
# Welcome to ISSCC 2013

# SESSION 18

## Advanced Embedded SRAM

# A 20nm 112Mb SRAM Design in High K/Metal Gate Technology with Assist Circuitry for Low Leakage and Low Vmin Applications

Jonathan Chang, Yen-Huei Chen, Hank Cheng, Wei-Min Chan, Hung-Jen Liao, Quincy Li, Stanley Chang, Sreedhar Natarajan, Robin Lee, Ping-Wei Wang, Shyue-Shyh Lin, Chung-Cheng Wu,

Kuan-Lun Cheng, Min Cao, George H. Chang

tsmc ®

# Outline

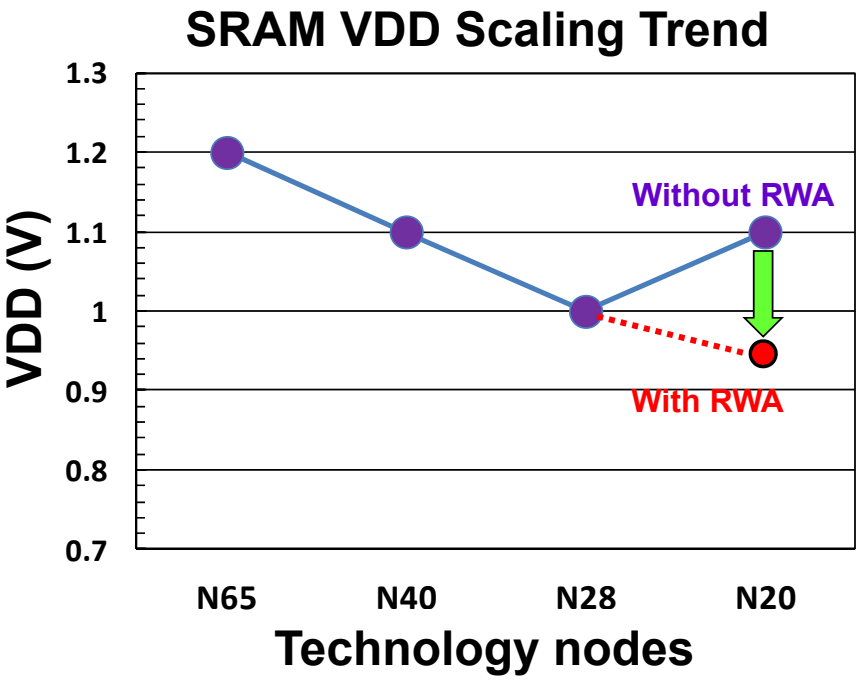
- Motivation
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - Partial Suppressed Word-Line (PSWL)
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- Power management
- Silicon results
- Summary

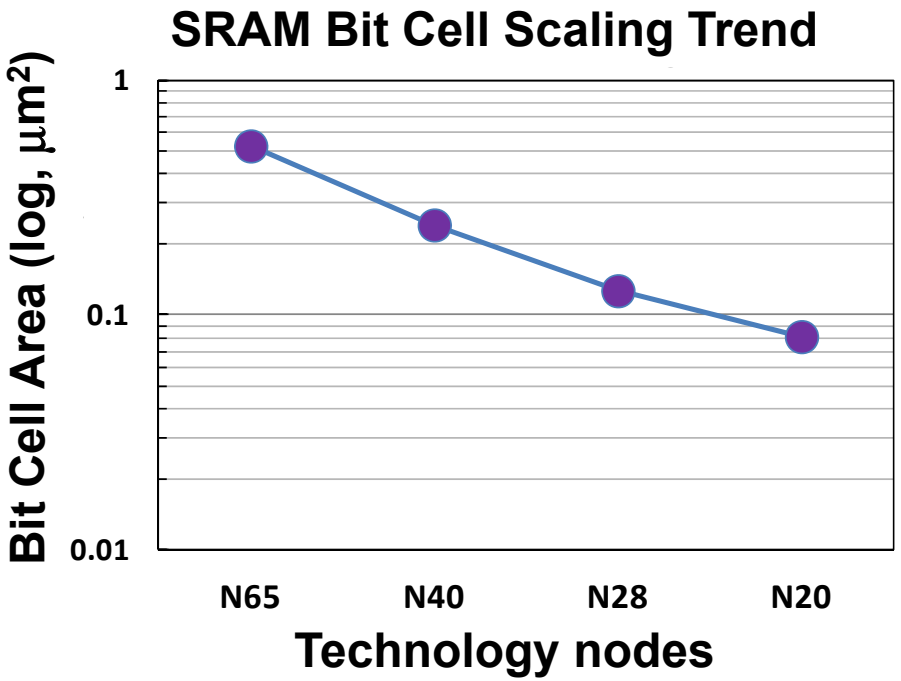
# Outline

- **Motivation**
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - Partial Suppressed Word-Line (PSWL)
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- Power management
- Silicon results
- Summary

# Technology Trend

- 50% cell scaling from 65nm to 28nm
- 60% cell scaling from 28nm to 20nm
- With assist circuitry, VDD can continue to be lowered for 20nm HD cell

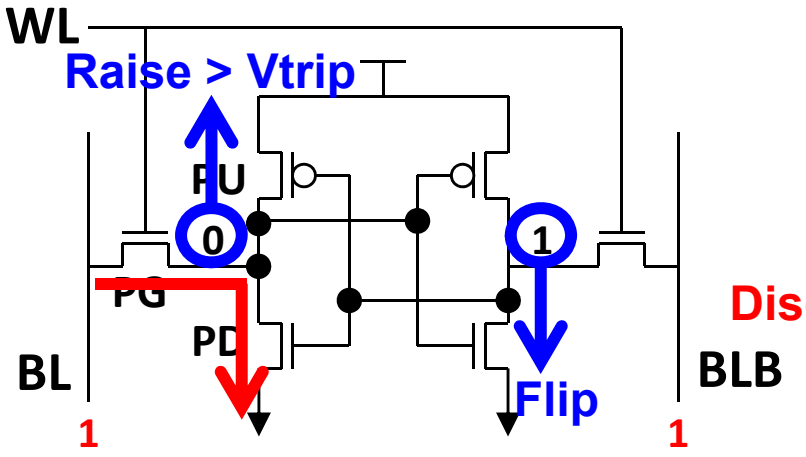


# Motivation (1)

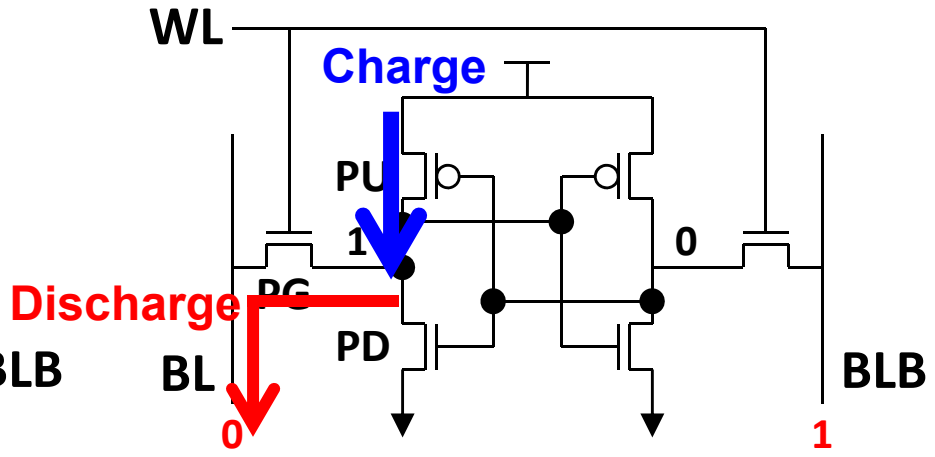
- **Low SRAM Vmin is critical for mobile applications**

***Read operation***




***Read failure: "0" node raises too high***
***Solution: Word-Line Under Drive (WLUD)***

***Write operation***

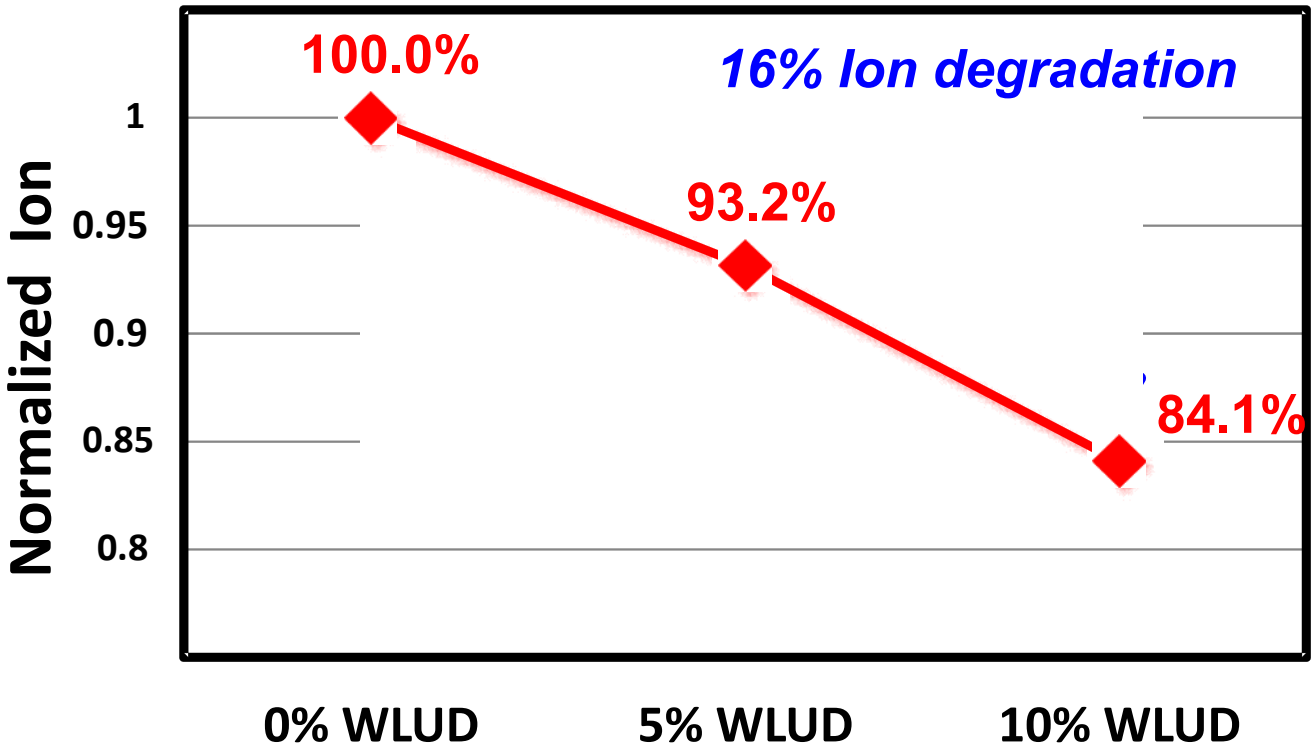



***Write failure: PG loses the fighting***
***Solution: Negative Bit-Line (NBL)***

**Key point: adjust PG driving strength by changing $V_{GS}$**

# Motivation (2)

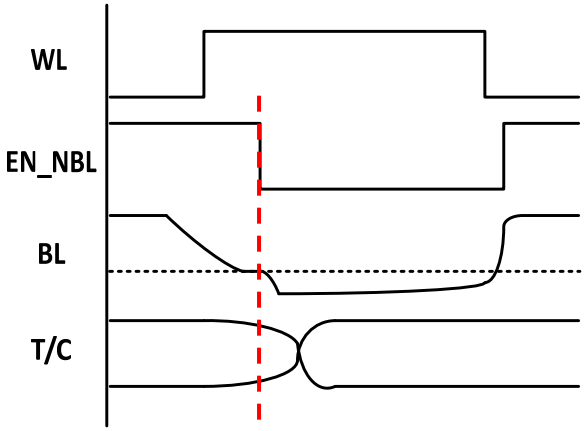
- **Speed penalty by WLUD**
  - **More than 10% SRAM Ion degradation**



# Motivation (3)

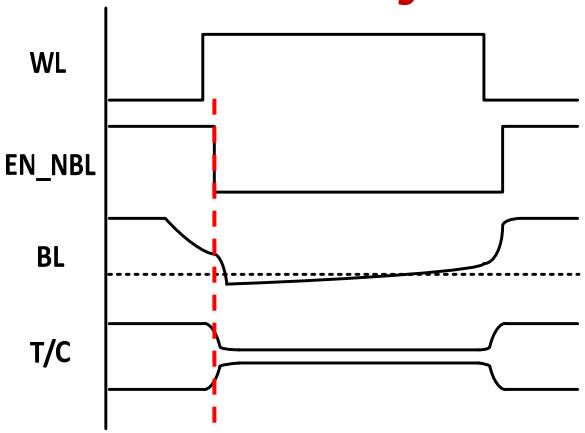
- **Coupling time crucial for the efficiency of NBL**

***Ideal***



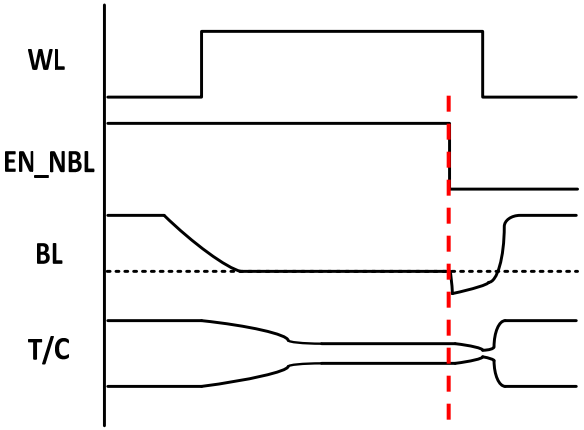

***BL coupled low after pulled to ground by write buffer***

***Too early***




***BL coupled low before pulled to ground by write buffer, smaller coupling voltage***

***Too late***




***BL coupled low near WL turned off, not enough time to flip bit cell content***

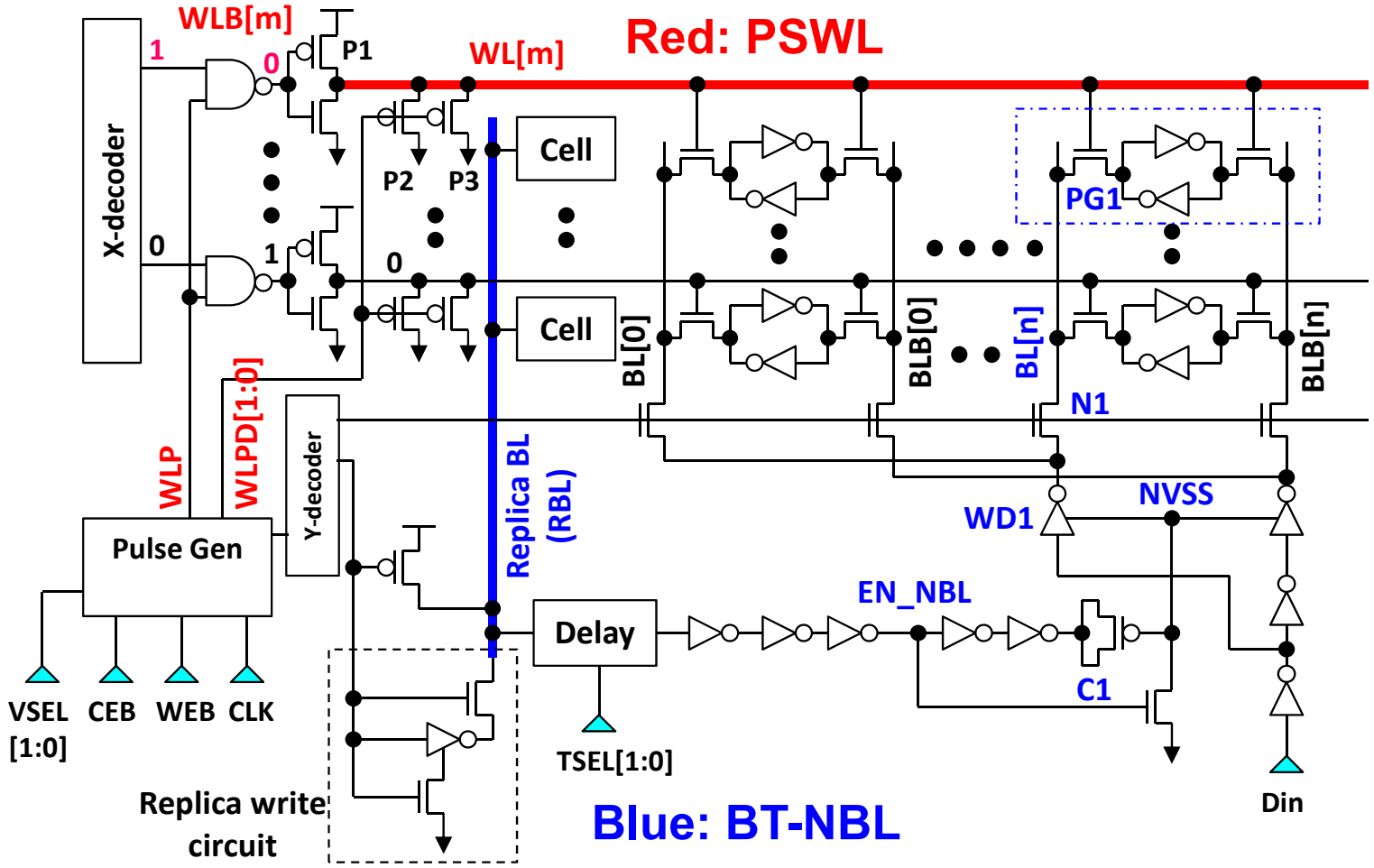
# Outline

- Motivation
- **Proposed low Vmin techniques**
  - **Proposed Read-Write-Assist Scheme**
  - Partial Suppressed Word-Line (PSWL)
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- Power management
- Silicon results
- Summary



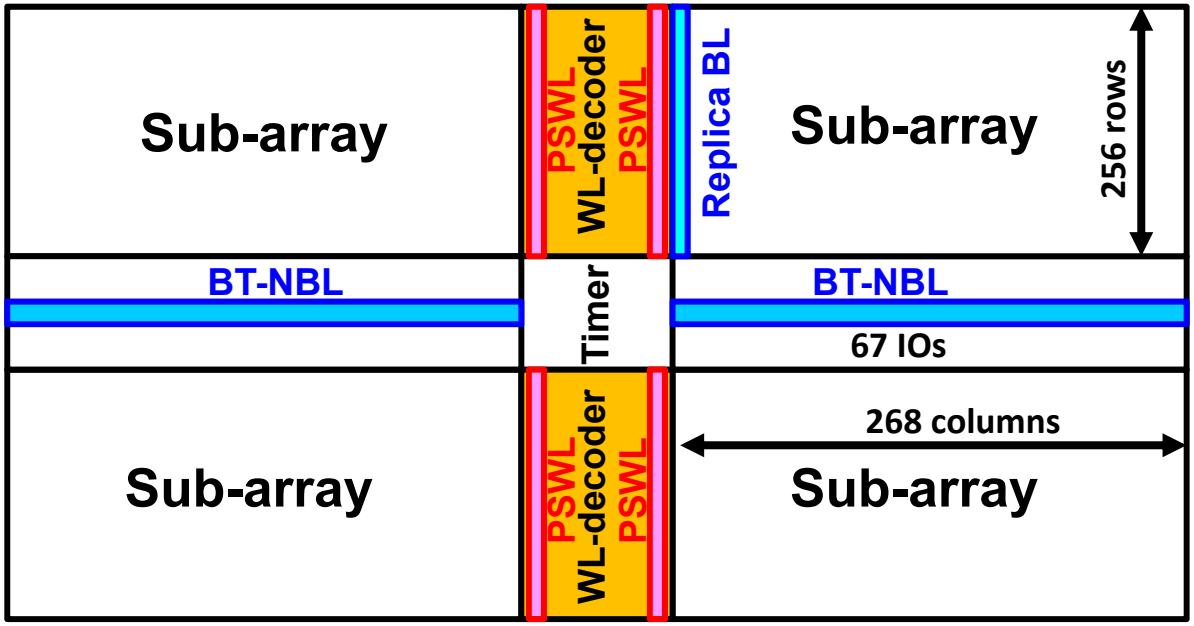
© 2013 IEEE

# Proposed Read-Write-Assist Scheme



# 268Kb SRAM Macro Floor Plan

- Array design
  - Self timed WL timing
  - 256bits per BL
  - 268bits per WL
  - ≈8.3Mb per mm$^2$
- Area cost for assist
  - 1.2% for PSWL
  - 3.7% for BT-NBL
    - Shrunk to 2.9% with optimized boost cap





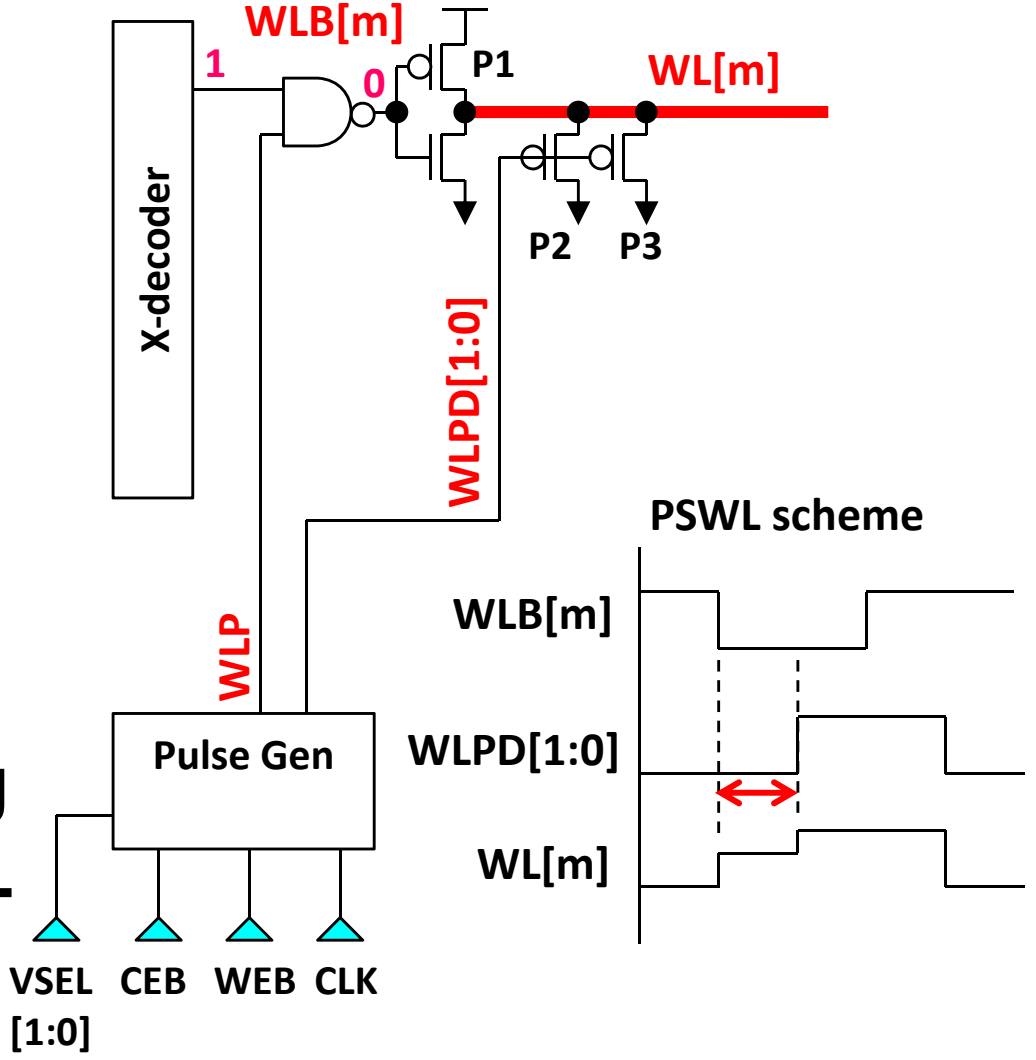

© 2013 IEEE

# Outline

- Motivation
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - **Partial Suppressed Word-Line (PSWL)**
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- Power management
- Silicon results
- Summary

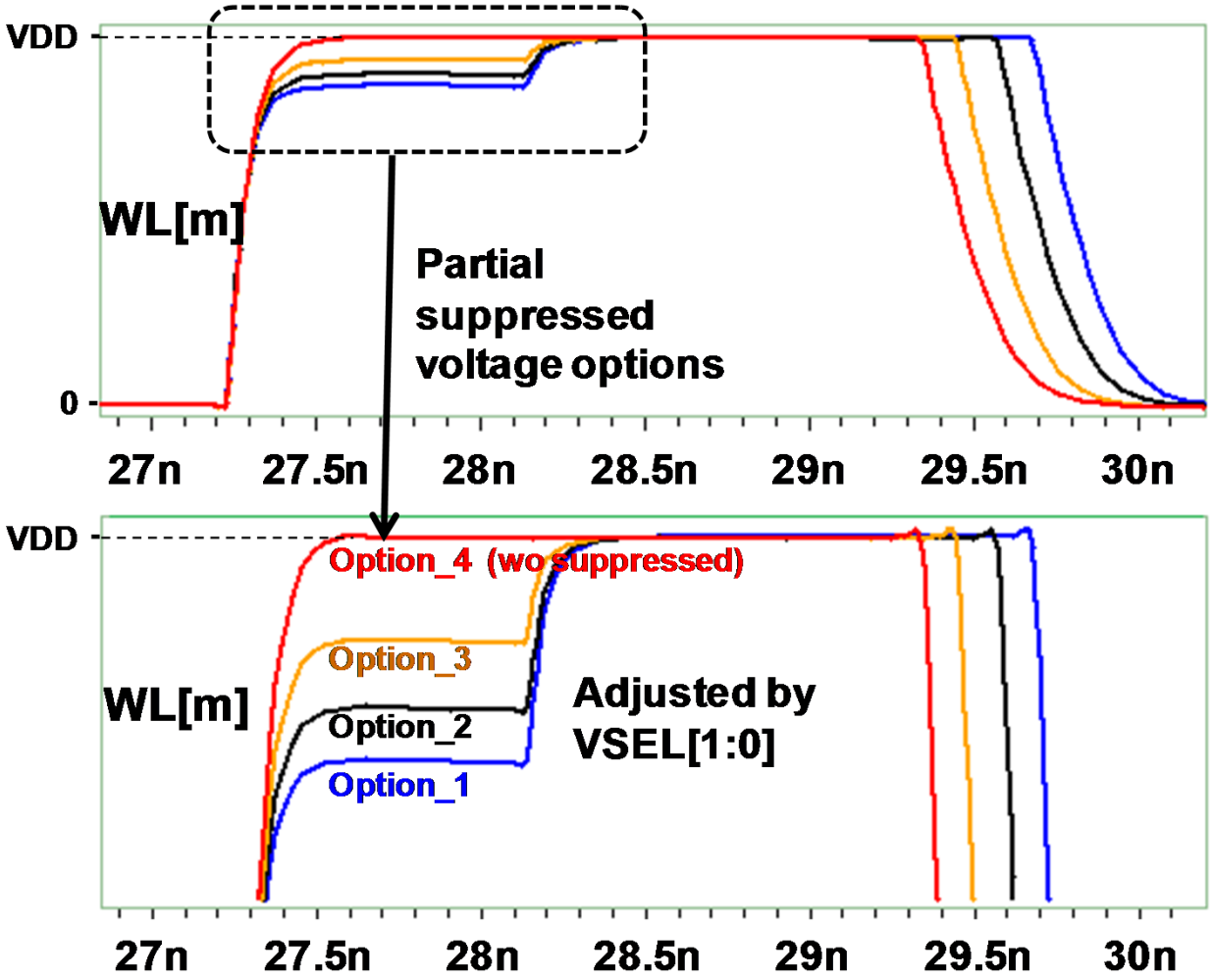
# PSWL Circuitry

- PMOS-PMOS voltage divider to minimize impact from process variation
- WL voltage level is determined by the resistance ratio of P1 and (P2, P3)
- Programmable timing options to control WL under drive duration



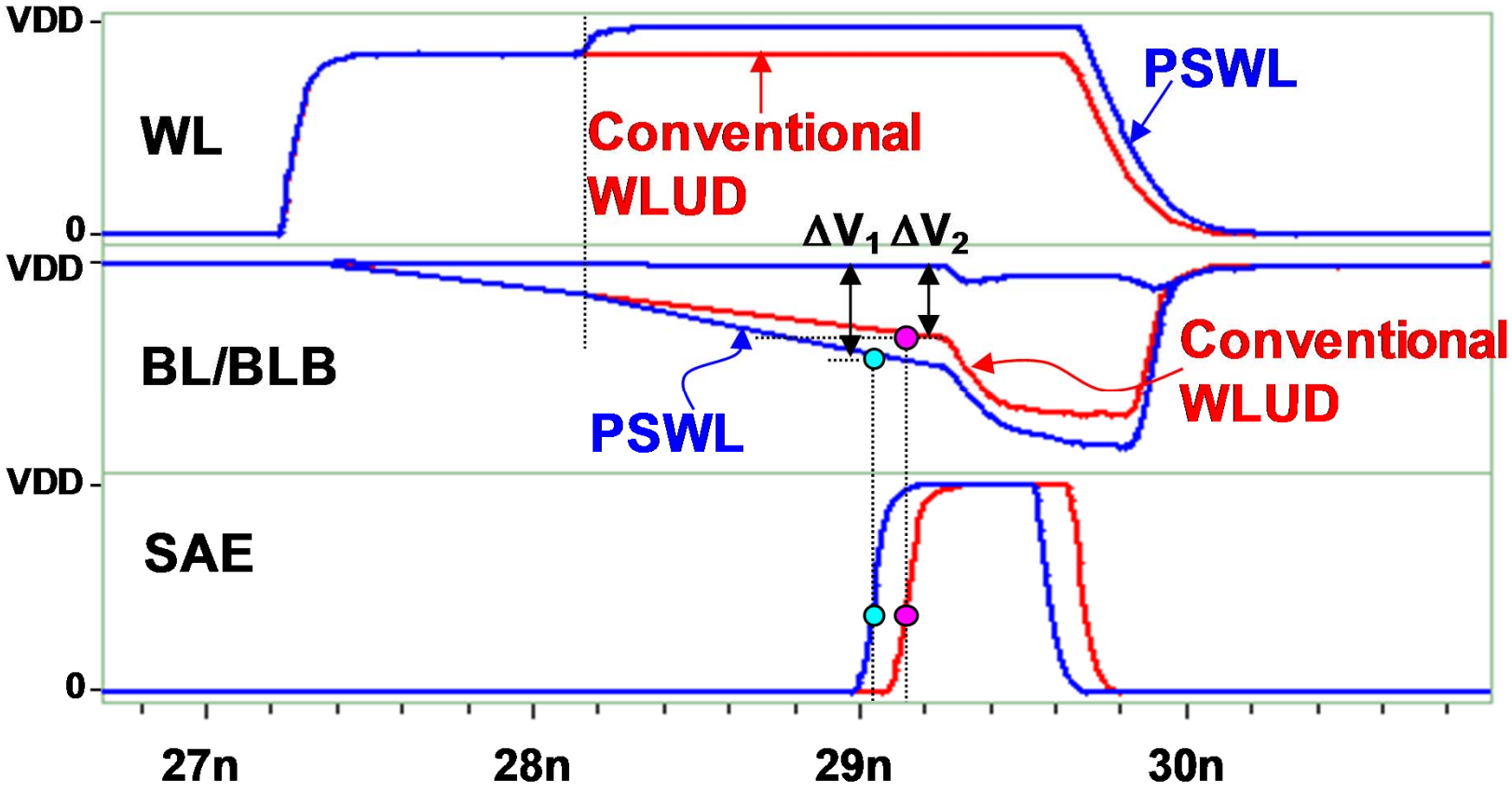
# PSWL Simulation Waveform (1)

- Programmable voltage options



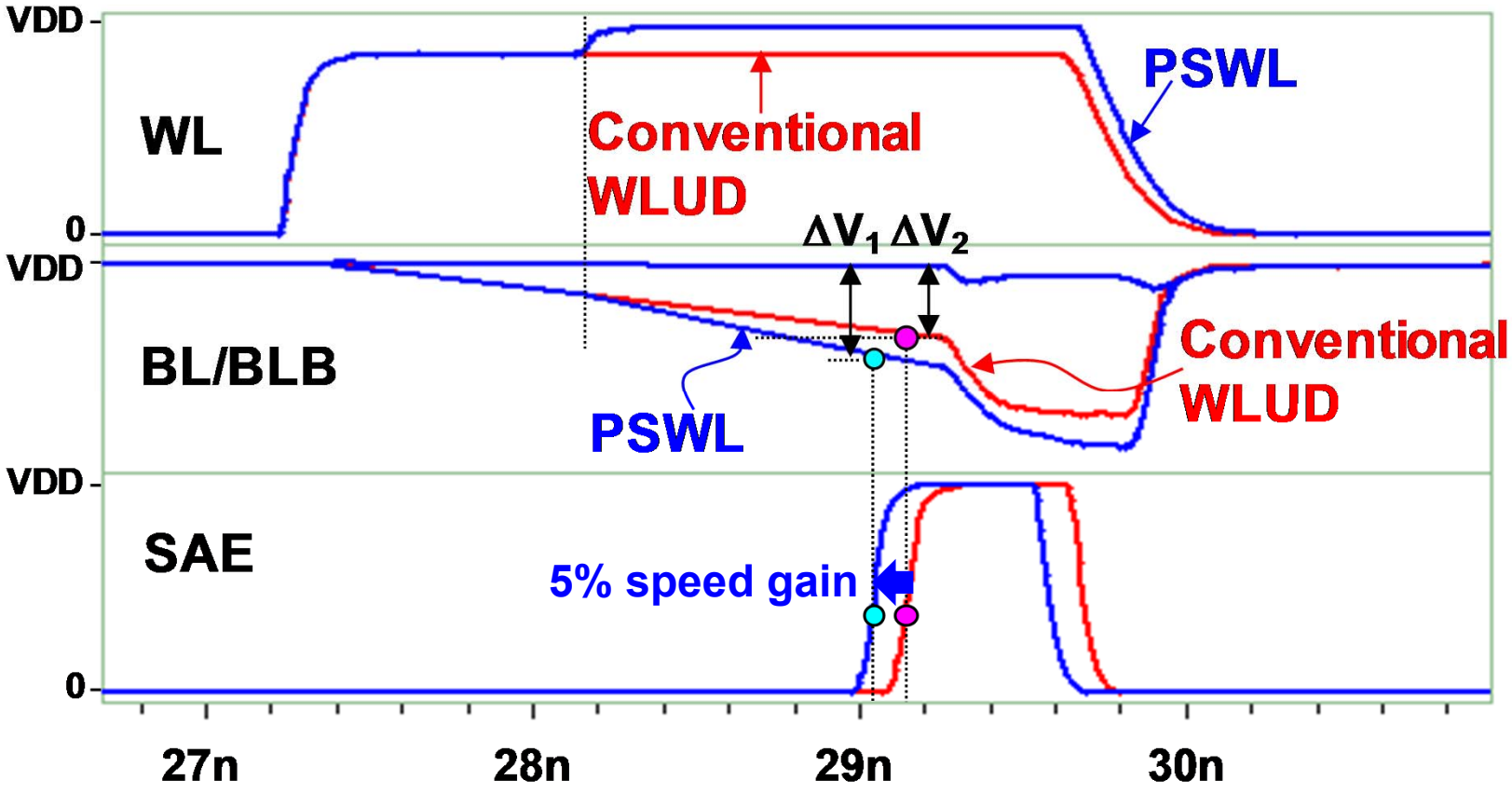
# PSWL Simulation Waveform (2)

- BL development is faster for PSWL
- Can pull in SAE to mitigate speed penalty



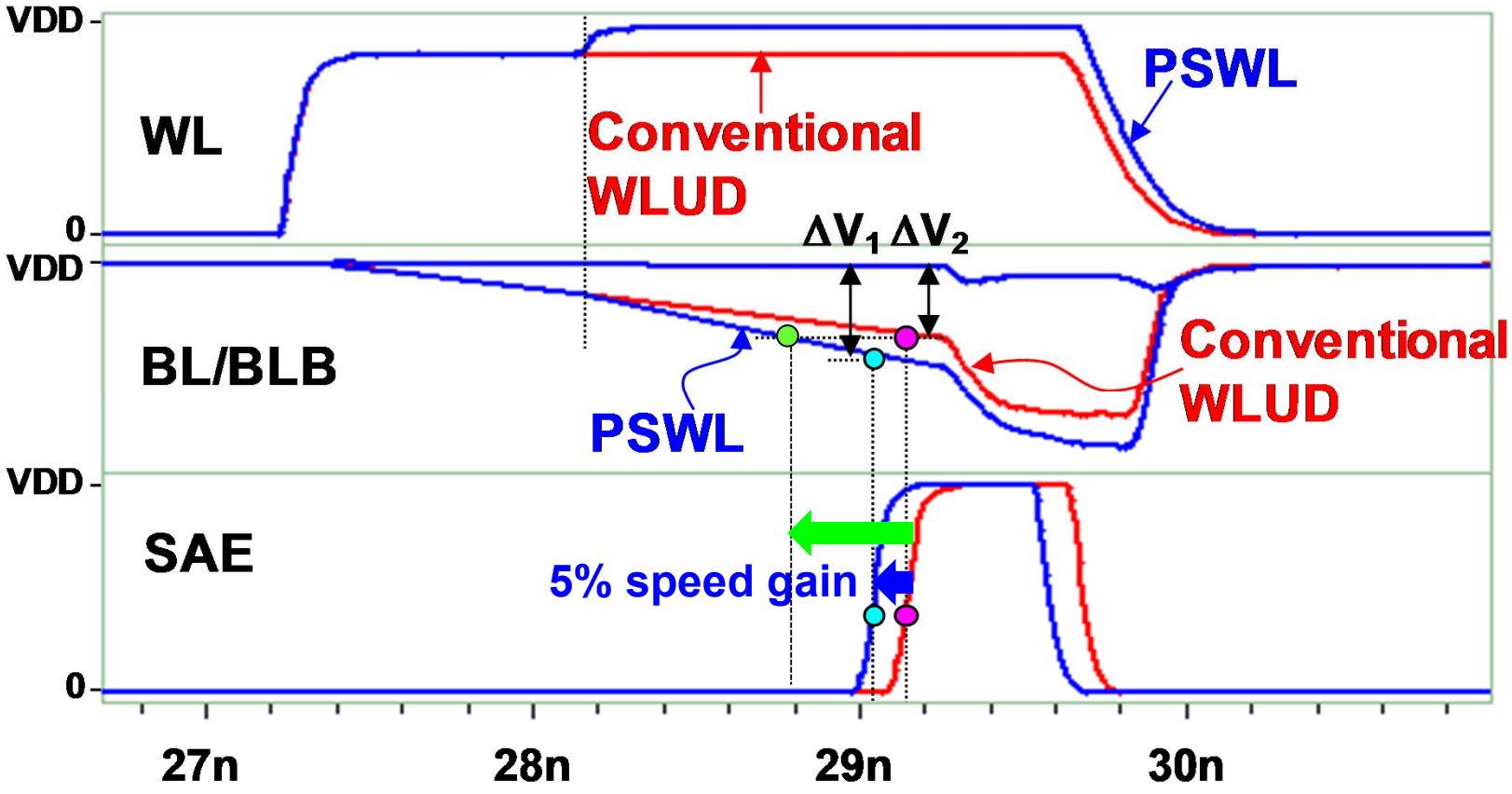
# PSWL Simulation Waveform (2)

