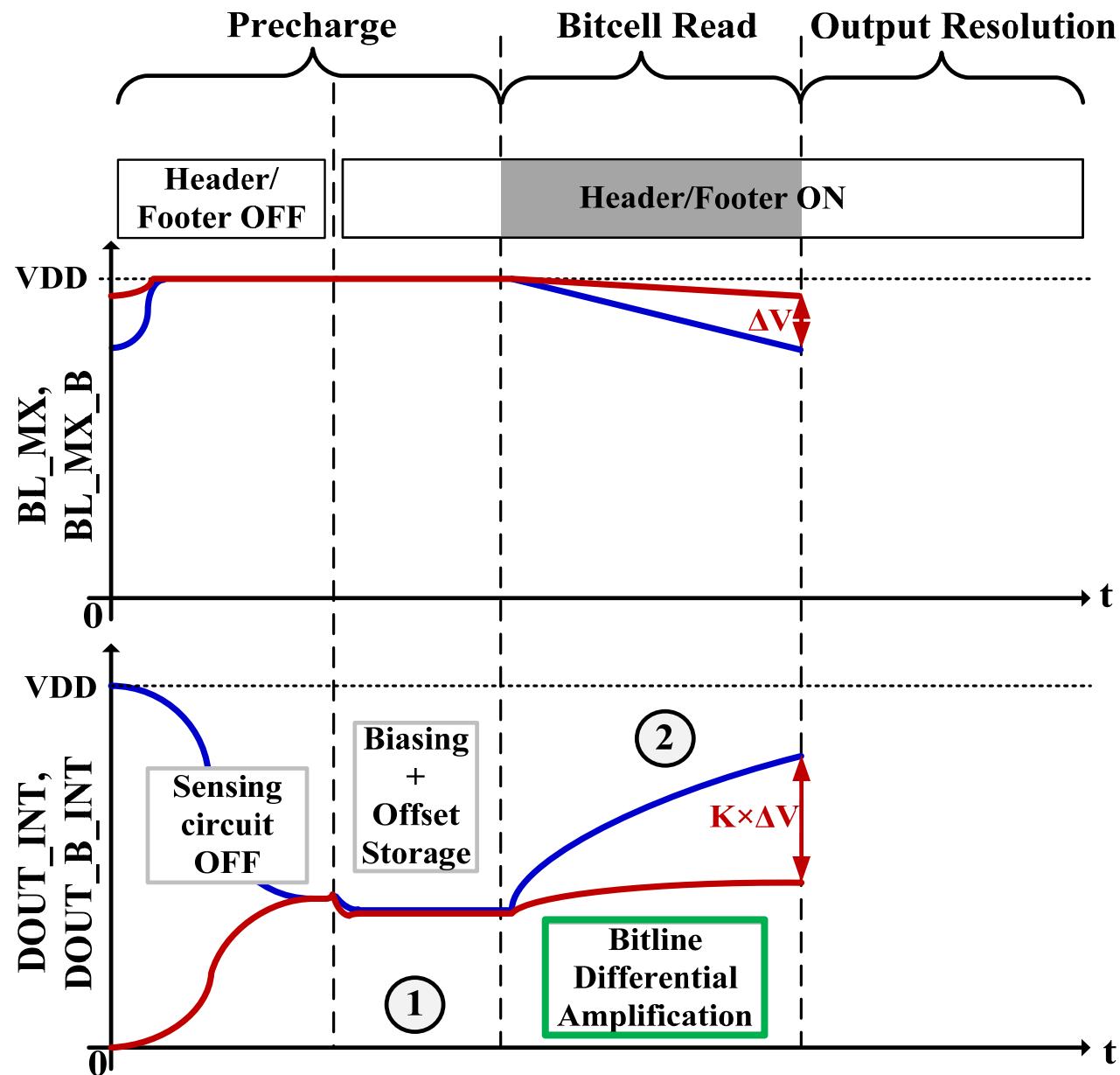
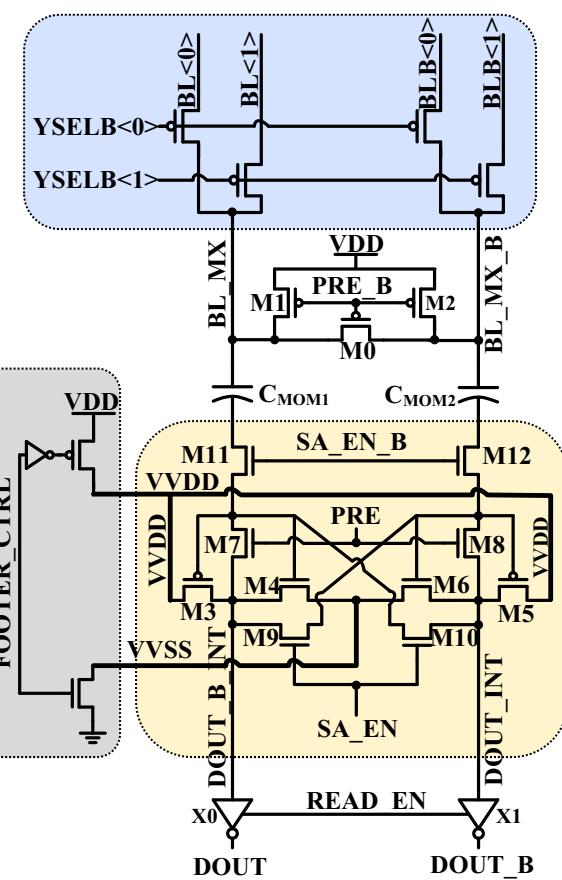
13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

VTS Operation Phases



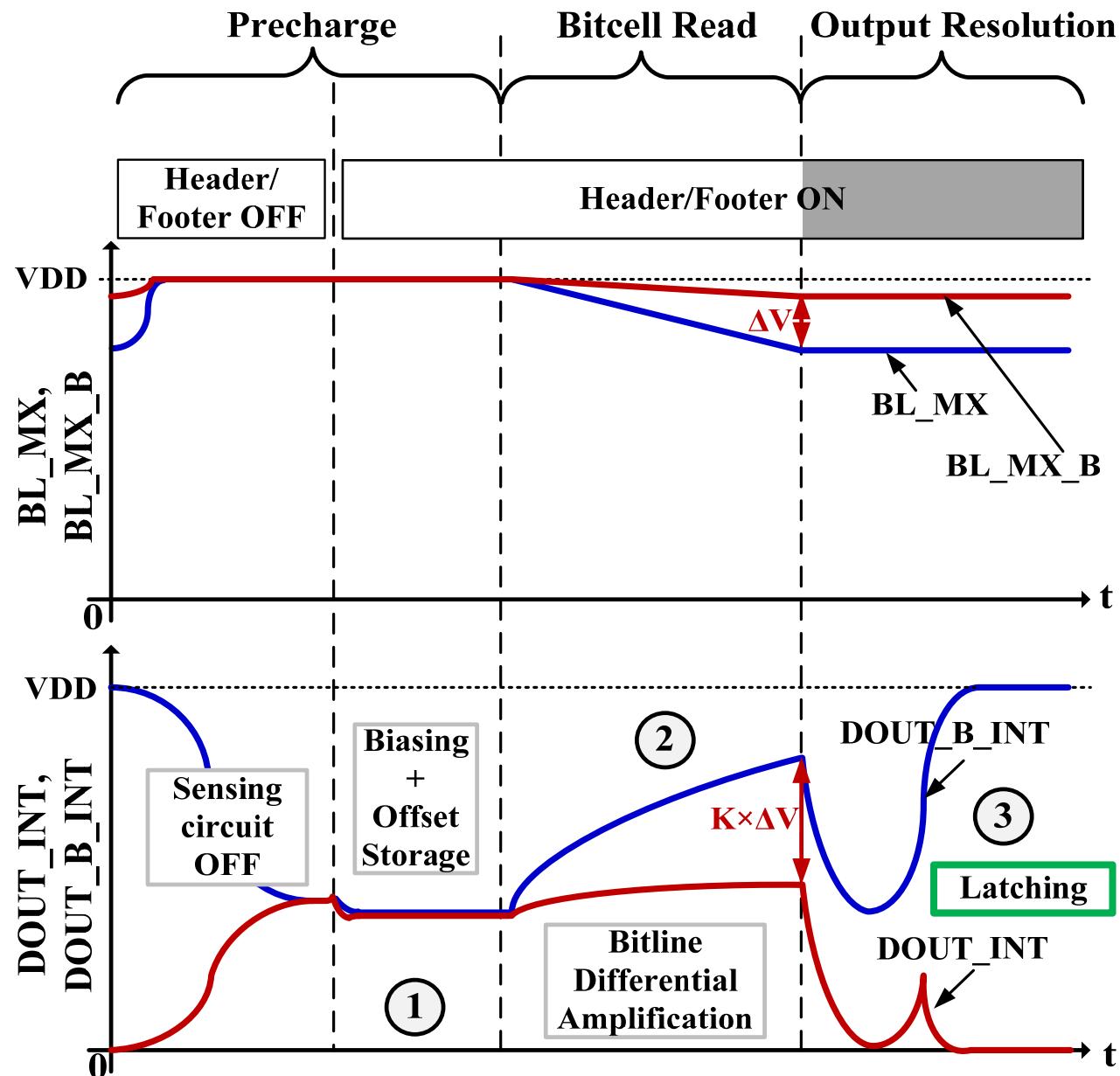
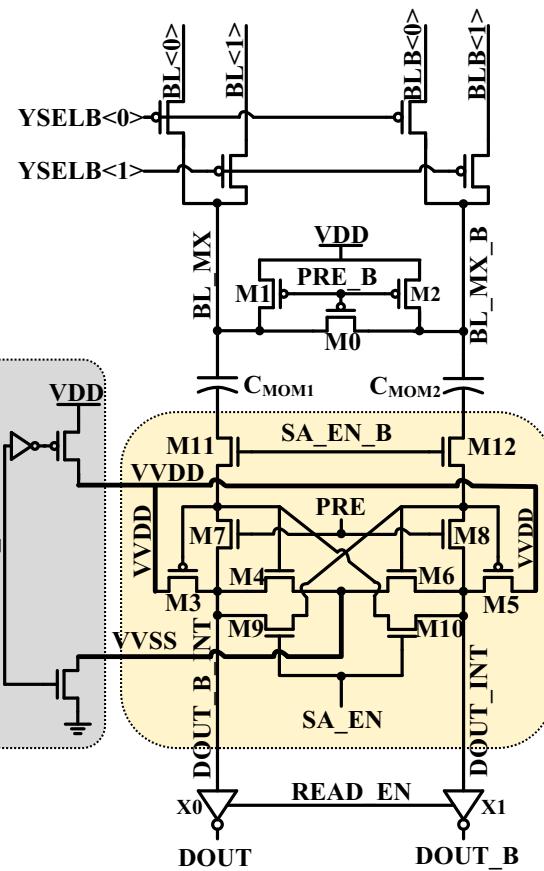
13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

VTS Operation Phases



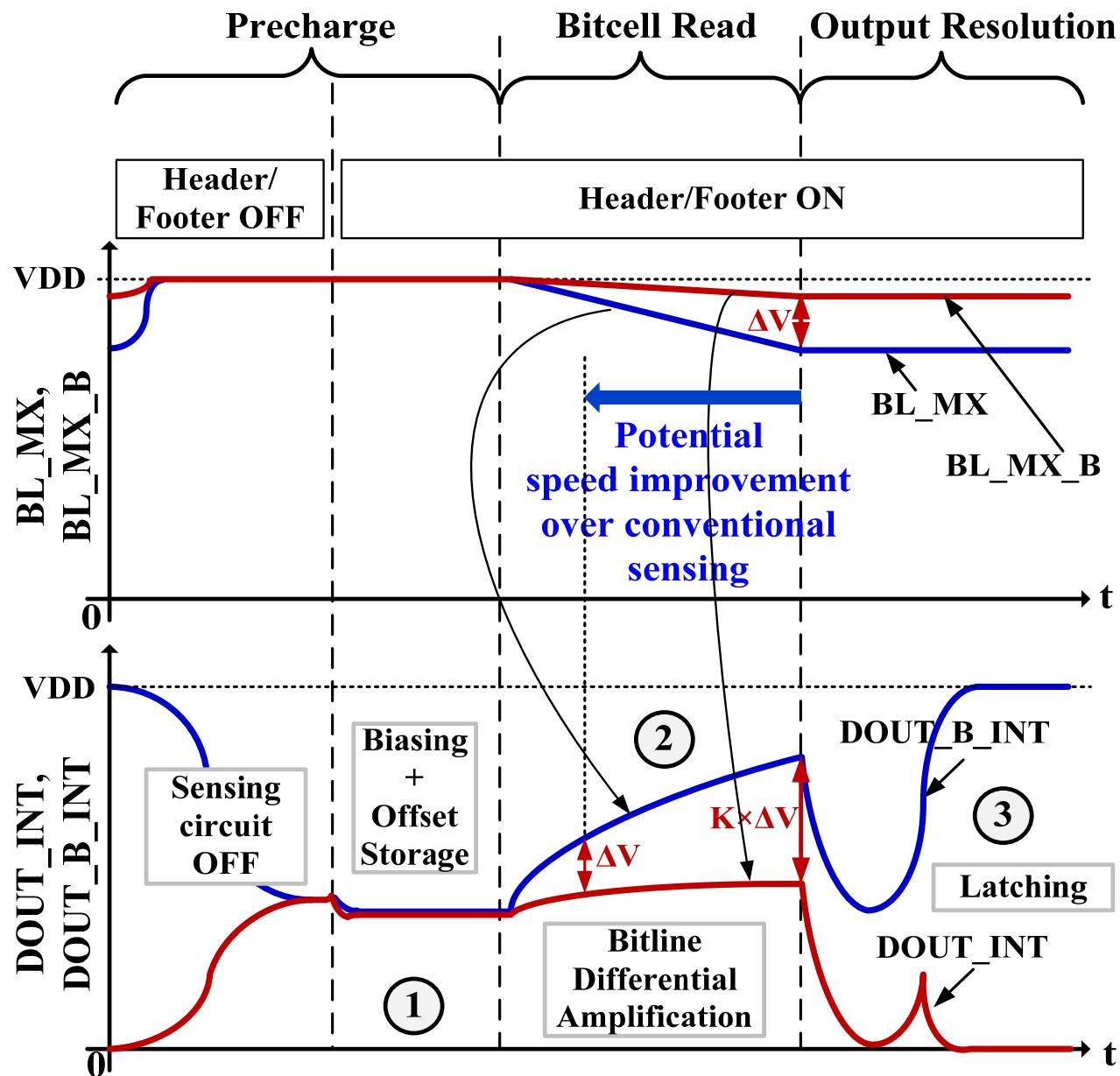
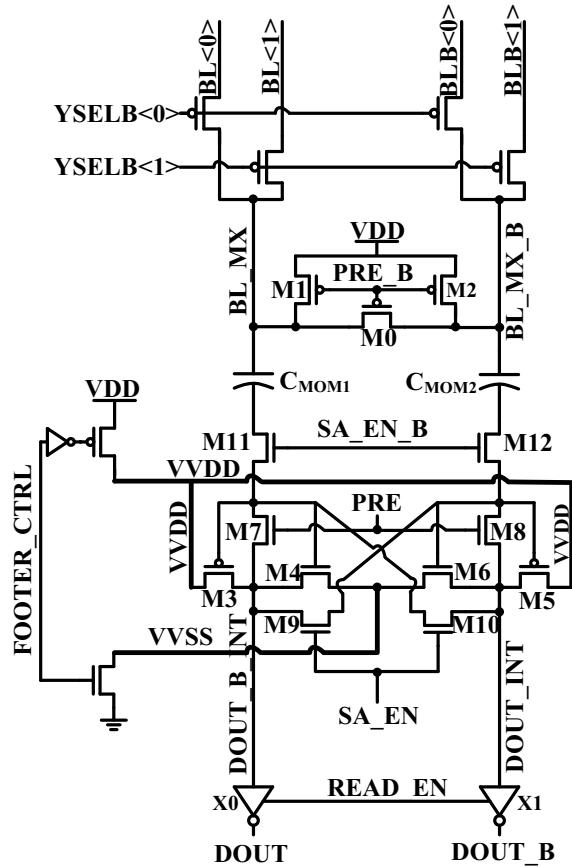
13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