- BL development is faster for PSWL
- Can pull in SAE to mitigate speed penalty



# PSWL Simulation Waveform (2)

- BL development is faster for PSWL
- Can pull in SAE to mitigate speed penalty

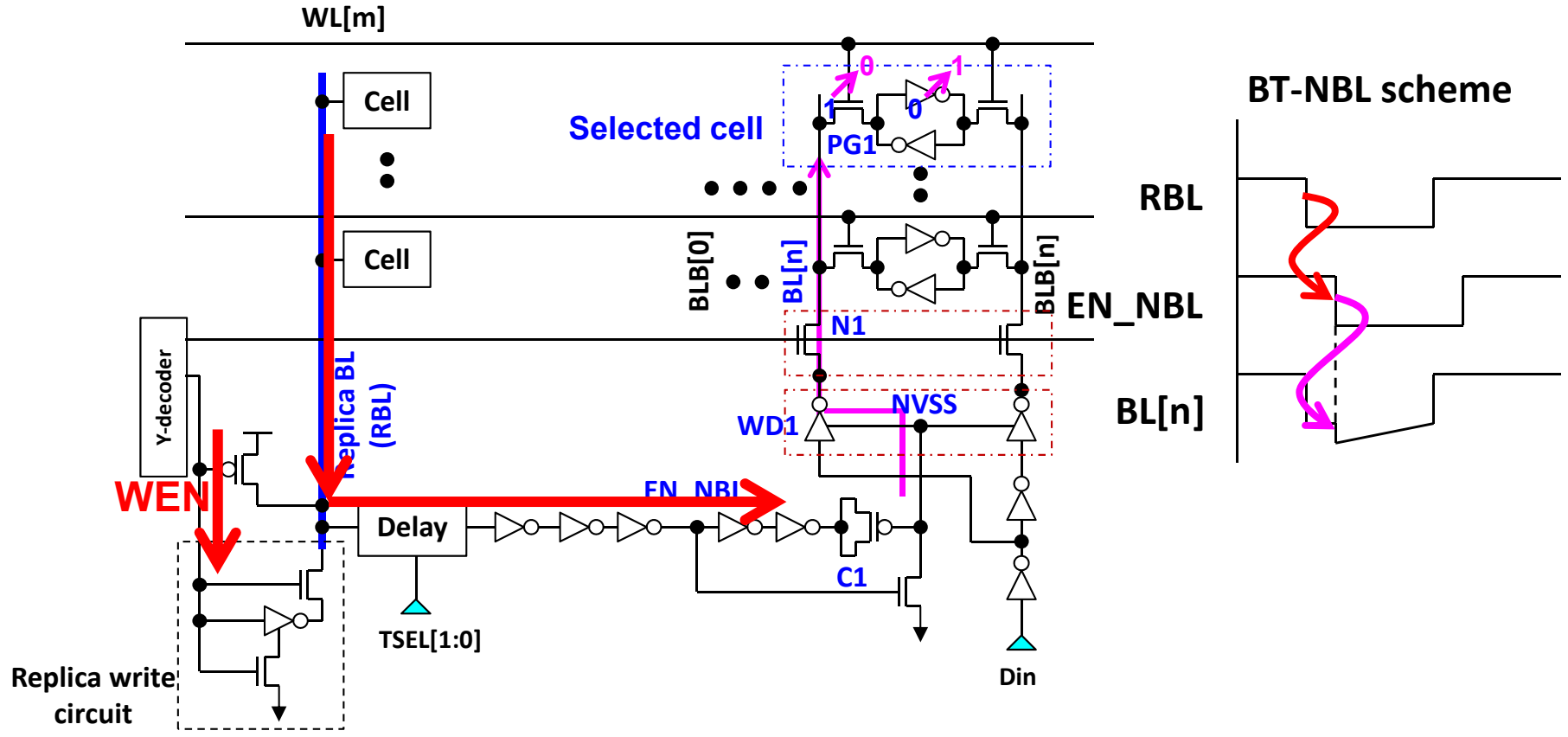


# Outline

- Motivation
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - Partial Suppressed Word-Line (PSWL)
  - **Bit-Line Length Tracked Negative Bit-Line (BT-NBL)**
- Power management
- Silicon results
- Summary

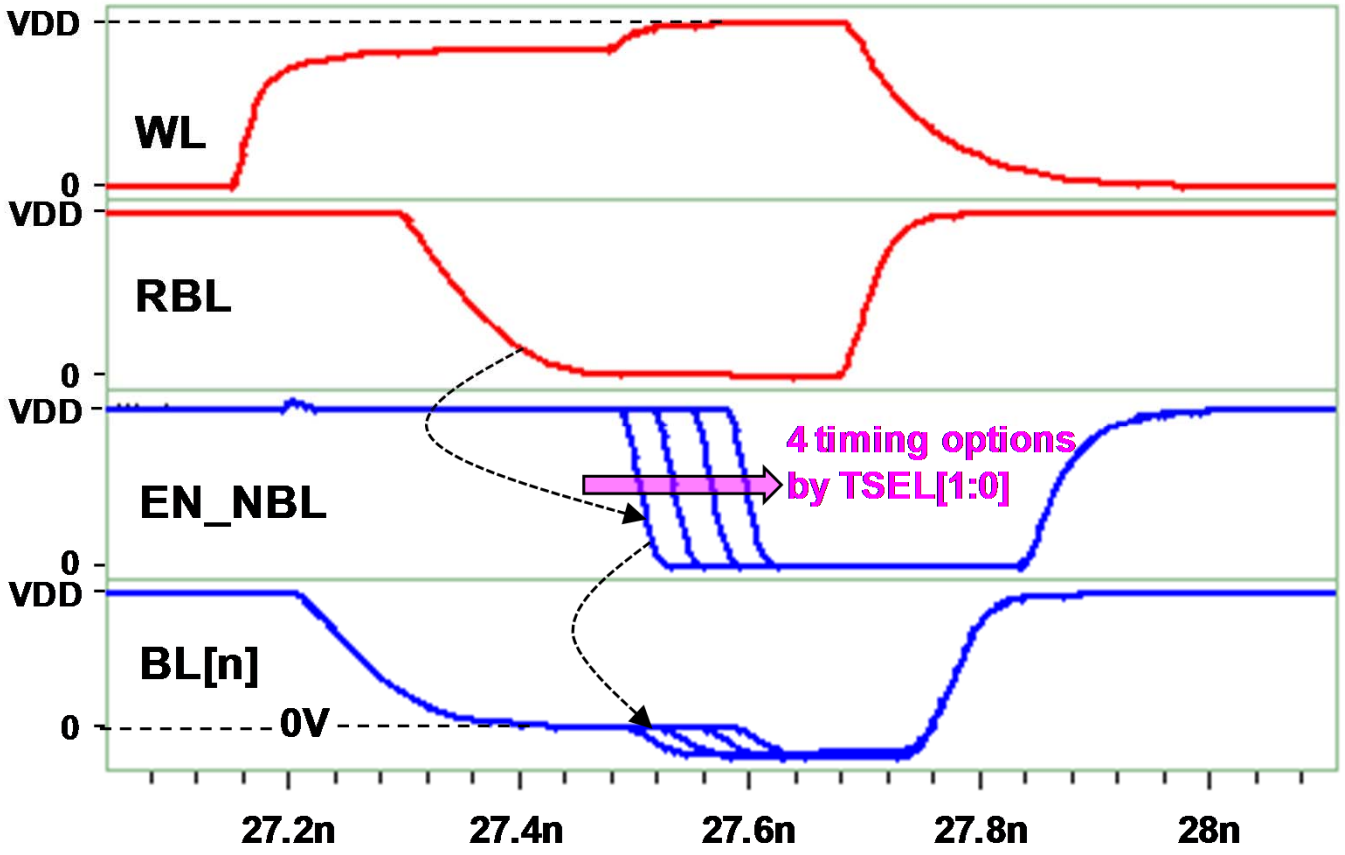
# BT-NBL Circuitry

- Key elements to decide suitable coupling time
  - Replica BL to track BL length
  - Replica write circuit to track write buffer strength



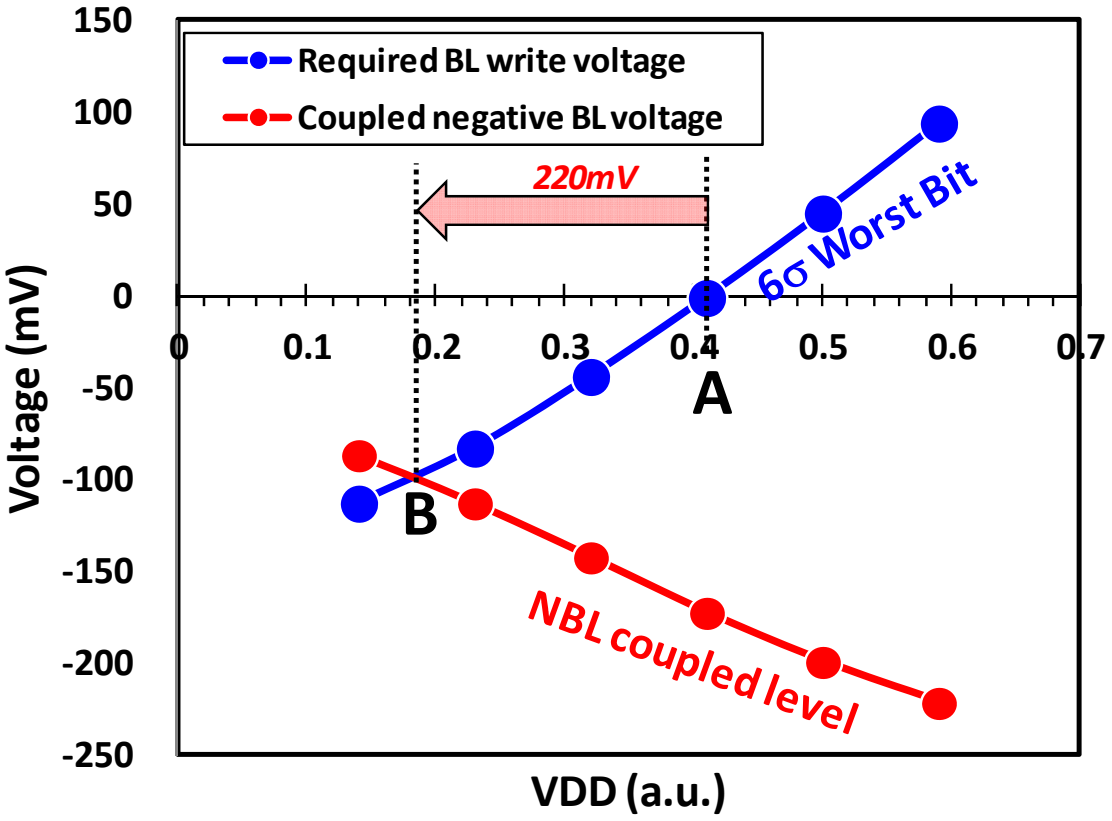
# BT-NBL Simulation Waveform

- Programmable timing options allows post silicon tuning to achieve optimal write Vmin



# Simulated Write Vmin Improvement

- Point A: intrinsic write Vmin
- Point B: improved write Vmin by BT-NBL

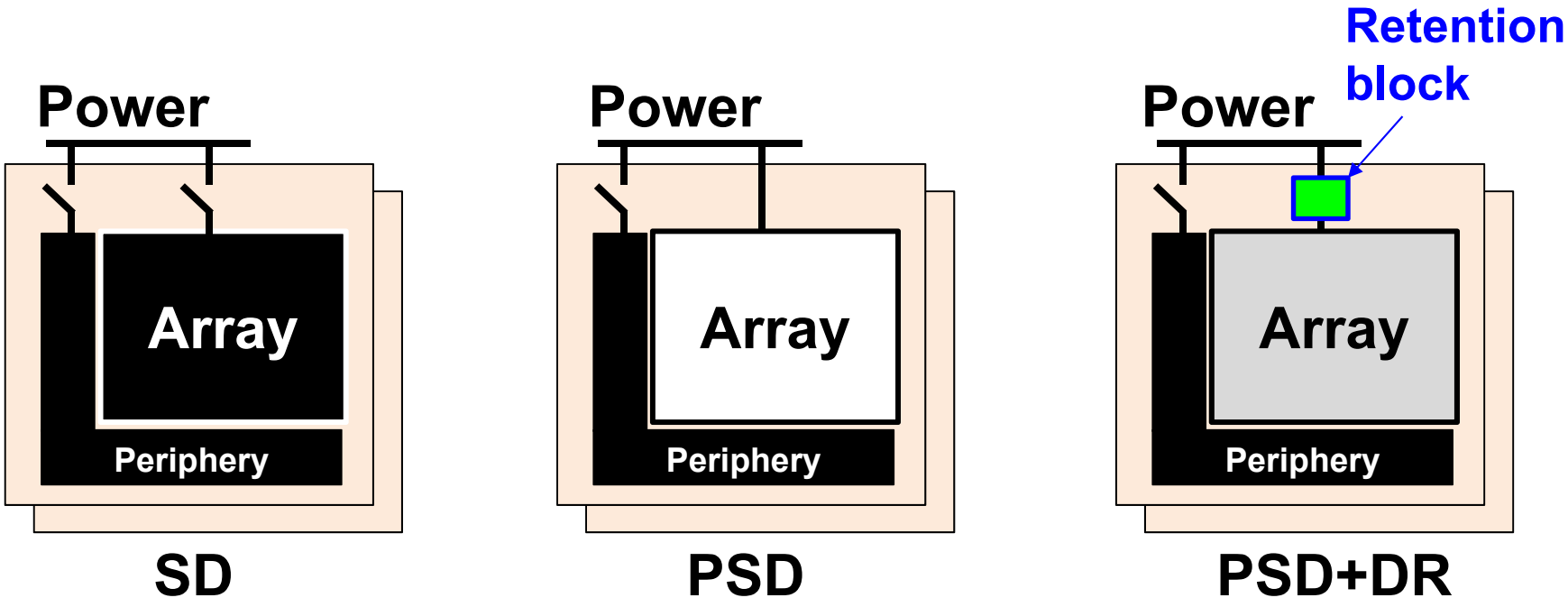


# Outline

- Motivation
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - Partial Suppressed Word-Line (PSWL)
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- **Power management**
- Silicon results
- Summary

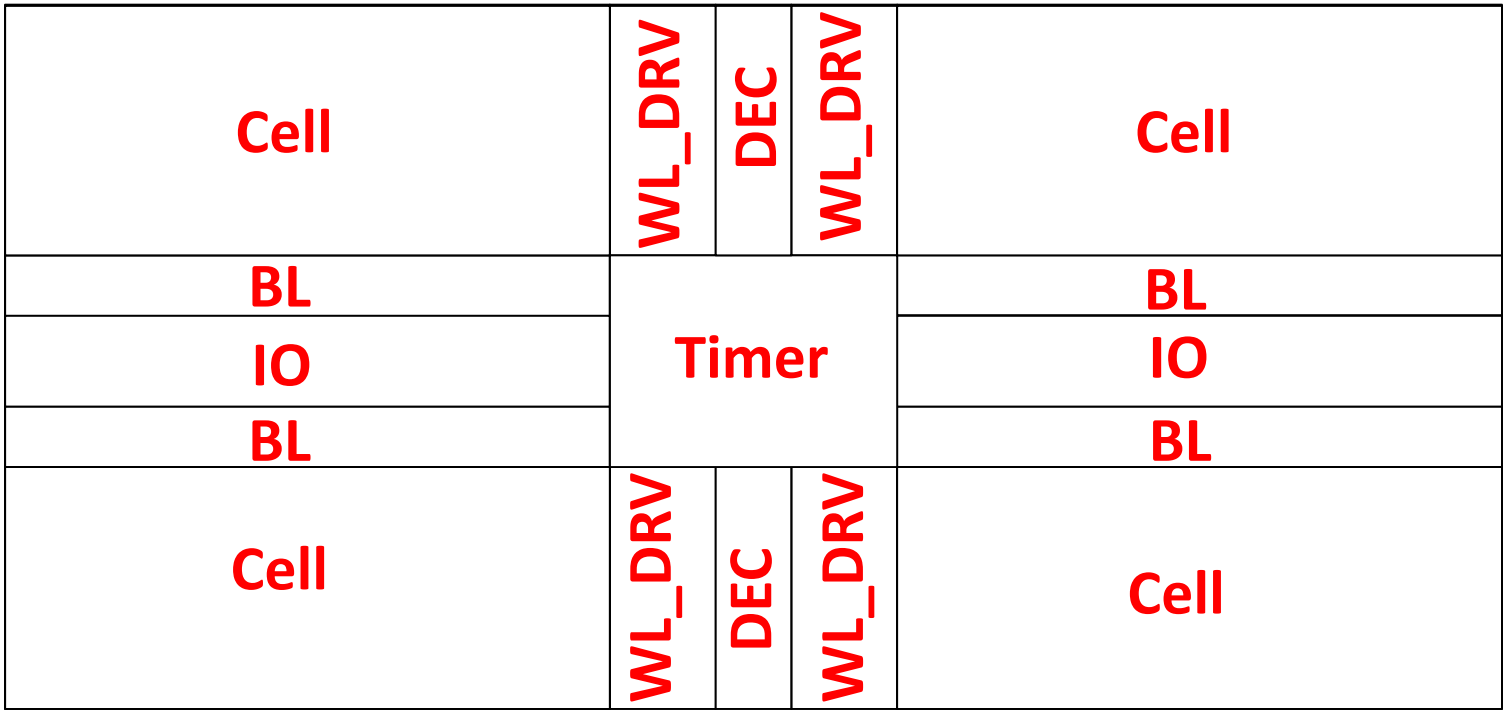
# Power Management

- Shutdown (SD)
- Peripheral shutdown (PSD)
- Data retention (DR)

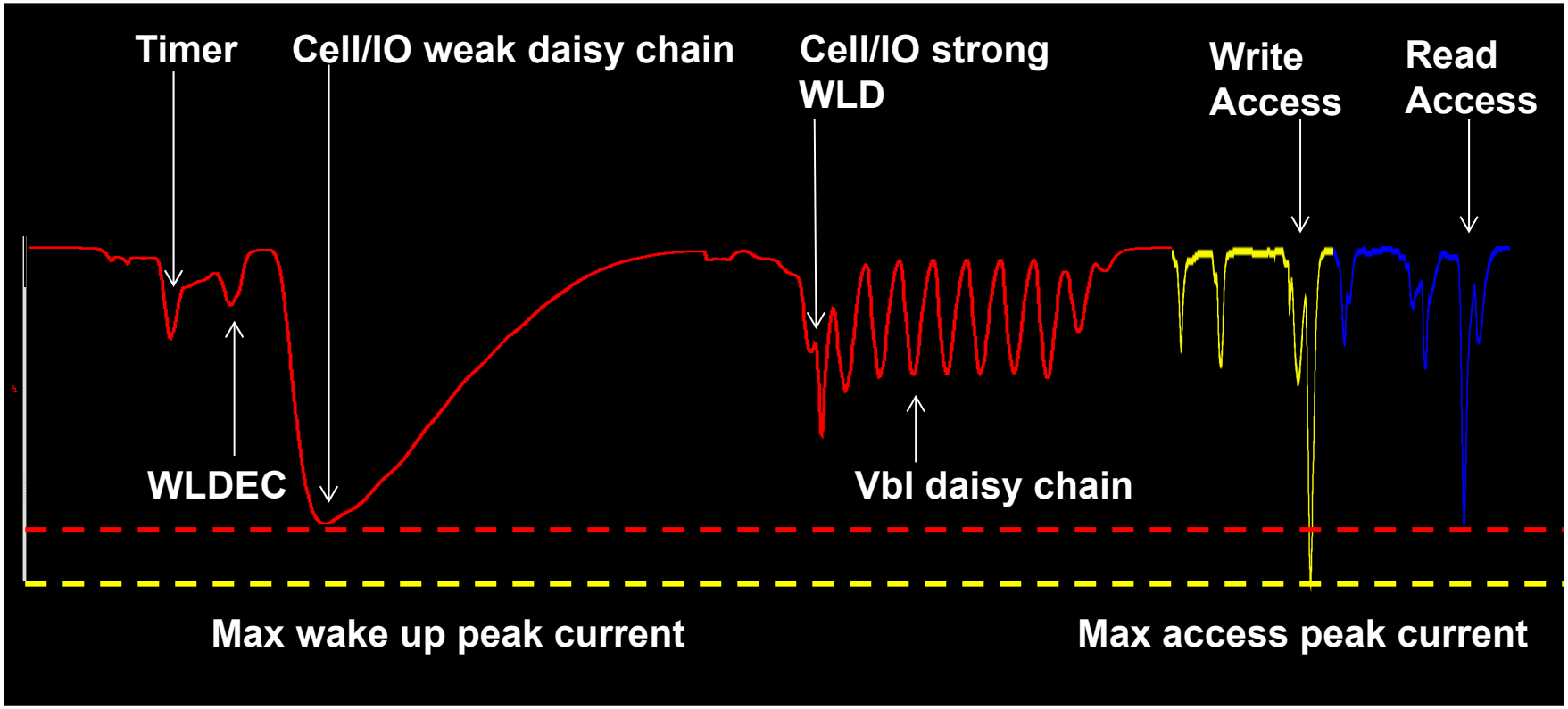


# Wake-up Sequence

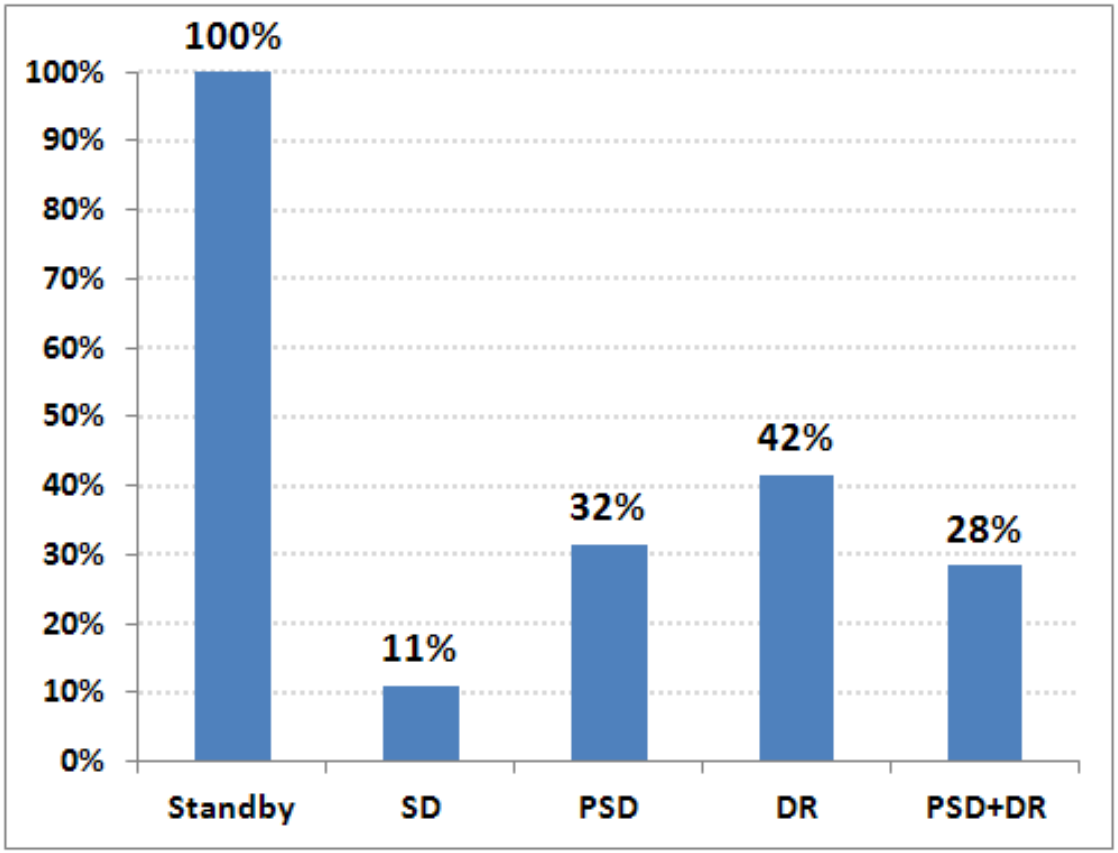
- Gradual wake-up sequence to eliminate di/dt noise



# Wake-up Peak Current



# Standby Power Reduction




Power Management Mode



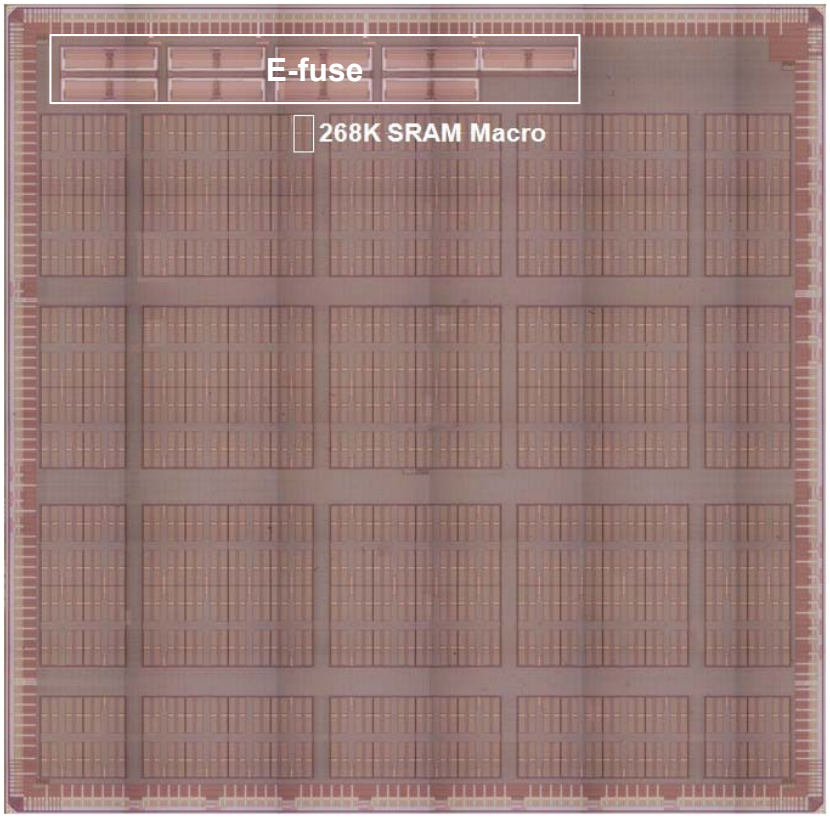
© 2013 IEEE

# Outline

- Motivation
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - Partial Suppressed Word-Line (PSWL)
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- Power management
- **Silicon results**
- Summary

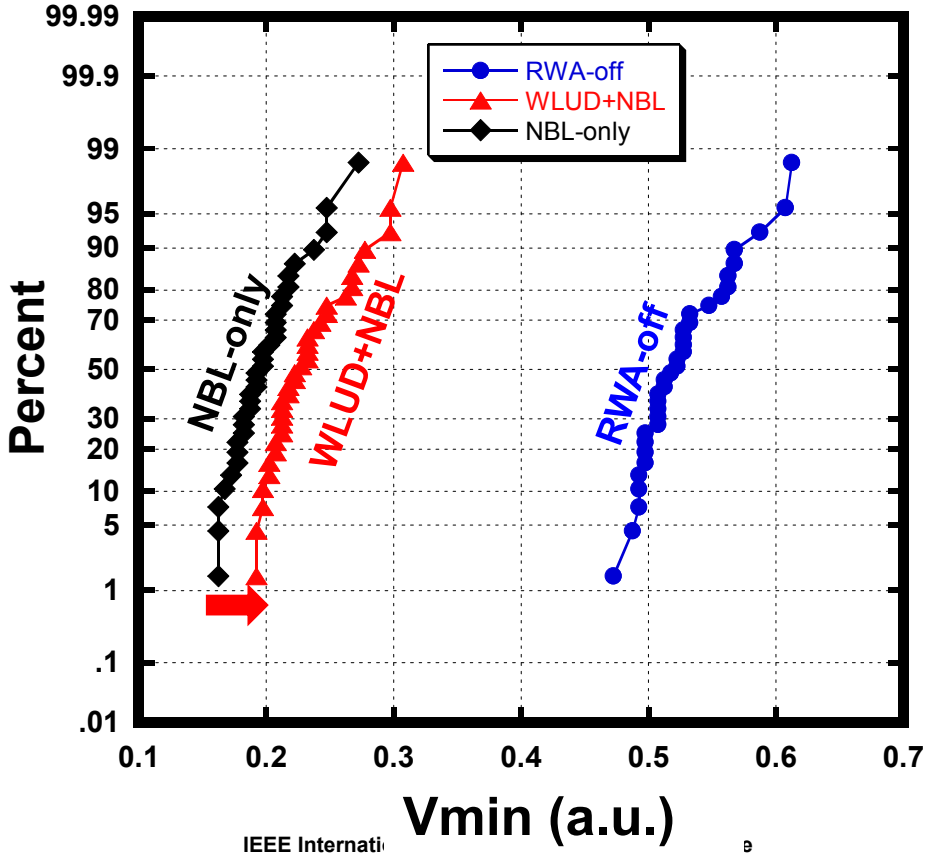
# Chip Information

| | |
|---|---|
| Technology | 20nm HK-MG SOC CMOS |
| Metal scheme | 1-poly / 7-metal |
| Supply | Core: 0.95V<br>IO: 1.8V |
| Bit cell size | 0.081μm$^2$ |
| SRAM macro configuration | 2048x134 MUX-4 |
| SRAM capacity | 112Mb |
| Chip size | 6400μm x 6300μm<br>=40.3mm$^2$ |



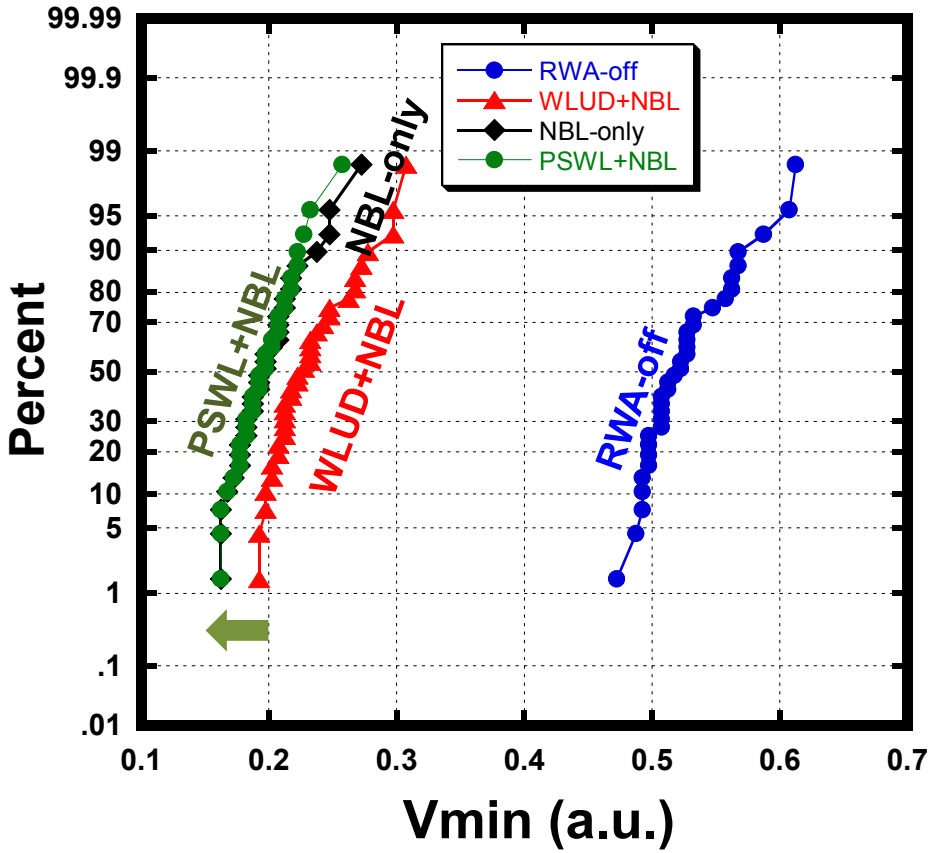
# Vmin Improvement from Write Assist

- NBL significantly improves write Vmin
- WLUD degrades the write assist capability due to under drive the pass-gate of SRAM cell



# Write Vmin Improvement from PSWL

- PSWL can mitigate the write capability degradation from WLUD

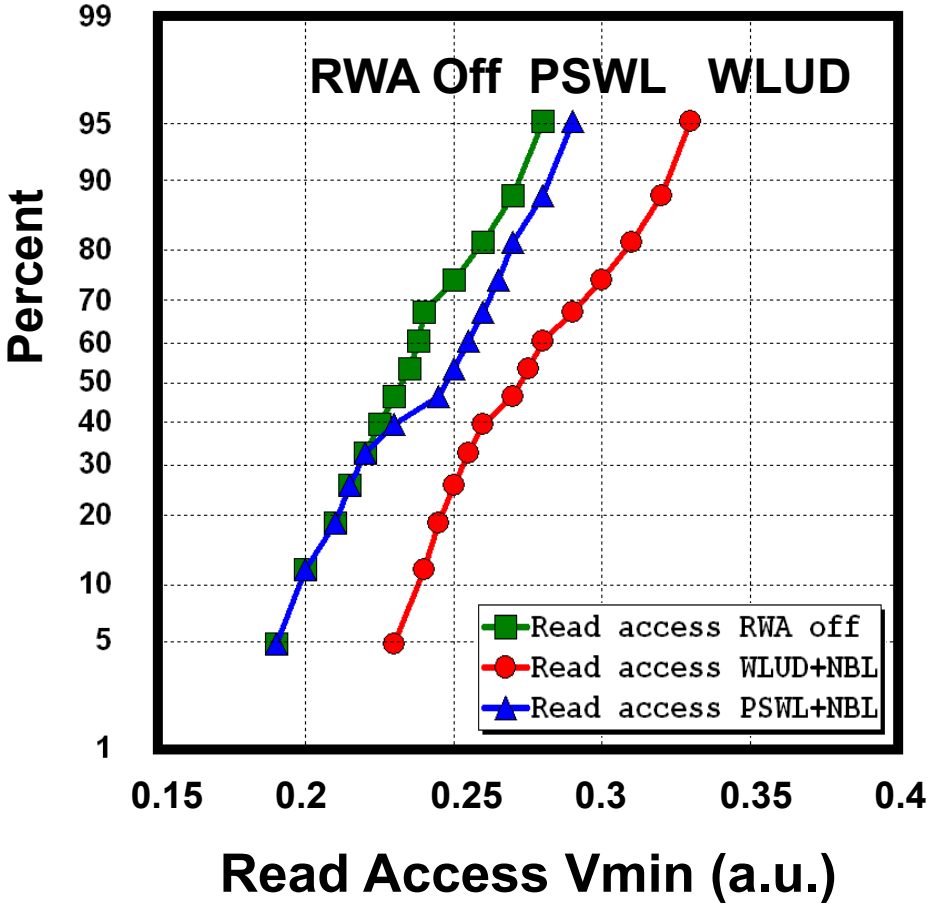




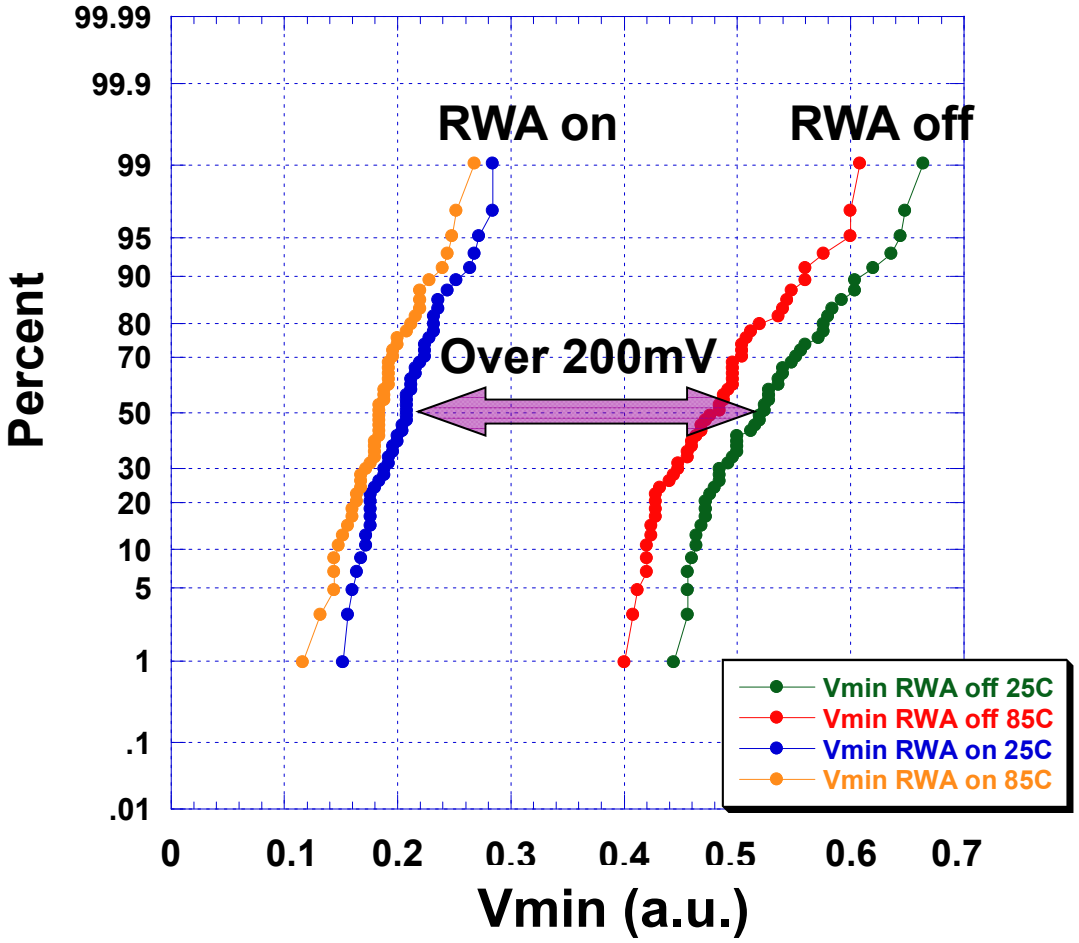

© 2013 IEEE

# Read Access Vmin Improvement from PSWL

- PSWL can reduce the read access Vmin impact from read assist (WLUD)



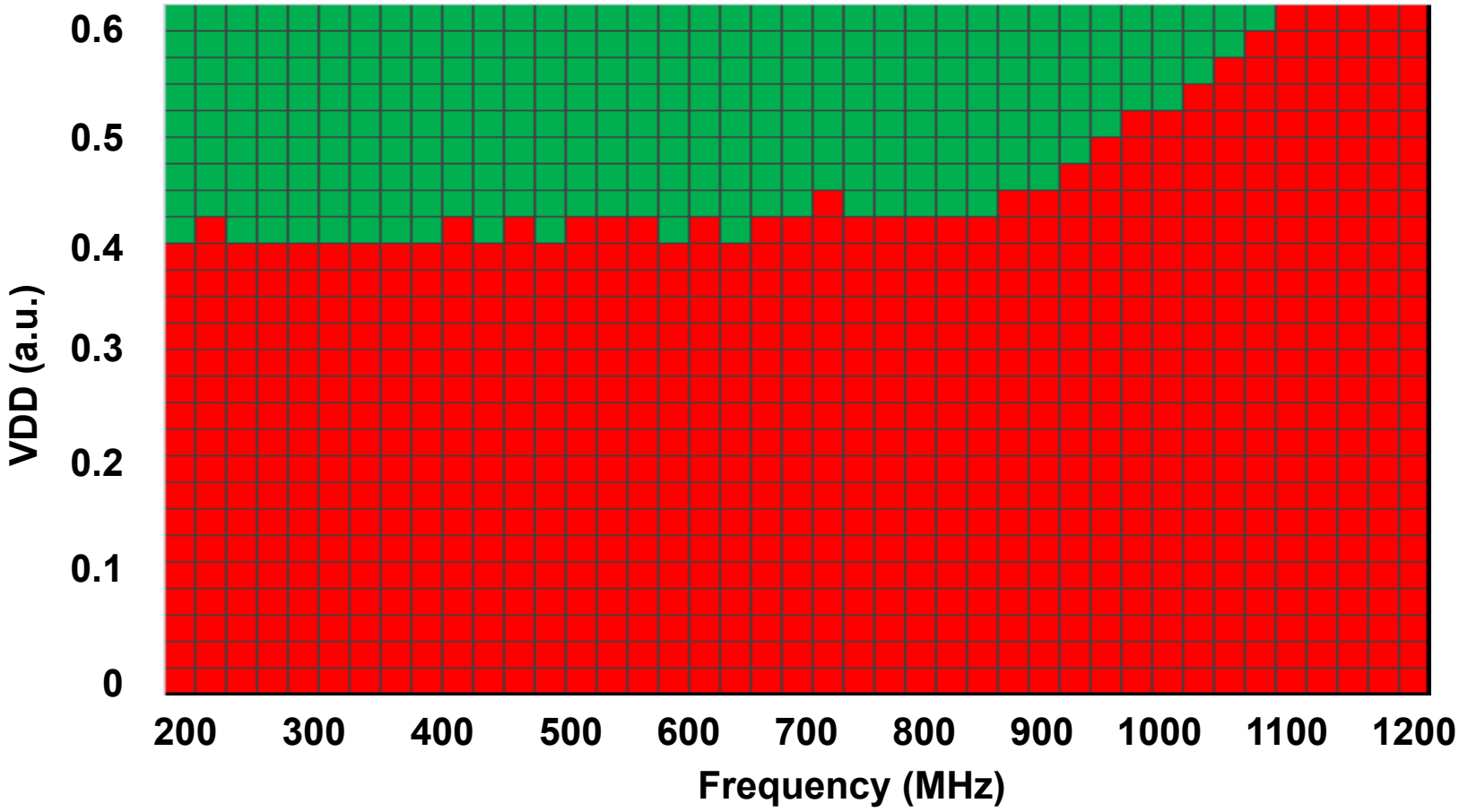
# Vmin Improvement for 112Mb SRAM



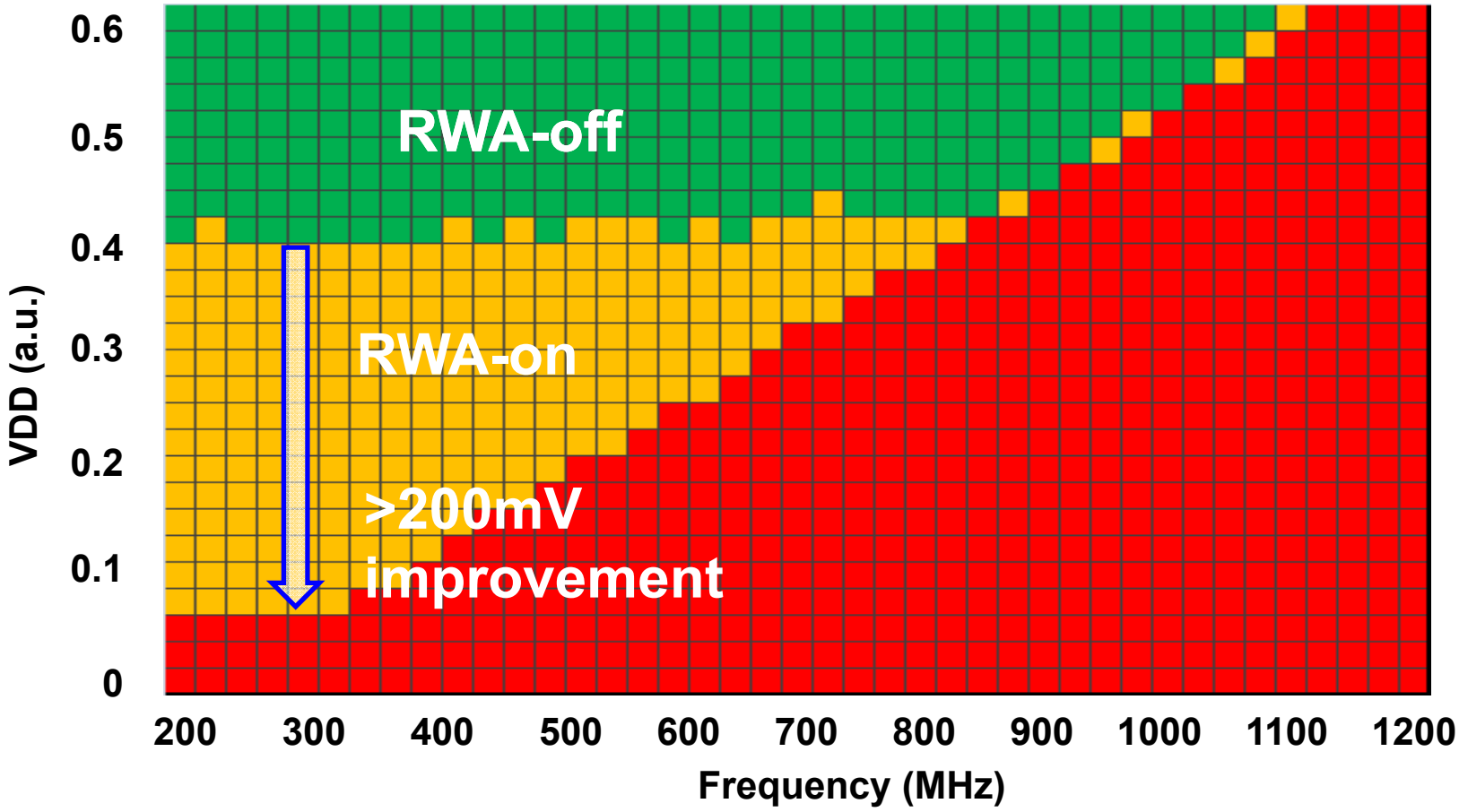



© 2013 IEEE

# Shmoo with RWA Off



# Shmoo with RWA On



# Outline

- Motivation
- Proposed low Vmin techniques
  - Proposed Read-Write-Assist Scheme
  - Partial Suppressed Word-Line (PSWL)
  - Bit-Line Length Tracked Negative Bit-Line (BT-NBL)
- Power management
- Silicon results
- **Summary**