VTS Operation Phases



13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

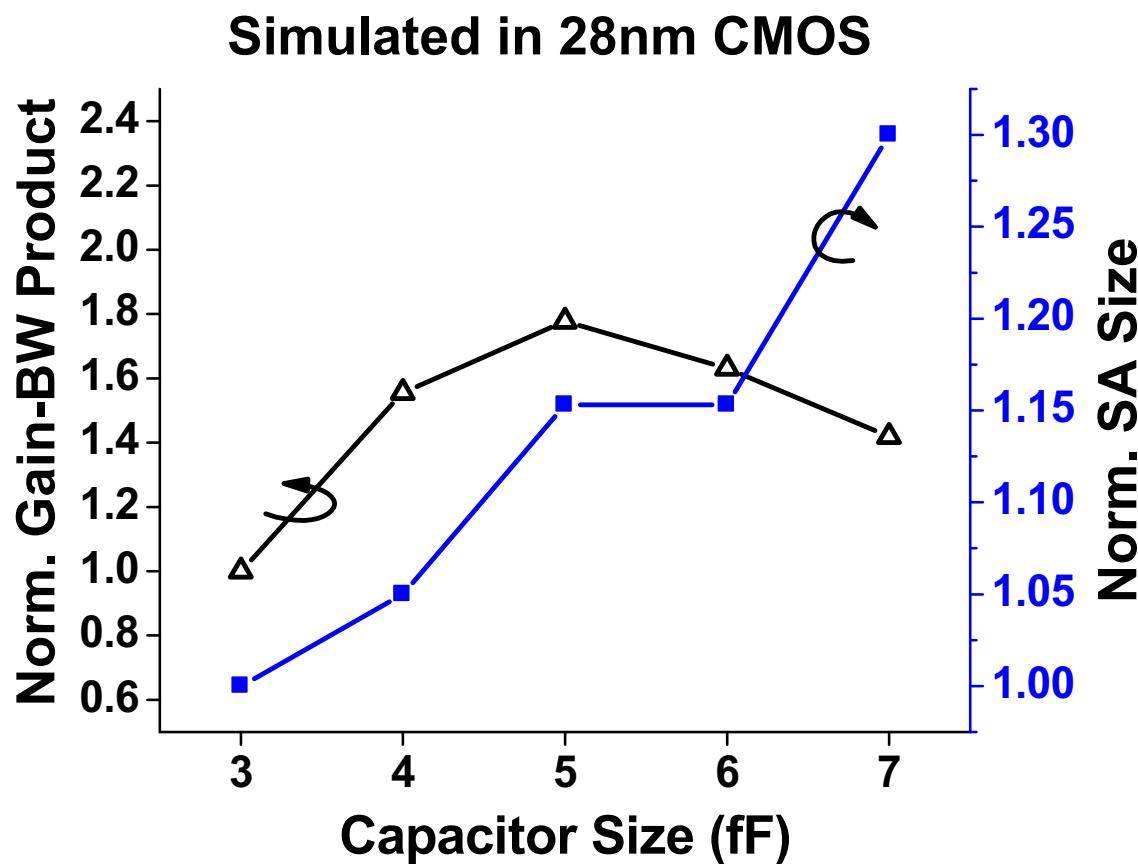
VTS Operation Phases



13.7: A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS

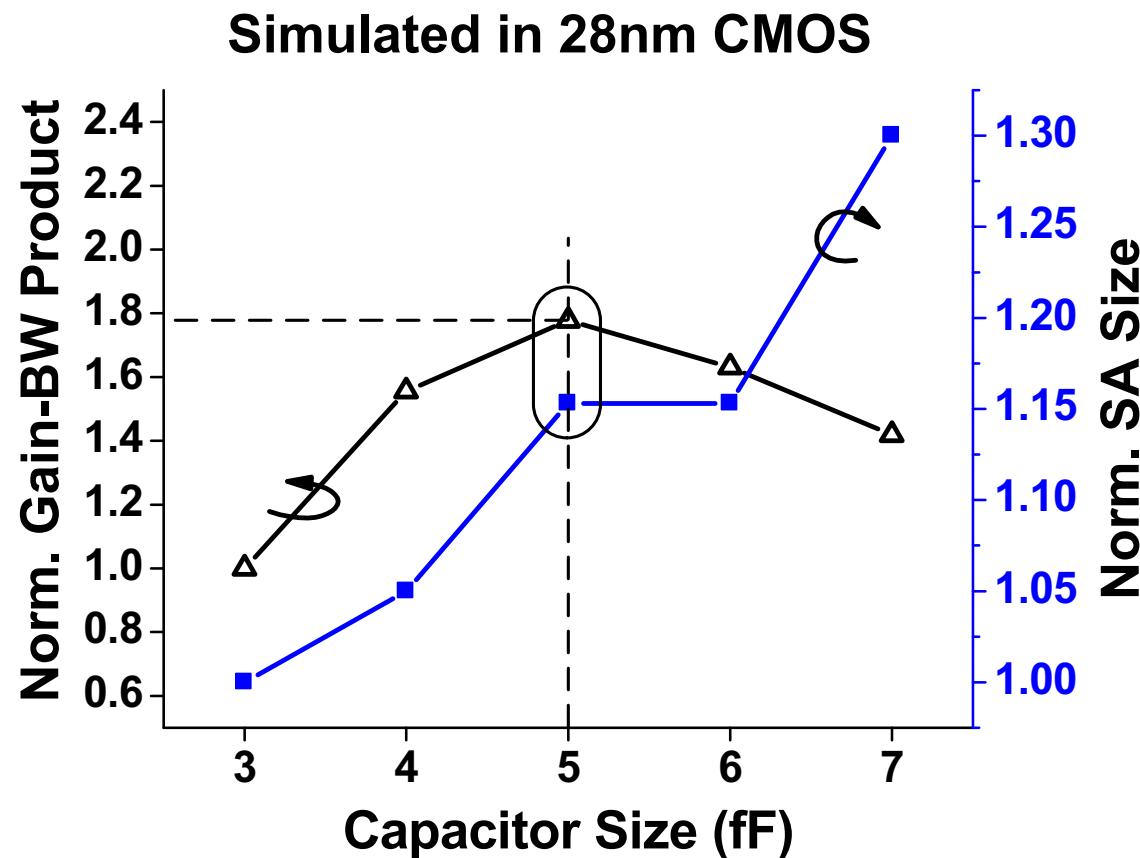
VTS Capacitor Design

- Increase cap
 - Requires upsizing devices to maintain precharge spec
 - Degrades sense time
- Decrease cap
 - Degrades coupling, negates pre-amp. benefit



VTS Capacitor Design

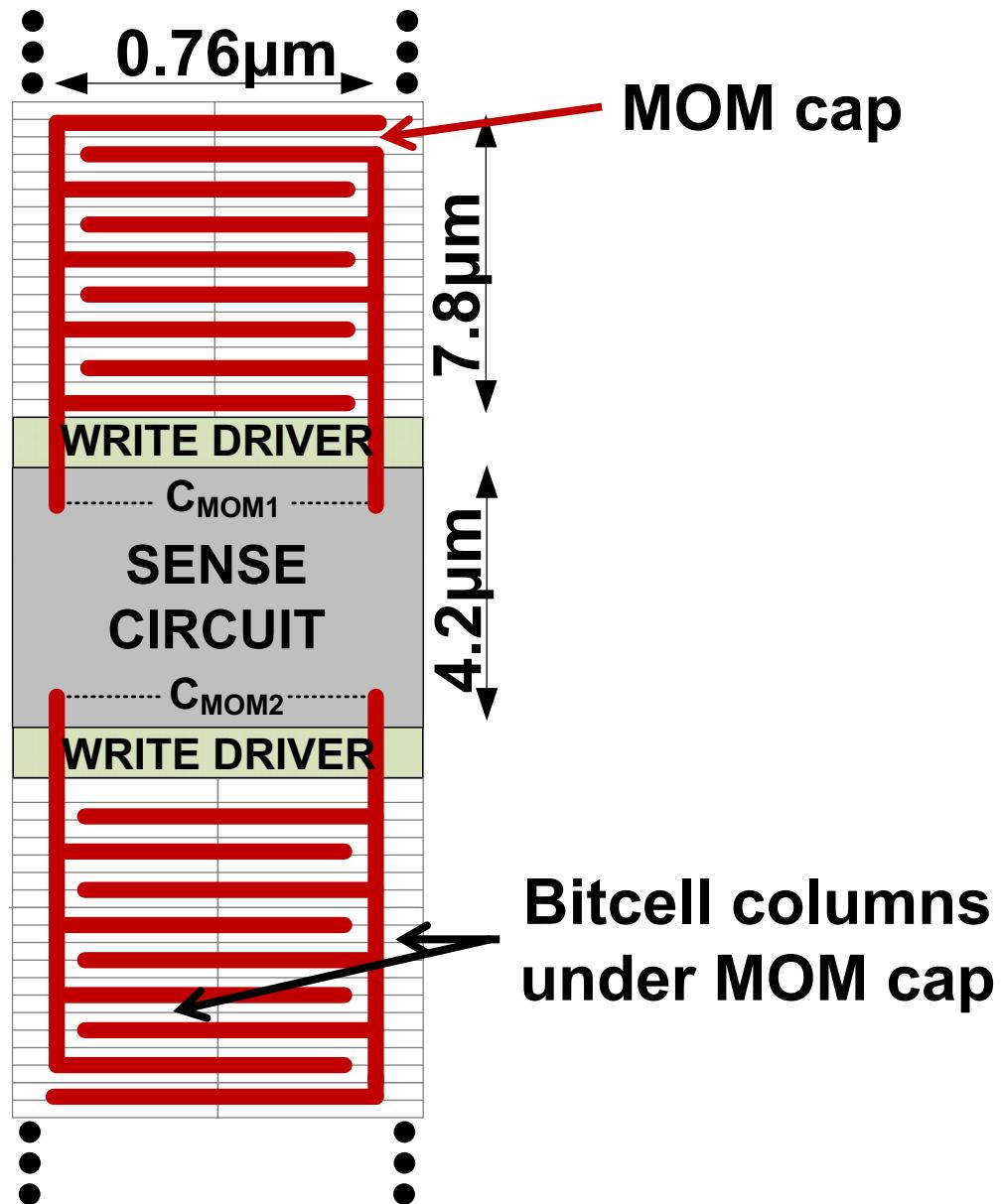
- Increase cap
 - Requires upsizing devices to maintain precharge spec
 - Degrades sense time
- Decrease cap
 - Degrades coupling, negates pre-amp. benefit



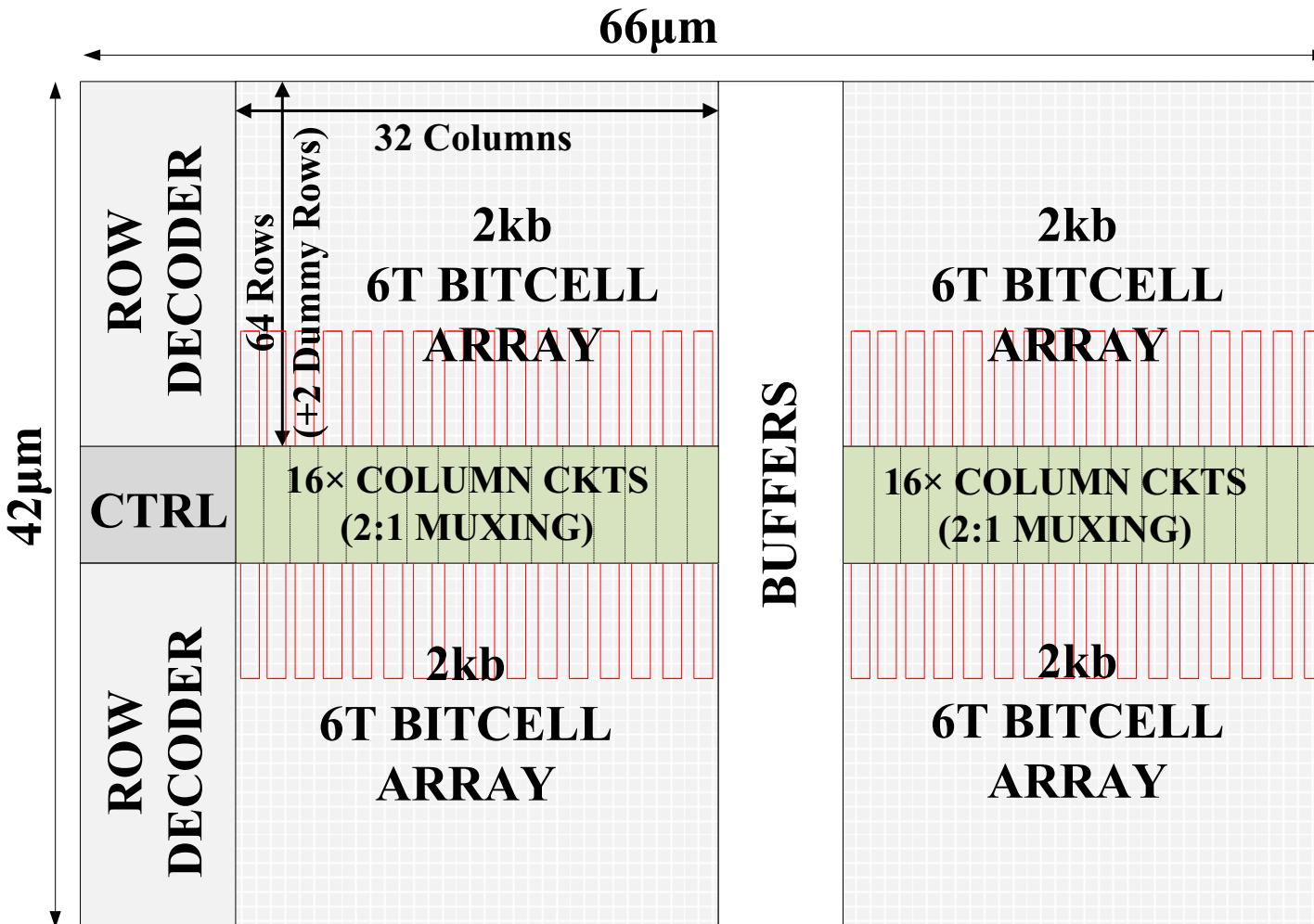
~5fF capacitors used for max. gain-bandwidth

VTS Capacitor Design

- Implemented as **MOM** capacitors
 - Over bitcell columns in M5-M6
- **MIM cap issues:** not viable due to larger min. size
- **MOS cap issues:** larger coupling losses, increase sensing circuit area