# Summary

- Demonstrated a fully functional 0.081$\mu m^2$ bit cell with proposed read-write assist scheme on 20nm 112Mb SRAM test chip
  - PSWL for read assist
  - BT-NBL for write assist
  - The area penalty is 1.2% for PSWL and 3.7% (2.9% with optimized cap) for BT-NBL
  - Overall Vmin improvement over 200mV by read/write assist circuitry
- The proposed fine-grained power management can improve 89% standby power reduction in Shutdown mode

# An SRAM Using Output Prediction to Reduce BL-Switching Activity and Statistically-Gated SA for up to 1.9× Reduction in Energy/Access

Mahmut E. Sinangil[1], Anantha P. Chandrakasan[2]

[1]NVIDIA, Bedford, MA

[2]Massachusetts Institute of Technology



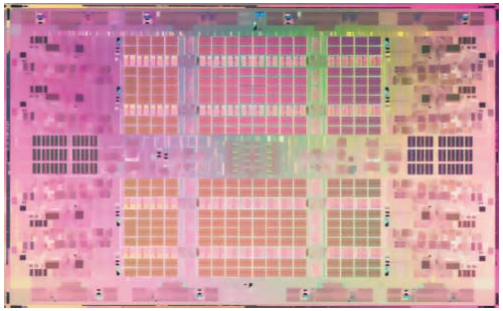


© 2013 IEEE

# Energy-Efficient SRAM Design

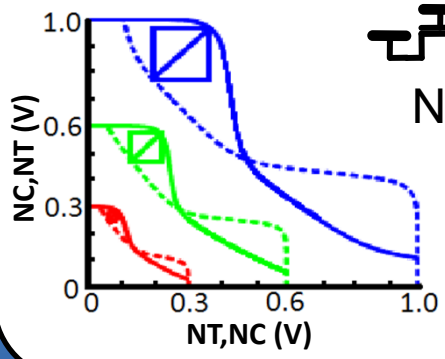
**Continuous increase in on-chip memory capacity necessitates energy-efficient SRAMs**

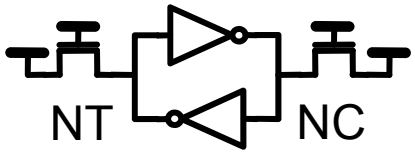
**54MB on-chip cache**



[ISSCC 2011]

## 1) *Voltage Scaling*



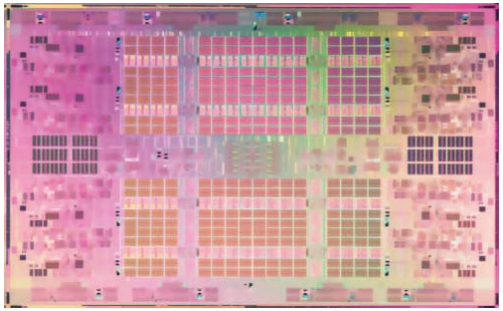

**Smaller noise margins as $V_{DD}$ is scaled down**

# Energy-Efficient SRAM Design

**Continuous increase in on-chip memory capacity necessitates energy-efficient SRAMs**

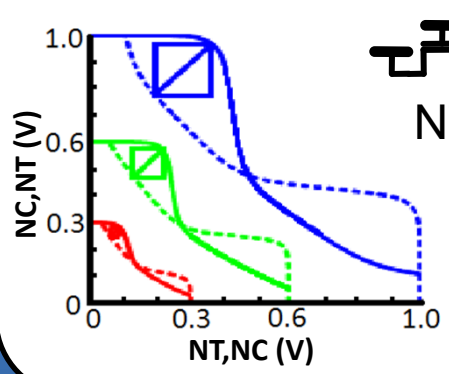
**App-specific features provide a new dimension for design and can complement voltage scaling for maximum energy efficiency**

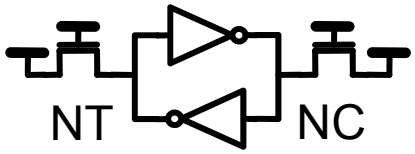
**54MB on-chip cache**



[ISSCC 2011]

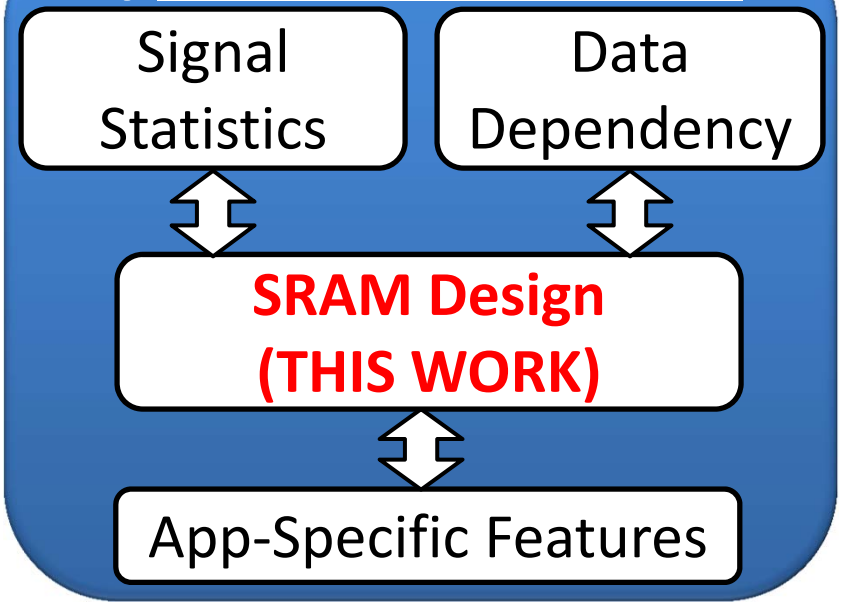
## 1) *Voltage Scaling*



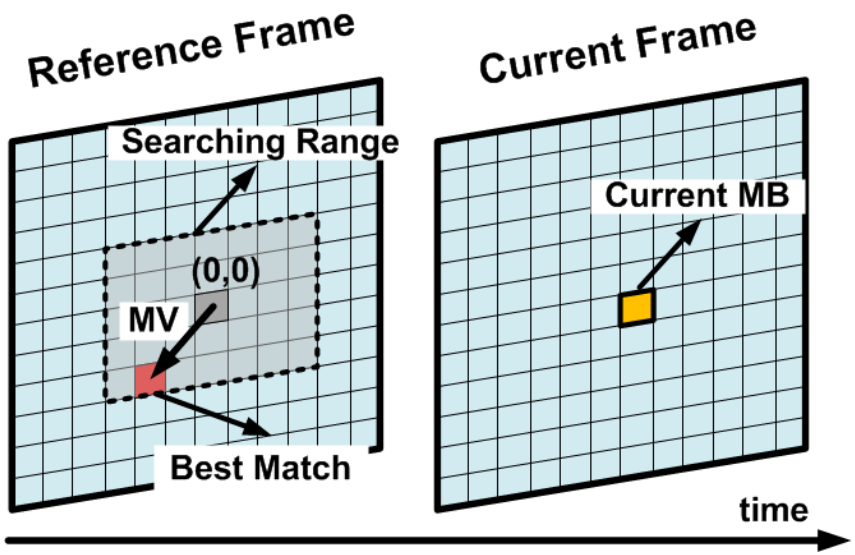




**Smaller noise margins as $V_{DD}$ is scaled down**

## 2) *App-Specific SRAMs*

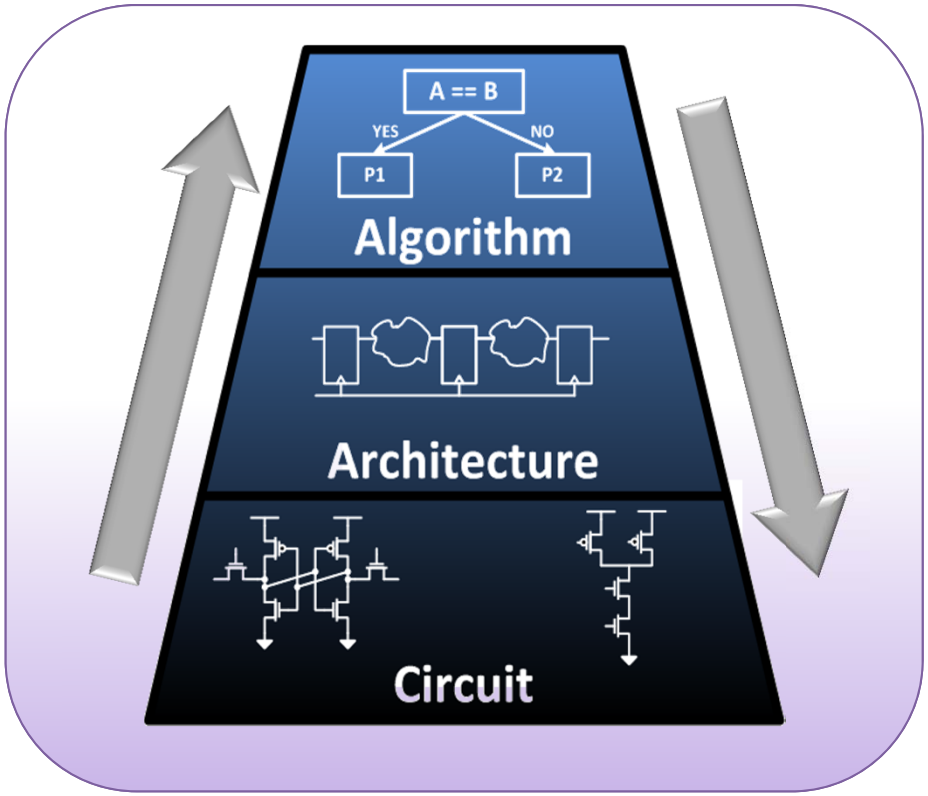


# Motion Estimation in Video Coding



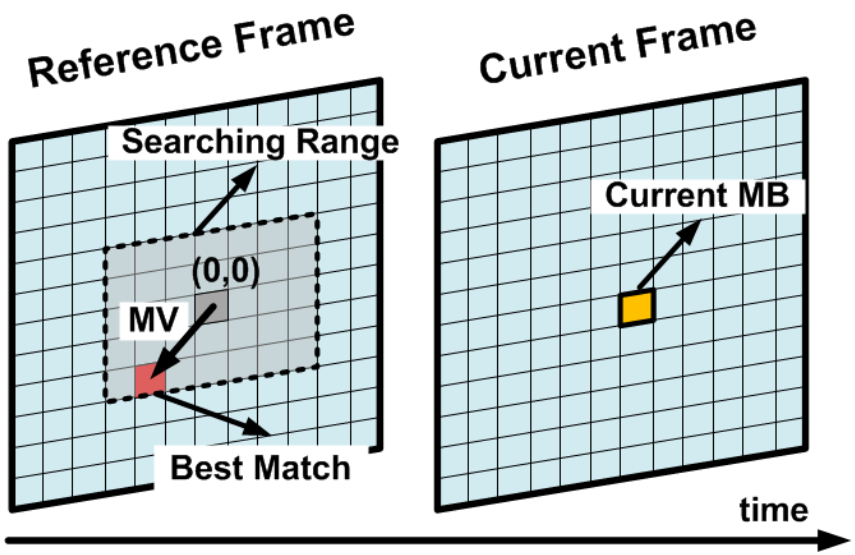

**Motion estimation requires consecutive memory accesses in a pattern that is determined by the search algorithm [1]**

[1]: Sinangil et. al. ICIP 2012



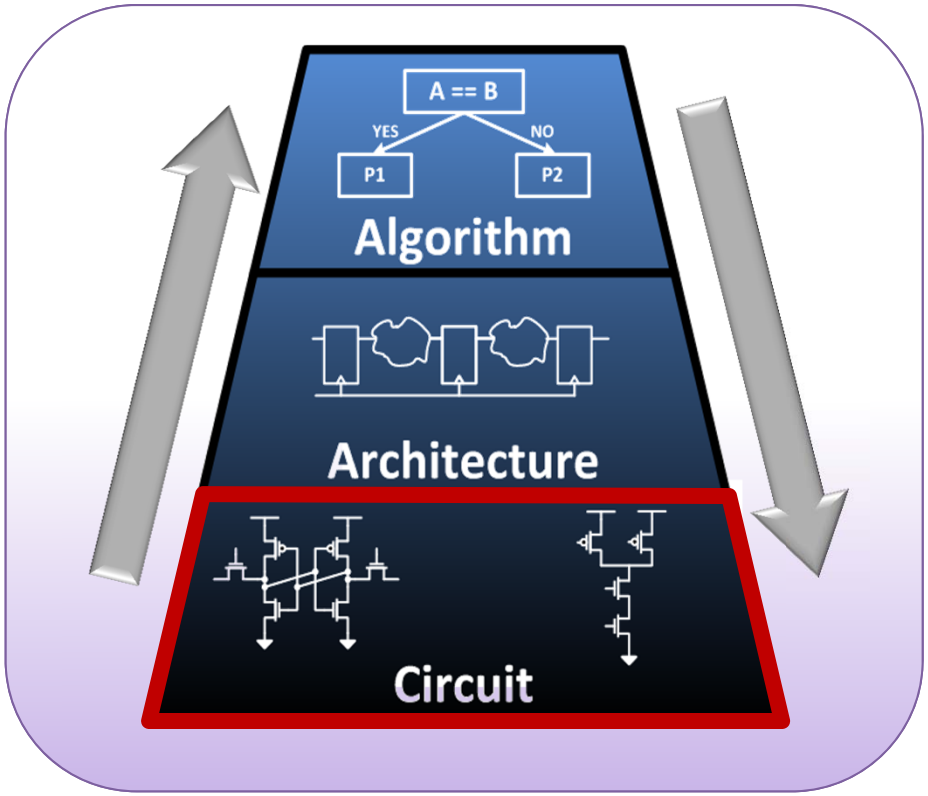

***Joint optimization of algorithms, architectures and circuits for maximum energy-efficiency***

# Motion Estimation in Video Coding




**Motion estimation requires consecutive memory accesses in a pattern that is determined by the search algorithm [1]**

[1]: Sinangil et. al. ICIP 2012




***Joint optimization of algorithms, architectures and circuits for maximum energy-efficiency***



© 2013 IEEE

# Outline

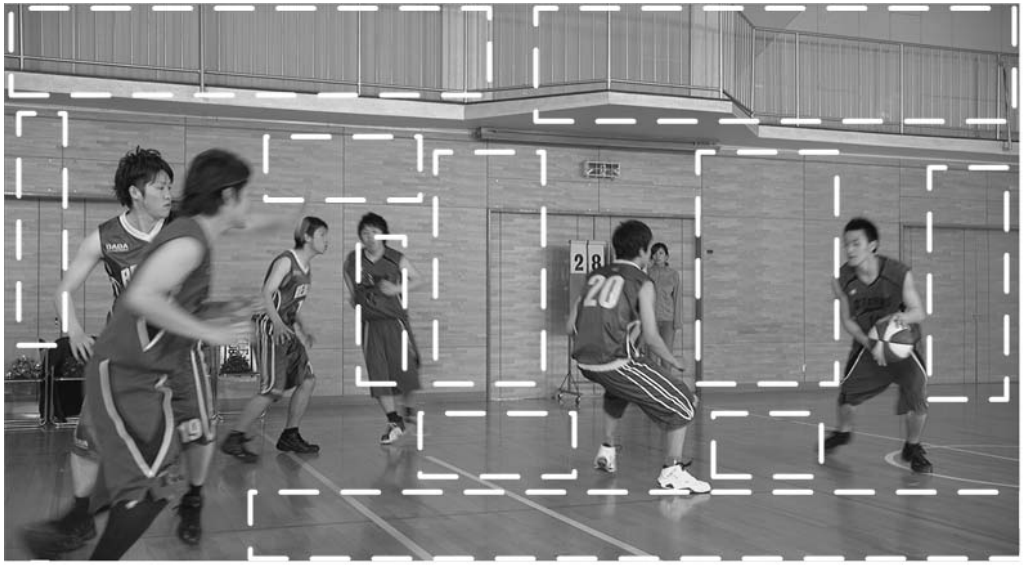
- **Application Specific Design Decisions**
  - Motion Estimation Specific Features
  - Bit-line Switching Activity
- **Prediction Based Reduced Bit-line Switching Activity (PB-RBSA) SRAM Design**
  - Bit-cell & Array Design
  - Prediction Generation
  - Statistically-Gated Sense-Amplifiers
- **Measurement Results**
- **Conclusions**

# Outline

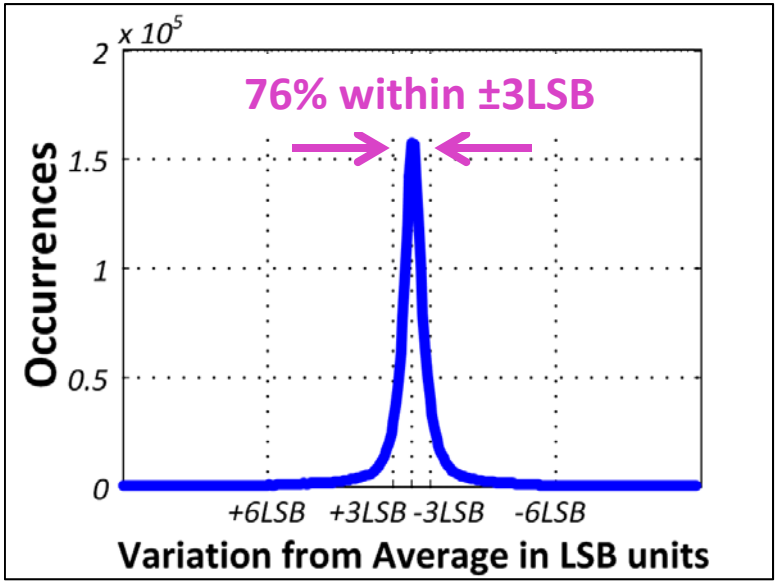
- **Application Specific Design Decisions**
  - Motion Estimation Specific Features
  - Bit-line Switching Activity
- **Prediction Based Reduced Bit-line Switching Activity (PB-RBSA) SRAM Design**
  - Bit-cell & Array Design
  - Prediction Generation
  - Statistically-Gated Sense-Amplifiers
- **Measurement Results**
- **Conclusions**

# Motion Estimation Specific Features

## 1. Correlation of Pixel Data

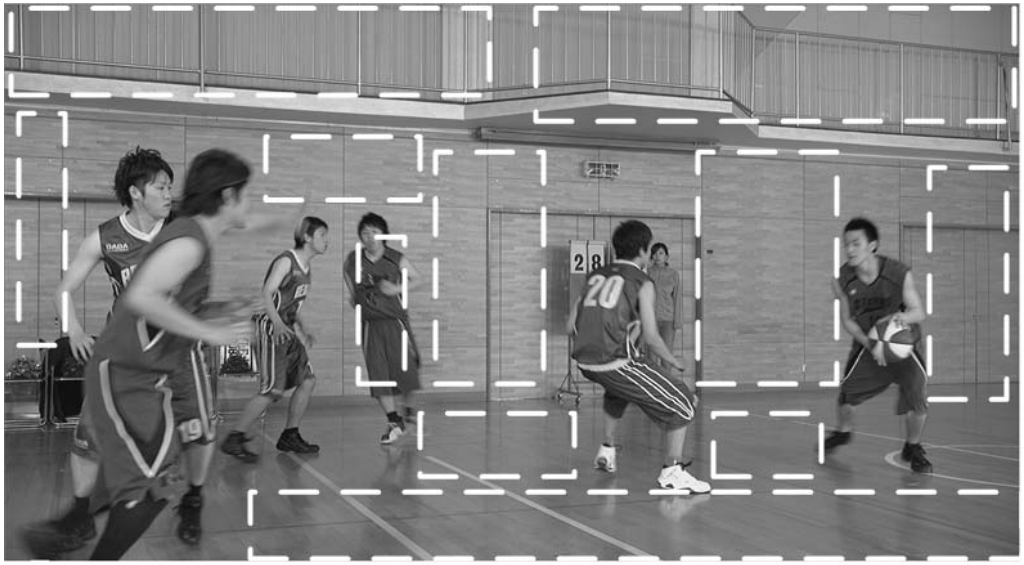


### Variation from a 16x16 Block average

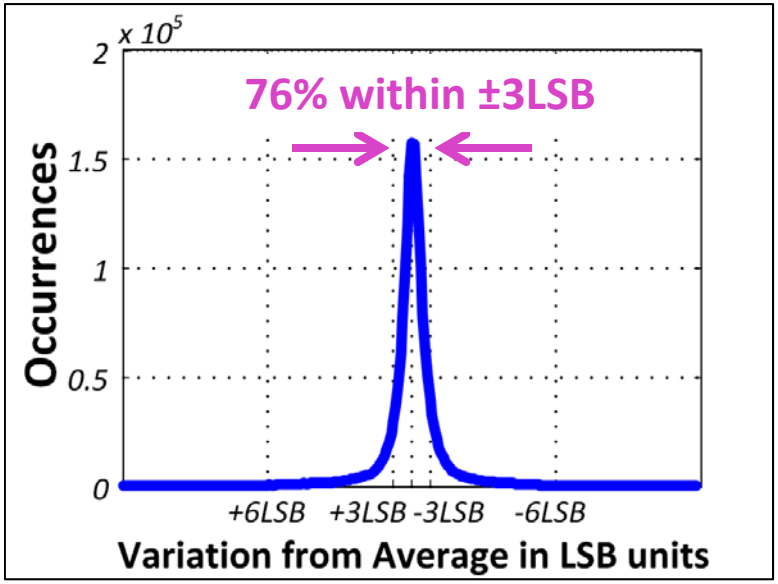


# Motion Estimation Specific Features

## 1. Correlation of Pixel Data



Variation from a 16x16 Block average

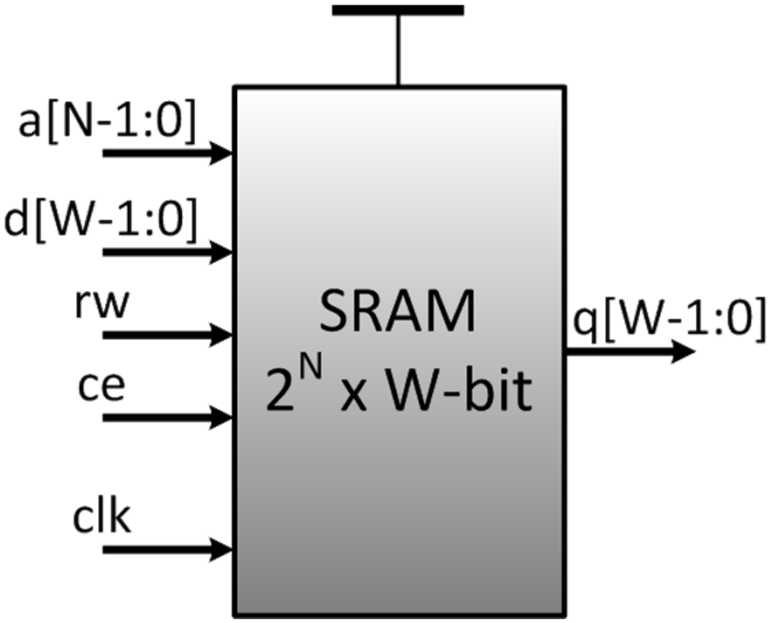


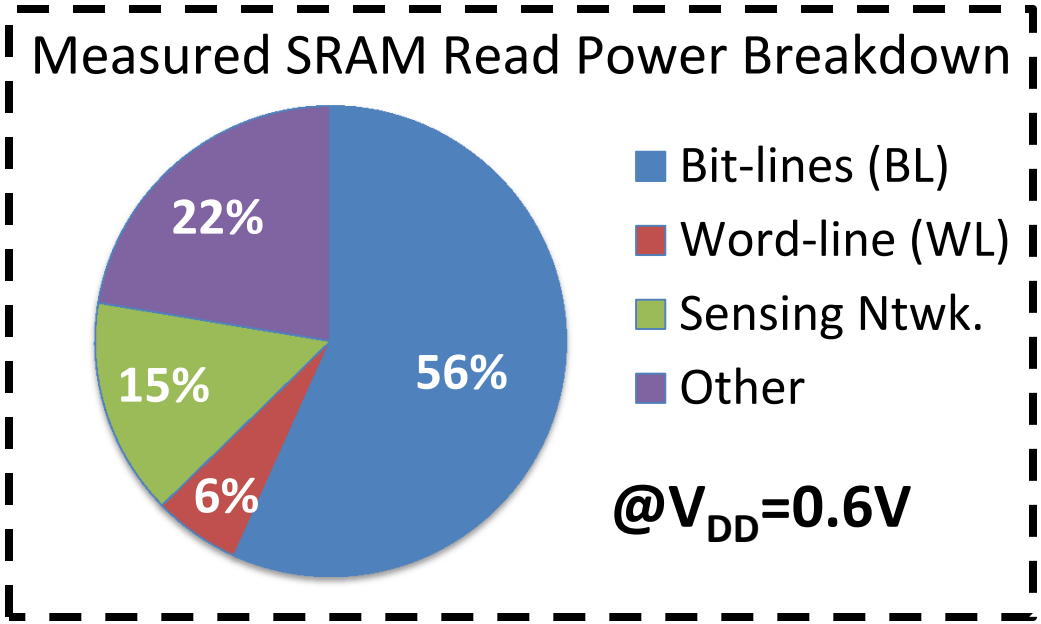

## 2. # of Read Accesses > # of Write Accesses

- Write once and read multiple times
  - Data reuse between consecutive blocks

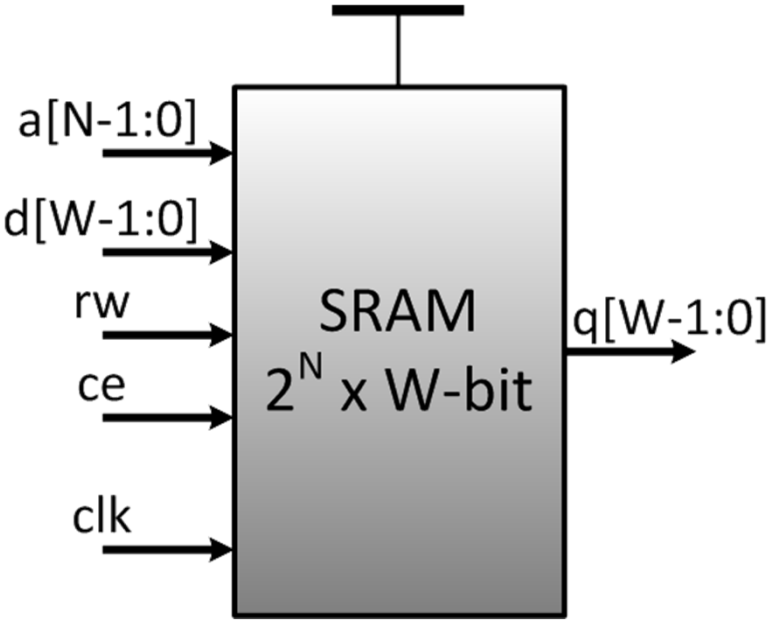
**Reduce energy/access in read accesses by utilizing correlation of pixel data**

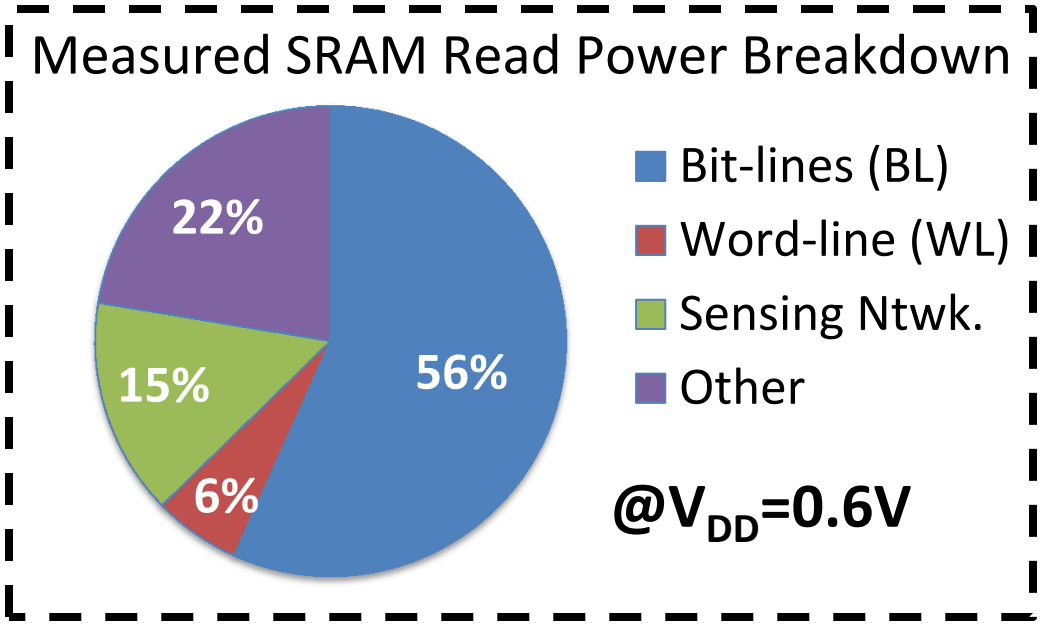
# SRAM Array Power Consumption



# SRAM Array Power Consumption

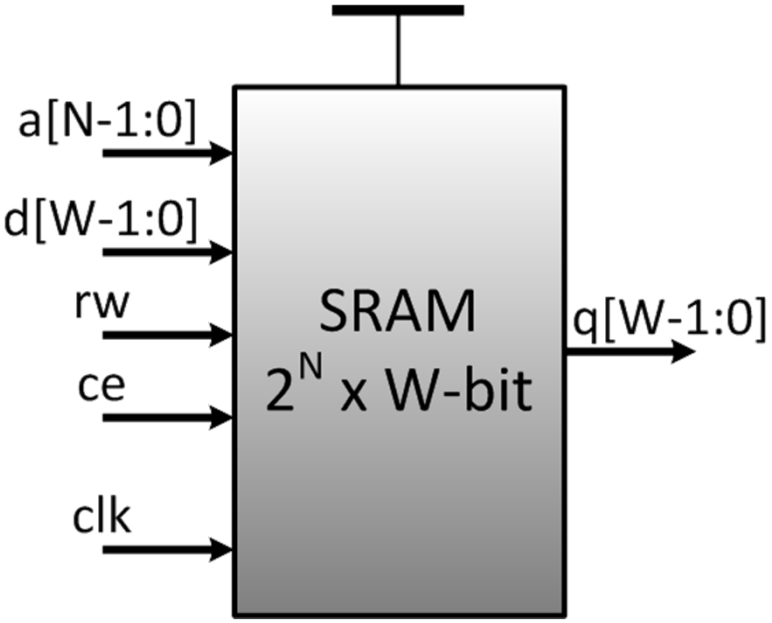


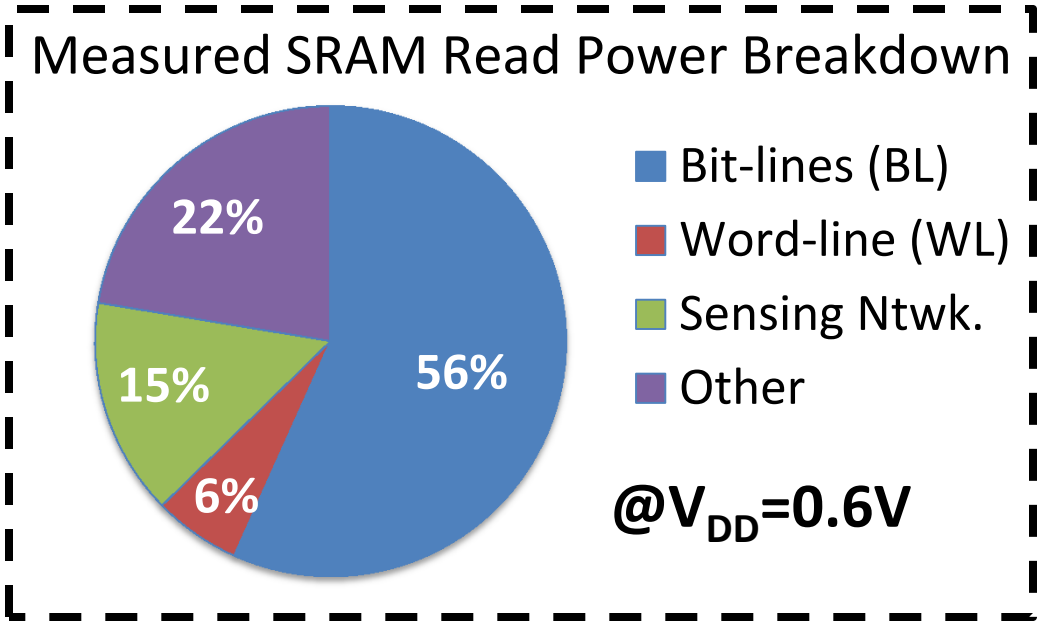




$$P_{BL,READ} = \alpha_{0\rightarrow1} \times C_{BL} \times V_{DD} \times \Delta V_{min} \times f$$

- $\alpha_{0\rightarrow1}$ is the activity factor
- $C_{BL}$ is the total switching BL capacitance
- $V_{DD}$ is the supply voltage
- $\Delta V_{min}$ is the minimum voltage swing on the BLs
- $f$ is the frequency of operation

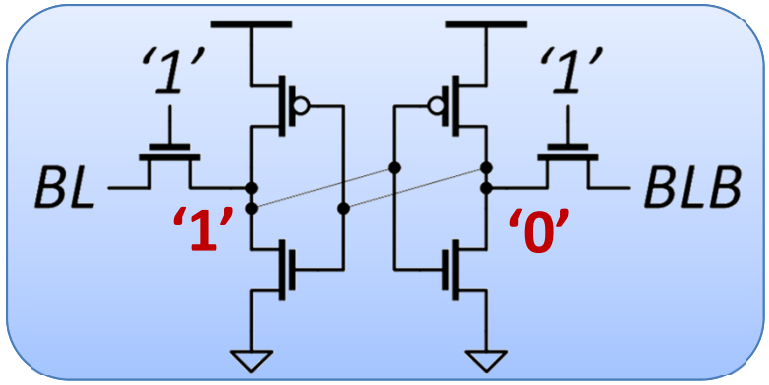
# SRAM Array Power Consumption



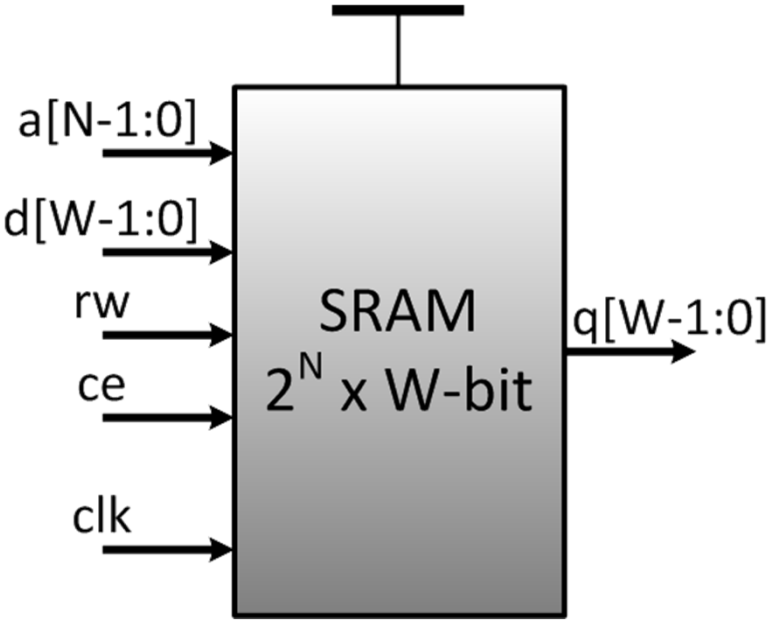



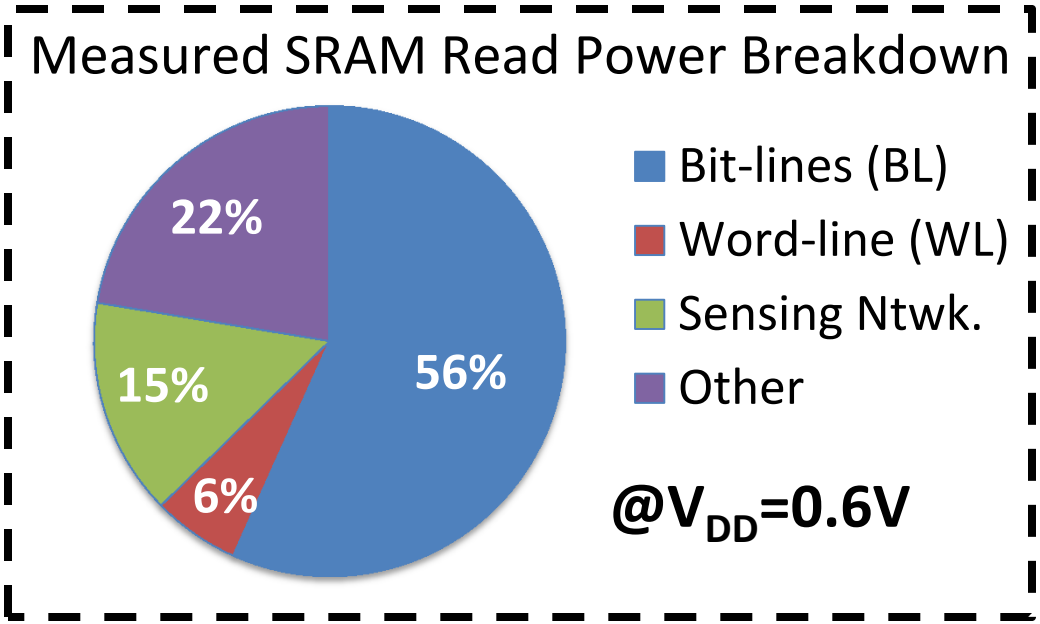

$$P_{BL,READ} = \alpha_{0\rightarrow1} \times C_{BL} \times V_{DD} \times \Delta V_{min} \times f$$

| Cell Type | $\alpha_{0\rightarrow1}$ |
|---|---|
| 6T | 1 |



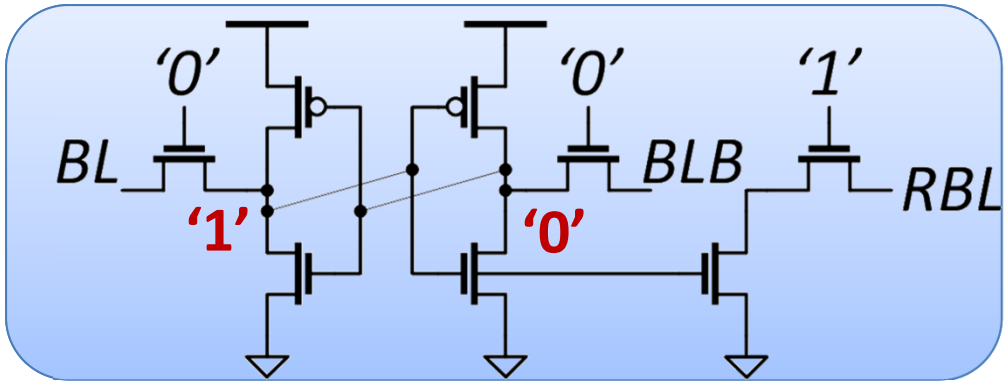
# SRAM Array Power Consumption



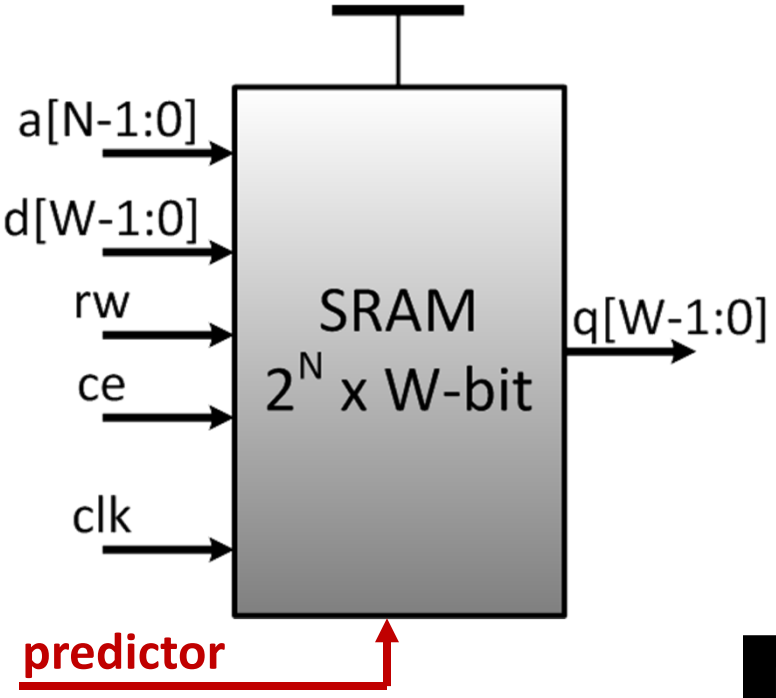



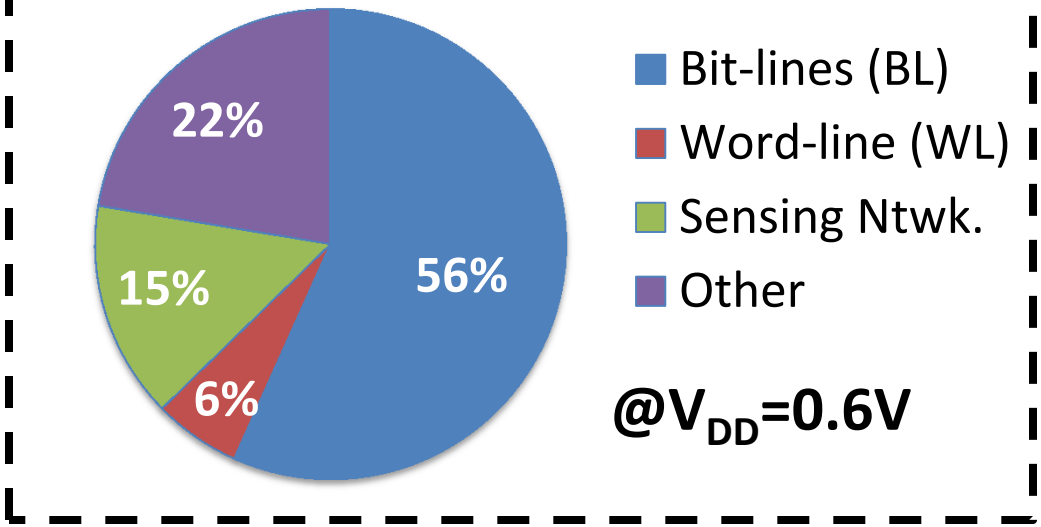

$$P_{BL,READ} = \alpha_{0\rightarrow1} \times C_{BL} \times V_{DD} \times \Delta V_{min} \times f$$

| Cell Type | $\alpha_{0\rightarrow1}$ |
|---|---|
| 6T | 1 |
| 8T | [0,1] |



# SRAM Array Power Consumption




$$P_{BL,READ} = \alpha_{0\to1} \times C_{BL} \times V_{DD} \times \Delta V_{min} \times f$$

| Cell Type | $\alpha_{0\to1}$ |
|---|---|
| 6T | 1 |
| 8T | [0,1] |

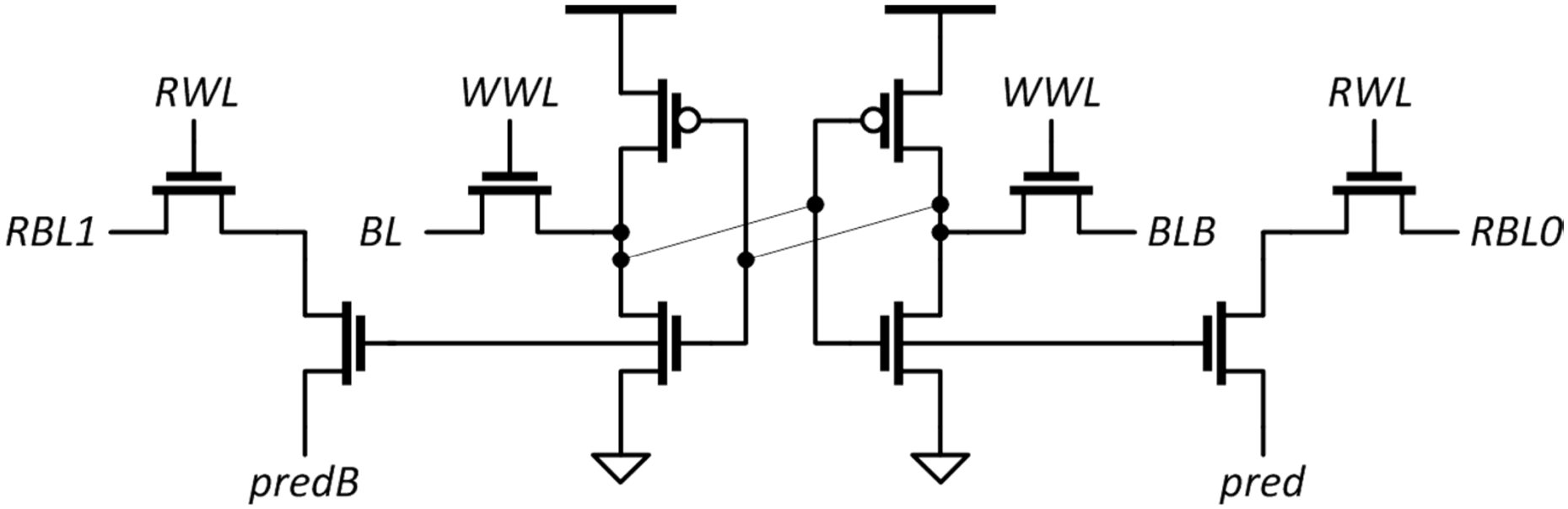
**Reduce $\alpha_{0\to1}$ by using correlation of pixel data**

# Outline

- **Application Specific Design Decisions**
  - Motion Estimation Specific Features
  - Bit-line Switching Activity
- **Prediction Based Reduced Bit-line Switching Activity (PB-RBSA) SRAM Design**
  - Bit-cell & Array Design
  - Prediction Generation
  - Statistically-Gated Sense-Amplifiers
- **Measurement Results**
- **Conclusions**

# PB-RBSA Scheme & Bit-cell

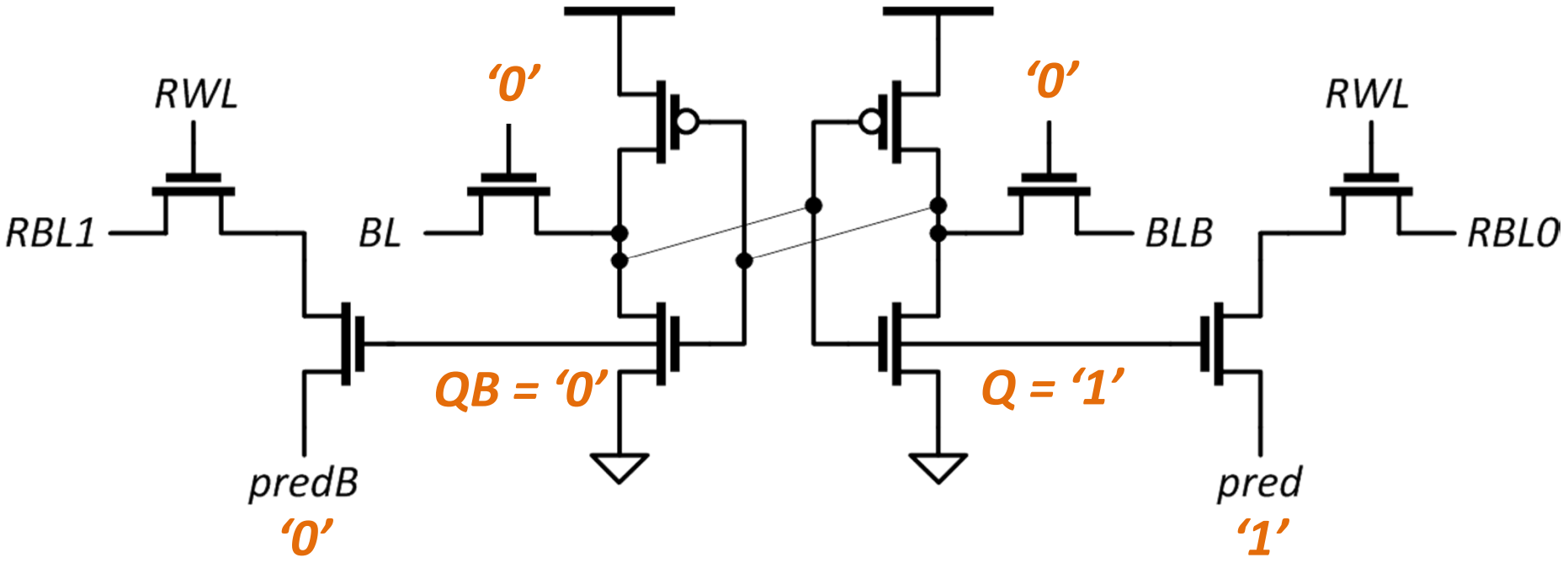
## Prediction Based – Reduced Bit-line Switching Activity (PB-RBSA) Scheme



# PB-RBSA Scheme & Bit-cell

## Prediction Based – Reduced Bit-line Switching Activity (PB-RBSA) Scheme

### Read Access – Correct Prediction (*Q*='1' and *pred* = '1')

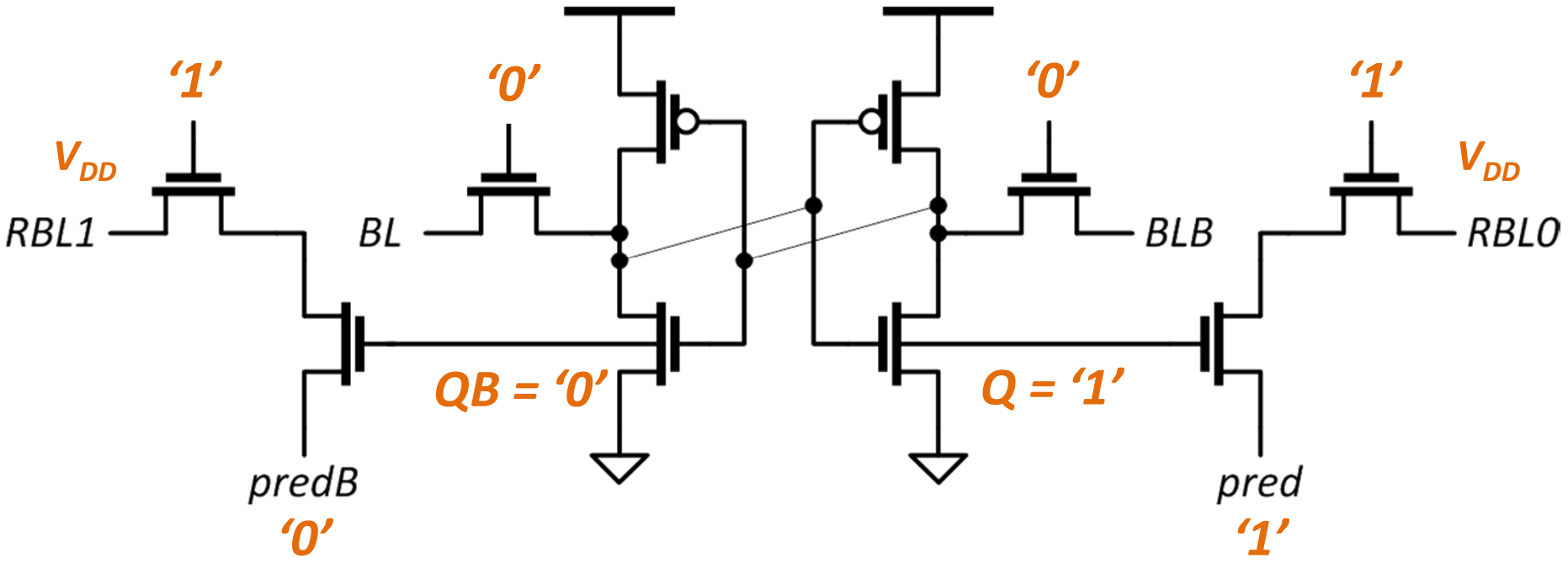



**No RBL Discharging – RBL0 and RBL1 stay at $V_{DD}$**

# PB-RBSA Scheme & Bit-cell

## Prediction Based – Reduced Bit-line Switching Activity (PB-RBSA) Scheme

### Read Access – Correct Prediction (*Q*='1' and *pred* = '1')

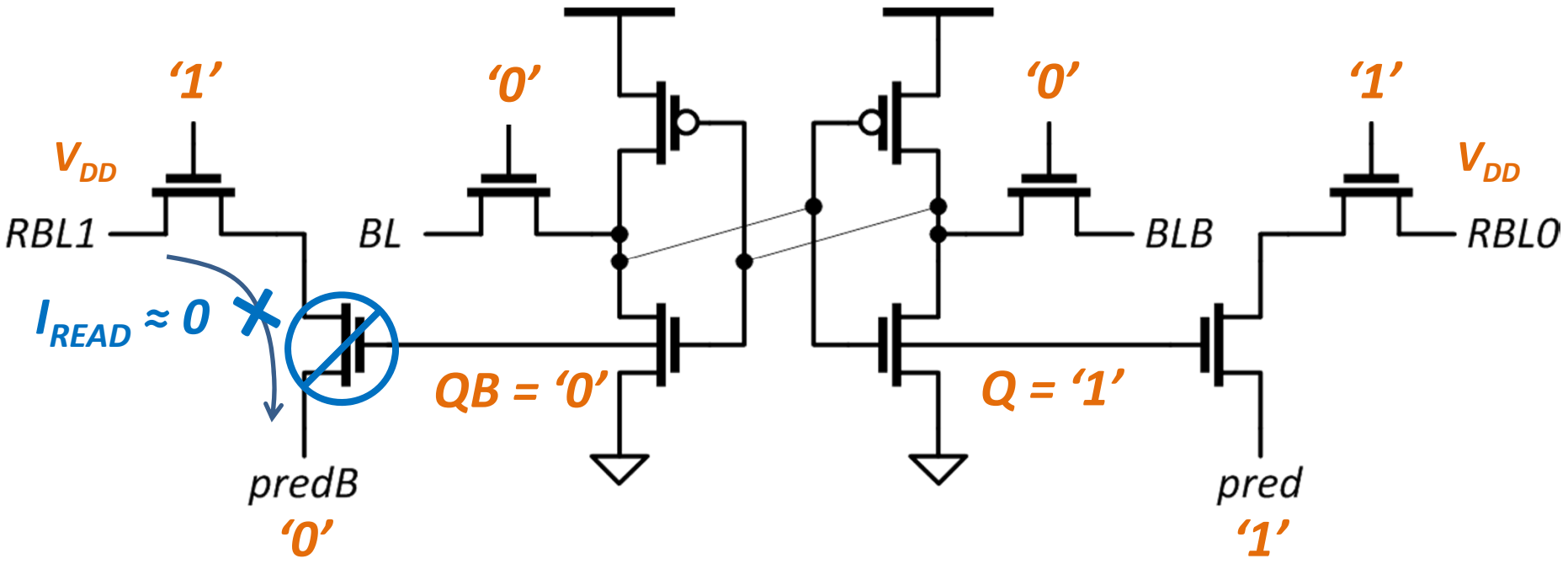



**No RBL Discharging – RBL0 and RBL1 stay at $V_{DD}$**

# PB-RBSA Scheme & Bit-cell

## Prediction Based – Reduced Bit-line Switching Activity (PB-RBSA) Scheme

### Read Access – Correct Prediction (*Q*='1' and *pred* = '1')

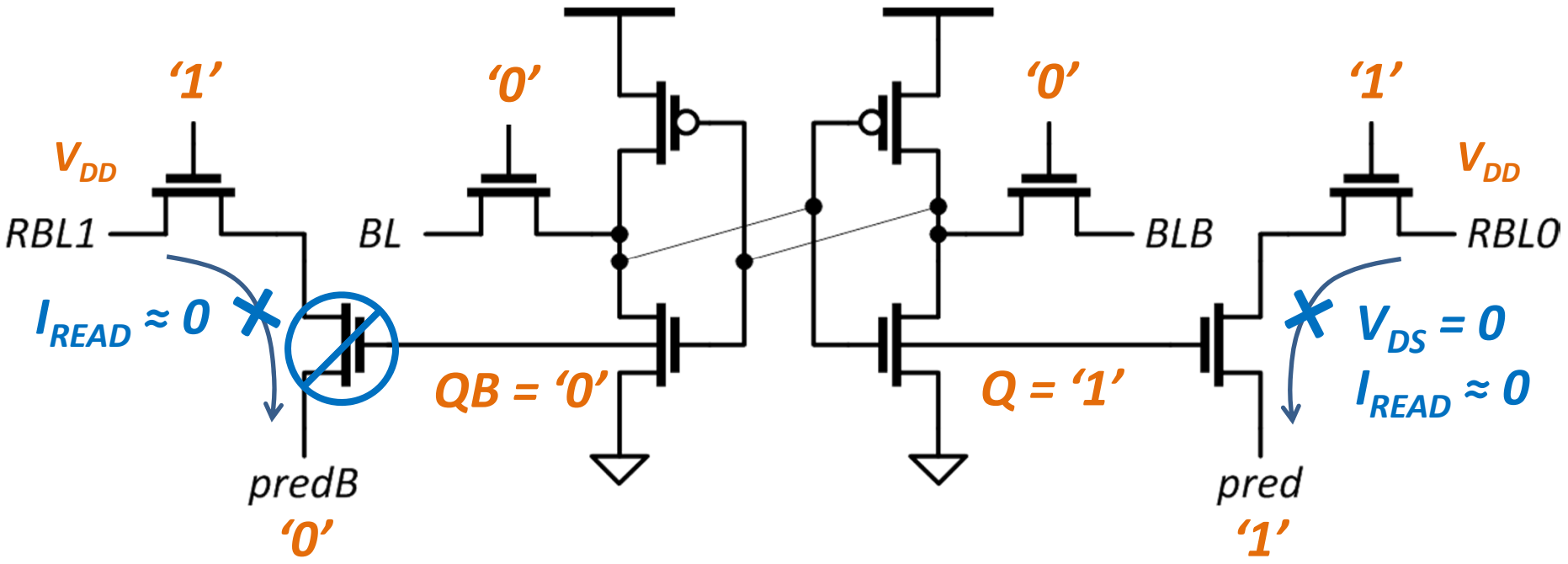



**No RBL Discharging – RBL0 and RBL1 stay at $V_{DD}$**

# PB-RBSA Scheme & Bit-cell

## Prediction Based – Reduced Bit-line Switching Activity (PB-RBSA) Scheme

### Read Access – Correct Prediction (*Q*='1' and *pred* = '1')

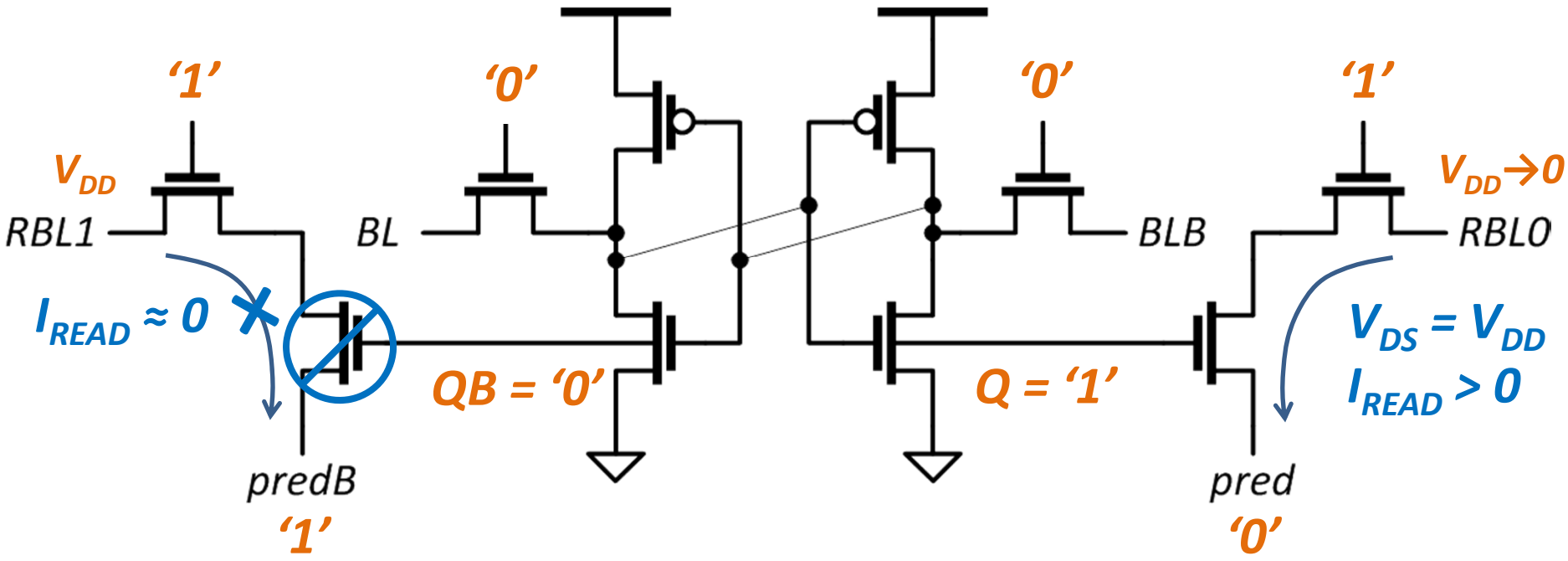



**No RBL Discharging – RBL0 and RBL1 stay at $V_{DD}$**

# PB-RBSA Scheme & Bit-cell

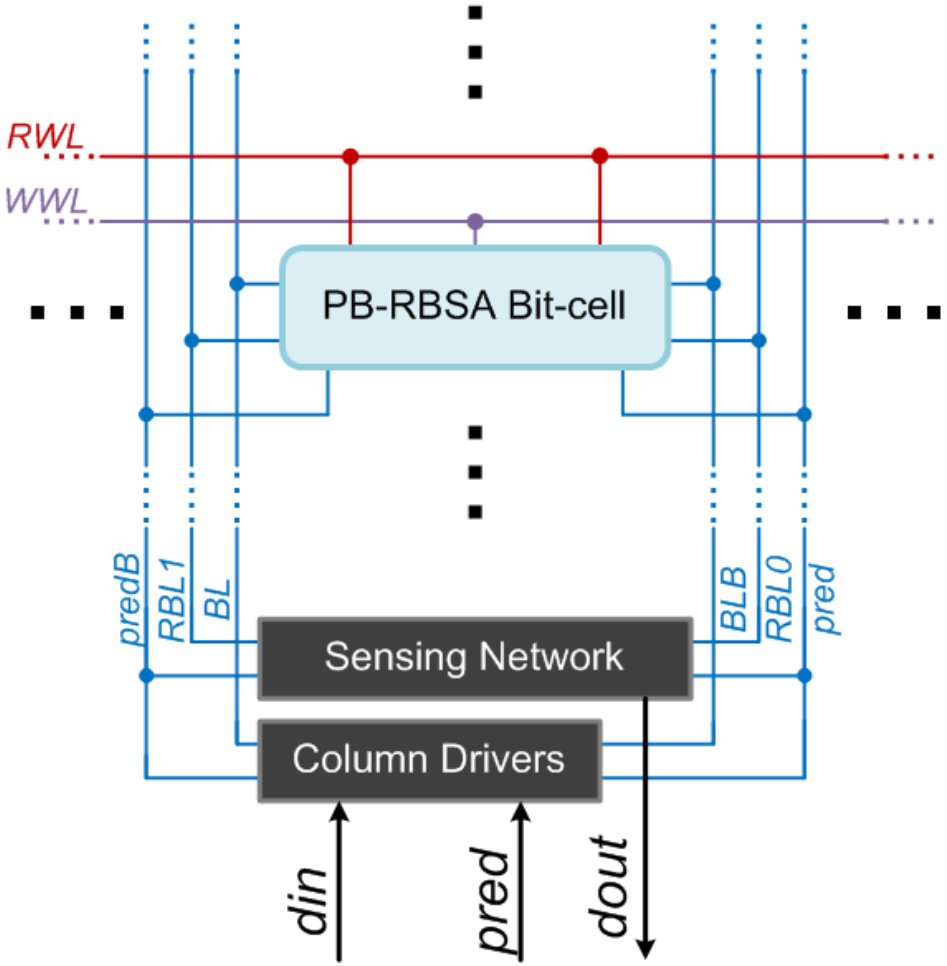
## Prediction Based – Reduced Bit-line Switching Activity (PB-RBSA) Scheme

### Read Access – Incorrect Prediction (*Q*='1' and *pred* = '0')

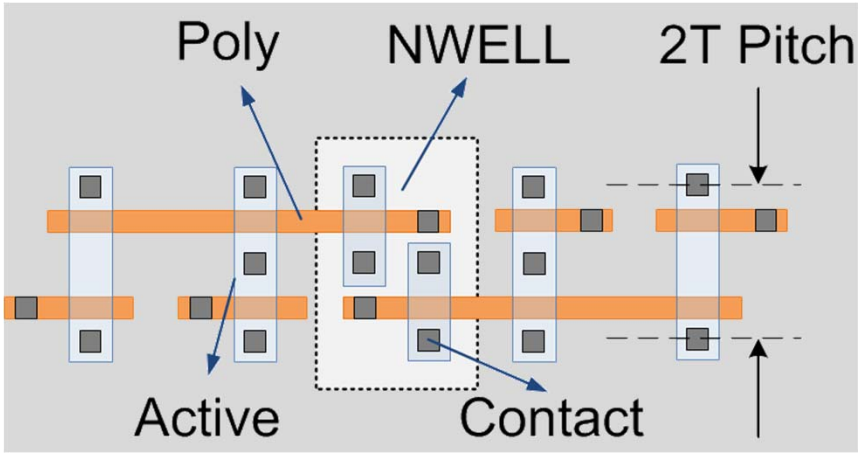



**RBL0 or RBL1 is discharged from $V_{DD}$ to ground**

# PB-RBSA Array Architecture




MET3 and MET5



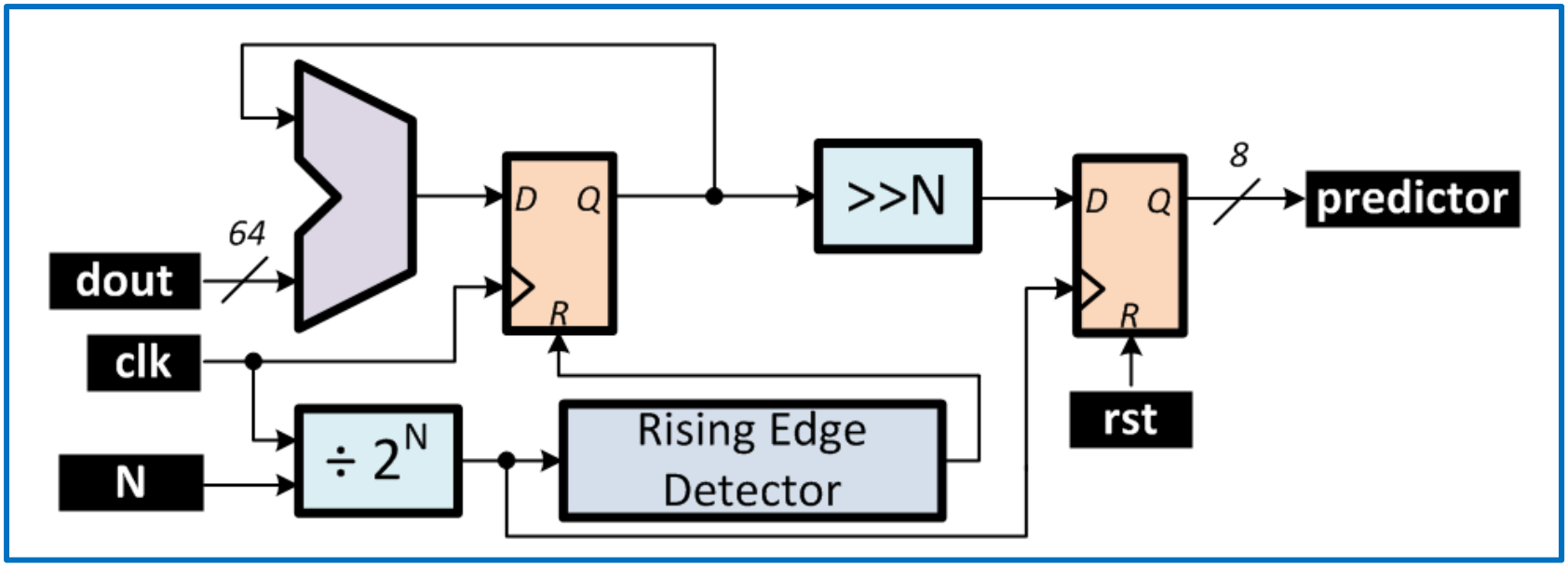

MET2

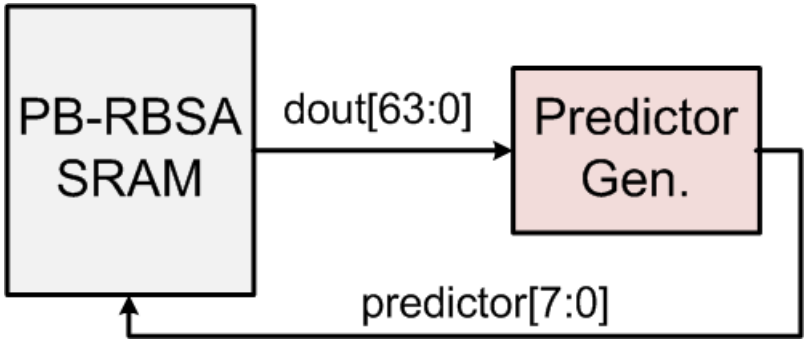
**PB-RBSA cell layout fit in 2 transistor poly pitch**

**pred & predB lines switch at a much lower rate**

# Prediction Generation

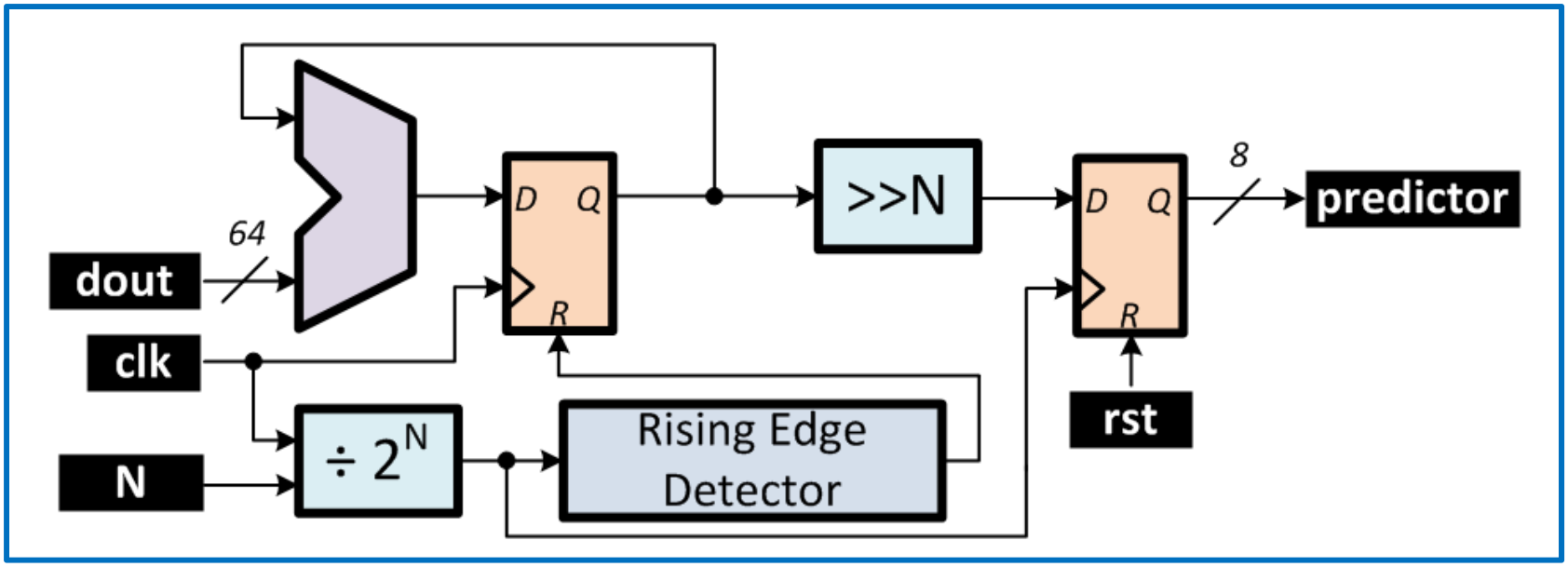
## Arithmetic average across $2^N$ cycles of 8 pixel wide inputs

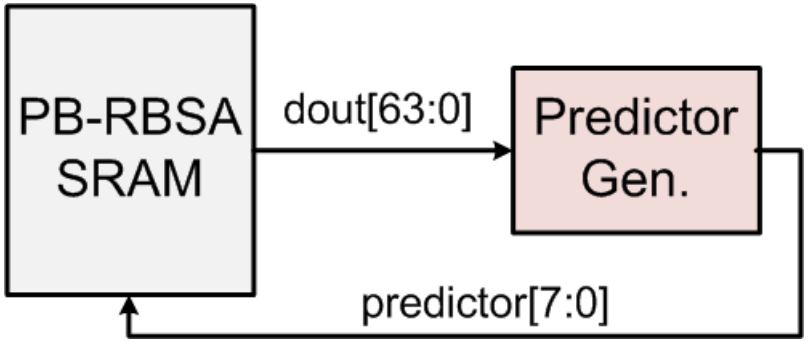


# Prediction Generation

## Arithmetic average across $2^N$ cycles of 8 pixel wide inputs

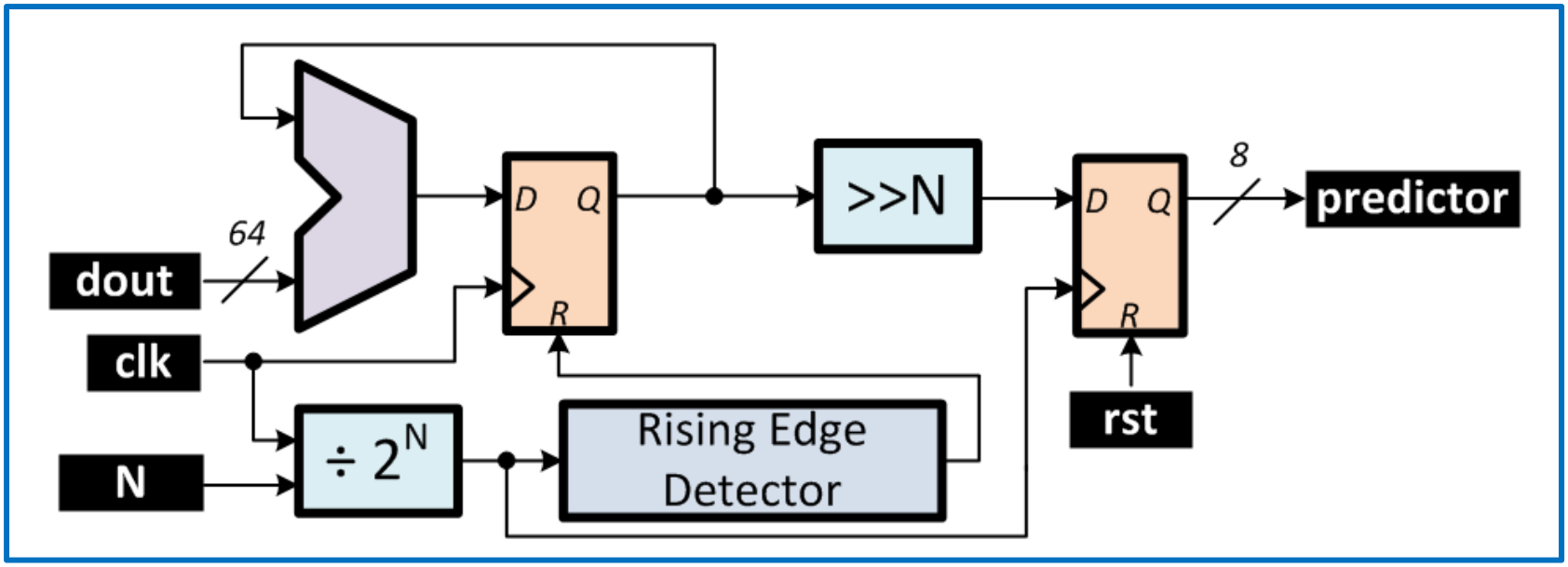


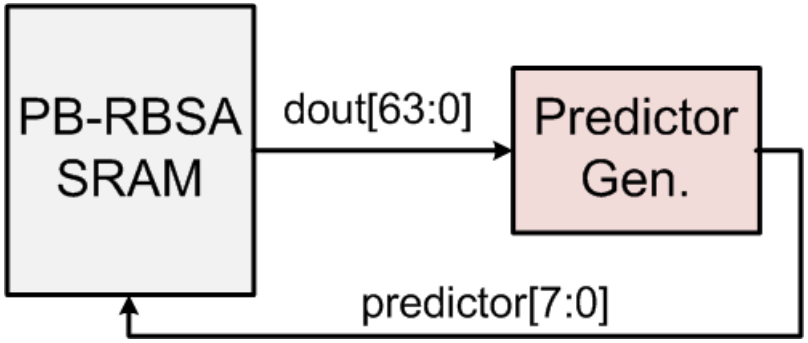




| Selection of N | Prediction Accuracy | pred/predB activity inside SRAM array |
|---|---|---|
| Smaller N | Higher | Higher |
| Larger N | Lower | Lower |