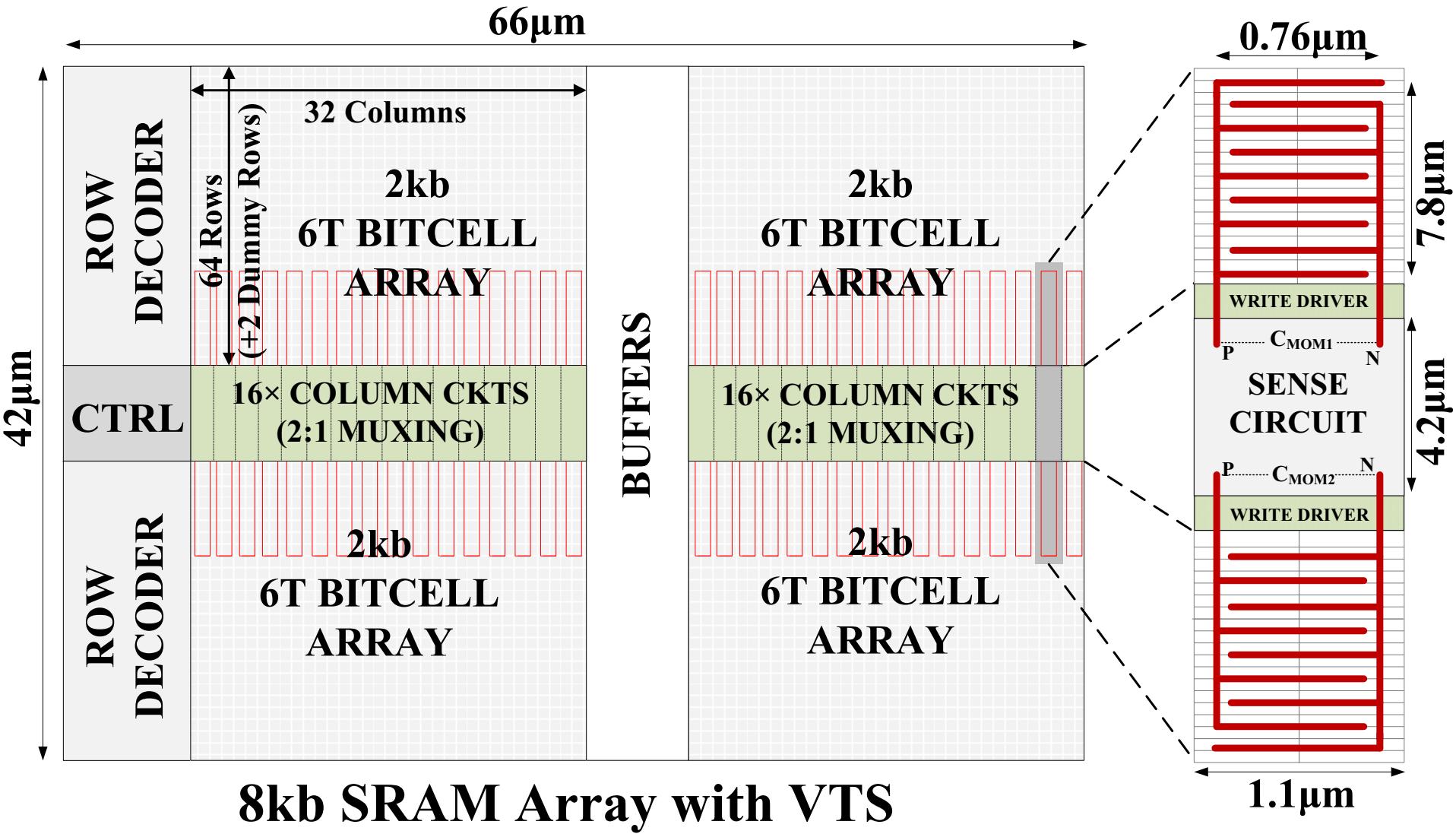


VTS Array Design



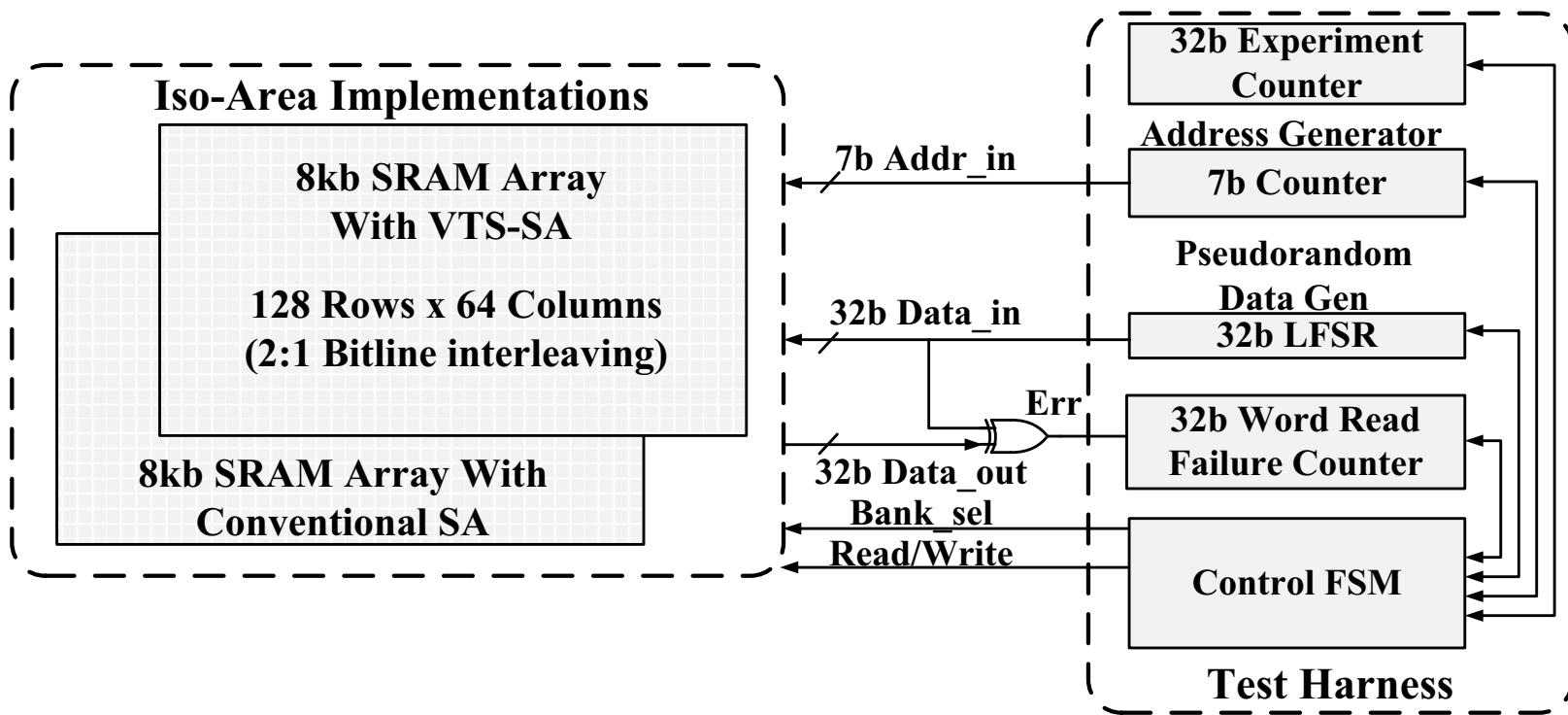
8kb SRAM Array with VTS

VTS Array Design

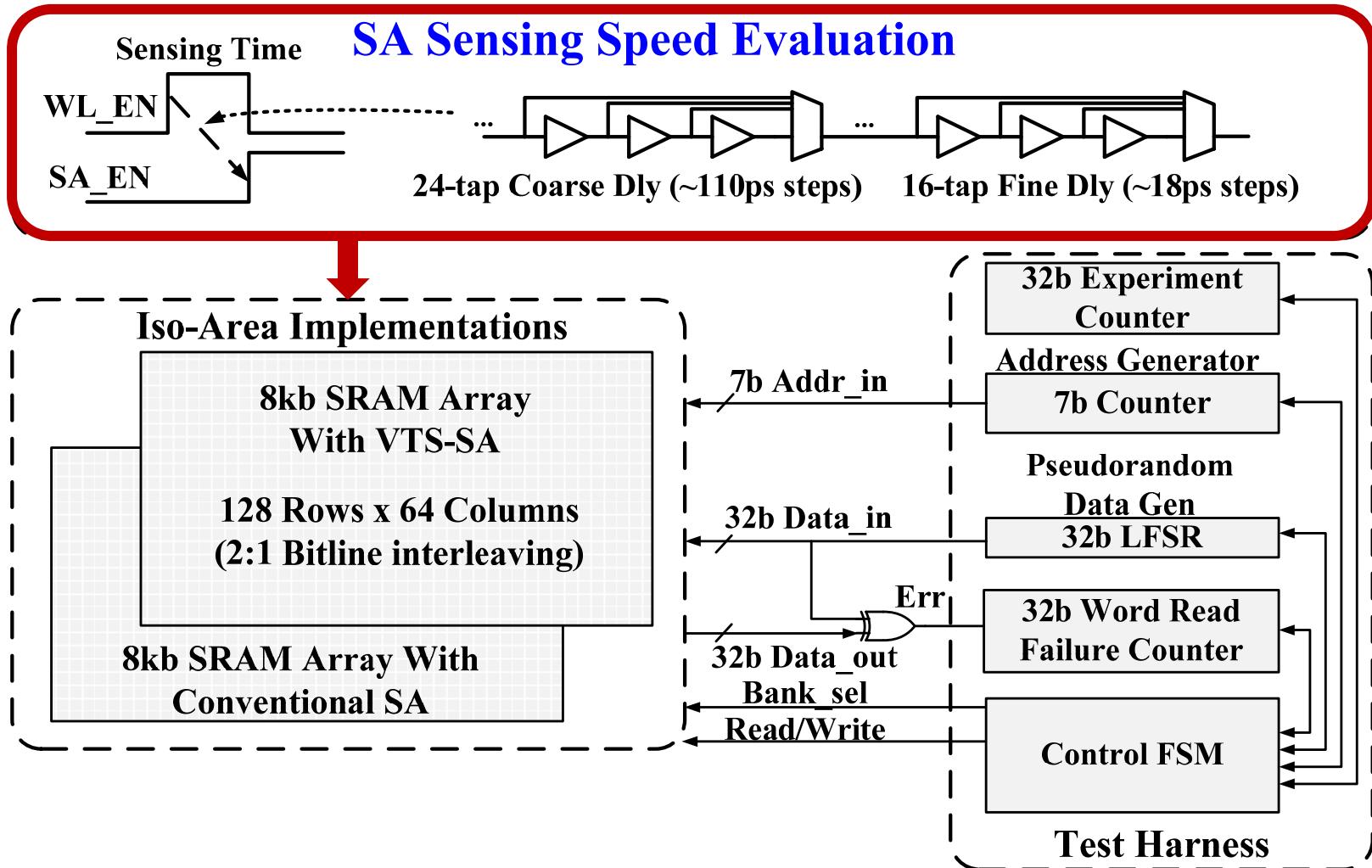


Test Chip in 28nm CMOS

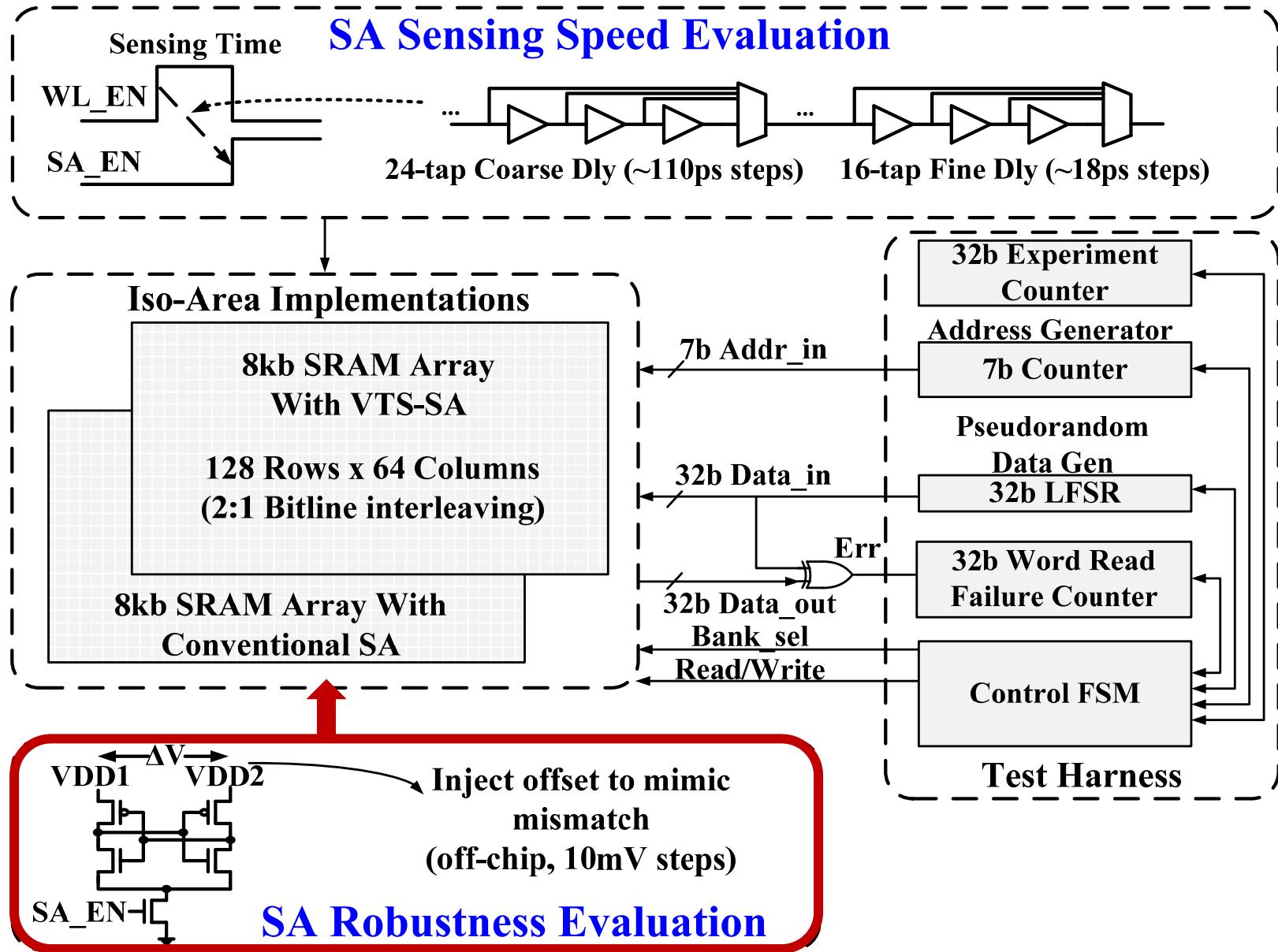
- Conventional SA sized for 4.5σ yield with $4.62\mu\text{m}^2$ area
- VTS-SA size-matched to conventional SA