# Prediction Generation

## Arithmetic average across $2^N$ cycles of 8 pixel wide inputs



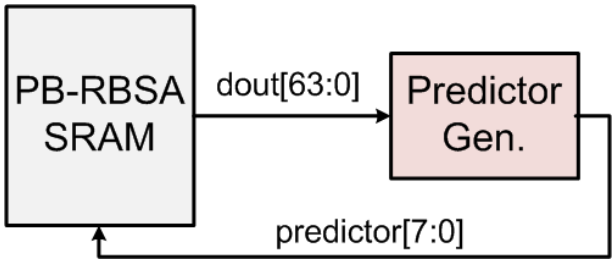



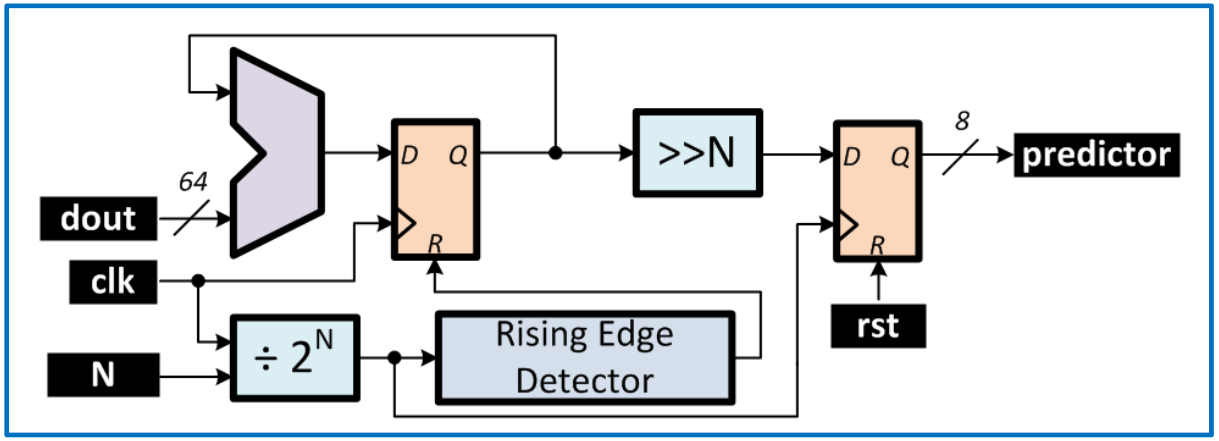

| Selection of N | Prediction Accuracy | pred/predB activity inside SRAM array |
|---|---|---|
| Smaller N | Higher | Higher |
| Larger N | Lower | Lower |



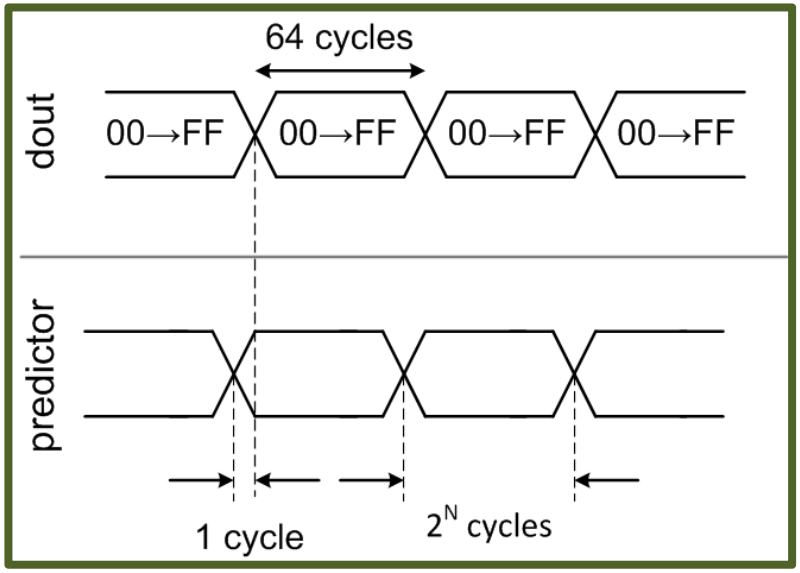
© 2013 IEEE

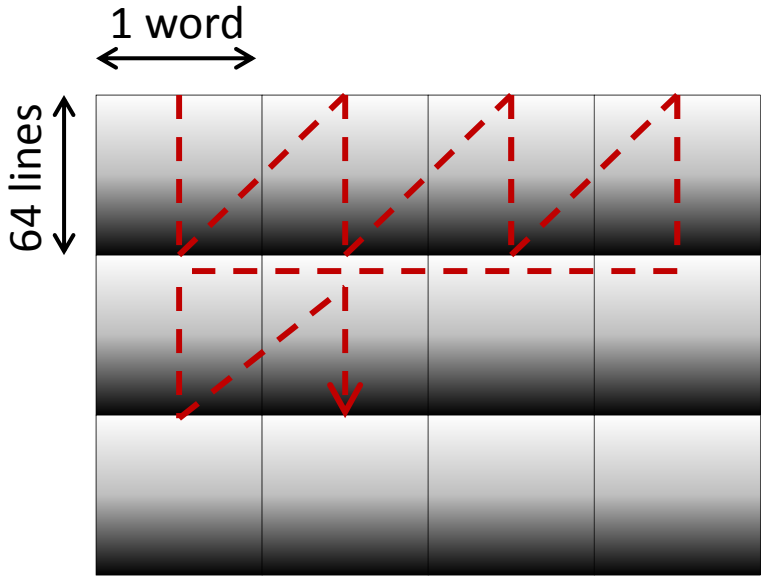
# Selection of N on Energy/Access



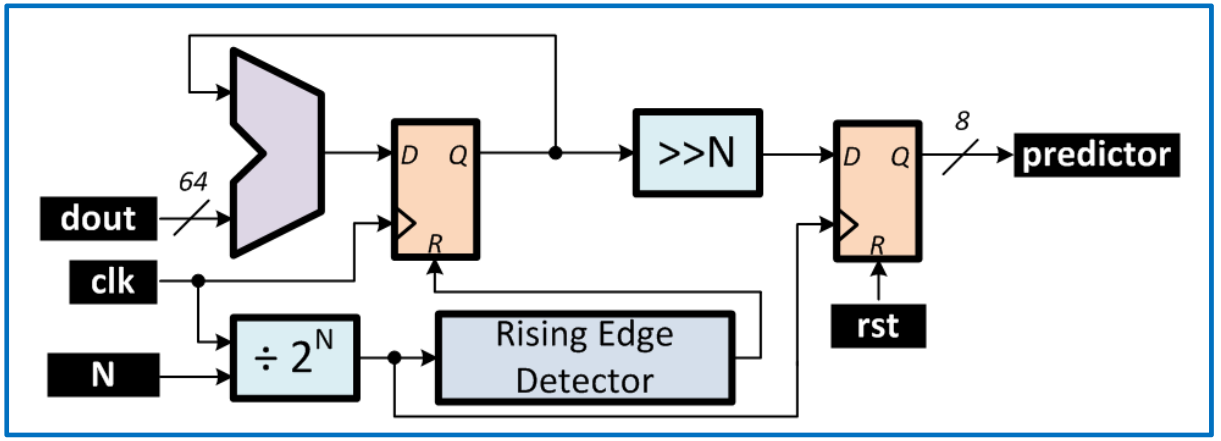

## Worst case access pattern



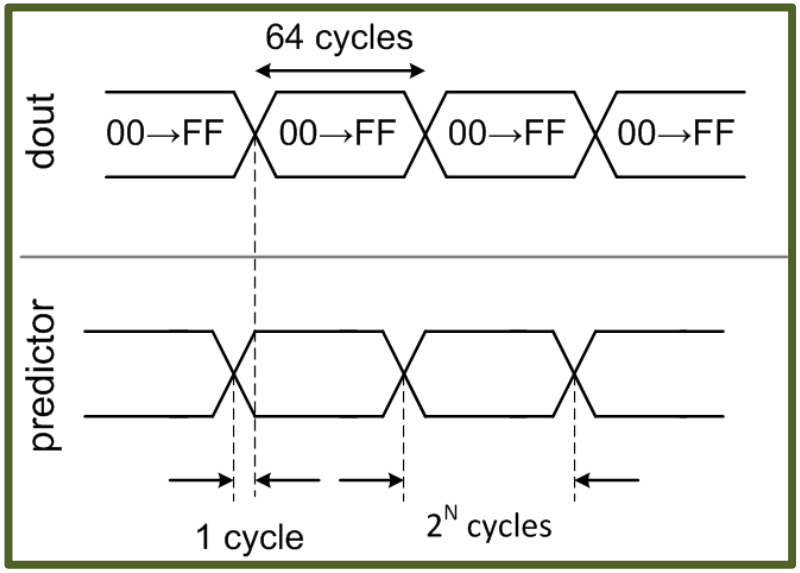



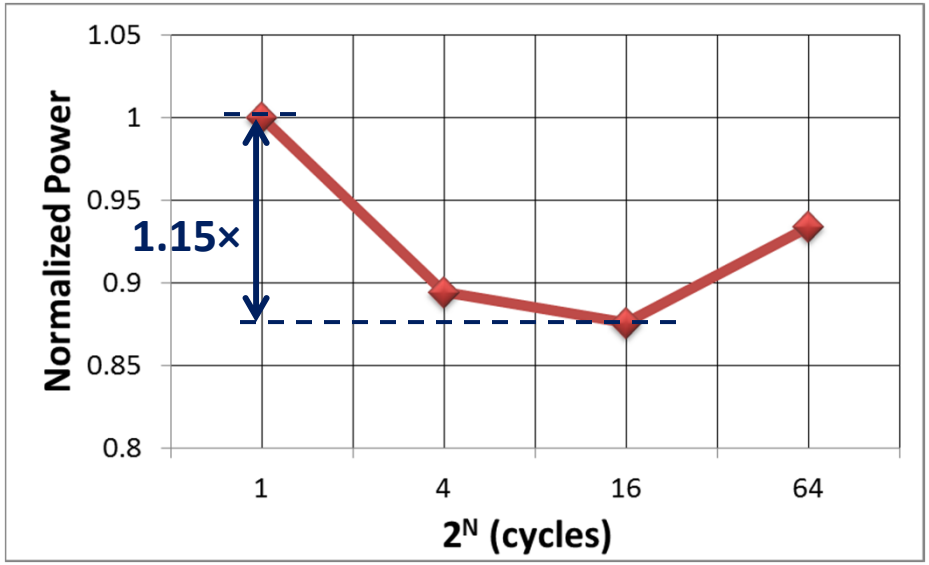
# Selection of N on Energy/Access

**Selection of N can provide up to 1.6× reduction in power consumption**




## Worst case access pattern



# Statistically Gated Sense-Amplifiers

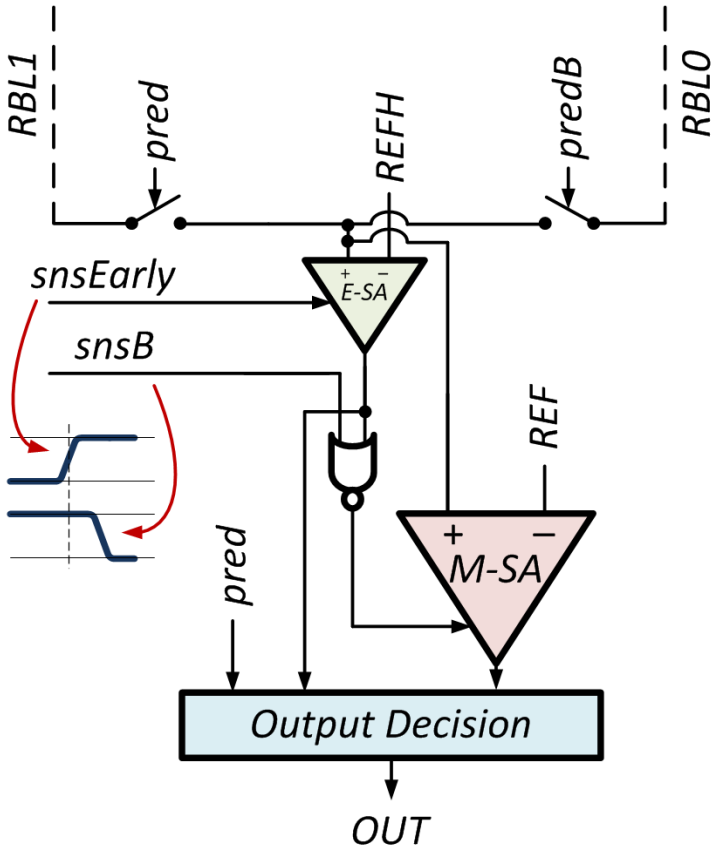
**Observation:** $\alpha_{0\rightarrow1}$ **on RBLs < 0.5**

**Approach: Sensing network consuming less energy when prediction is correct**

# Statistically Gated Sense-Amplifiers

**Observation:** $\alpha_{0\rightarrow1}$ **on RBLs < 0.5**

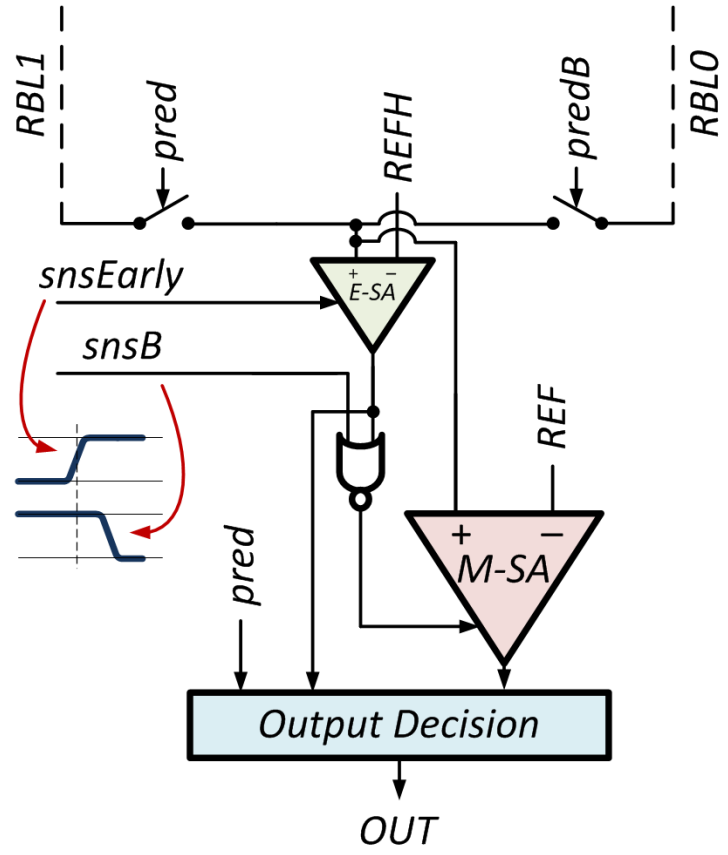
**Approach: Sensing network consuming less energy when prediction is correct**

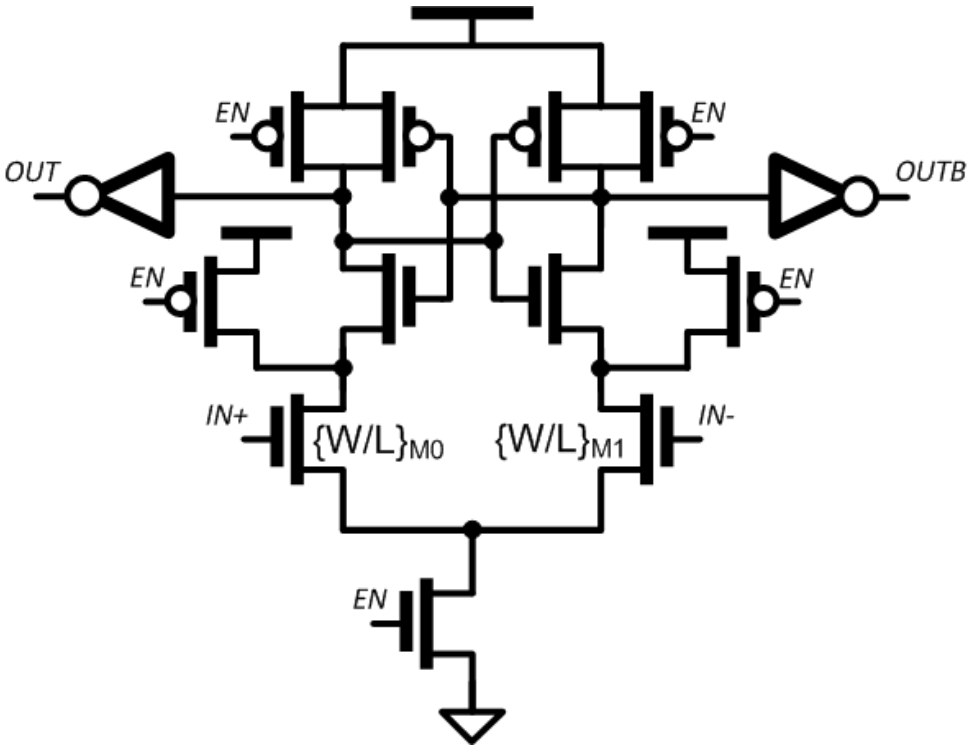


# Statistically Gated Sense-Amplifiers

**Observation:** $\alpha_{0\rightarrow1}$ **on RBLs < 0.5**

**Approach: Sensing network consuming less energy when prediction is correct**

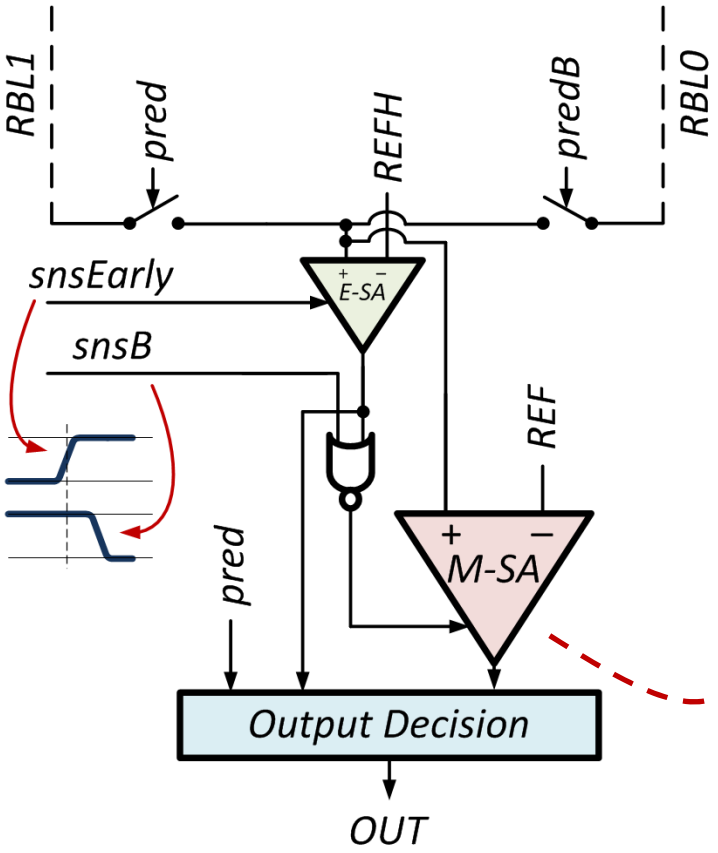


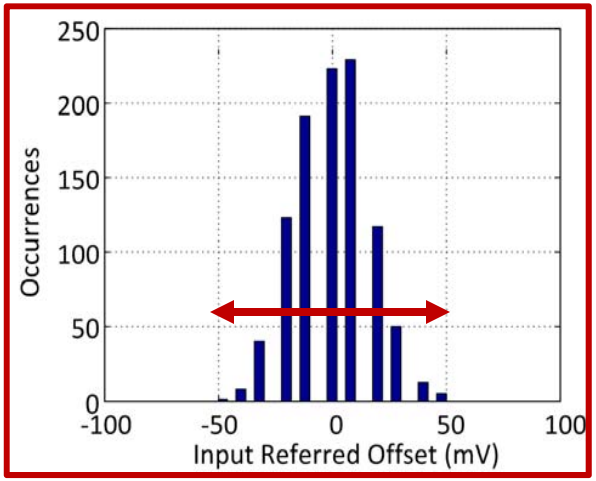
# Statistically Gated Sense-Amplifiers

**Observation:** $\alpha_{0\rightarrow1}$ **on RBLs < 0.5**

**Approach: Sensing network consuming less energy when prediction is correct**




M-SA: Larger area & Smaller offset

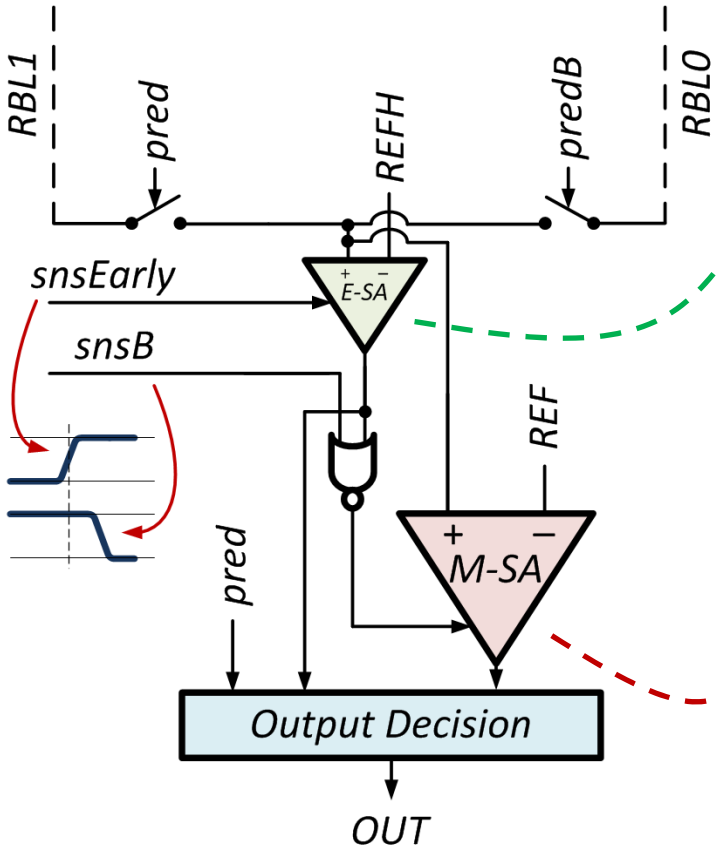



A minimum of 100mV difference necessary between "High" and "Low" signals

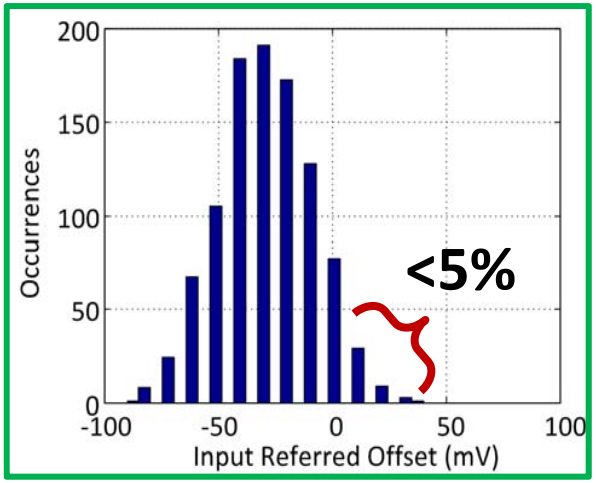
# Statistically Gated Sense-Amplifiers

**Observation:** $\alpha_{0\to1}$ **on RBLs < 0.5**

**Approach: Sensing network consuming less energy when prediction is correct**

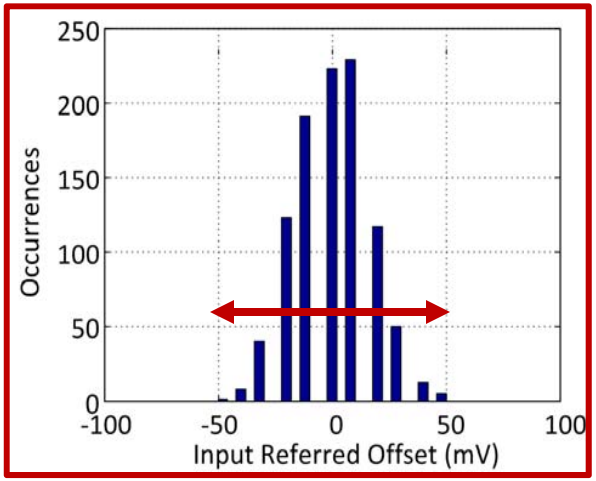



E-SA: Smaller area & Larger offset




Skewed input offset through input transistor sizing

M-SA: Larger area & Smaller offset



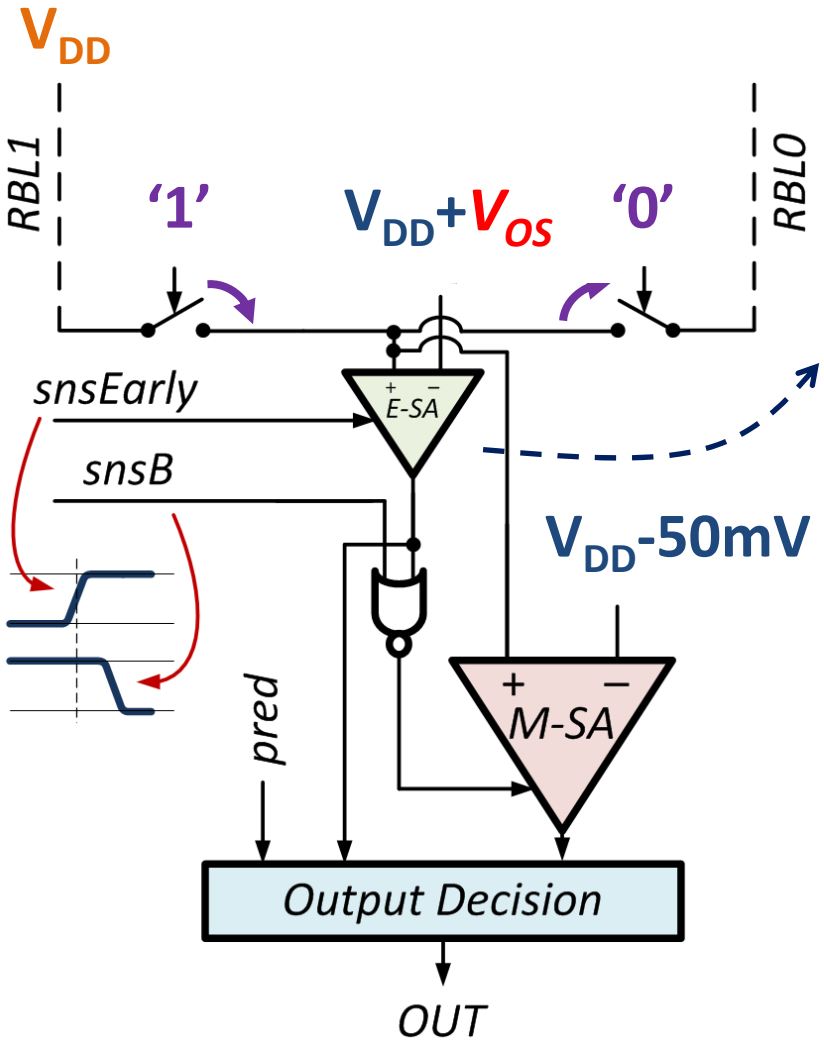

A minimum of 100mV difference necessary between "High" and "Low" signals

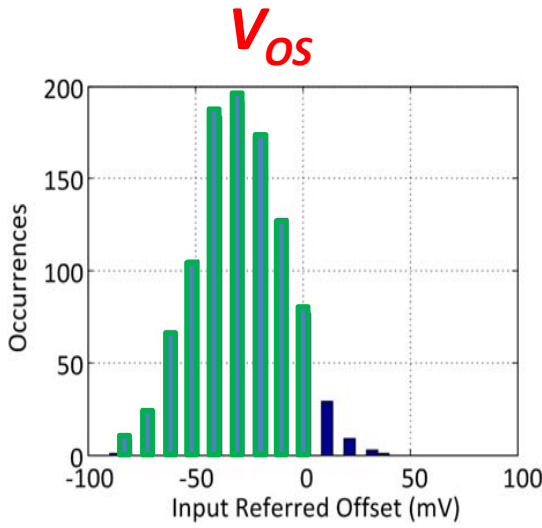
# Statistically Gated Sense-Amplifiers

## Correct Prediction & E-SA with Negative Input Offset

RBL1 stay at $V_{DD}$

E-SA resolves RBL1 correctly and outputs '1'



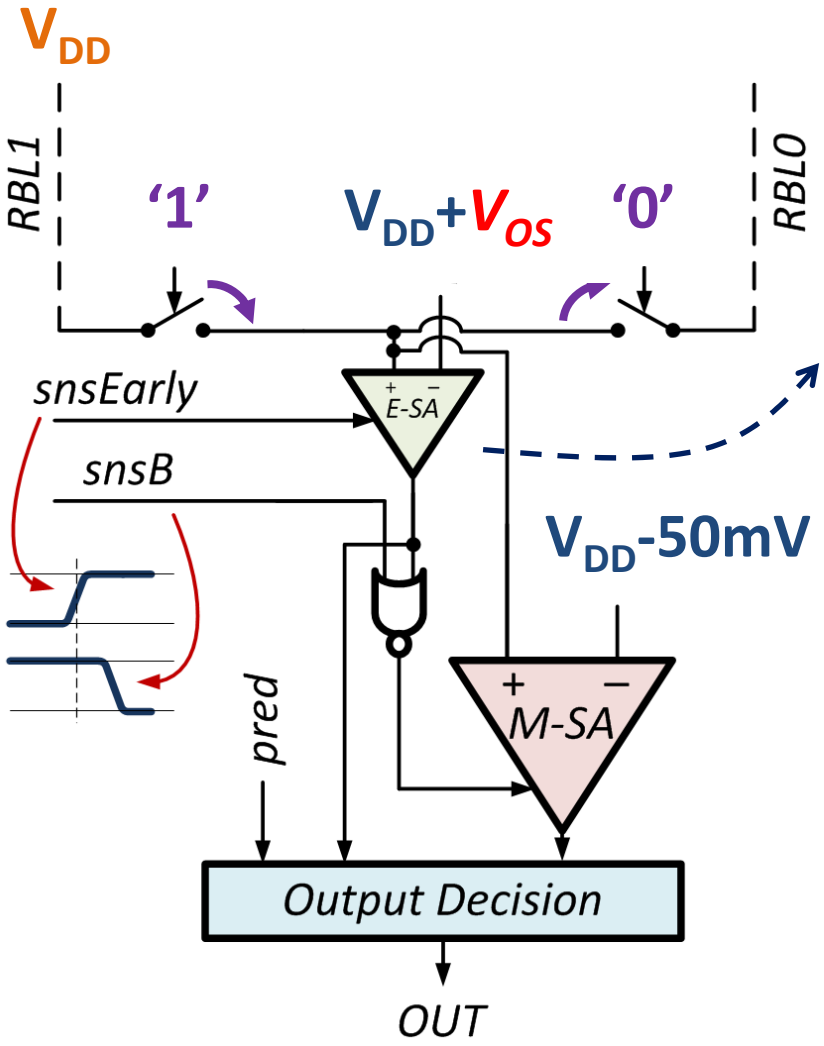

M-SA is gated and not activated

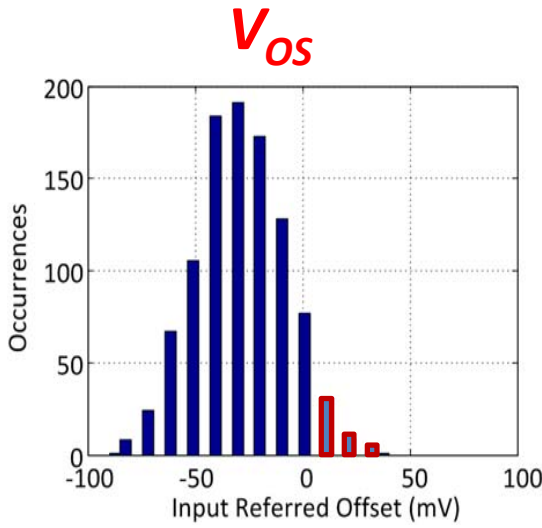
# Statistically Gated Sense-Amplifiers

## Correct Prediction & E-SA with Positive Input Offset

RBL1 stay at $V_{DD}$

E-SA resolves RBL1 incorrectly and outputs '0'



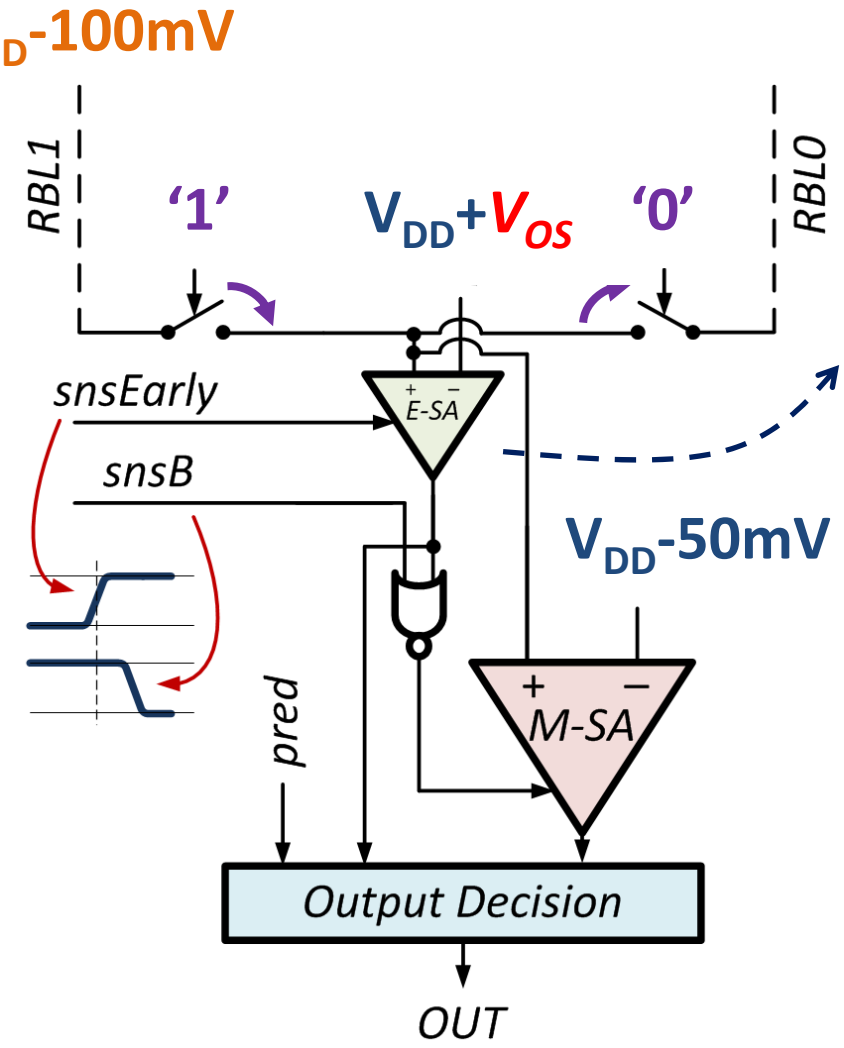

M-SA is activated and resolves RBL1 correctly

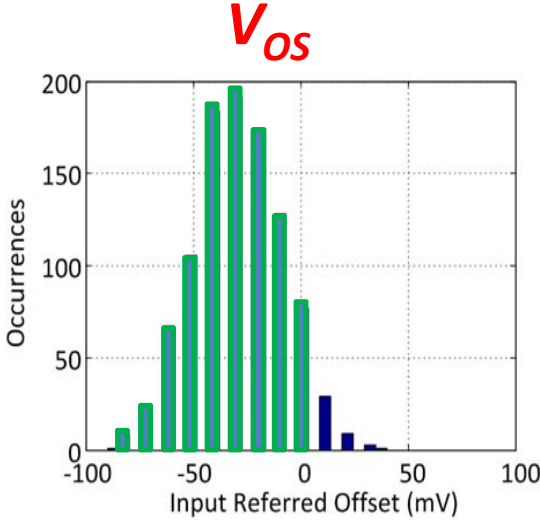
# Statistically Gated Sense-Amplifiers

## Incorrect Prediction & E-SA with Negative Input Offset

RBL1 discharged to $V_{DD}$-100mV

E-SA resolves RBL1 correctly and outputs '0'



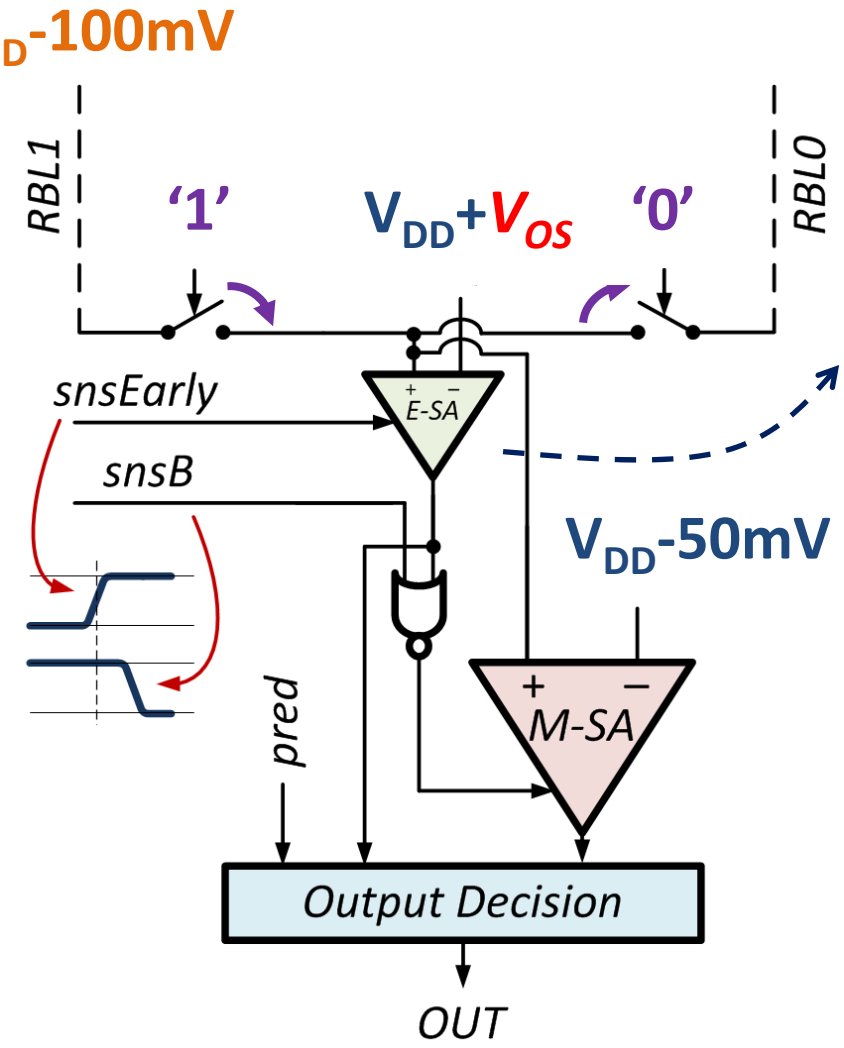

M-SA is activated

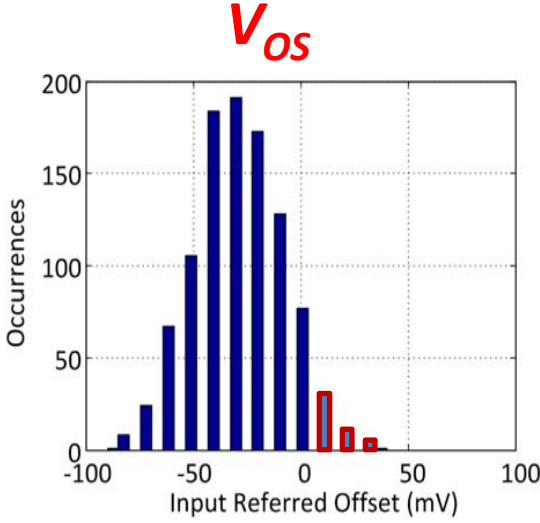
# Statistically Gated Sense-Amplifiers

## Incorrect Prediction & E-SA with Positive Input Offset

RBL1 discharged to $V_{DD}$-100mV

E-SA resolves RBL1 correctly and outputs '0'



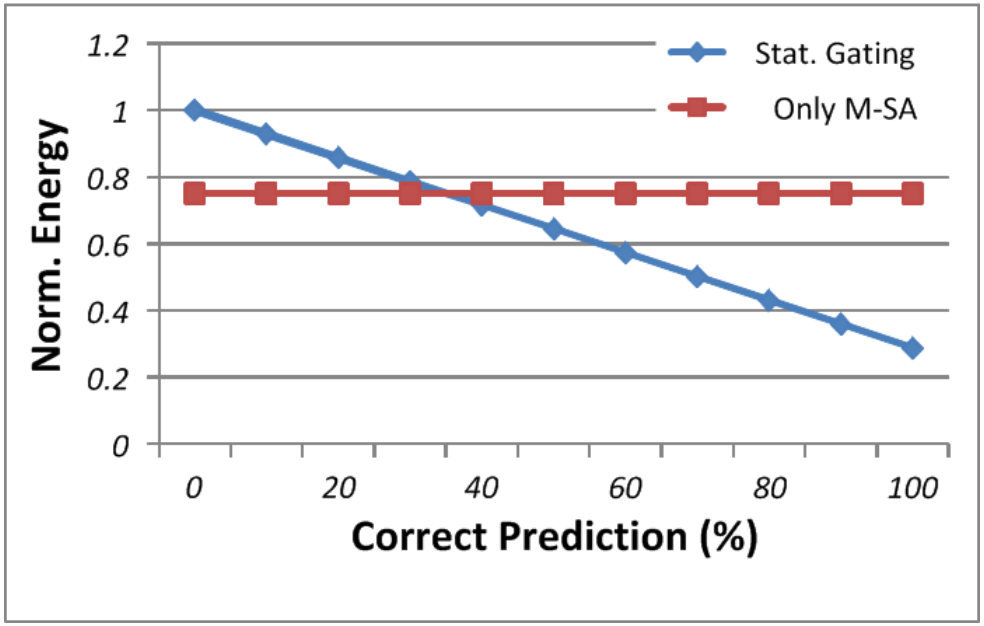




M-SA is activated

# Statistically Gated Sense-Amplifiers

| | | *pred - Correct* | *pred - Incorrect* |
|---|---|---|---|
| **E-SA w/ (-) offset** | *M-SA gated?* | Yes | No |
| | *Energy Consumed* | $C_{E\text{-}SA}V_{DD}^2$ | $C_{E\text{-}SA}V_{DD}^2 + C_{M\text{-}SA}V_{DD}^2$ |
| **E-SA w/ (+) offset** | *M-SA gated?* | No | No |
| | *Energy Consumed* | $C_{E\text{-}SA}V_{DD}^2 + C_{M\text{-}SA}V_{DD}^2$ | $C_{E\text{-}SA}V_{DD}^2 + C_{M\text{-}SA}V_{DD}^2$ |

Energy Savings with Statistically Gated Sense-Amps

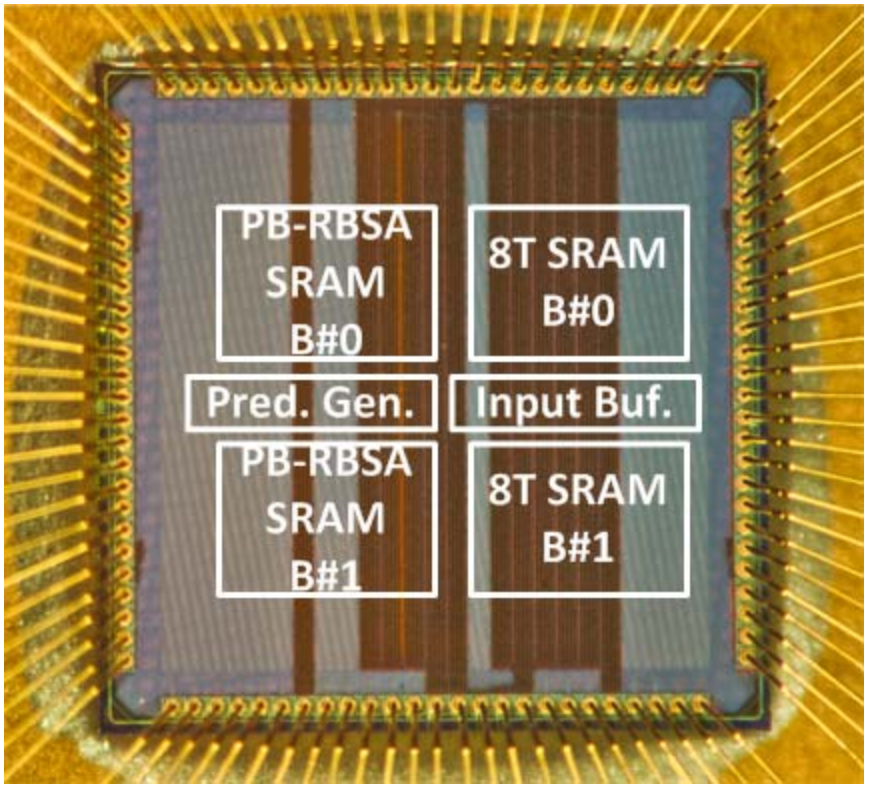



**Statistically Gated S/A provides energy savings when prediction is correct > 40% of the time**

# Outline

- **Application Specific Design Decisions**
  - Motion Estimation Specific Features
  - Bit-line Switching Activity
- **Prediction Based Reduced Bit-line Switching Activity (PB-RBSA) SRAM Design**
  - Bit-cell & Array Design
  - Prediction Generation
  - Statistically-Gated Sense-Amplifiers
- **Measurement Results**
- **Conclusions**

# Test Chip Features

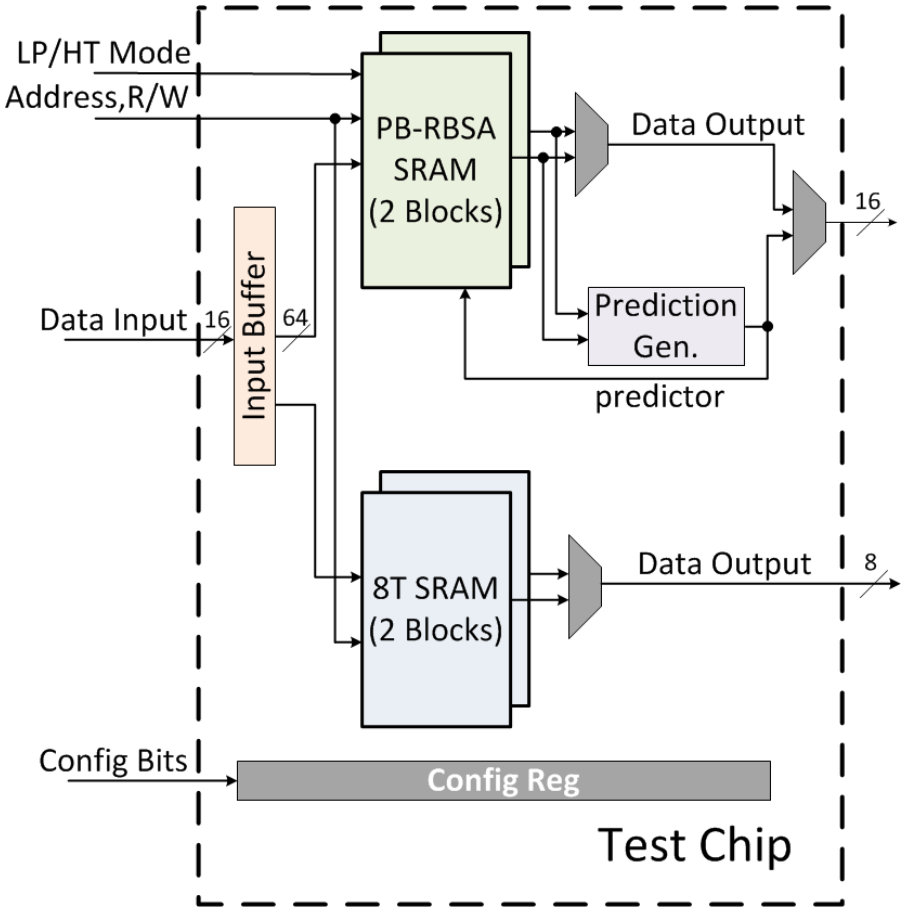



| Summary of Features | |
|---|---|
| Technology | 65nm LP CMOS |
| Die Size | 2.3mm x 2.3mm |
| PB-RBSA Organization | 32Kb [256 x 64bits x 2blk] |
| 8T SRAM Organization | 32Kb [256 x 64bits x 2blk] |
| Voltage Range | 0.52V – 1.2V |
| Frequency Range | 0.11MHz – 145Mhz |



© 2013 IEEE

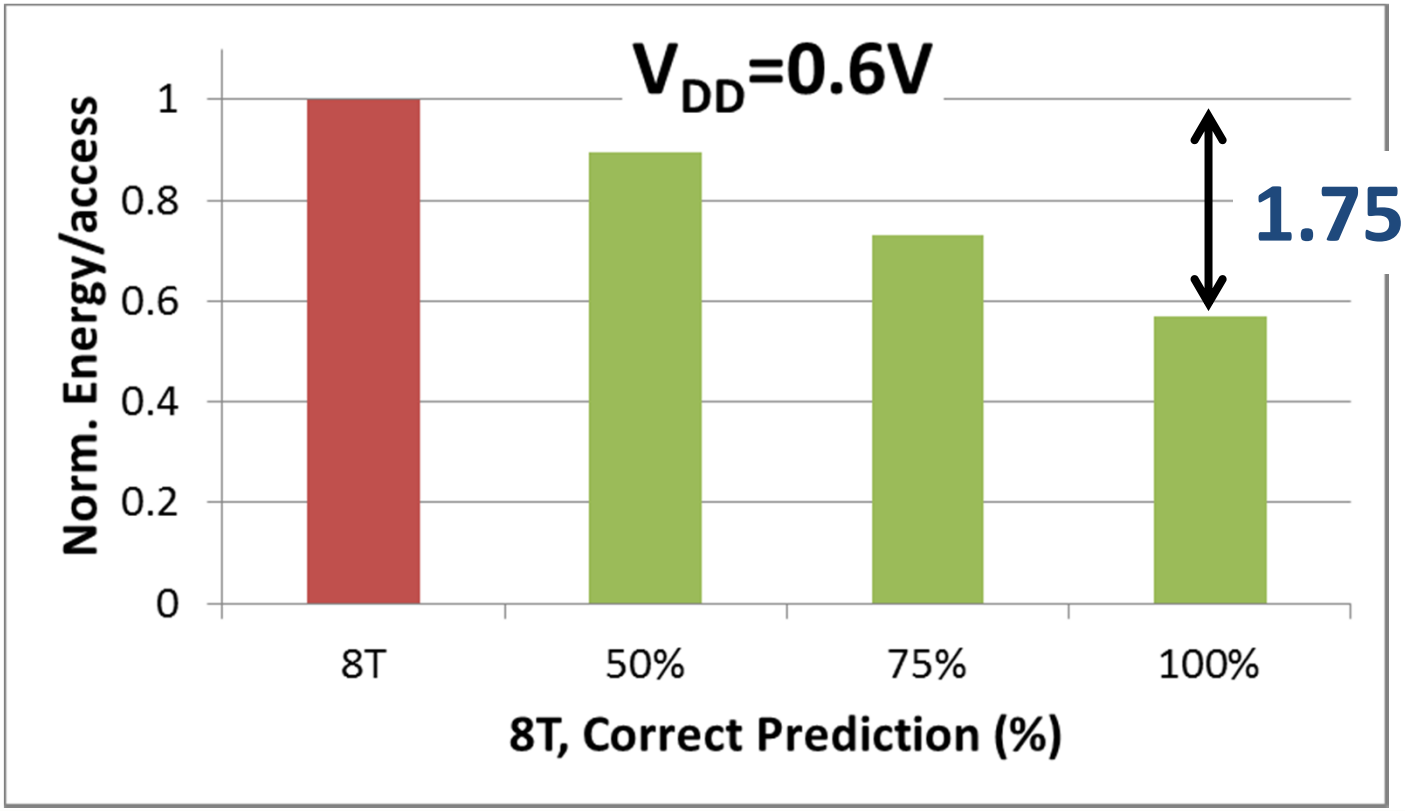
# Test Chip Architecture




| Summary of Features | |
|---|---|
| Technology | 65nm LP CMOS |
| Die Size | 2.3mm x 2.3mm |
| PB-RBSA Organization | 32Kb [256 x 64bits x 2blk] |
| 8T SRAM Organization | 32Kb [256 x 64bits x 2blk] |
| Voltage Range | 0.52V – 1.2V |
| Frequency Range | 0.11MHz – 145Mhz |

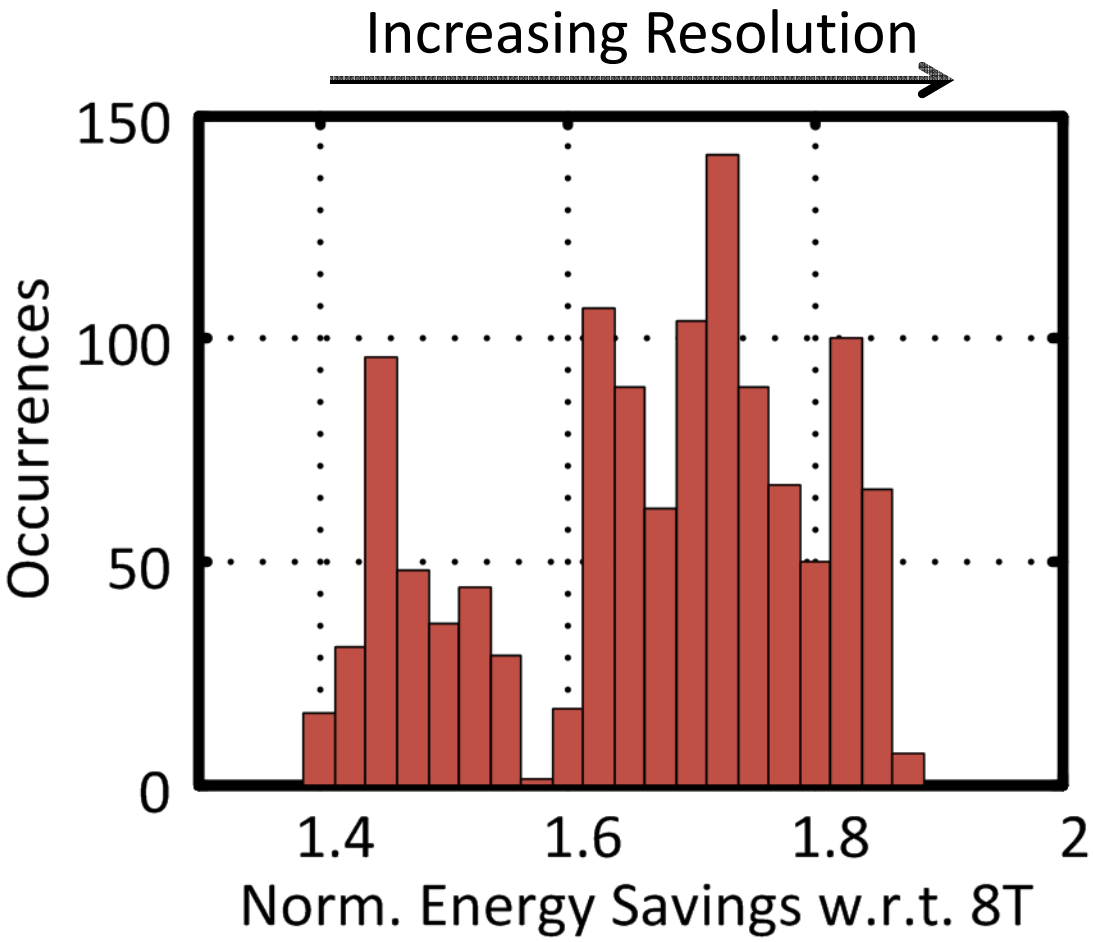
# Measured Read Energy/access

Array holds 50% '0' and 50% '1'




**Up to 1.75× energy/access reduction compared to a conventional 8T design ($\alpha_{0\rightarrow1,8T}$ = 0.5)**

# Measured Energy/access




*# of Video Sequences*

*11*

*# of Frames*

*1100*

*Resolution:*

*1280x720 to 2560x1600*

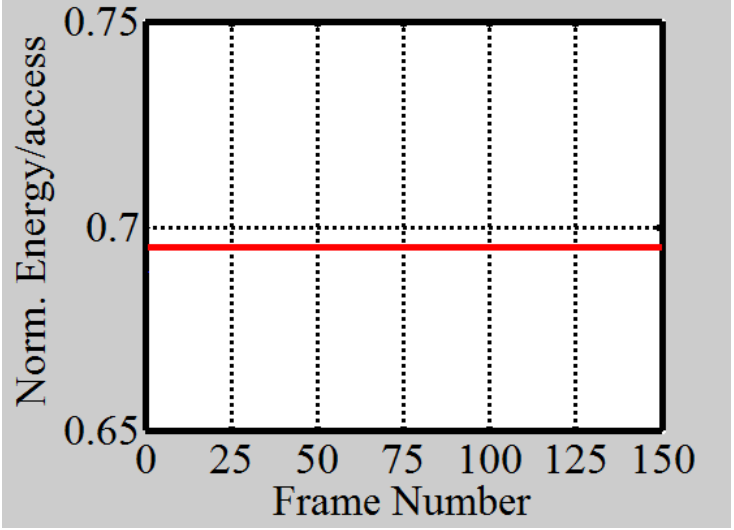
**Up to 1.9× energy savings with PB-RBSA SRAM**

# PB-RBSA vs. 8T Comparison

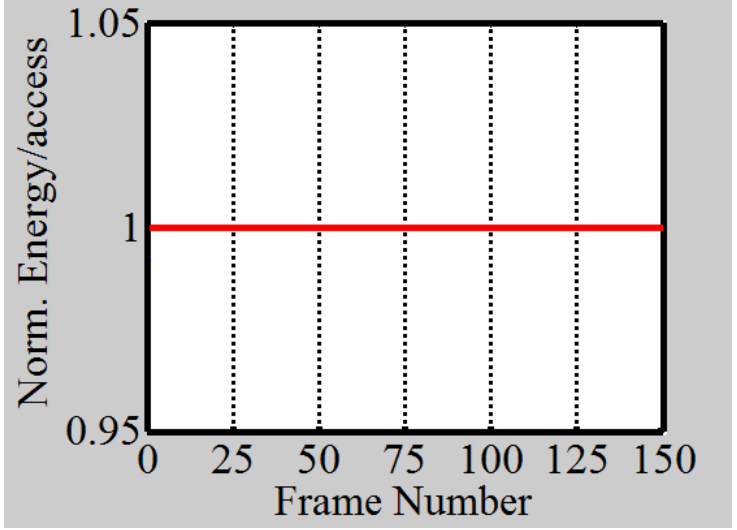
*Sequence: RaceHorseD*
*Resolution: 416x240*
*# of Frames: 150*

**Average Energy/access savings is 1.45×**

**PB-RBSA SRAM**




**8T SRAM**

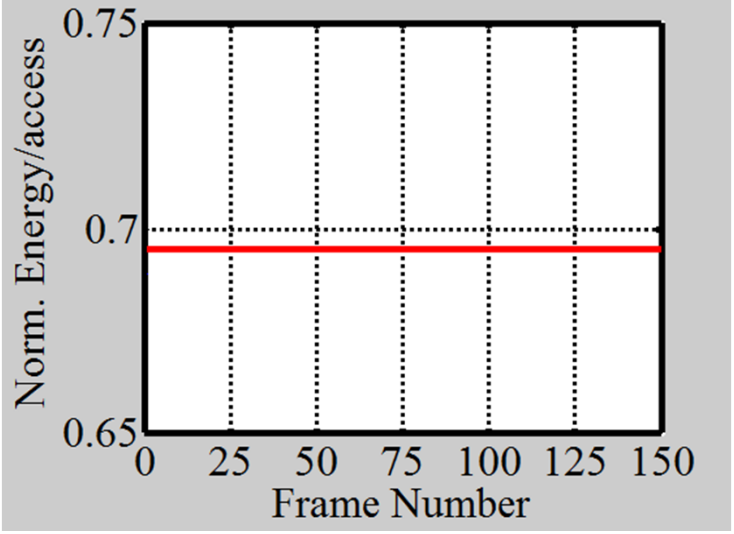


# PB-RBSA vs. 8T Comparison

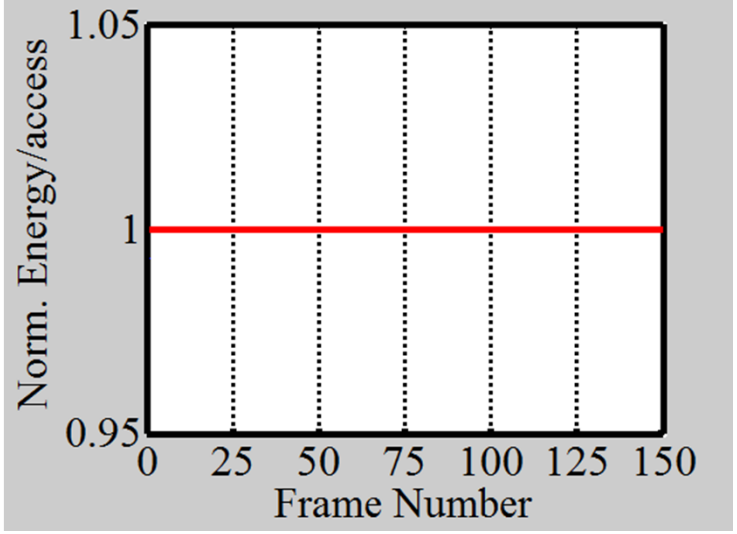
*Sequence: RaceHorseD*
*Resolution: 416x240*
*# of Frames: 150*

**Average Energy/access savings is 1.45×**

**PB-RBSA SRAM**




**8T SRAM**



# Outline

- **Application Specific Design Decisions**
  - Motion Estimation Specific Features
  - Bit-line Switching Activity
- **Prediction Based Reduced Bit-line Switching Activity (PB-RBSA) SRAM Design**
  - Bit-cell & Array Design
  - Prediction Generation
  - Statistically-Gated Sense-Amplifiers
- **Measurement Results**
- **Conclusions**



© 2013 IEEE

# Conclusions

- Application-specific features provide a new dimension for energy-efficient SRAM design
- PB-RBSA SRAM provides up to 1.9× reduction in energy/access by utilizing correlation of data and output prediction
- Incorporating signal statistics into circuit design can provide further improvements in energy efficiency
- Acknowledgements
  - Funding by Texas Instruments
  - Chip fabrication by the TSMC University Shuttle Program

# A 27% Active and 85% Standby Power Reduction in Dual-Power-Supply SRAM Using BL Power Calculator and Digitally Controllable Retention Circuit

**F. Tachibana**, O. Hirabayashi, Y. Takeyama, M. Shizuno, A. Kawasumi, K. Kushida, A. Suzuki, Y. Niki, S. Sasaki, T. Yabe and Y. Unekawa

Toshiba Corporation
Kawasaki, Japan

# Outline

- Background
  - Active Mode
  - Standby Mode
- Power Management Unit
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- Measurement Results
- Conclusion

# Outline

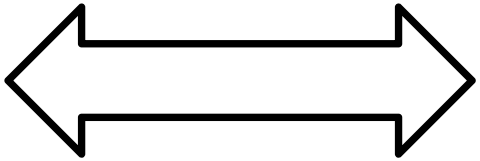
- **Background**
  - Active Mode
  - Standby Mode
- Power Management Unit
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- Measurement Results
- Conclusion

# Background

Low Performance
- Low Freq.
- Standby >> Active

Low Power

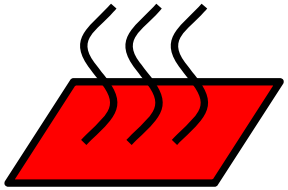
Cool



Our target

High Performance
- High Freq.
- Active >> Standby

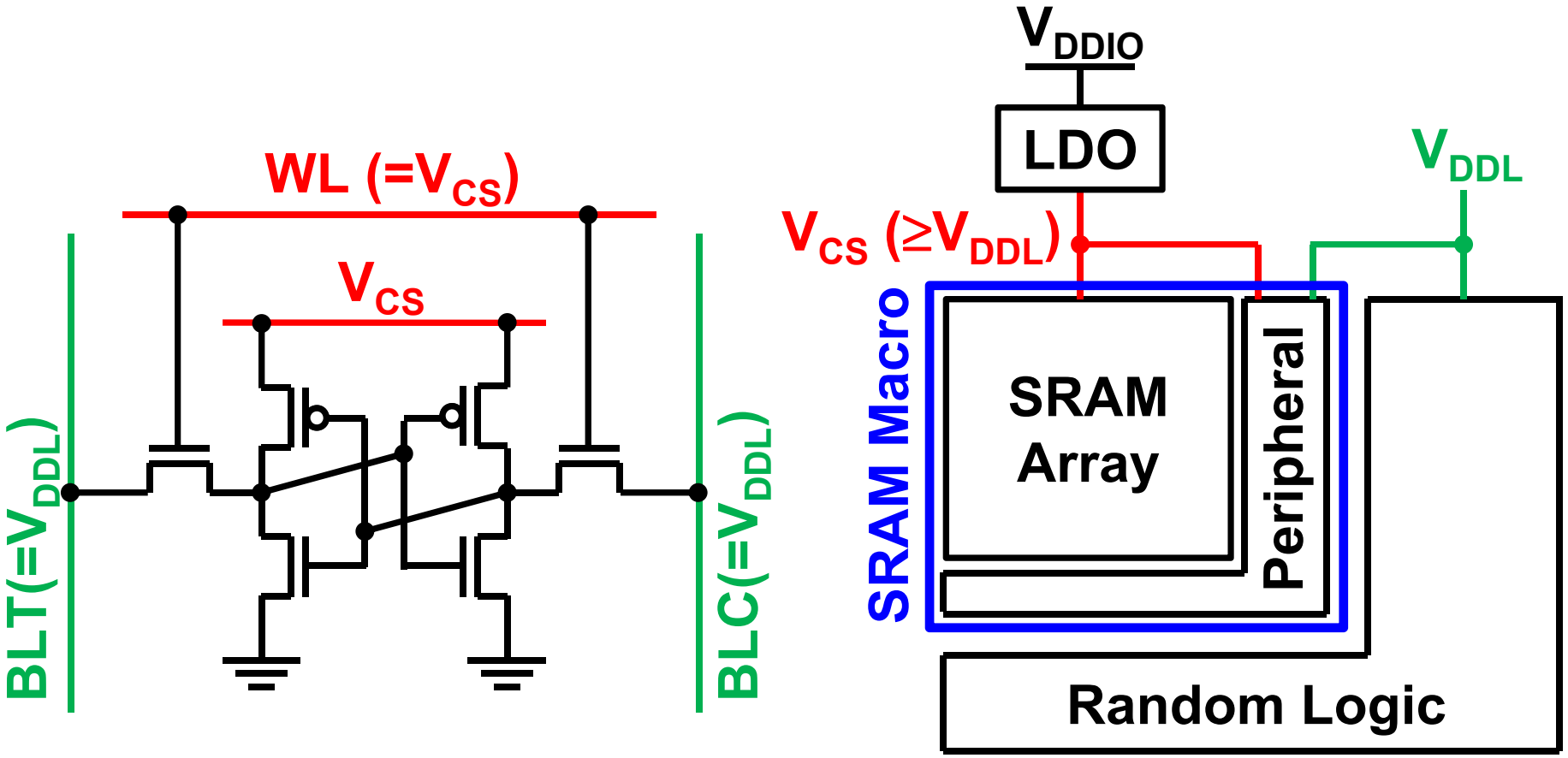
High Power



Hot

- Multimedia SoC is required to cover low to high performance mode.
- For a longer battery life, power reduction at room temperature (RT) is also important.

# Conventional Dual-Power-Supply SRAM

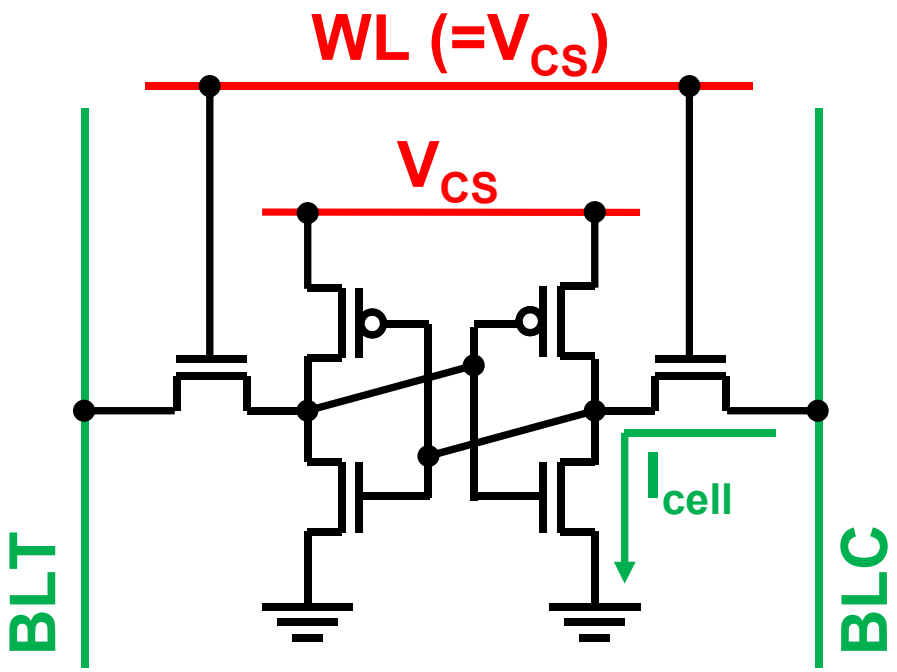


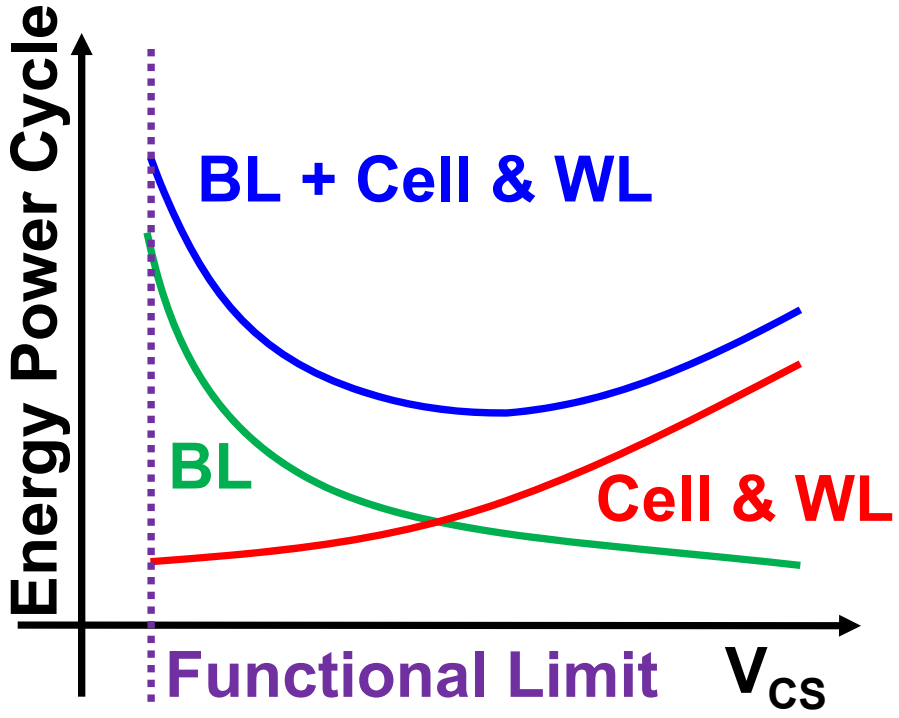

**$V_{CS}$ is controlled to the minimum functional voltage, which is NOT power optimum point.**

# Outline

- **Background**
  - **Active Mode**
  - Standby Mode
- Power Management Unit
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- Measurement Results
- Conclusion

# Power Consumption in Active Mode



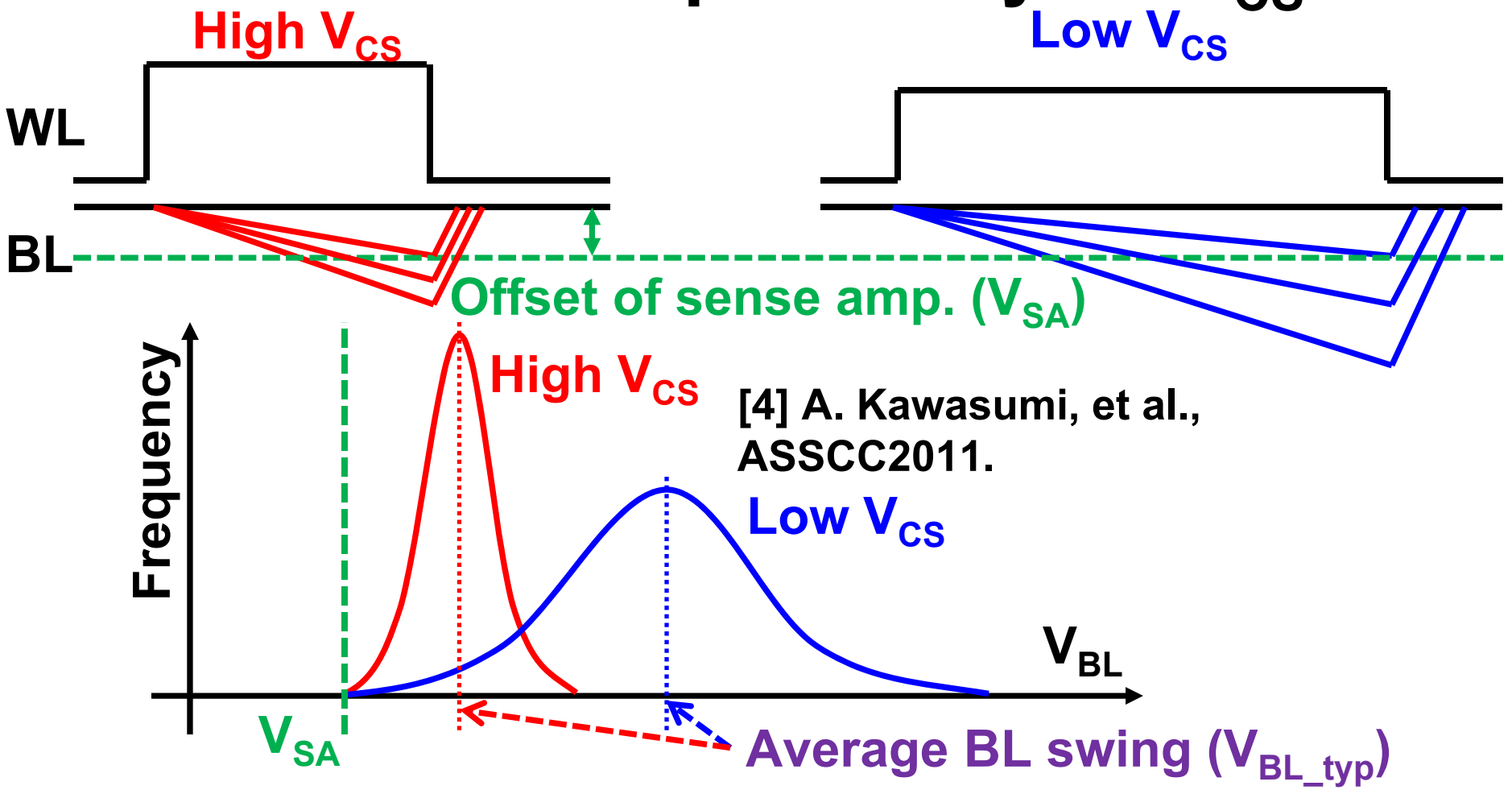




**Because of the BL power, optimum $V_{CS}$ point for the SRAM power is NOT usually minimum $V_{CS}$.**



© 2013 IEEE

# BL Power Dependency on $V_{CS}$

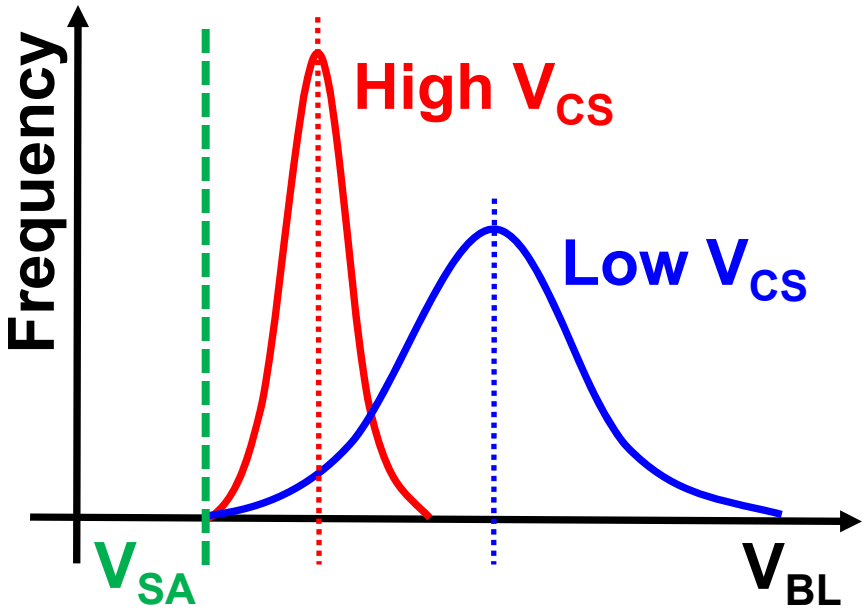


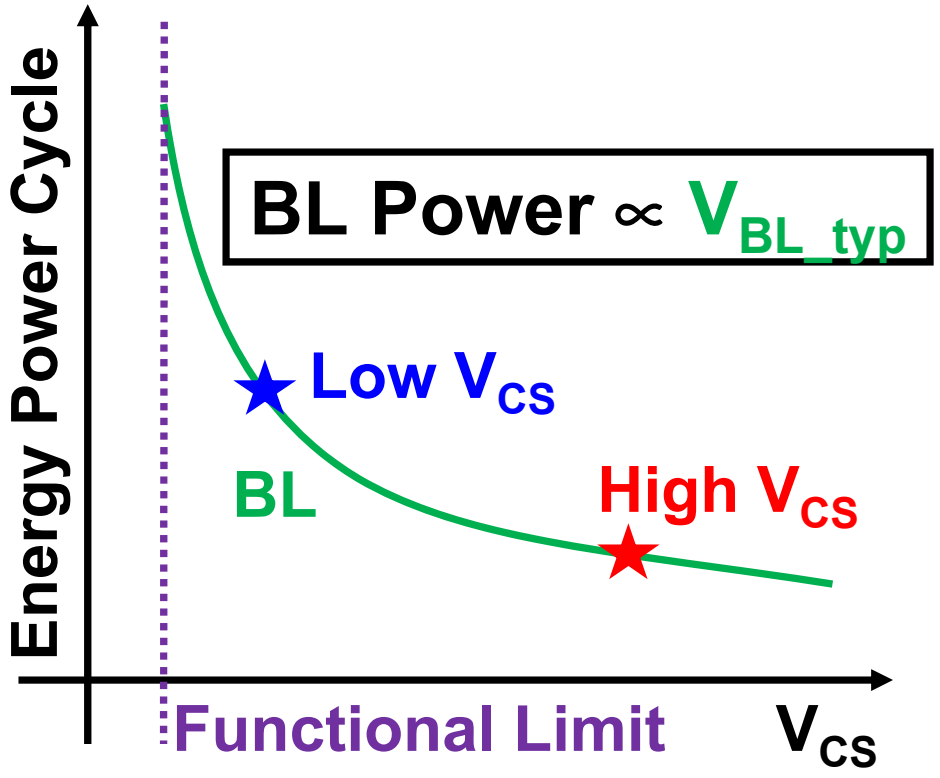

- Due to random variation, average BL swing becomes larger at lower $V_{CS}$.



© 2013 IEEE

# BL Power Dependency on $V_{CS}$



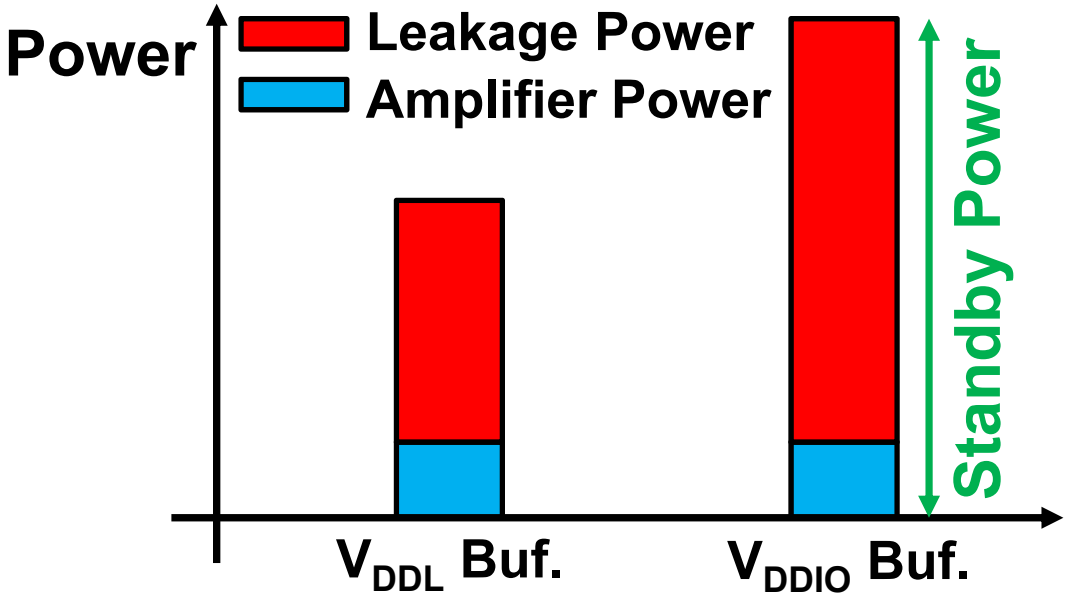



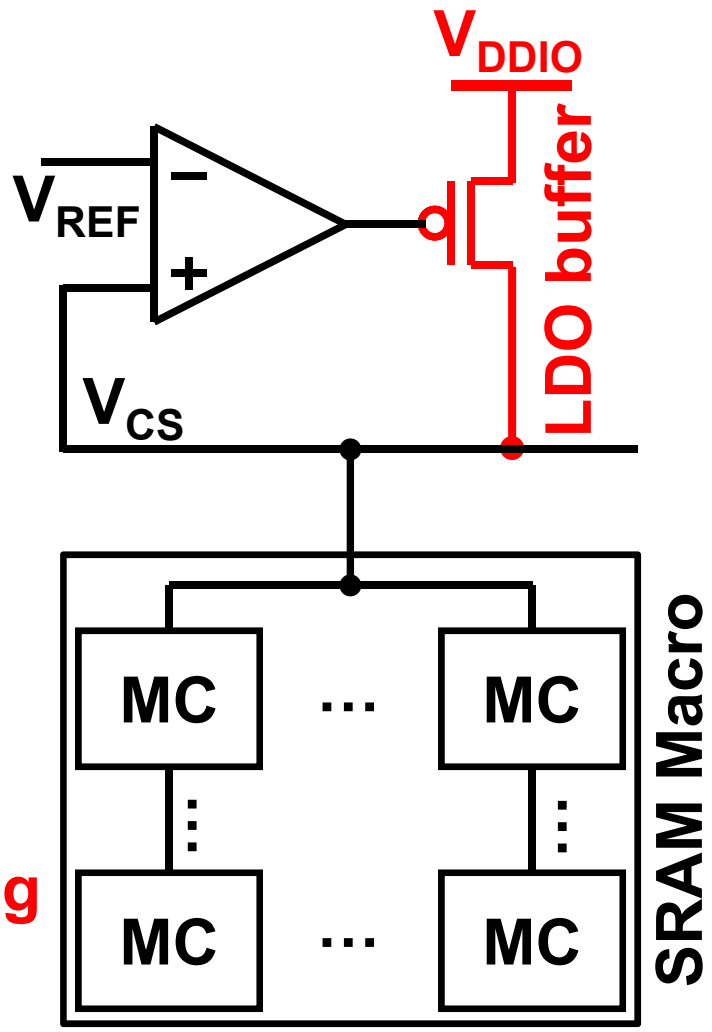

- BL power is largest at minimum functional $V_{CS}$.
- Scheme to calculate BL power information is required.