Test Chip in 28nm CMOS

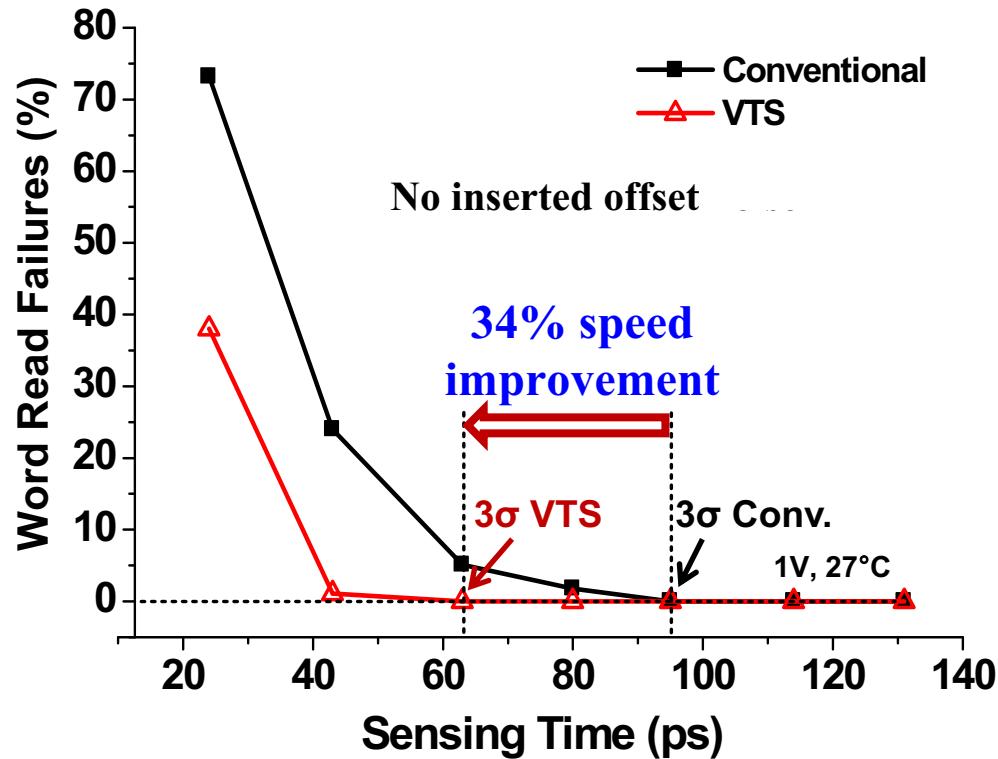


Test Chip in 28nm CMOS



Measured Results – Typical Die

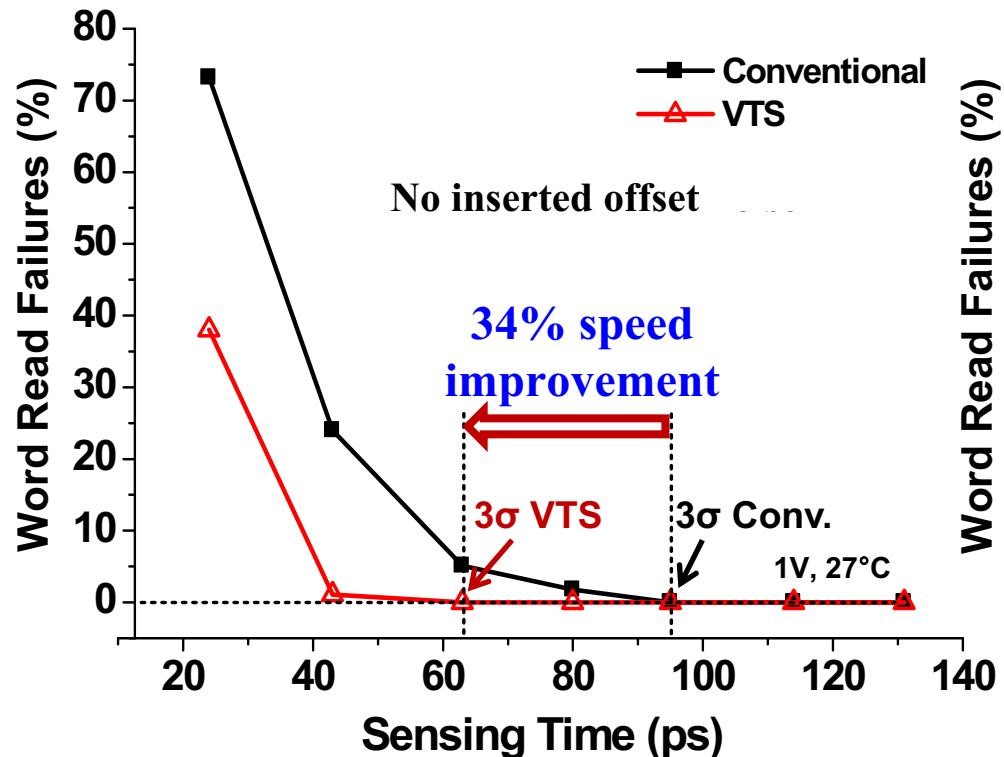
SA Sensing speed



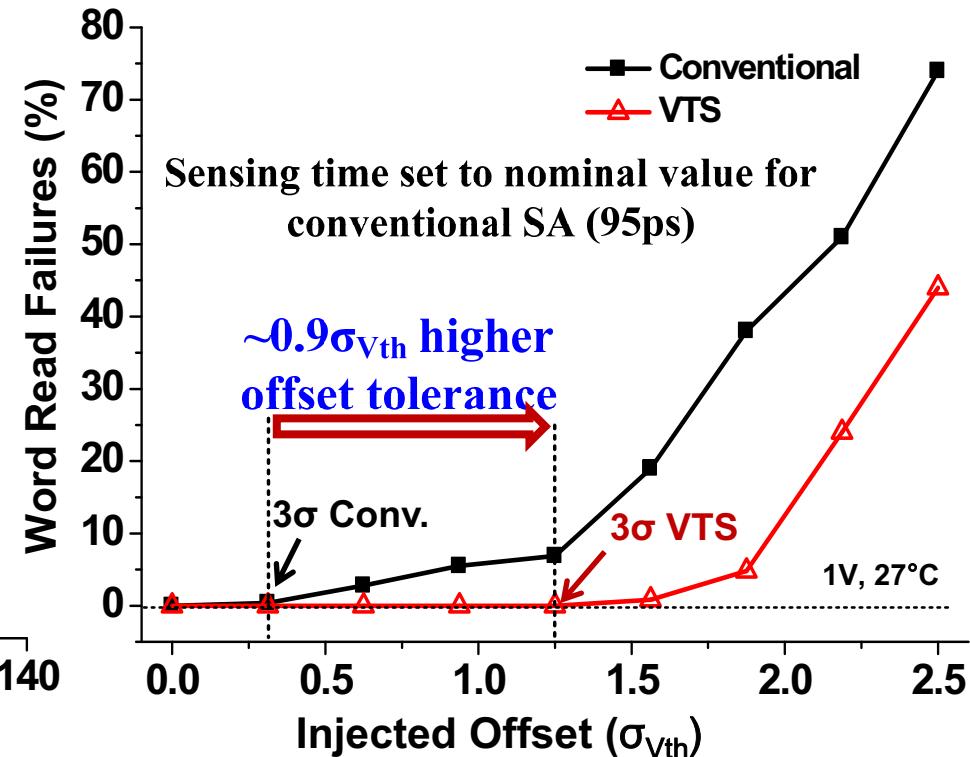
@Iso-robustness
(with 3σ yield)

Measured Results – Typical Die

SA Sensing speed



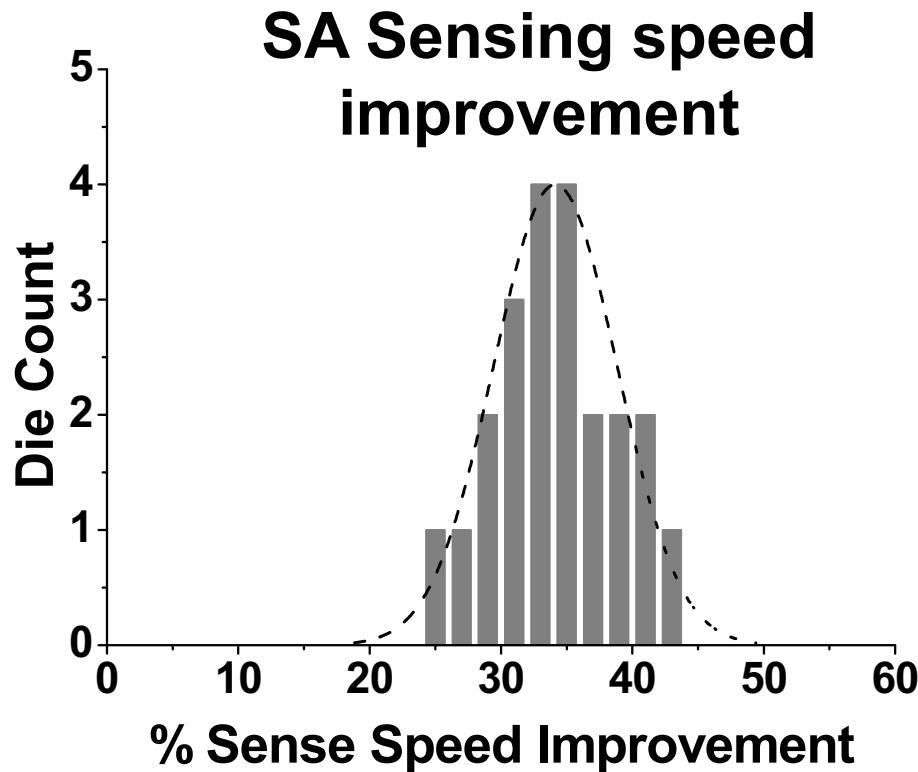
SA Robustness



@Iso-robustness
(with 3 σ yield)

@Iso-sensing time
(with 3 σ yield)