# Outline

- **Background**
  - Active Mode
  - **Standby Mode**
- Power Management Unit
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- Measurement Results
- Conclusion

# Conventional Scheme in Standby Mode



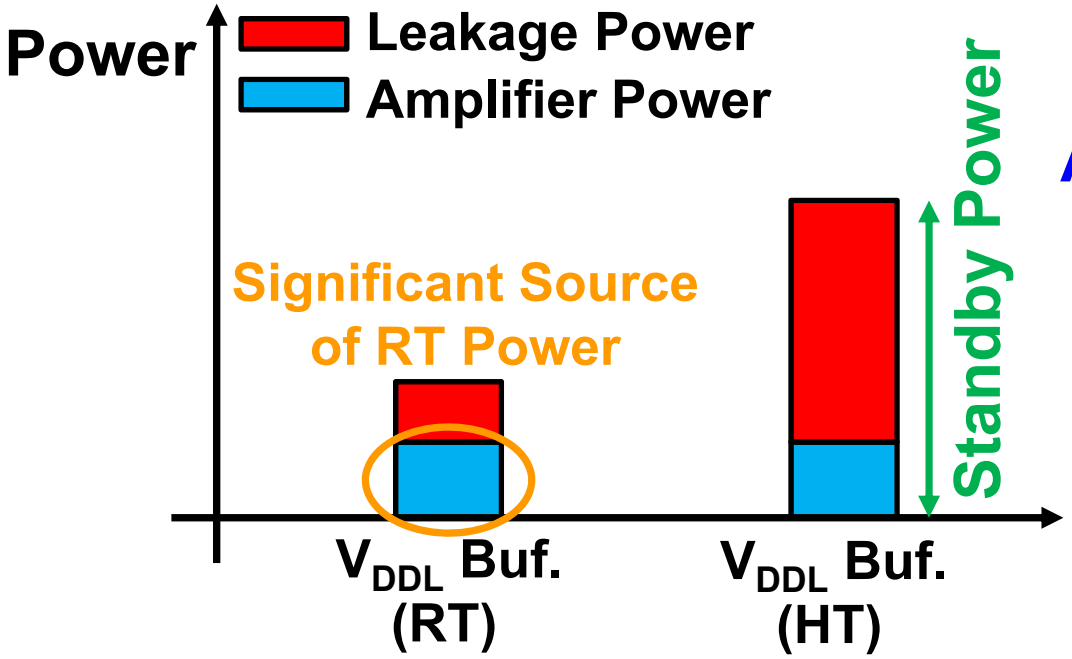

**Problems in leakage power reduction for high and low performance mode**

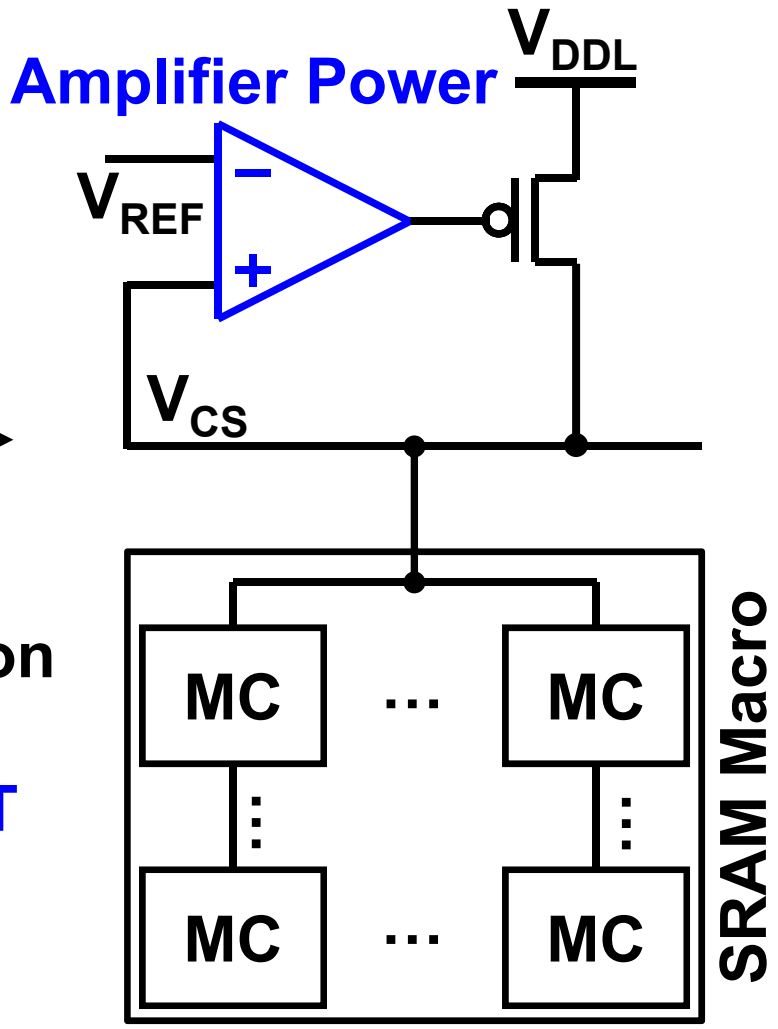
**•Leakage power (P=V*I) can only be reduced in retention mode by reducing source voltage.**



© 2013 IEEE

# Conventional Scheme in Standby Mode



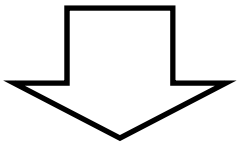

**Problems in leakage power reduction for low performance mode**

- **Constant Power of amplifier is NOT negligible at RT.**

# Objectives of This Work

## Active Mode Power Reduction

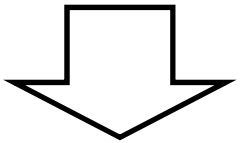
- Calculate the information about BL power



**Replica BL based BL power calculator**

## Standby Mode Power Reduction

- Use $V_{DDL}$ as buffer power supply
- Reduce $V_{CS}$ control power



**Digitally controllable retention circuit**

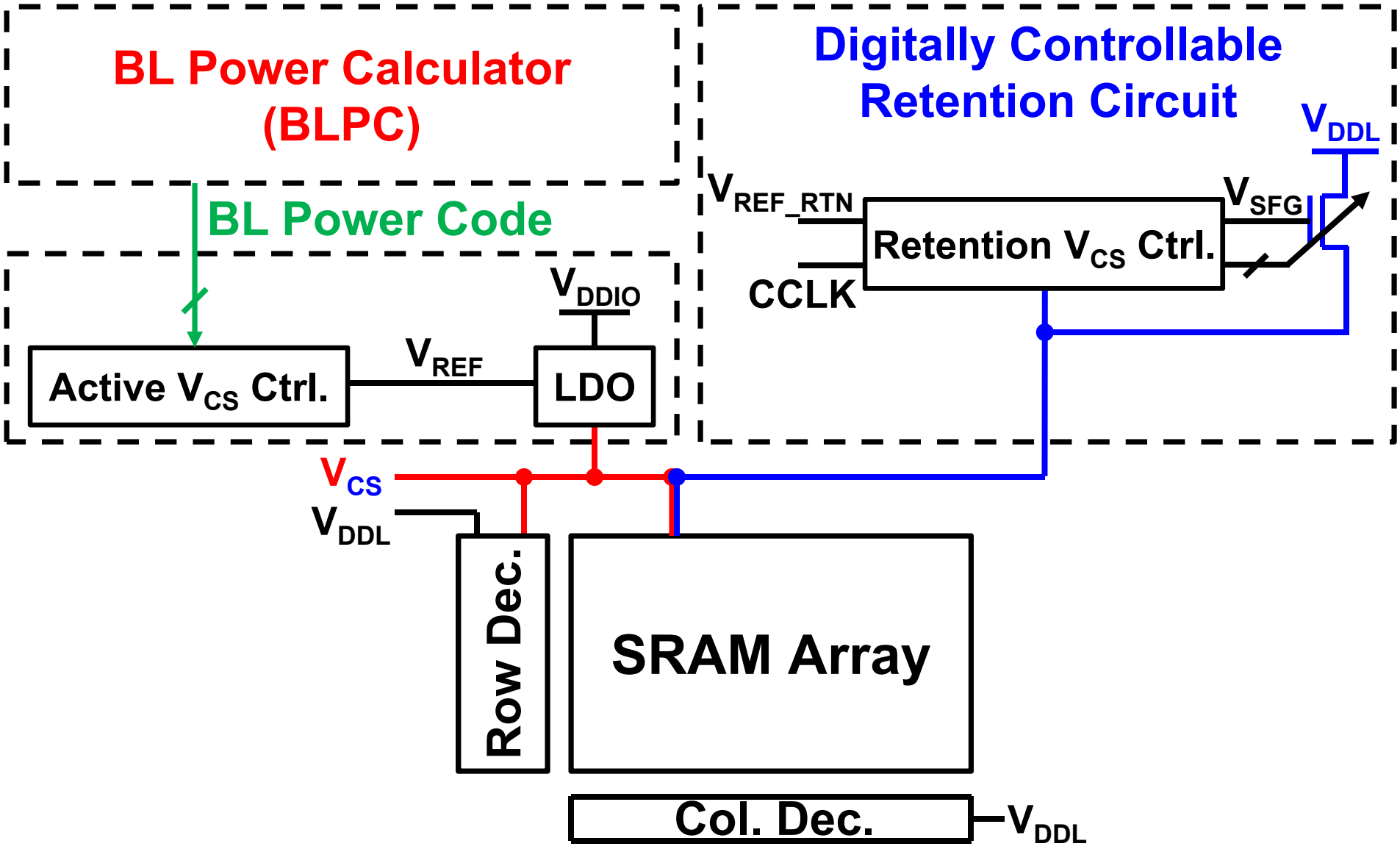


© 2013 IEEE

# Outline

- Background
  - Active Mode
  - Standby Mode
- **Power Management Unit**
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- Measurement Results
- Conclusion

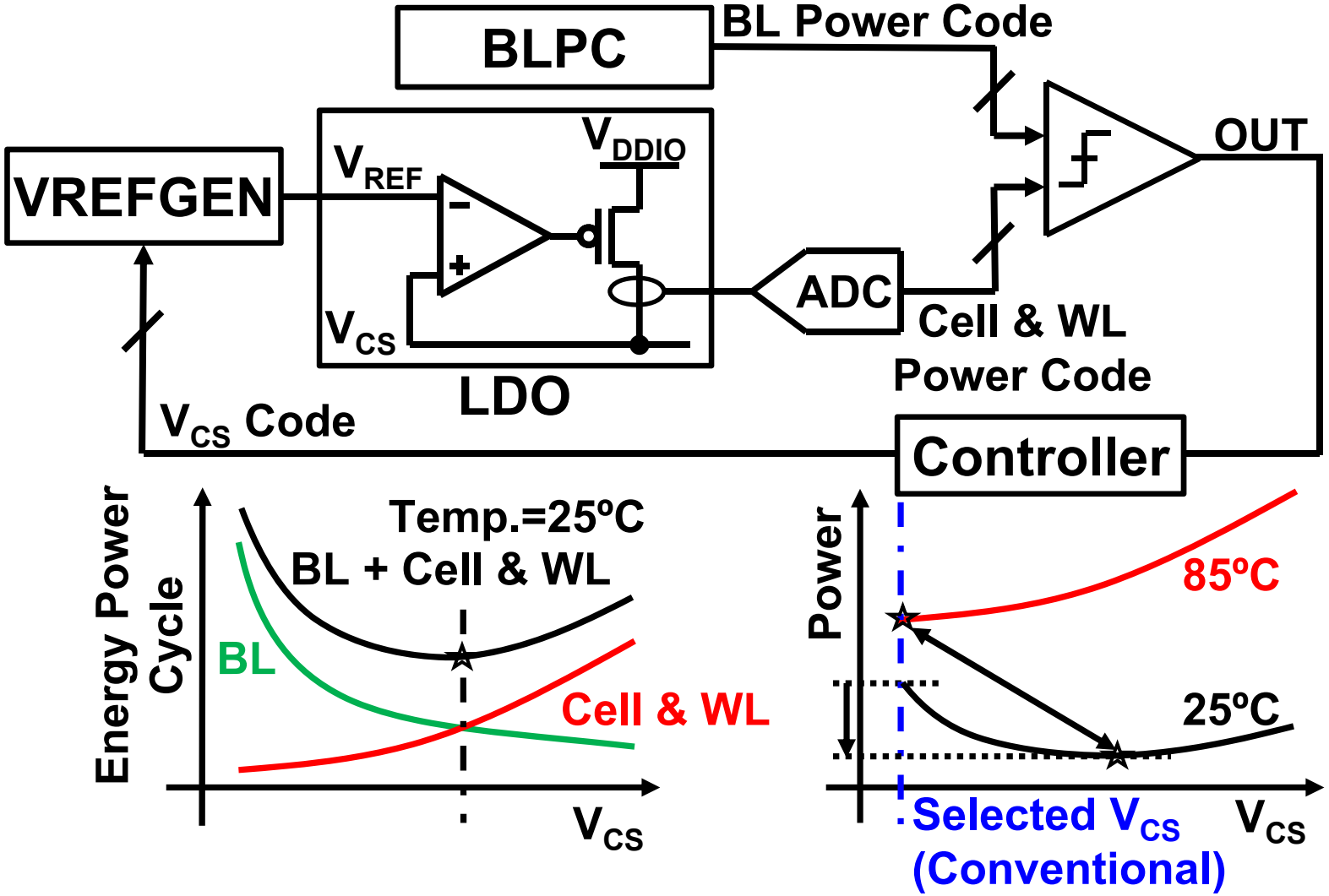
# Block Diagram of Proposed SRAM



# Outline

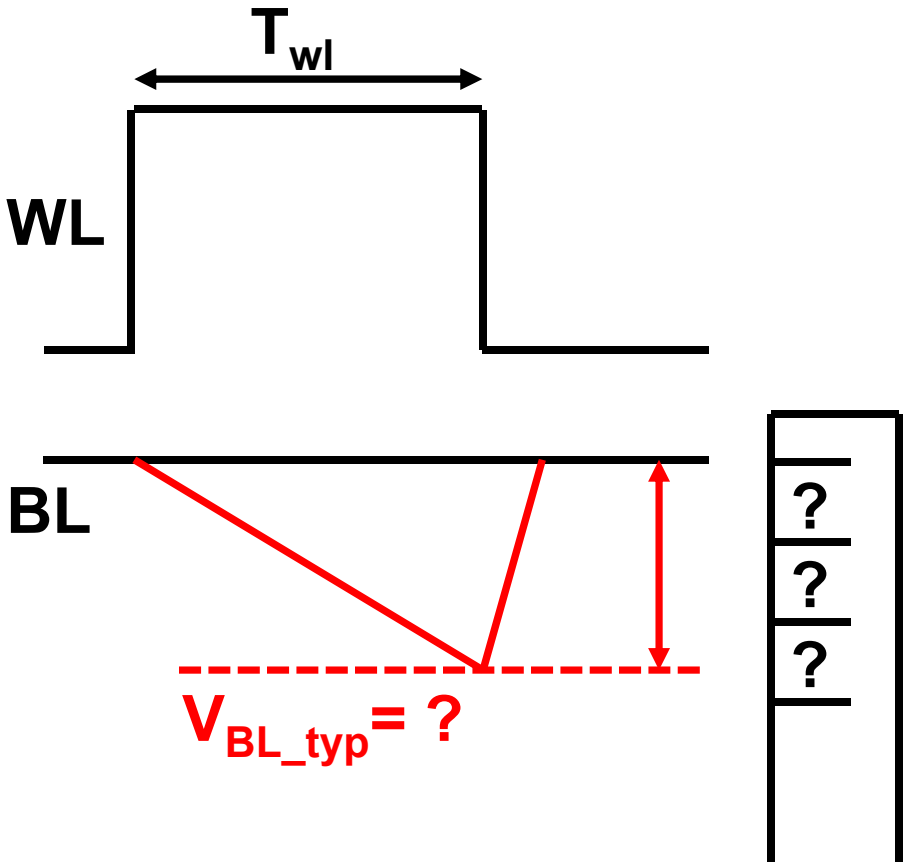
- Background
  - Active Mode
  - Standby Mode
- **Power Management Unit**
  - **BL Power Calculator**
  - Digitally Controllable Retention Circuit
- Measurement Results
- Conclusion

# Active Mode Power Control



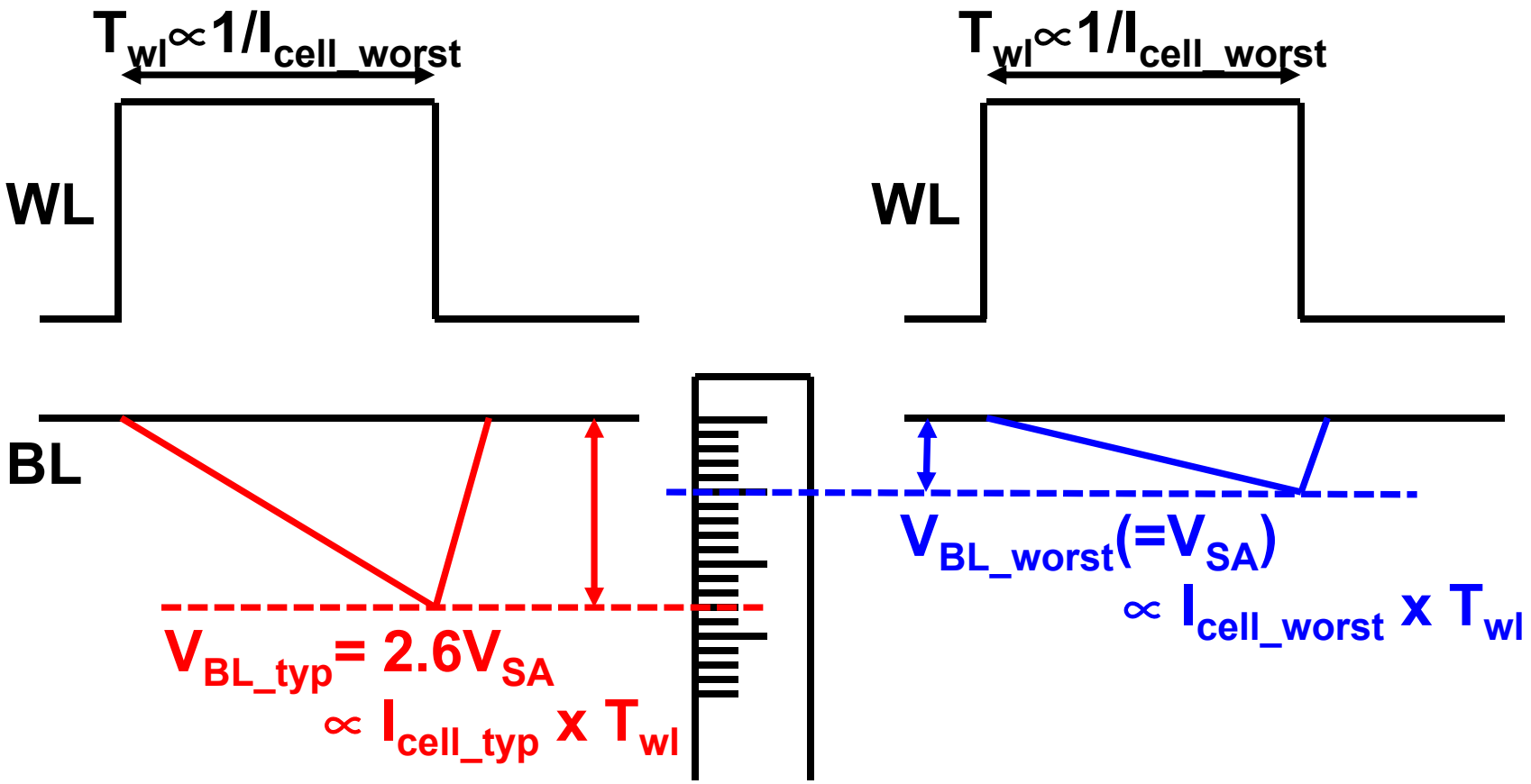

- $V_{CS}$ is selected to balance BL and Cell & WL power.
- Selected point changes with PVT variation.

# Basic Concept of BL Power Calculator



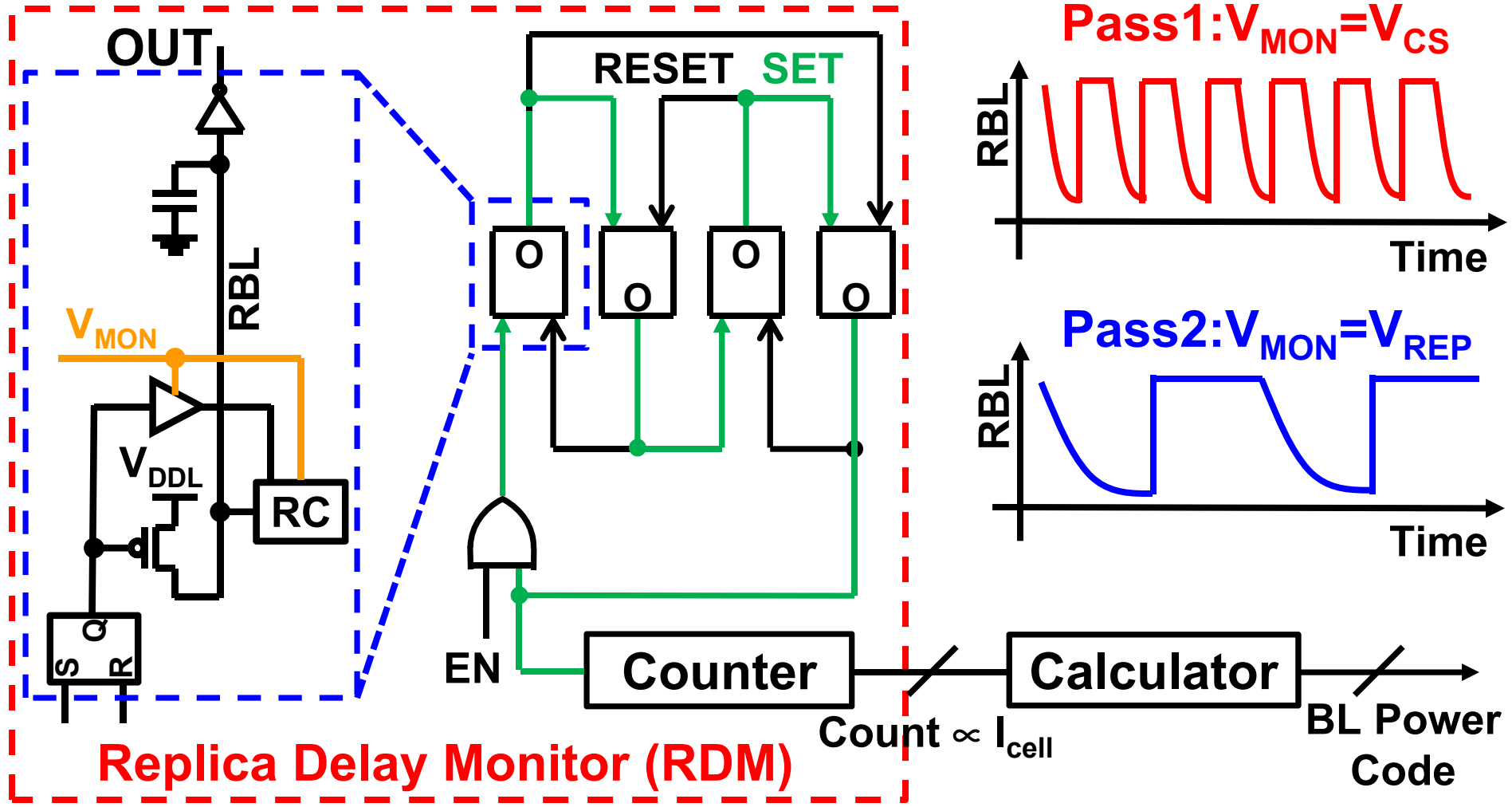

- To measure $V_{BL\_typ}$, reference ruler is required.

# Basic Concept of BL Power Calculator



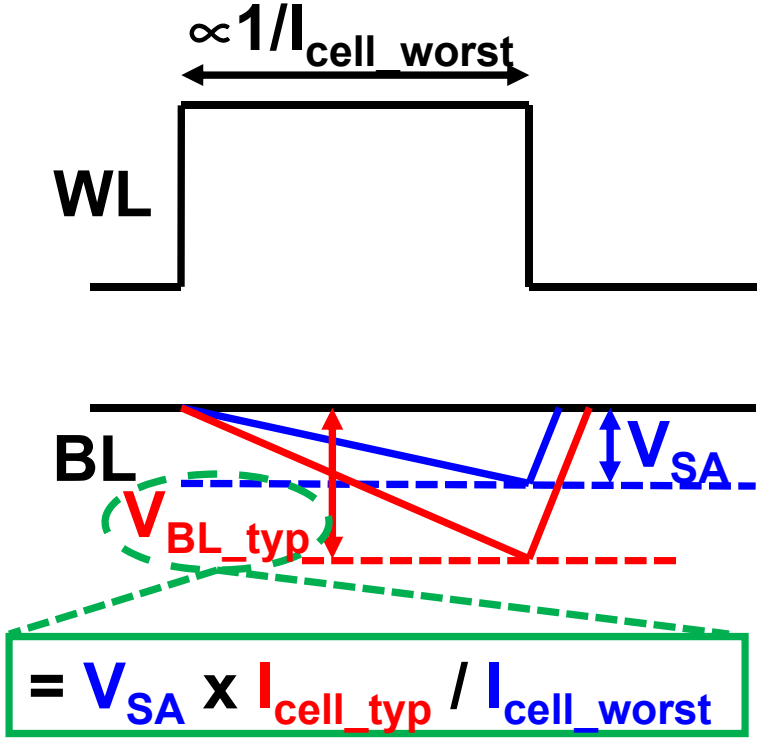

- When $V_{BL\_worst}$ is controlled to $V_{SA}$, $V_{SA}$ can be used as a reference.
- With $V_{SA}$ and the ratio of $I_{cell}$, $V_{BL\_typ}$ can be obtained.

# BL Power Calculator




- RDM consists of replica BL based ring oscillator.
- Calc. outputs BL power using the count from RDM.

# How to Calculate BL power

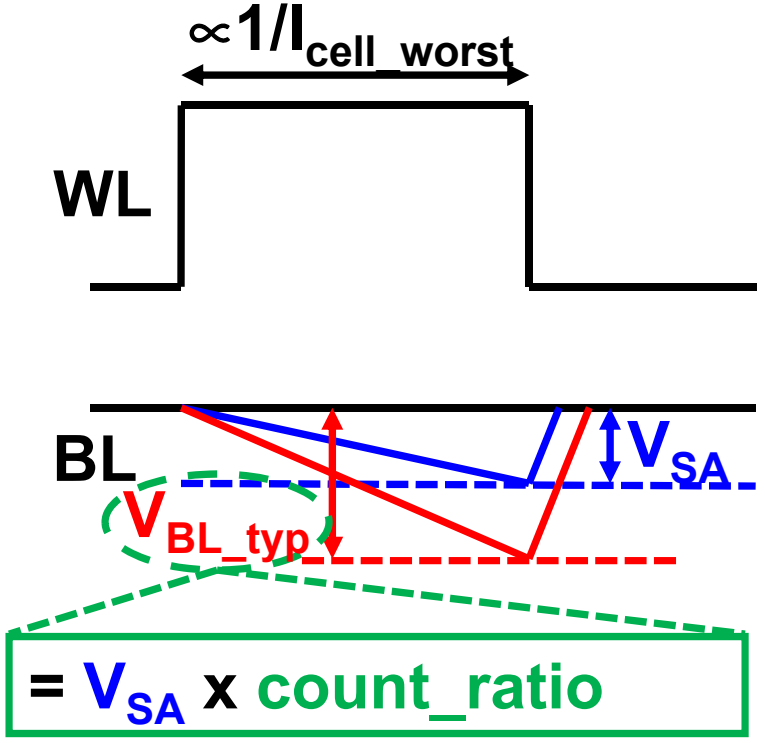


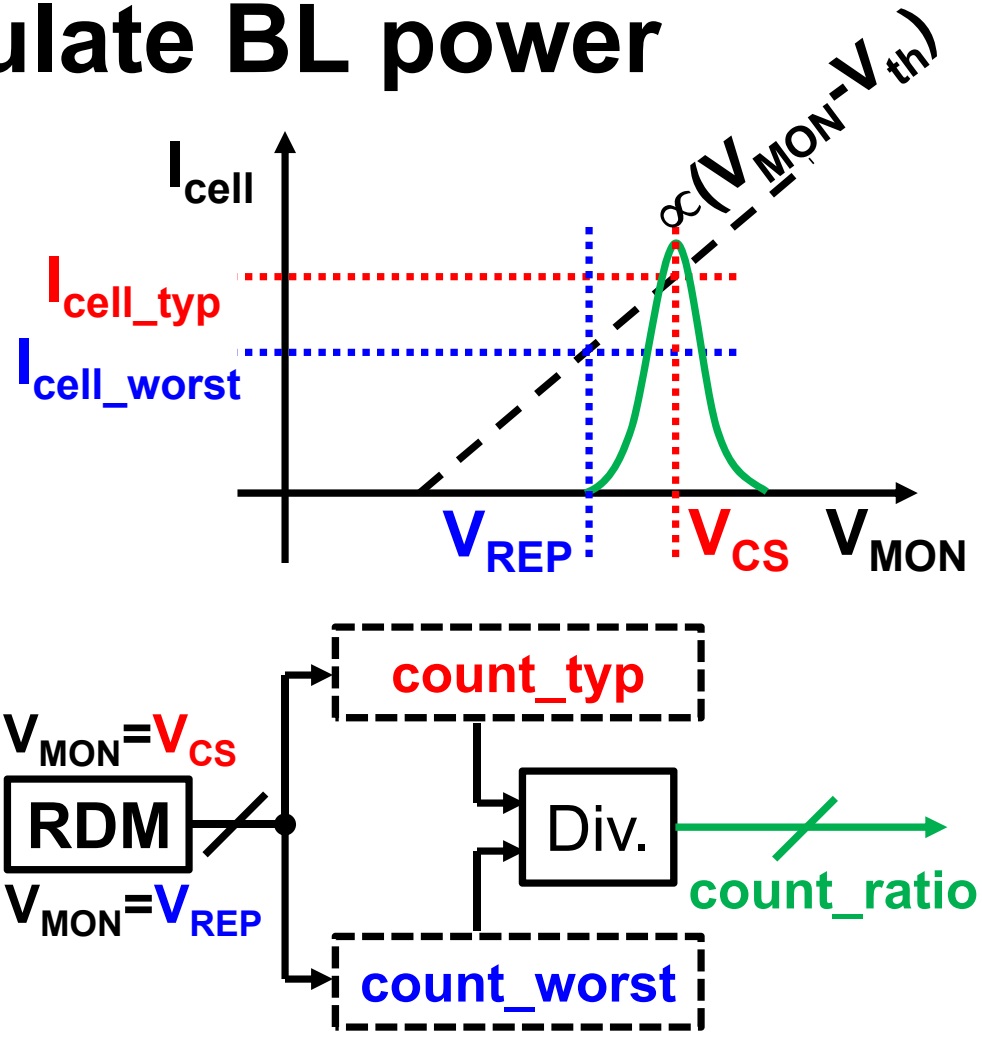

- Since BL slope is proportional to cell current, $V_{BL\_typ}$ is obtained from $V_{SA}$ and the ratio between $I_{cell\_typ}$ and $I_{cell\_worst}$.



© 2013 IEEE

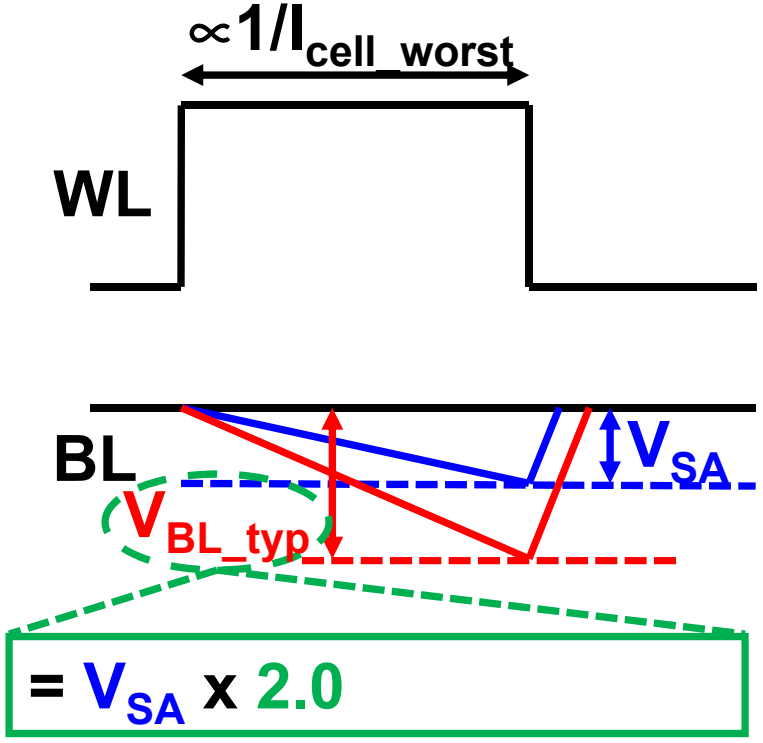
# How to Calculate BL power

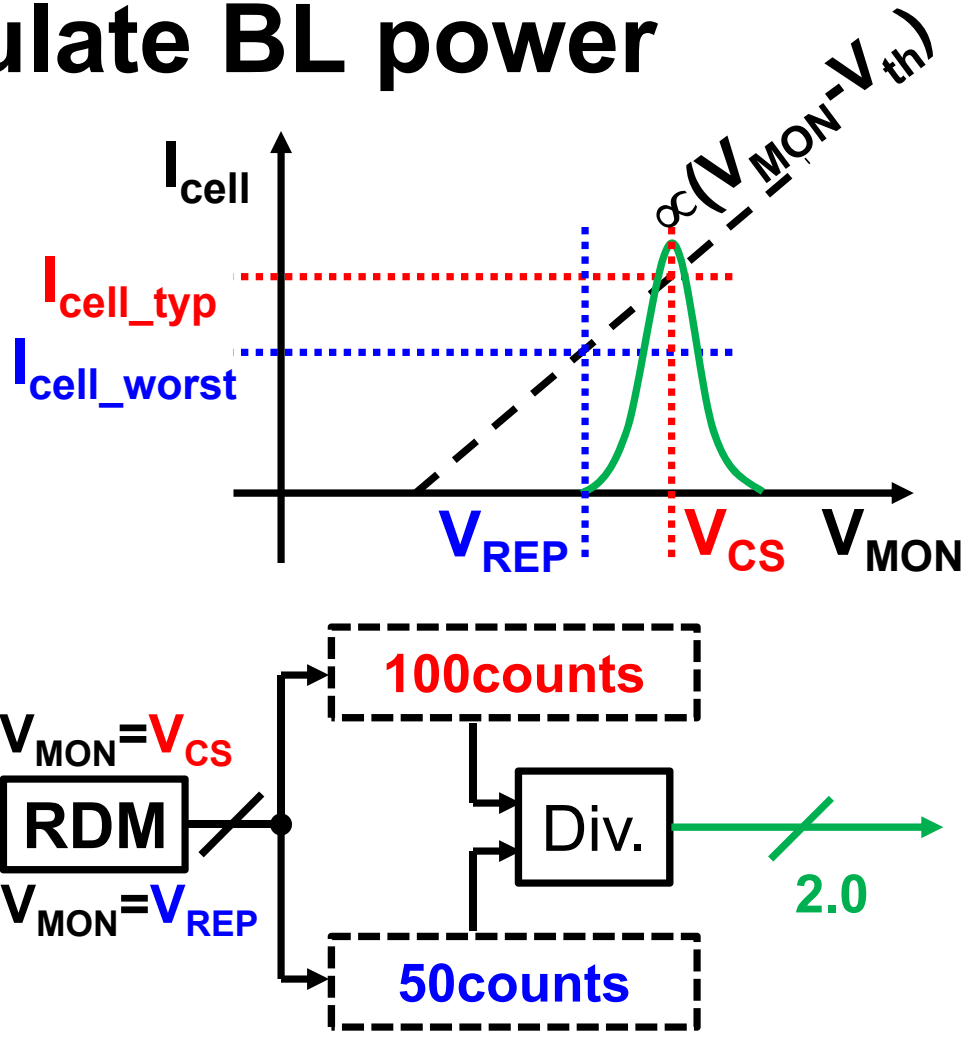



- Operating RDM twice, the ratio between $I_{cell\_typ}$ and $I_{cell\_worst}$ is obtained as count_ratio, and $V_{BL\_typ}$ is calculated.
- Since BL power ∝ $V_{BL\_typ}$, BL power is also calculated.

# How to Calculate BL power



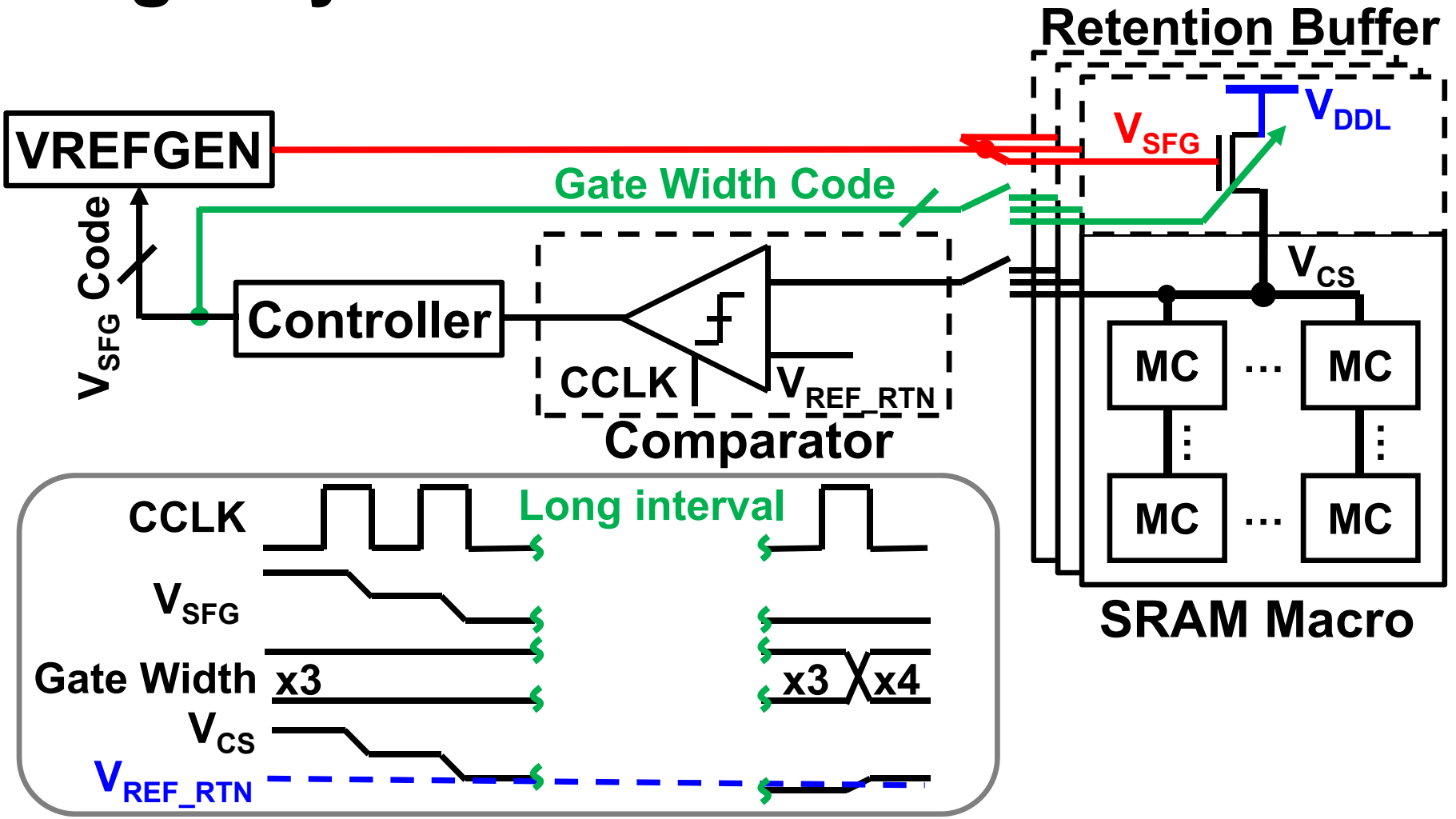




- Operating RDM twice, the ratio between $I_{cell\_typ}$ and $I_{cell\_worst}$ is obtained as count_ratio, and $V_{BL\_typ}$ is calculated.
- Since BL power ∝ $V_{BL\_typ}$, BL power is also calculated.