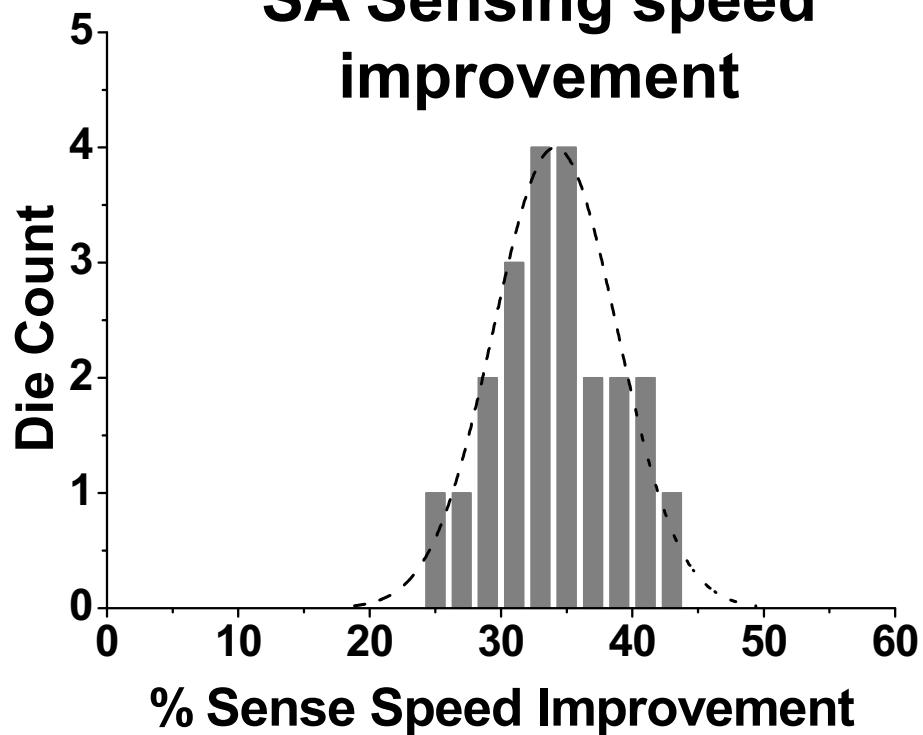
Measured Results – Across 22 Dies



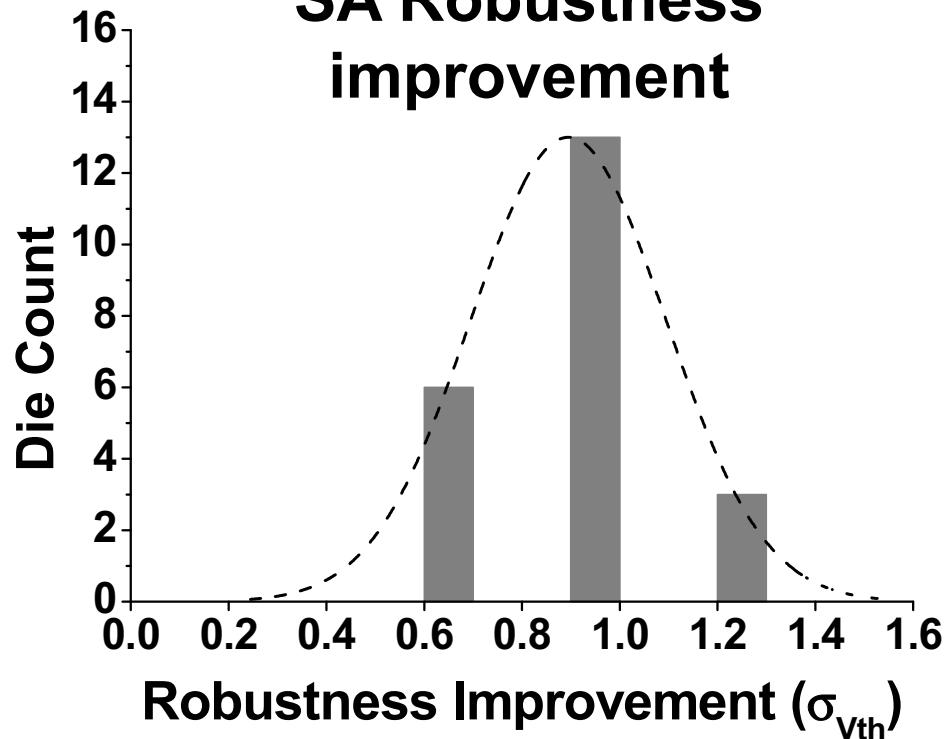
Sensing speedup
Mean: 34%
Max 42%

Measured Results – Across 22 Dies

SA Sensing speed improvement



SA Robustness improvement

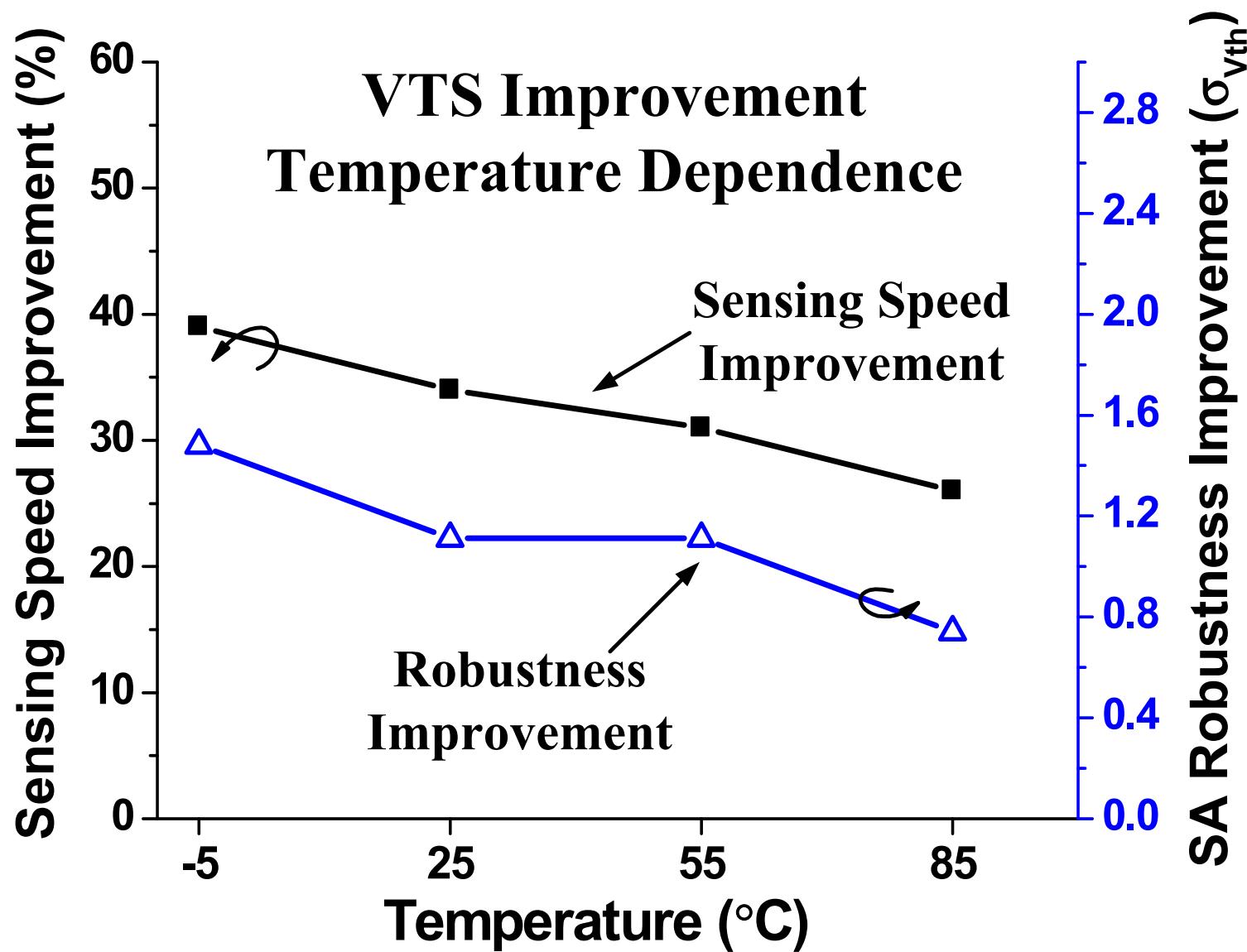


Sensing speedup
Mean: 34%
Max 42%

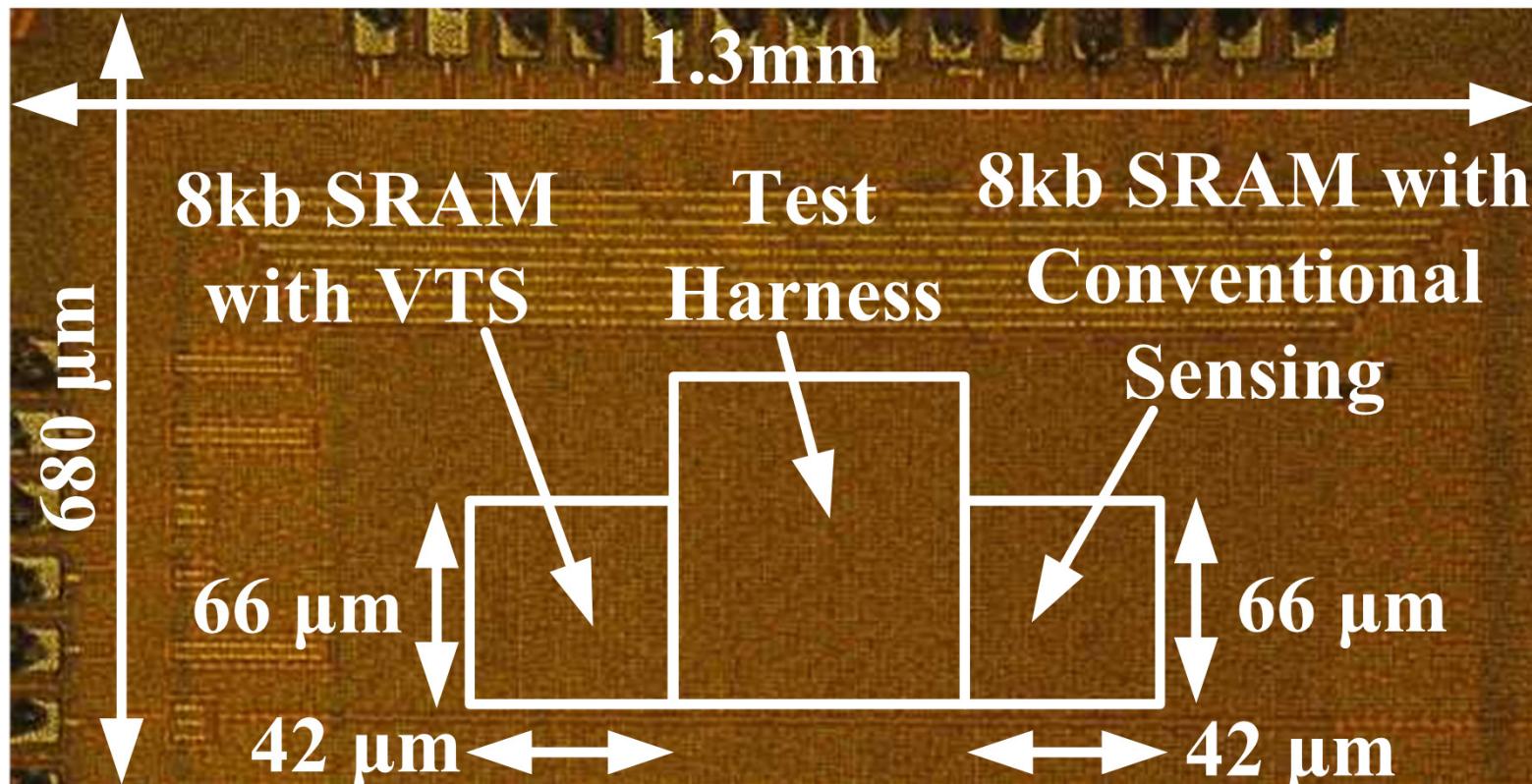
Robustness improvement
Mean: $0.9\sigma_{V_{th}}$ V
Max: $1.2\sigma_{V_{th}}$ V



Measured Results – Temp. Dependence



Die Micrograph



Test chip in 28nm bulk CMOS

Performance Summary

	Conventional SA	VTS-SA (Proposed)
Technology	28nm Bulk CMOS	
Supply Voltage	1V	
Power @ 1.8GHz	5.6 μ W	5.1 μ W
Sensing Circuit Area		4.62 μ m ²
Mean Sensing Speed (Improvement @ Iso-Robustness)	95ps	63ps (26-42%)
Mean Robustness (Improvement @ Iso-Sensing Speed)	0.31 $\sigma_{V_{th}}$ V	1.25 $\sigma_{V_{th}}$ V (0.6-1.2 $\sigma_{V_{th}}$)

Performance Summary

	Conventional SA	VTS-SA (Proposed)
Technology	28nm Bulk CMOS	
Supply Voltage	1V	
Power @ 1.8GHz	5.6 μ W	5.1 μ W
Sensing Circuit Area	4.62 μ m ²	
Mean Sensing Speed (Improvement @ Iso-Robustness)	95ps	63ps (26-42%)
Mean Robustness (Improvement @ Iso-Sensing Speed)	0.31 $\sigma_{V_{th}}$ V	1.25 $\sigma_{V_{th}}$ V (0.6-1.2 $\sigma_{V_{th}}$)

Performance Summary

	Conventional SA	VTS-SA (Proposed)
Technology	28nm Bulk CMOS	
Supply Voltage	1V	
Power @ 1.8GHz	5.6 μ W	5.1 μ W
Sensing Circuit Area	4.62 μ m ²	
Mean Sensing Speed (Improvement @ Iso-Robustness)	95ps	63ps (26-42%)
Mean Robustness (Improvement @ Iso-Sensing Speed)	0.31 $\sigma_{V_{th}}$ V	1.25 $\sigma_{V_{th}}$ V (0.6-1.2 $\sigma_{V_{th}}$)

Performance Summary

	Conventional SA	VTS-SA (Proposed)
Technology	28nm Bulk CMOS	
Supply Voltage	1V	
Power @ 1.8GHz	5.6 μ W	5.1 μ W
Sensing Circuit Area		4.62 μ m ²
Mean Sensing Speed (Improvement @ Iso-Robustness)	95ps	63ps (26-42%)
Mean Robustness (Improvement @ Iso-Sensing Speed)	0.31 $\sigma_{V_{th}}$ V	1.25 $\sigma_{V_{th}}$ V (0.6-1.2 $\sigma_{V_{th}}$)

Conclusion

- Reconfigurable sense amplifier
 - Allows amp to latch circuit reconfiguration
- Three Operation Phases:
 - **Auto-zeroing + Pre-amplification** for offset comp.
 - **Latching** for data regeneration
- Up to **42%** sensing speedup at iso-robustness
- Up to **$1.2\sigma_{V_{th}}$** improved tolerance to injected offset
- Footprint and power comparable to conventional SA

A 32kb SRAM for Error-Free and Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS

Fabio Frustaci*, Mahmood Khayatzadeh**, David Blaauw**, Dennis Sylvester**, Massimo Alioto***

* DIMES, University of Calabria (Italy)

** EECS, University of Michigan, Ann Arbor (MI)

*** ECE, National University of Singapore (Singapore)

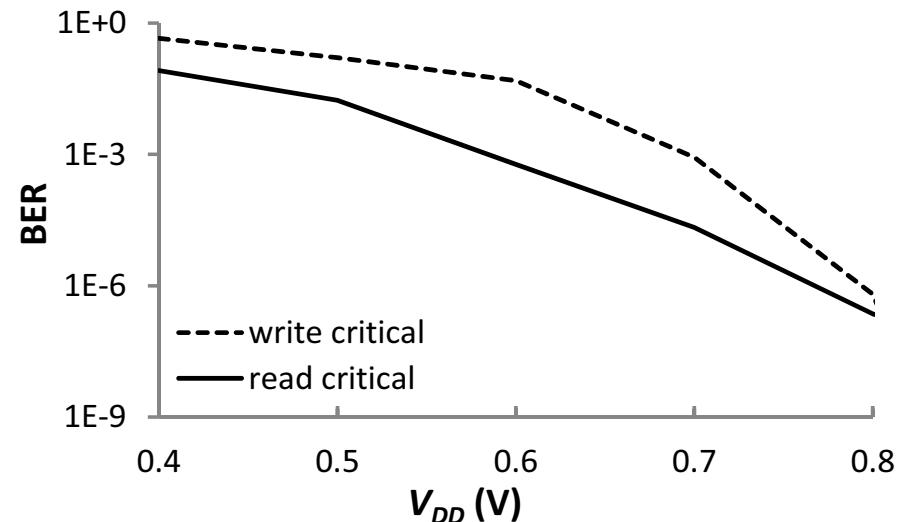
Outline

- ◆ Errors in aggressively scaled SRAMs and impact on signal quality vs bit position
- ◆ Dynamic quality-energy tradeoff at bit level
- ◆ Approach, architecture and adopted techniques
- ◆ Testchip and measurement results
 - energy vs. quality, energy/voltage reduction
- ◆ Broader View on Energy-Quality Tradeoff
- ◆ Conclusion

Errors in Aggressively Scaled SRAMs

- ◆ Energy efficiency is key in many applications
- ◆ Effectively improved through voltage scaling
 - Degraded bitcell read/write margin
- ◆ Max. energy reduction limited by V_{MIN}

- For $V_{DD} < V_{MIN}$, errors in bitcells with read/write failures



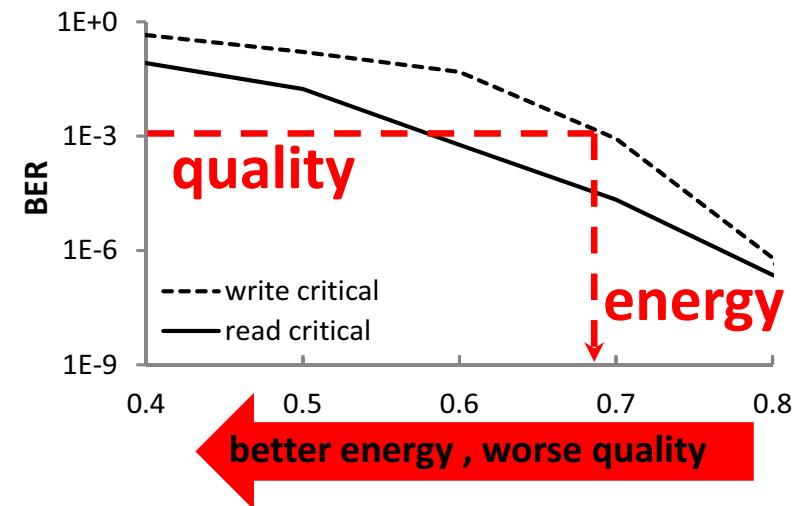
- Technology scaling
 - V_{MIN} and benefits of voltage scaling do not improve
 - SRAM V_{MIN} limits system voltage scaling

Error-Free vs Error-Tolerant Applications

- ◆ Error-tolerant apps: errors are acceptable [Shanbhag2004]
 - Ex.: multimedia/graphics, sensor fusion

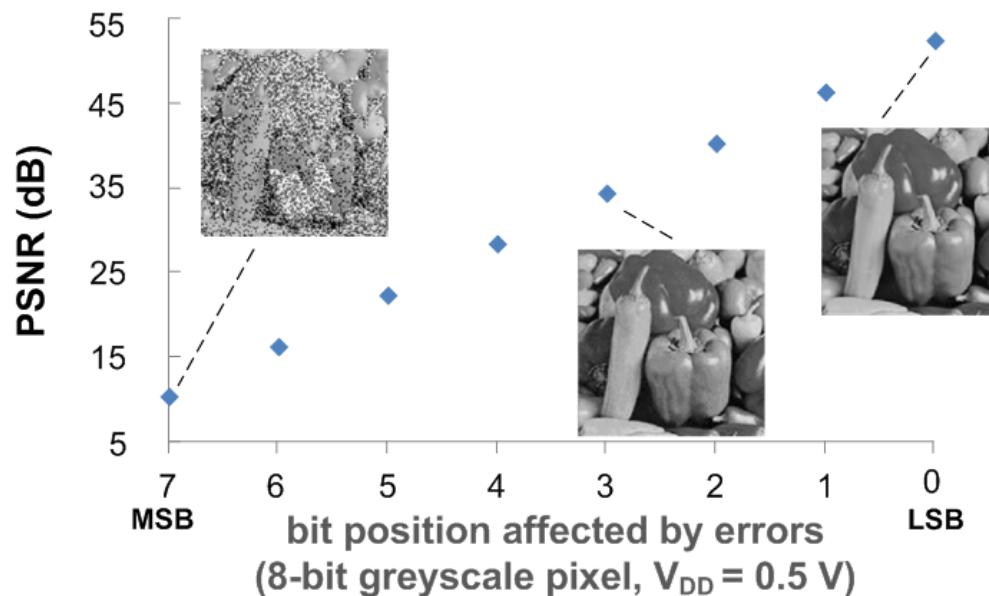
error-free	error-tolerant
$V_{DD} > V_{MIN}$	$V_{DD} < V_{MIN}$ allowed

- Potentially larger energy savings than error-free
- ◆ Energy savings limited by exp degradation of Bit(cell) Error Rate
 - Graceful degradation
⇒ lower V_{DD}



Impact of Error on Quality vs Bit Position

- ◆ Focus on image processing (general)
- ◆ Errors = noise added to processed data
 - Metrics: Peak SNR (the higher, the better)
 - Errors in different places affect quality in different ways
 - Ex.: 8-bit gray-scale, errors at one bit position leads to:

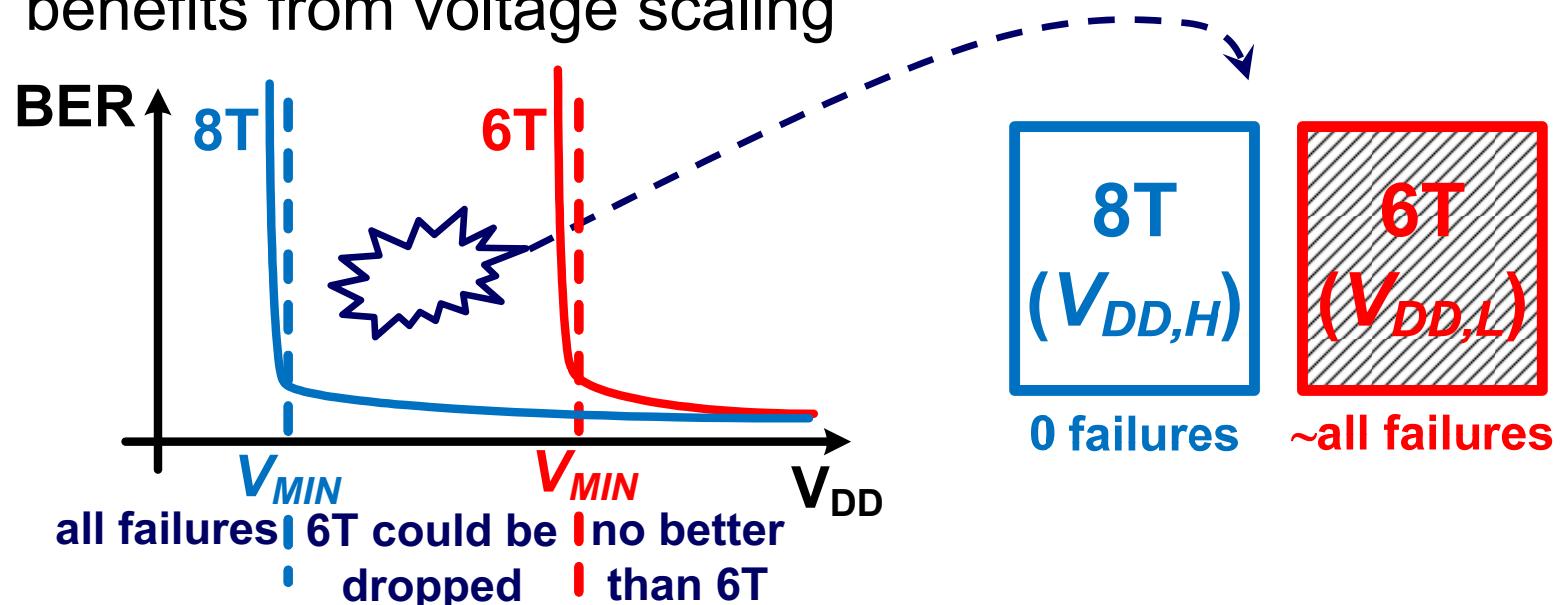


Previous Work

◆ Previous work on error-tolerant SRAMs

hybrid 6T-8T array [Roy2011]	dual VDD columns [Wolf2011]
8T bitcell for more important bits (6T for others)	$V_{DD,H}$ for more important bits ($V_{DD,L}$ for others)

- Energy-quality set at design time, application specific
- Ungraceful quality degradation \Rightarrow limited energy benefits from voltage scaling



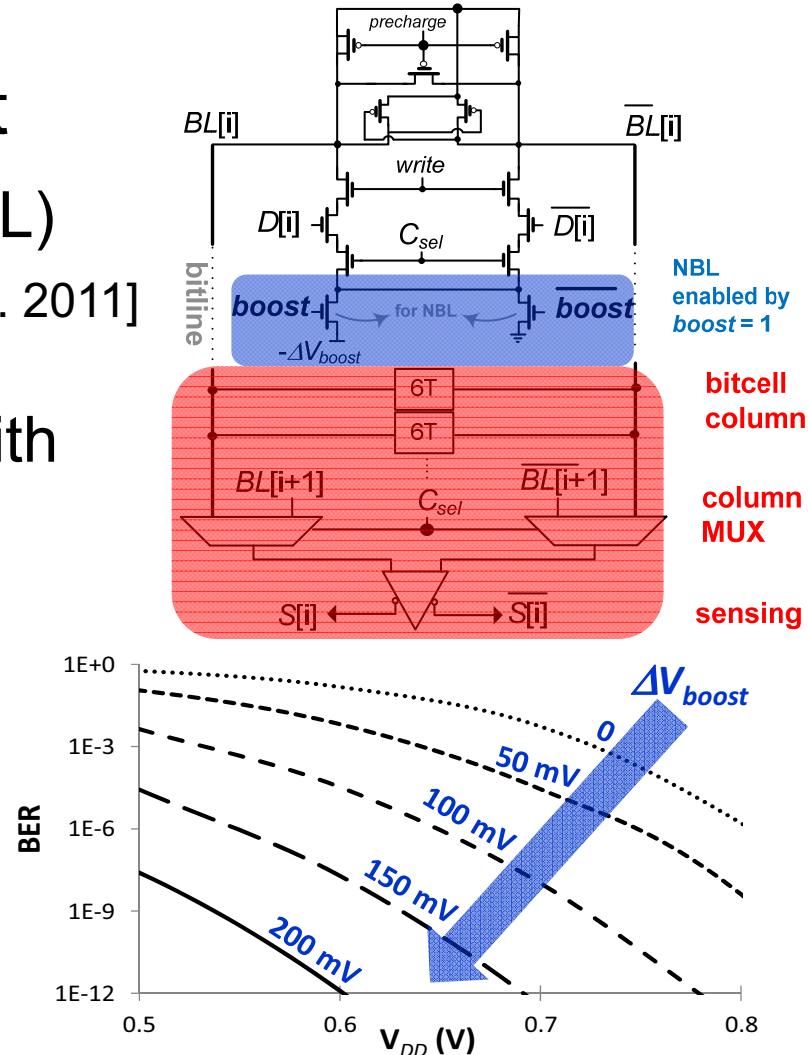
13.8: A 32kb SRAM for Error-Free and Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS

Approach

- ◆ Run-time adjustment of energy-quality tradeoff
- ◆ Application-independent design (general purpose)
 - Wide energy scalability (error-free → error-tolerant)
 - Standard 6T bitcell (non-customized)
- ◆ Graceful **quality** degradation, low energy
 - ↑ resiliency: spend energy to mitigate errors
 - Allocate energy non-uniformly (only where errors affect quality) ⇒ improve resiliency at bit level
 - Fix write/read failures (graceful degradation at any corner)

Techniques to Adjust Quality-Energy Tradeoff at Bit Level

- ◆ Write margin improvement
 - Negative bitline boosting (NBL)
[Shibata et al. 2006], [Mukhopadhyay et al. 2011]
 - **Selective**: only in columns with $boost=1$
 - Boosting voltage set by write BER target

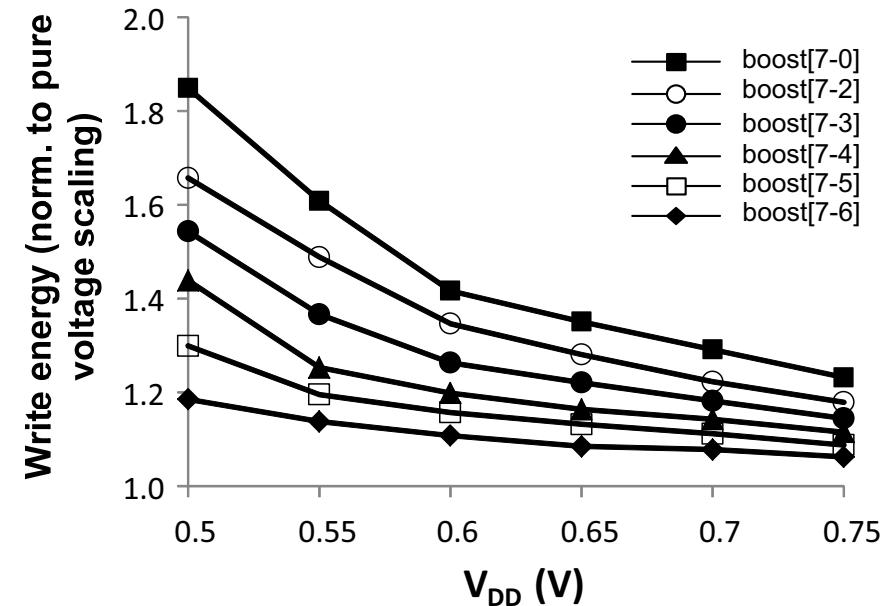


13.8: A 32kb SRAM for Error-Free and Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS

Techniques to Adjust Quality-Energy Tradeoff at Bit Level

◆ Write margin improvement

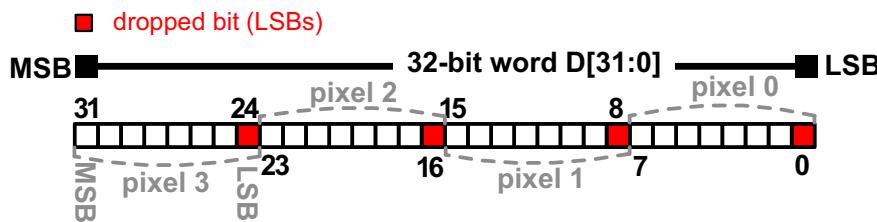
- $\Delta V_{boost} = 200 \text{ mV}$, $V_{DD} = 500 \text{ mV}$: BL energy increased by $\sim 2X$ w.r.t. no NBL
- Selective NBL on MSBs
 - Improves quality (**strongly**)
 - Low energy cost (few MSBs)
- Energy-quality tradeoff (**favorable**)
- NBL applied to MSBs:
more graceful quality degradation at lower energy cost



Techniques to Adjust Quality-Energy Tradeoff at Bit Level

- ◆ Read margin improvement: **selective ECC**
 - Traditional LSB dropping reduces activity [Kaul2013]
 - Unused bits (columns) save BL energy

LSB dropping



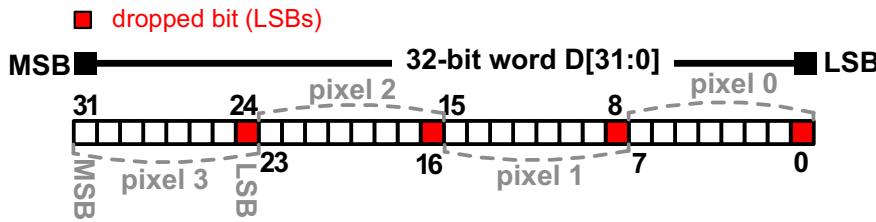
- unused LSBs
- ⇒ **linear energy reduction**

- Different approach: use “unused” LSBs to have more graceful quality degradation
- Higher activity, but more aggressive voltage scaling

Techniques to Adjust Quality-Energy Tradeoff at Bit Level

- ◆ Read margin improvement: **selective ECC**
 - More graceful quality degradation at linear energy cost
 - Energy savings **well beyond LSB dropping**

LSB dropping

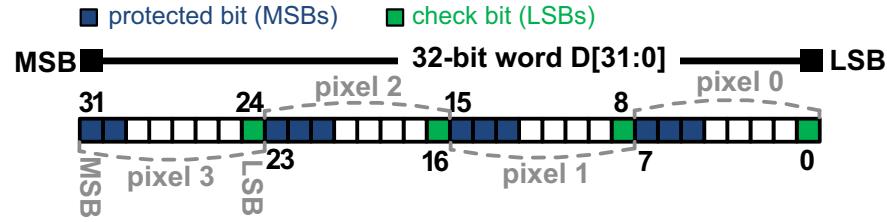


- unused LSBs

⇒ **linear energy reduction**

selective ECC encoder

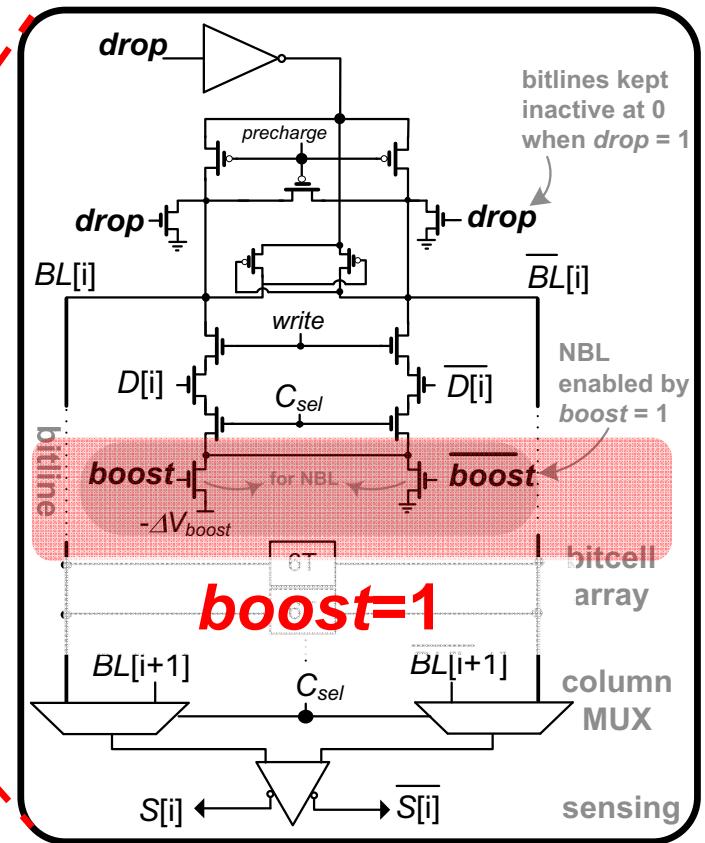
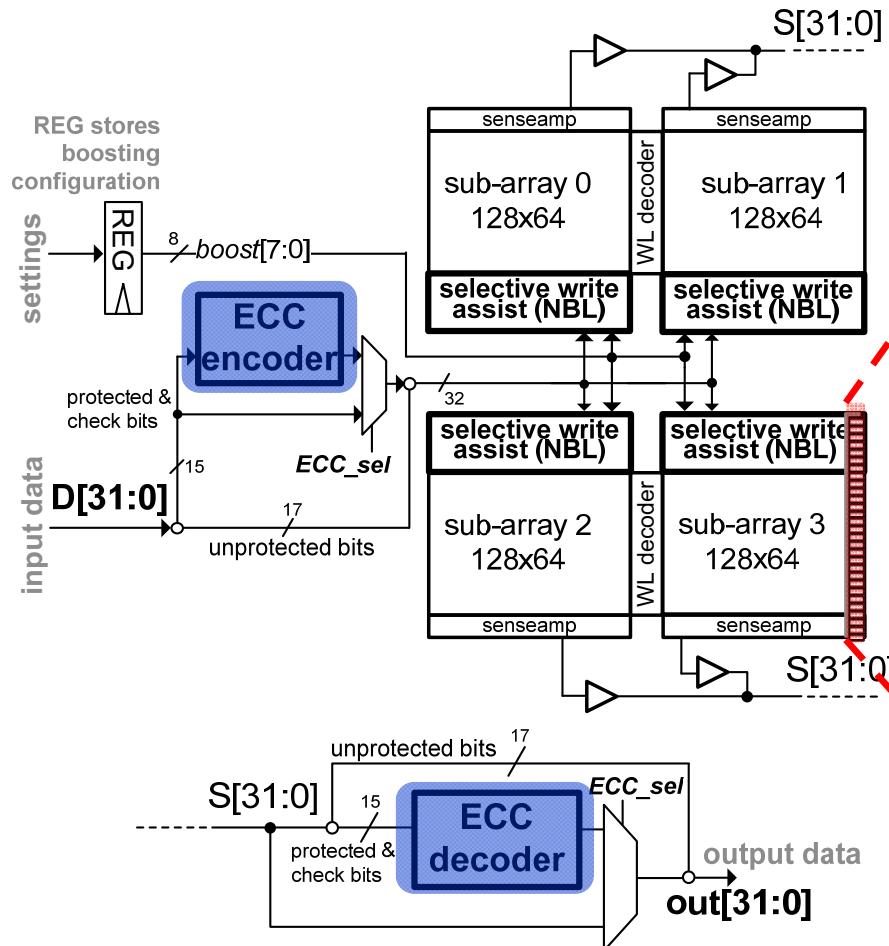
(LSBs protect MSBs via Hamming(15,11))



- unused LSBs protect MSBs

⇒ **quadratic energy reduction**

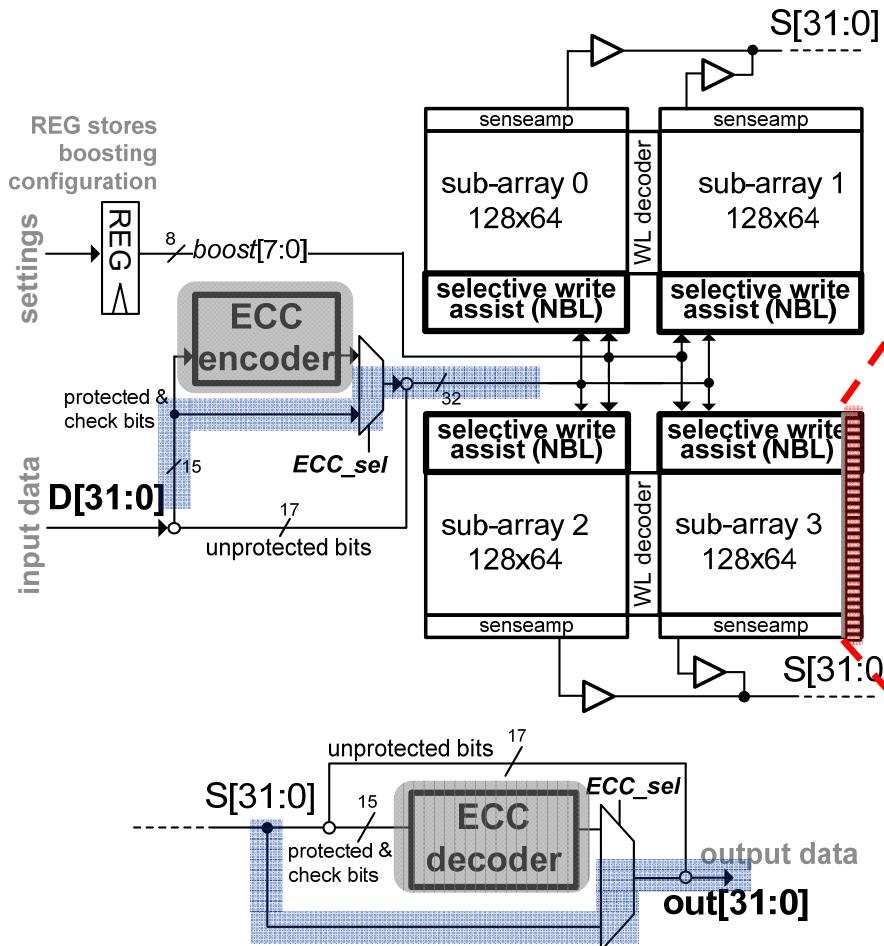
Architecture (Error-Tolerant Mode)



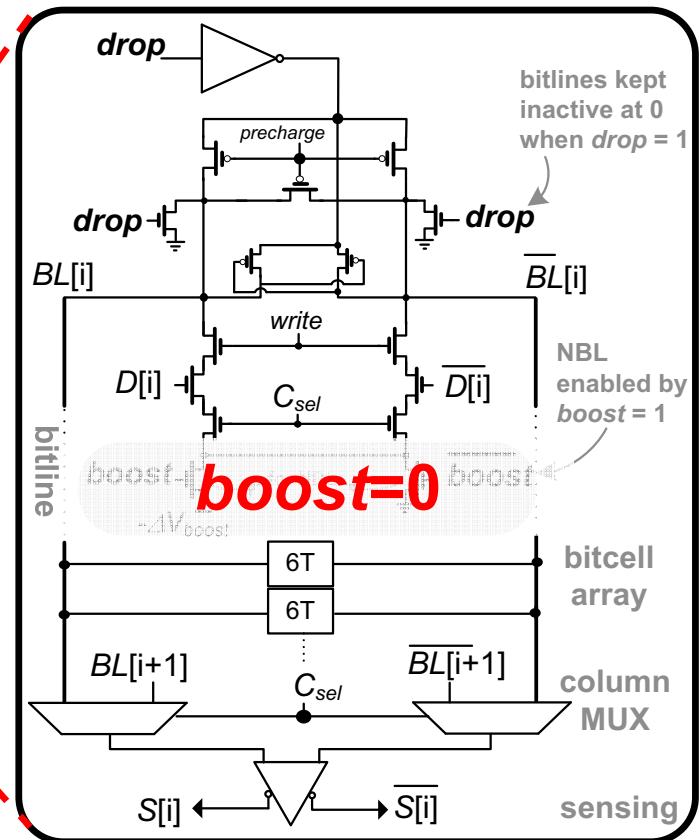
**ECC encoder/decoder
selectively protect MSBs**

**selective negative bitline
boosting (NBL)**

Architecture (Error-Free Mode)



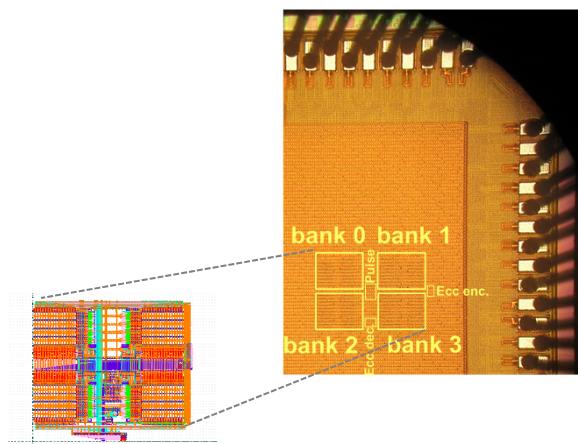
ECC encoder/decoder bypassed



selective negative bitline boosting disabled

Testchip

- ◆ Error-free/tolerant SRAM in 28 nm CMOS
 - 32kb (four 128×64 subarrays, 6T)
 - 32-bit word (4 X 8-b pixel), 2:1 column MUXing
 - $\Delta V_{boost} = -130\text{mV}$ (writeability over 5σ)
 - Reconfigurable (error-free, different NBL/ECC configurations, bit dropping, pure voltage scaling)
 - Mimic different corners through tunable WL voltage



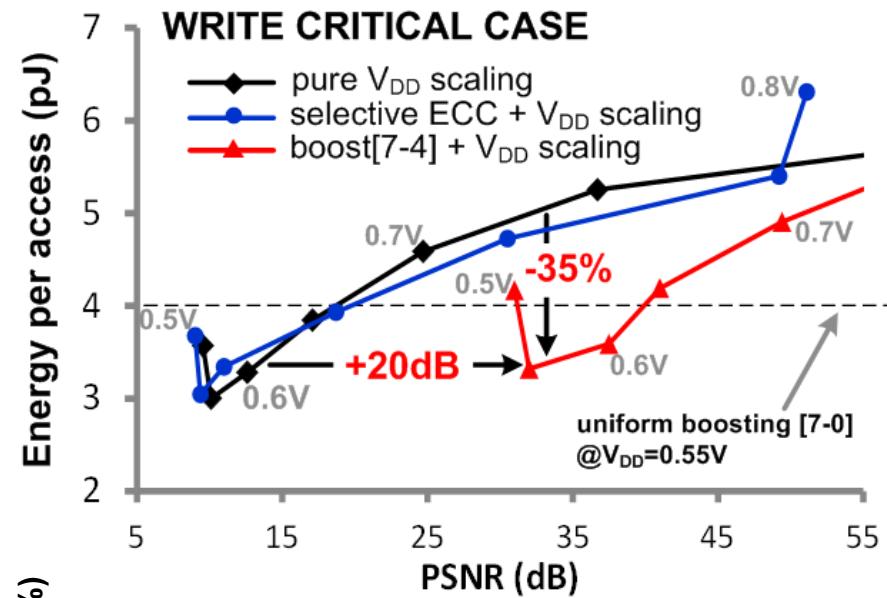
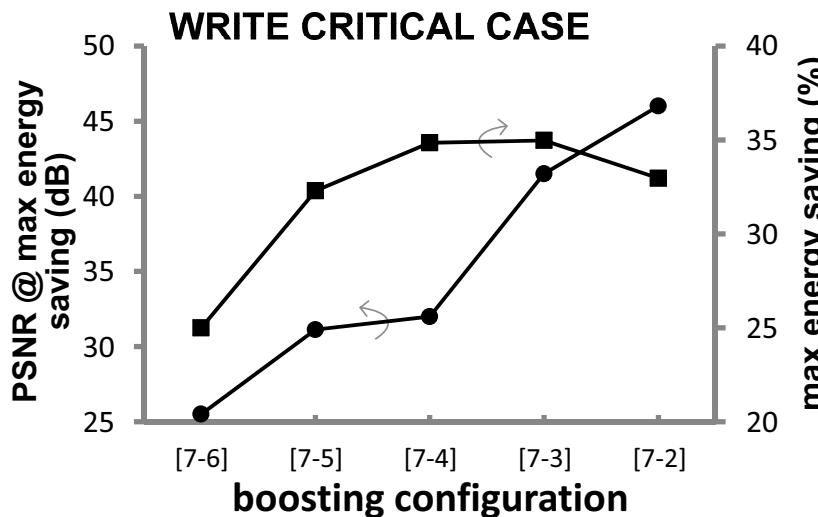
technology	28nm CMOS
area	252×202 μm^2
operating voltage	0.5V – 1V
min. data retention voltage @ 22°C	325mV
leakage @ V_{min} 22°C	11uA

13.8: A 32kb SRAM for Error-Free and Error-Tolerant Applications with Dynamic Energy-Quality Management in 28nm CMOS

Energy vs Quality (Write Critical)

- ◆ Selective NBL has **lower energy at iso-quality**

- Example: boosting [7-4] reduces energy by 35% w.r.t. pure V_{DD} scaling

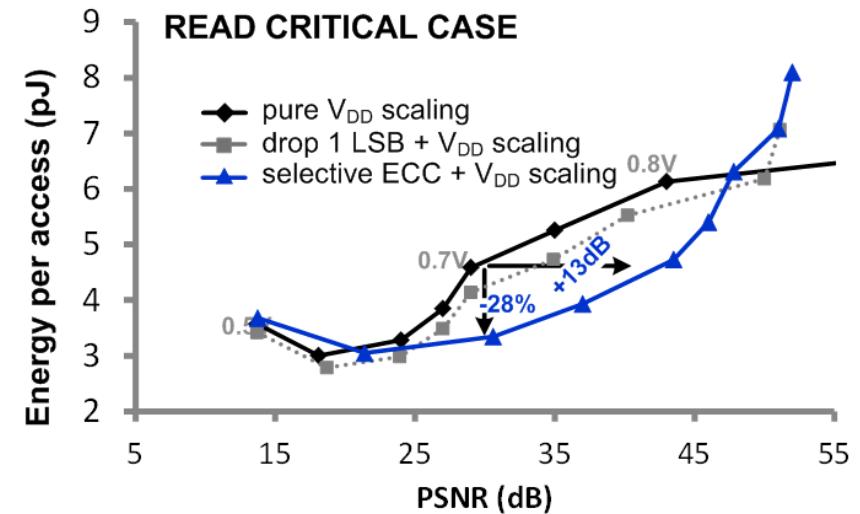


- Energy-optimal boosting configuration vs PSNR target

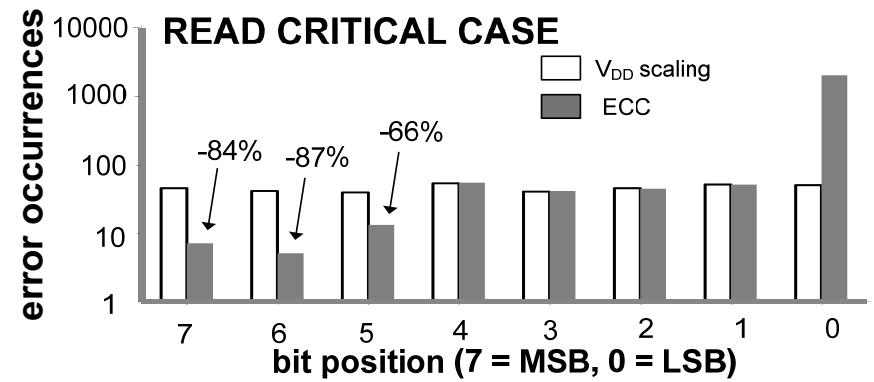
Energy vs Quality (Read Critical)

- ◆ Selective ECC has **lower energy at iso-quality**

- Reduces energy by 28% in read critical case w.r.t. pure V_{DD} scaling

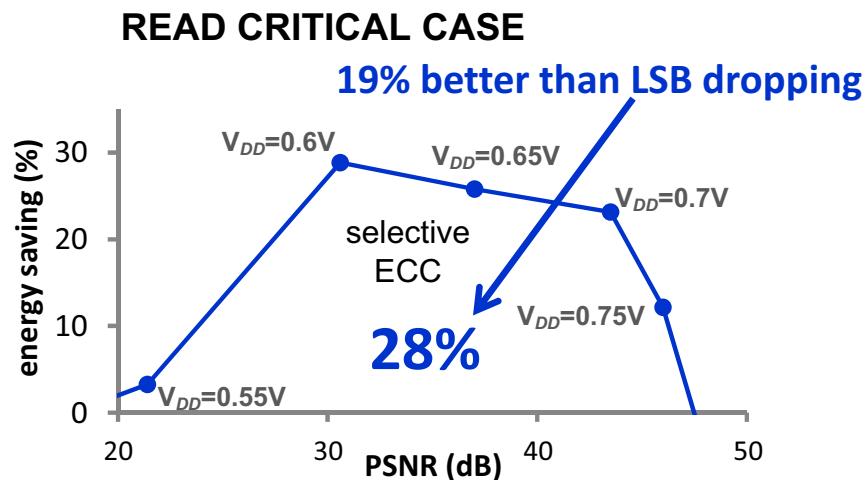
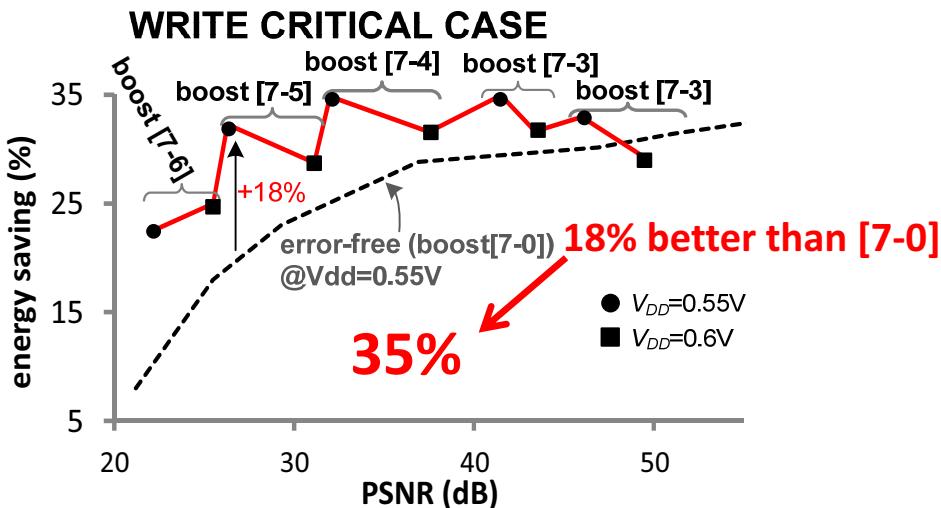


- Selective ECC (SEC) corrects most of errors in MSBs

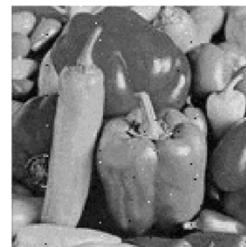


Energy Savings at Iso-Quality

- ◆ Energy optimal configuration vs quality target:
 - More aggressive voltage scaling \Rightarrow lower energy



PSNR=20dB



PSNR=30dB



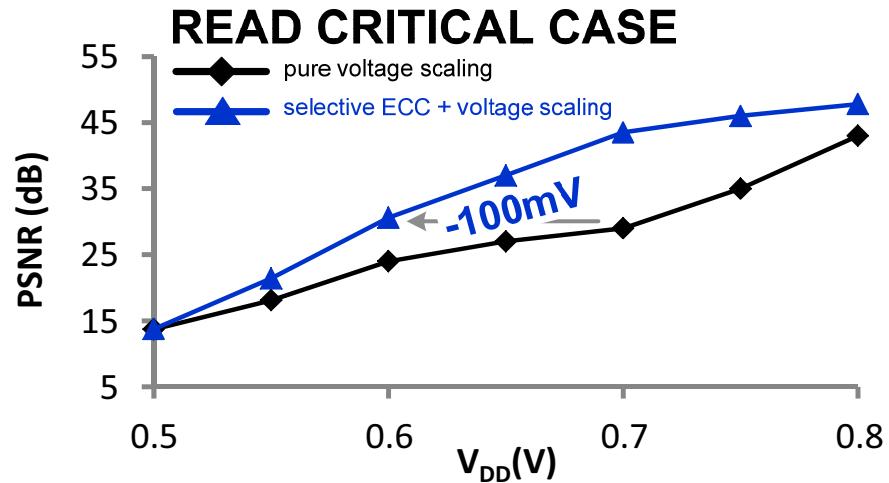
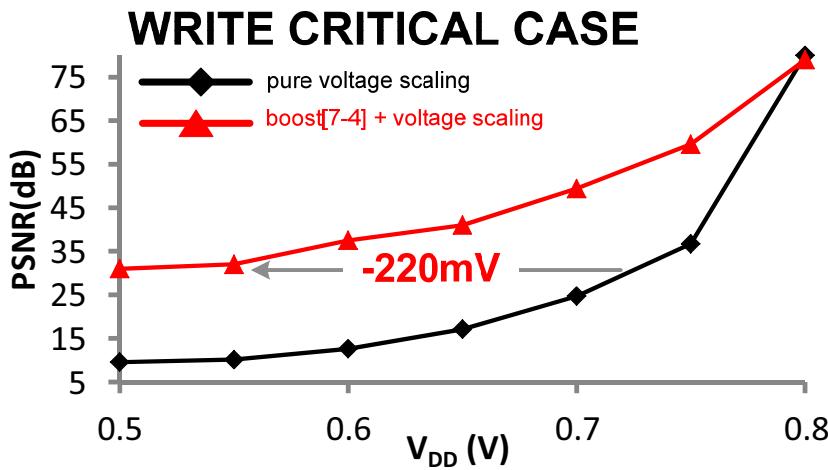
PSNR=40dB



PSNR=+ ∞

Voltage Scaling and V_{MIN} reduction

- ◆ Proposed approach enables more aggressive voltage scaling at iso-quality
 - Graceful error degradation + optimal energy allocation
 - Mitigates V_{MIN} gap between logic and SRAM

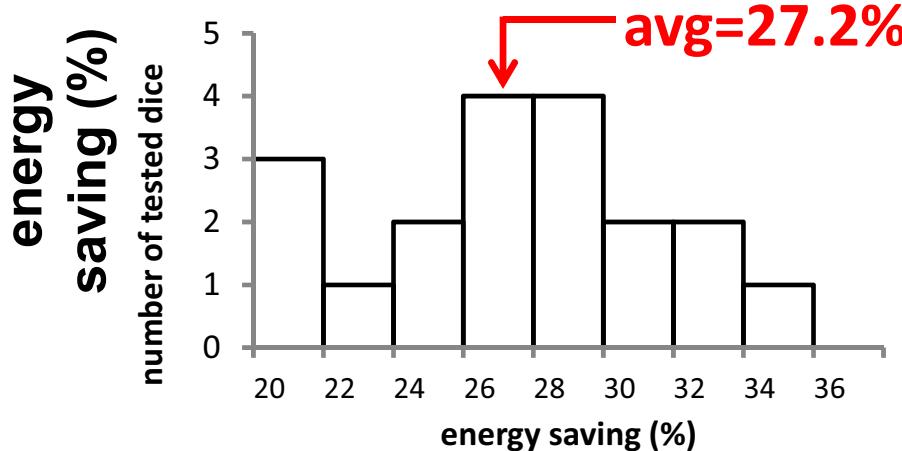


- @ PSNR=30 dB: V_{DD} is **reduced by 220 mV** at iso-quality in write critical corner (**100 mV in read**)

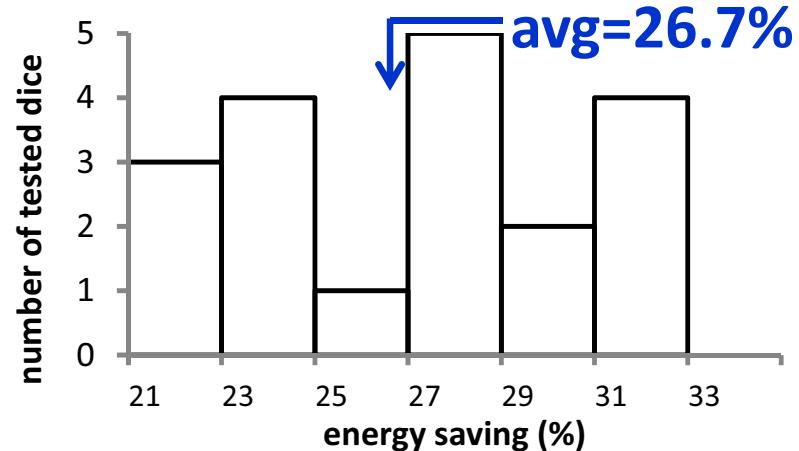
Measurements across Multiple Dice

- ◆ Energy saving across 19 dice
 - Quality target: PSNR=30 dB

WRITE CRITICAL CASE



READ CRITICAL CASE

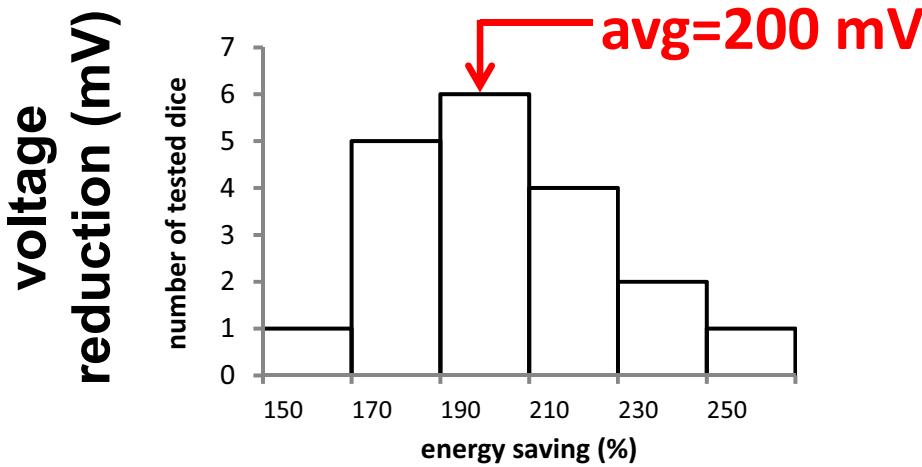


- Average energy saving @ PSNR=30 dB close to 27%
- Max energy saving is higher (energy-optimal PSNR)

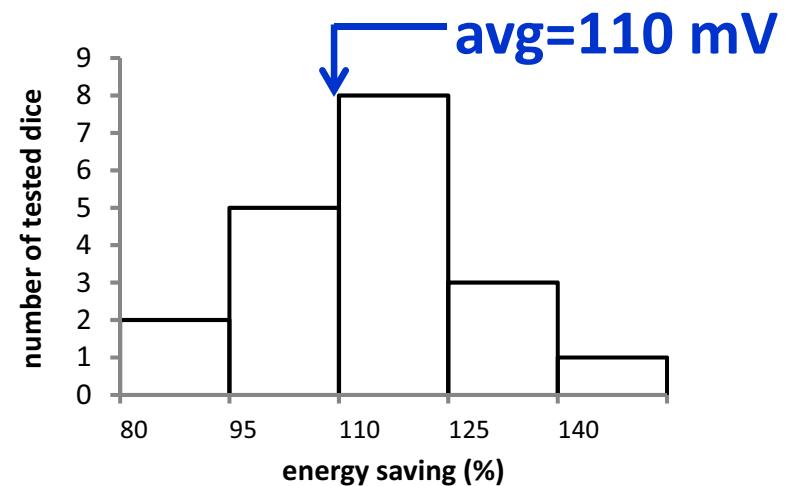
Measurements across Multiple Dice

- ◆ Voltage reduction across 19 dice
 - Quality target: PSNR=30 dB

WRITE CRITICAL CASE



READ CRITICAL CASE

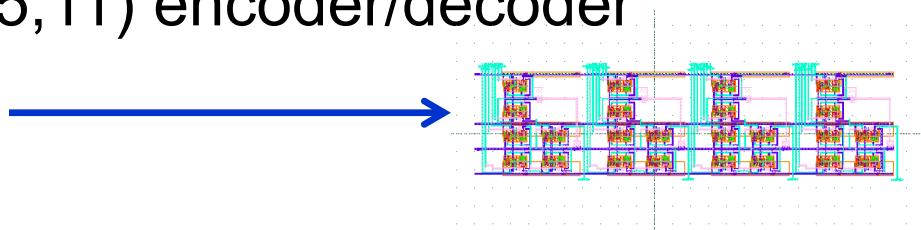


- V_{DD} scaled down by 200 mV (110 mV) at iso-quality
- More aggressive scaling @ energy-optimal PSNR

Overhead

- ◆ Selective NBL overhead
 - Latches to store NBL configuration (*boost* for 8 columns) + 2 transistors per column
- ◆ Selective ECC overhead
 - 15-b ECC Hamming(15,11) encoder/decoder
 - 24 XOR gates each
- ◆ Area overhead:

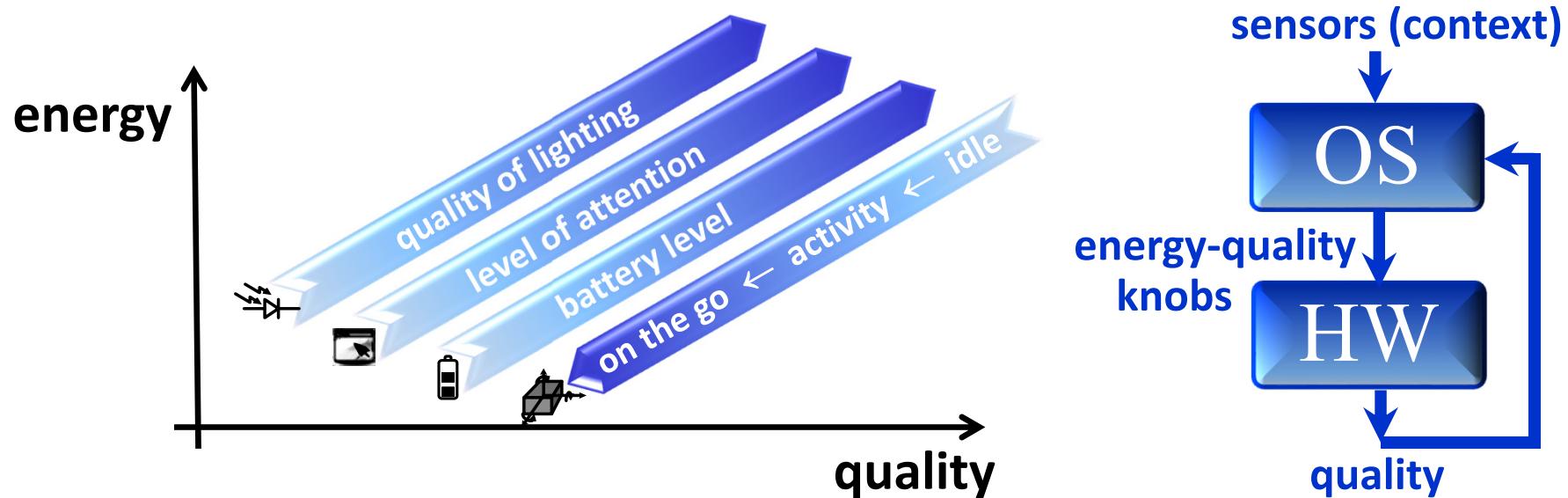
	area [$\mu\text{m} \times \mu\text{m}$] (%)
32-kb SRAM	252 X 202 (100%)
selective NBL	166 (0.3%)
ECC encoder	27.8 X 9.4 (0.5%)
ECC decoder	24.2 X 14.7 (0.7%)



**area overhead:
1.5%**

Broader View on Dynamic Energy-Quality Tradeoff

- ◆ Context-aware dynamic adjustment of quality
 - Quality target set by the context (Operating System)



- Minimize energy for quality target at ckt level
- Expose measured quality (or bit-level BER) to OS

Conclusion

- ◆ SRAM for both error-free and error-tolerant applications
 - Flexible, low design effort (6T)
 - Area overhead: 1.5% (scalable IP for error-free/tolerant apps)
- ◆ Energy-quality tradeoff dynamically managed
 - Non-uniform (bit-level) knobs: use energy only in important bits
 - More graceful degradation enables more V_{DD} scaling
 - Selective NBL and ECC (better than bit dropping)
- ◆ Testchip in 28nm (19 dice)
 - 35% energy reduction at iso-quality (error-tolerant mode)
 - Voltage downscaled by 220 mV at iso-quality (error-tolerant mode)

Acknowledgements

- ◆ The authors wish to thank STMicroelectronics for chip fabrication
- ◆ Work supported in part by
 - the NSF Variability Expedition
 - DARPA (agreement HR0011-13-2-0006)