# Outline

- Background
  - Active Mode
  - Standby Mode
- **Power Management Unit**
  - BL Power Calculator
  - **Digitally Controllable Retention Circuit**
- Measurement Results
- Conclusion

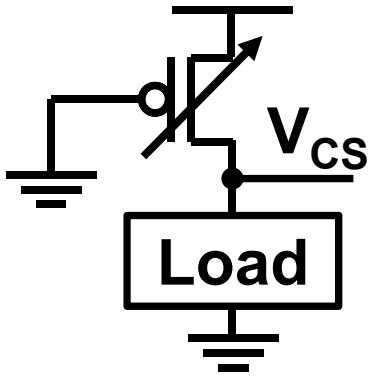
# Digitally Controllable Retention Circuit

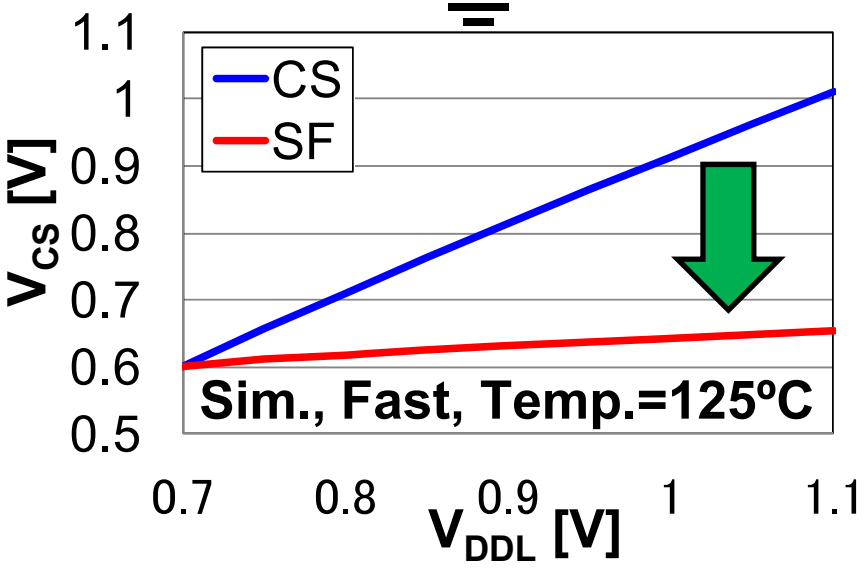



- $V_{CS}$ is controlled by digital code(Amp. removed from macro).
- Control BLK can be shared by different macros.

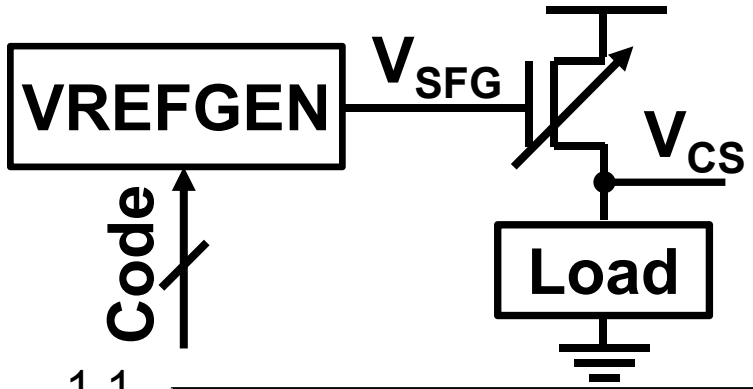
# SF based Digital LDO vs CS based digital LDO

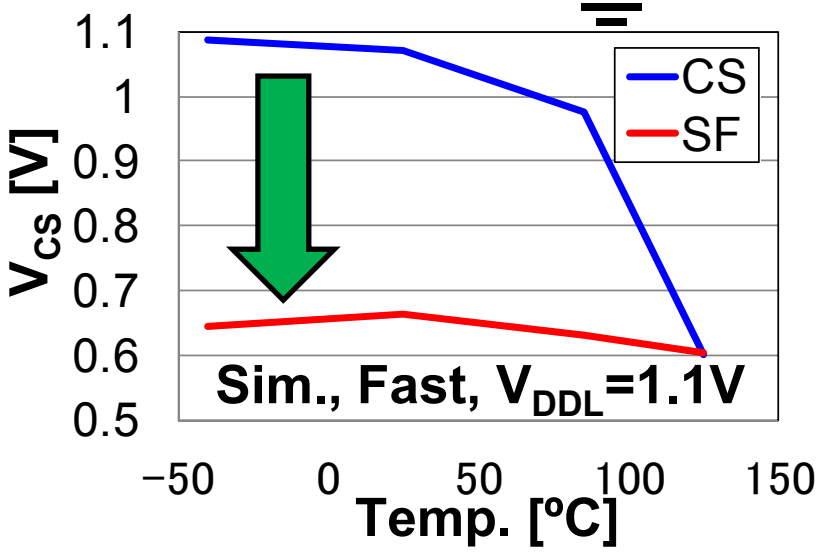
## Common Source (CS) Buffer (Conventional)



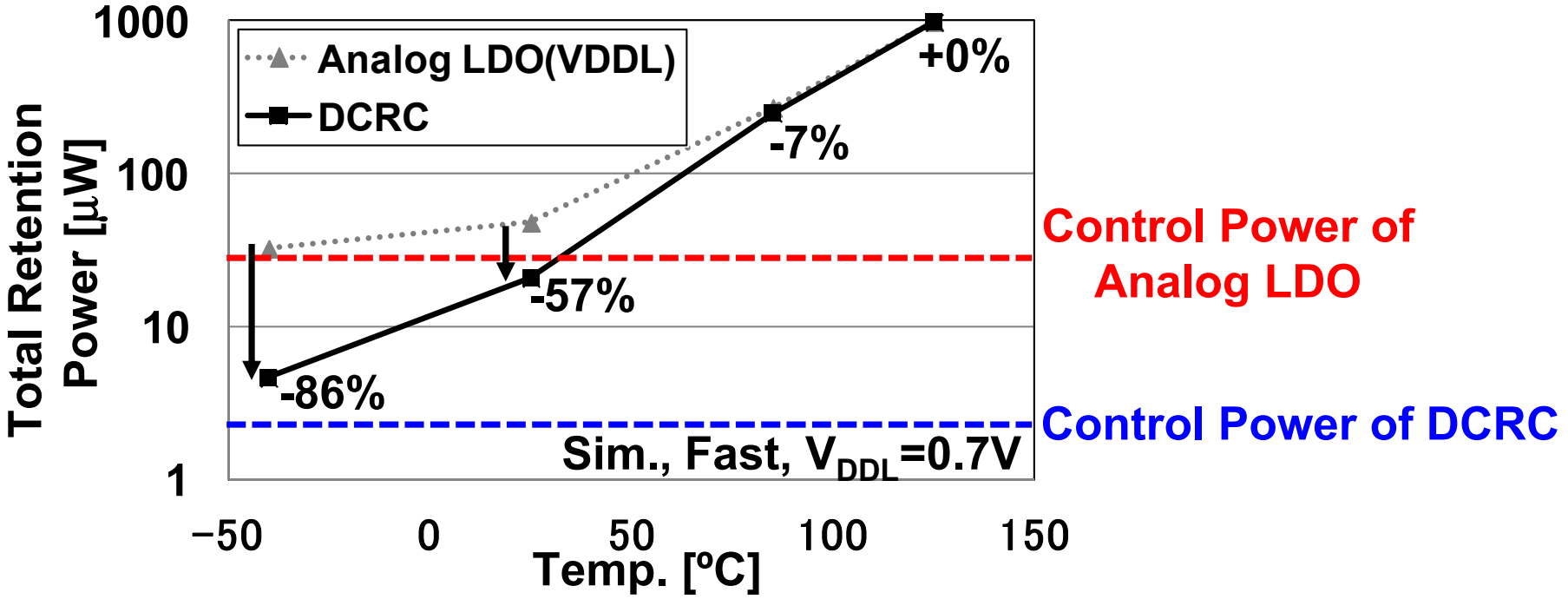

## Source Follower (SF) Buffer (Proposed)




- Compared with CS, SF is immune to VT change.
- SF buffer does NOT need frequent update.

# Retention Power Comparison



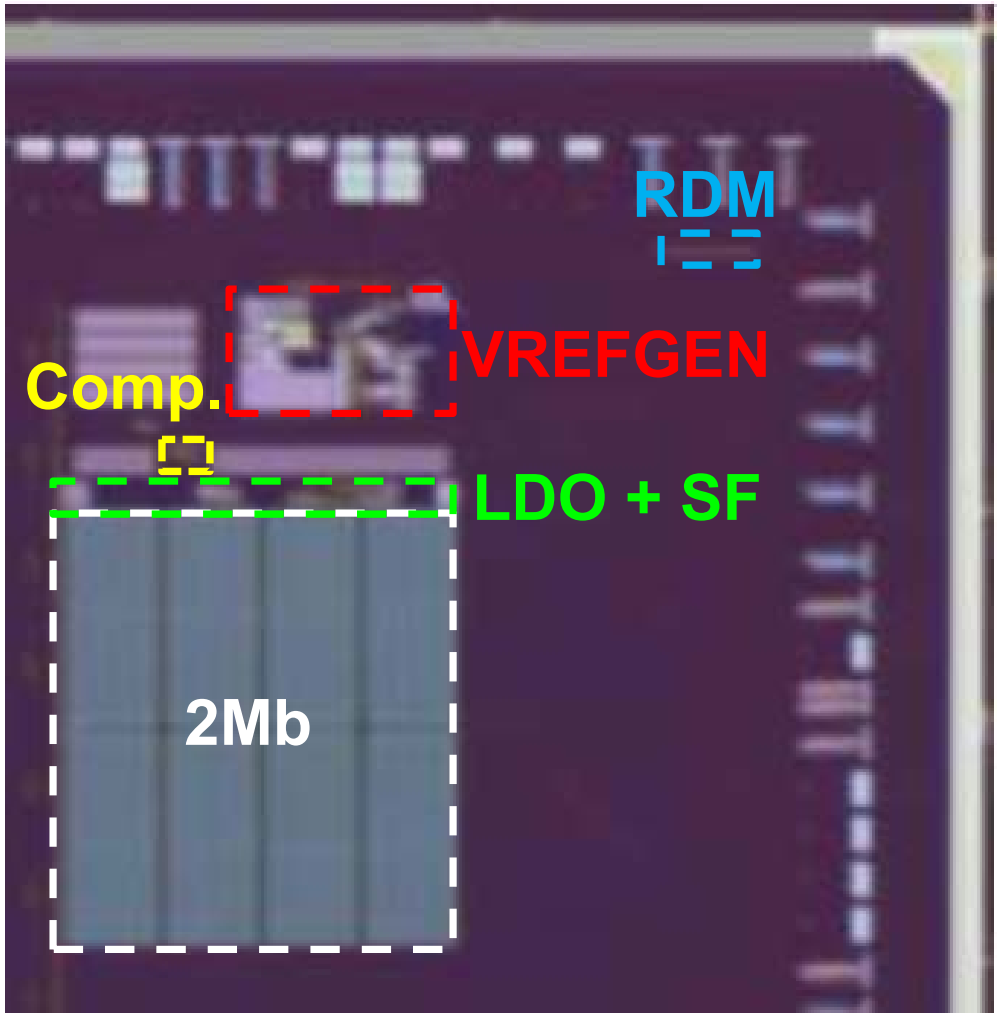

| Type | Control Power | Total Power (w/ VT) | Area Overhead |
|---|---|---|---|
| Analog LDO($V_{DDL}$) | ~30mW | 32m~**1.5mW** | 1~2% |
| DCRC | **~2mW** | **4.6m~1.5mW** | **<1%** |

**Since control power is reduced, retention power with DCRC can reduce power at low temperature.**

# Outline

- Background
  - Active Mode
  - Standby Mode
- Power Management Unit
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- **Measurement Results**
- Conclusion

# Chip Micrograph

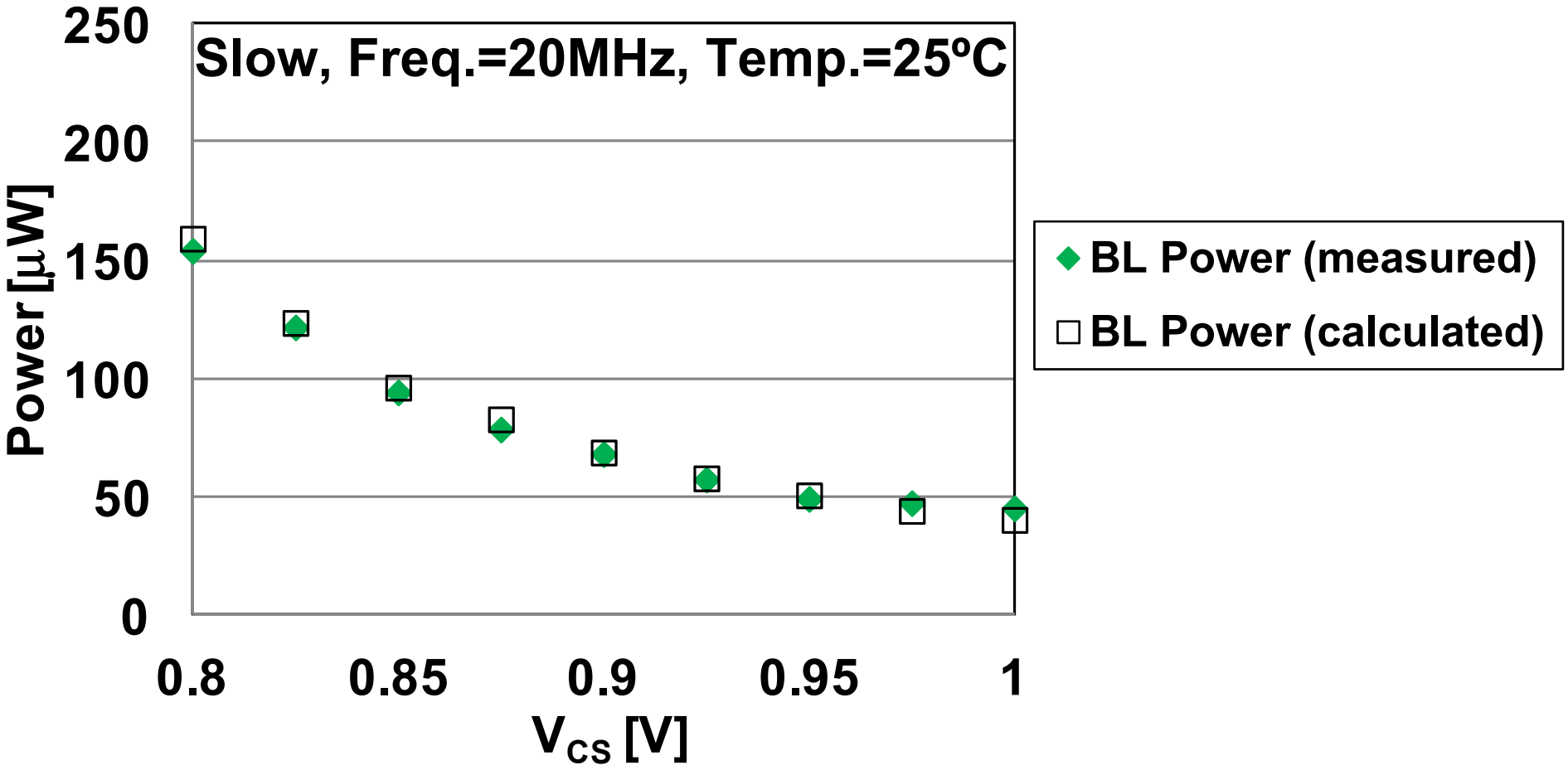



| | |
|---|---|
| Organization | 2Mb (1k X 512 X 4) |
| Power Supply | 0.7-1.1V, 1.5V |
| Technology | 28nm HK/MG CMOS |
| Cell Size | 0.120μm$^2$ |
| Macro Size w/ LDO + SF | 0.322mm$^2$ (Area +6.5%) |



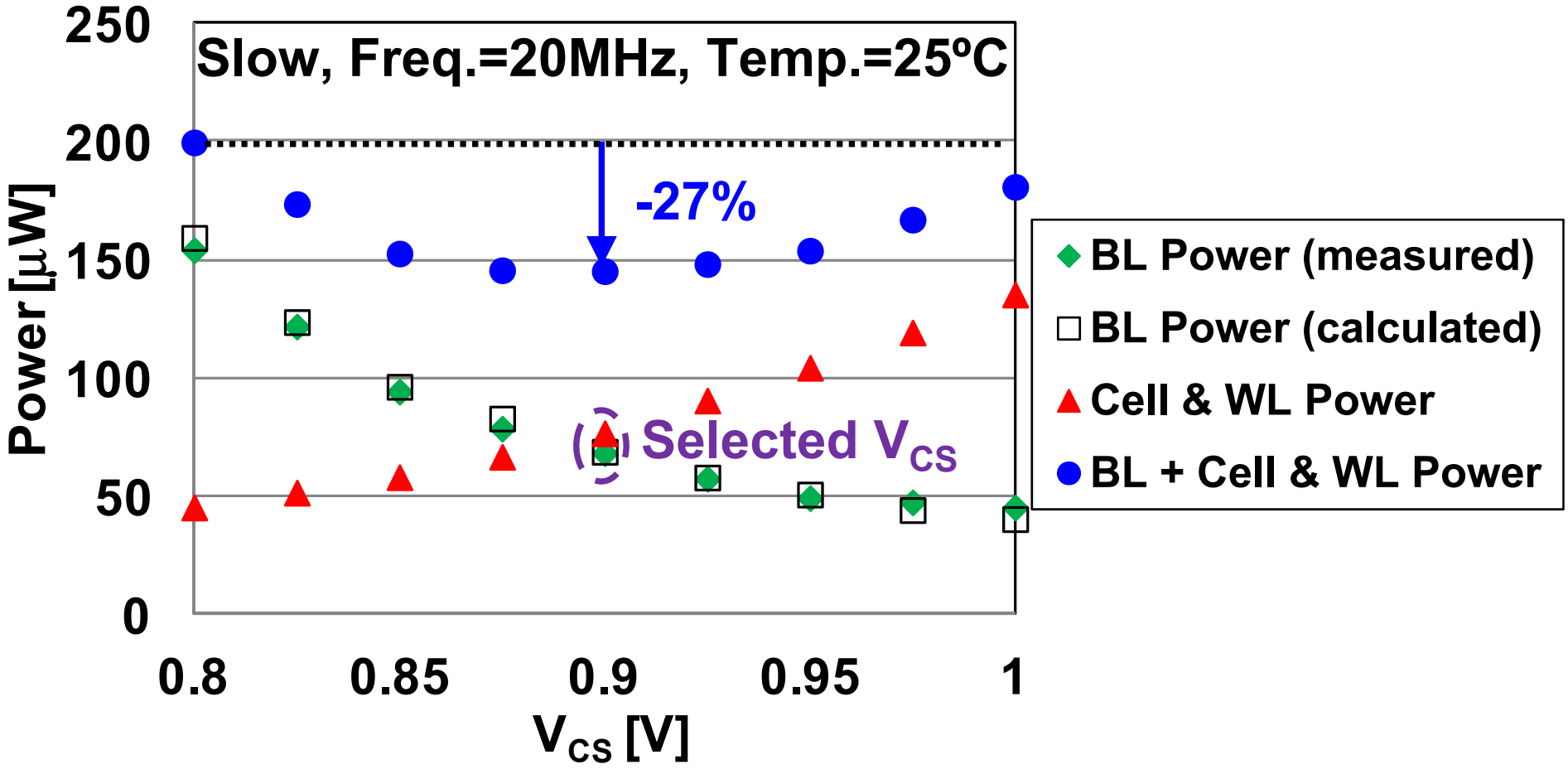
© 2013 IEEE

# Active Mode Power Control



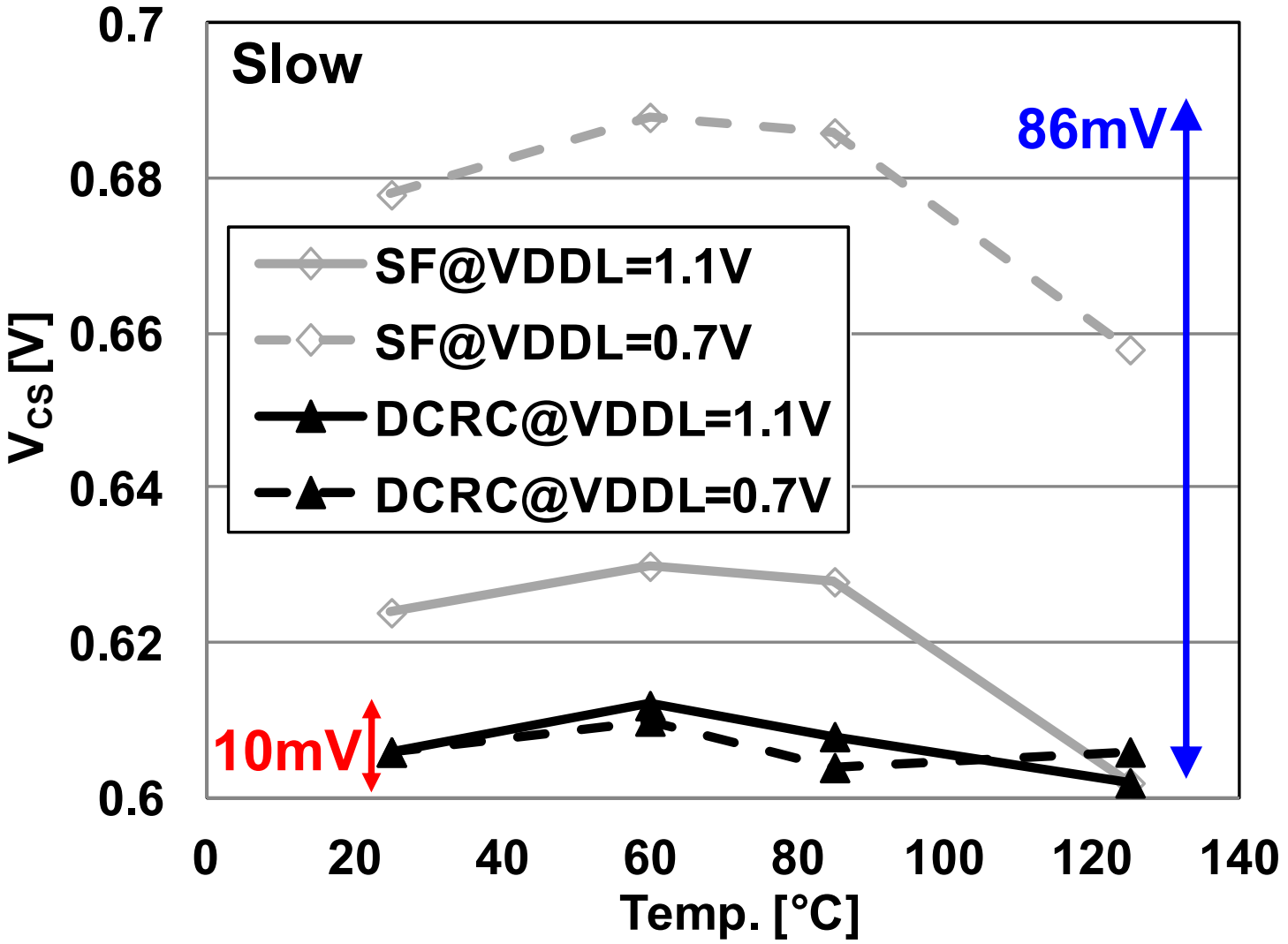

- Excellent correlation between calculated and measured BL power

# Active Mode Power Control



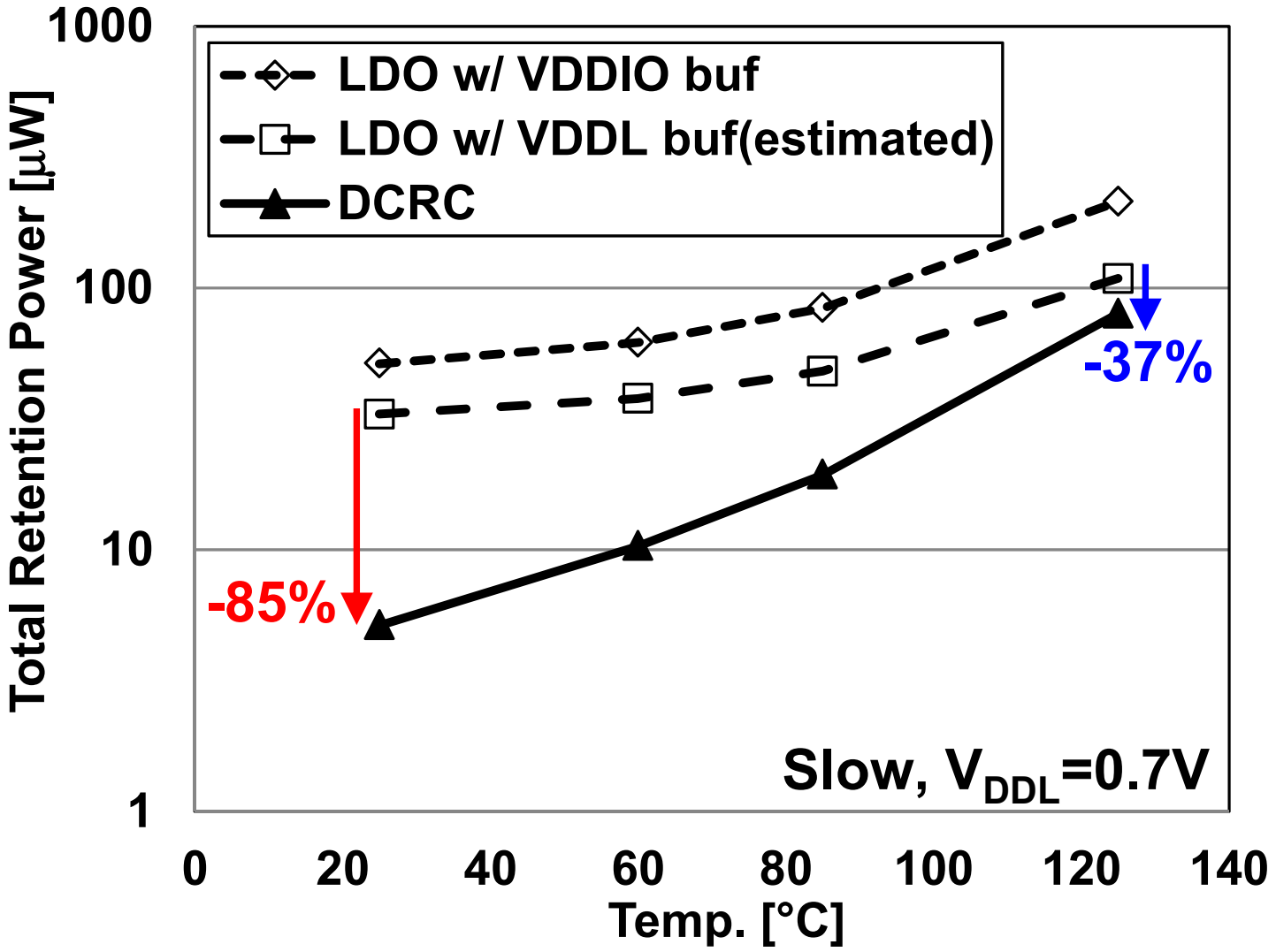

- New $V_{CS}$ selection based on this work results in 27% power reduction versus min. functional $V_{CS}$

# Cell Voltage Control in Standby Mode




**$V_{CS}$ variation is reduced to 10mV with digital control.**

# Standby Mode Power Control




## 85% power reduction at 25ºC from analog LDO

# Outline

- Background
  - Active Mode
  - Standby Mode
- Power Management Unit
  - BL Power Calculator
  - Digitally Controllable Retention Circuit
- Measurement Results
- **Conclusion**

# Conclusion

- Active and standby mode power reduction scheme has been implemented in a 28nm dual power supply SRAM.
- BL power calculator is proposed to reduce the active mode power.
  - Power reduction of 27% at 25ºC compared with minimum functional $V_{CS}$
- Digitally controllable retention circuit is proposed to reduce the standby mode power.
  - Power reduction of 85% at 25ºC compared with analog LDO

# A 64Mb SRAM in 22nm SOI Technology Featuring Fine-Granularity Power Gating and Low-Energy Power-Supply-Partition Techniques for 37% Leakage Reduction

Harold Pilo[1], Chad Adams[2], Igor Arsovski[1], Robert Houle[1], Steve Lamphier[1], Michael Lee[1], Frank Pavlik[1], Sushma Sambatur[3], Adnan Seferagic[1], Richard Wu[1], Imran Younus[4]

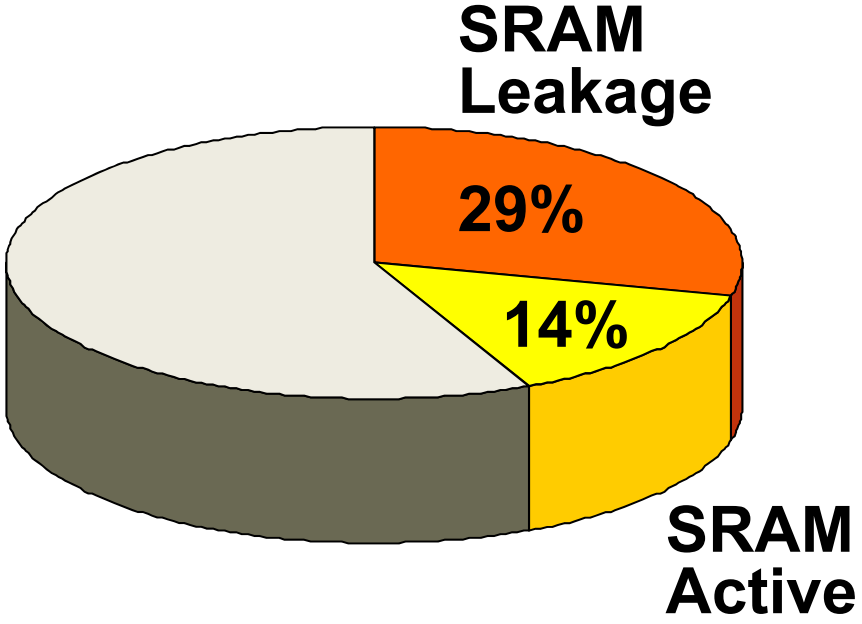
[1]IBM, Essex Junction, VT,
[2]IBM, Rochester, MN,
[3]IBM, Bangalore, India,
[4]IBM, Hopewell Junction, NY

# Outline

- Introduction
- Fine-Granularity Power Gating
- Low-Energy Power-Supply Partition
- Measurement Results
- Conclusion

# Motivation

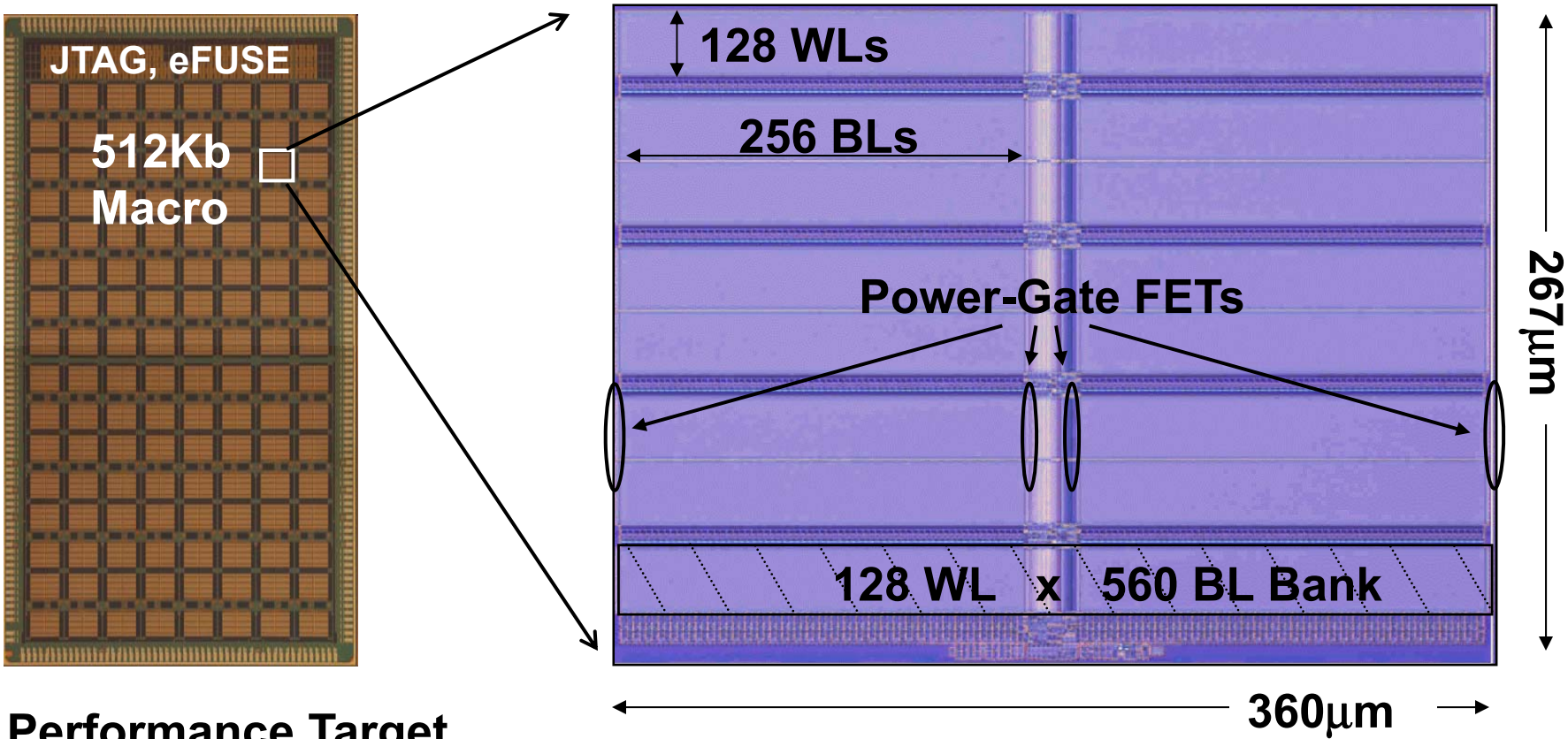
## Average power breakdown for 10 SoC products




**Of the Total SoC Power, on average, SRAM=43%**

**This work describes SRAM leakage savings of 37% to achieve >10% power reduction at chip level**

# 64Mb Test Chip Features




- Performance Target
  - 1.6GHz (Slow process, 0.85V, 0C)
- Operating Voltage
  - VDD: 0.70V - 1.10V (0.85V Typ.)
  - VCS: 0.75V - 1.10V (0.95V Typ.)
- Technology
  - 22nm PD SOI w/ HKMG
- Bitcell Size:
  - 0.128µm$^2$

# Outline

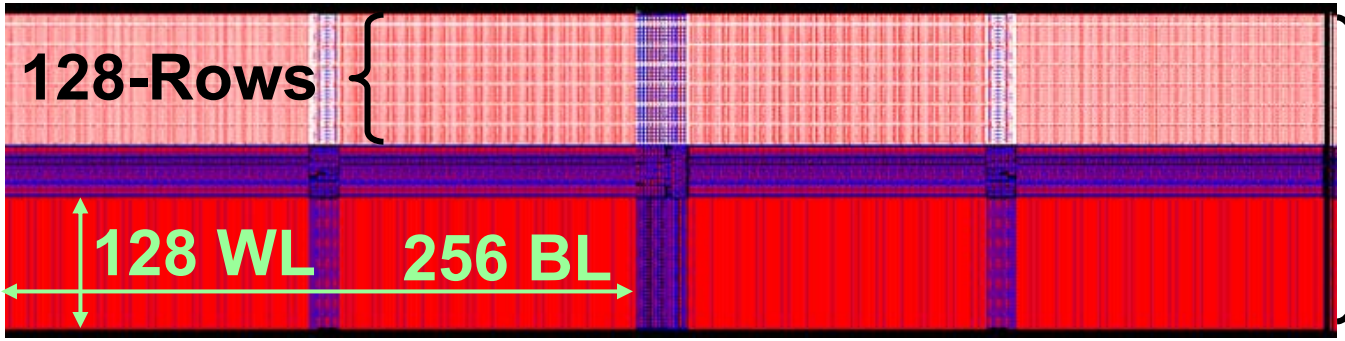
- Introduction
- **Fine-Granularity Power Gating**
- Low-Energy Power-Supply Partition
- Measurement Results
- Conclusion



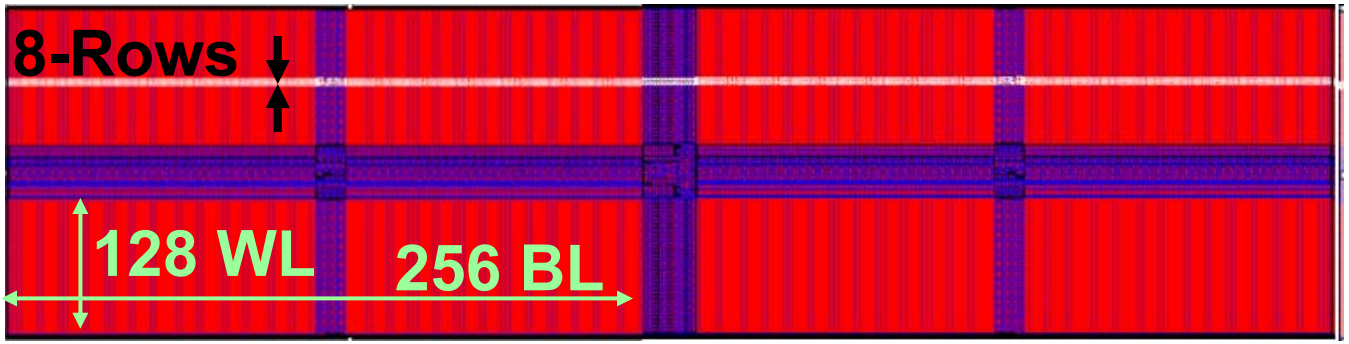
© 2013 IEEE

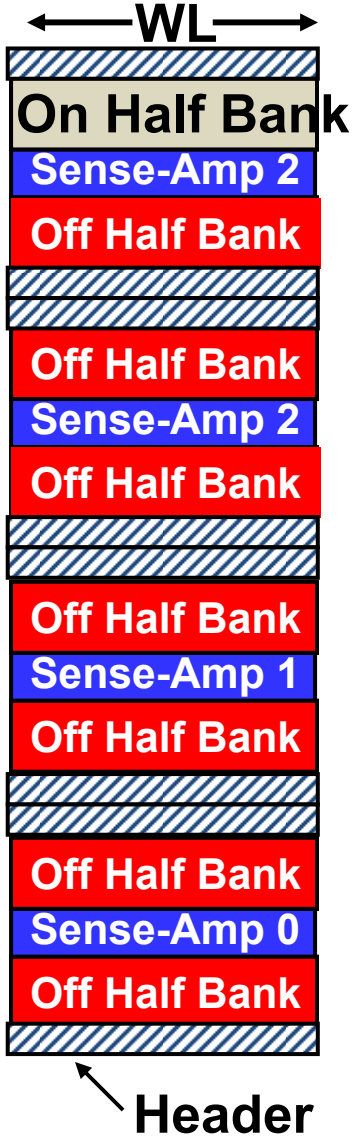
# SRAM Power Gating

## Bank-Based Power-Gating (Ramadurai, ISSCC'08)



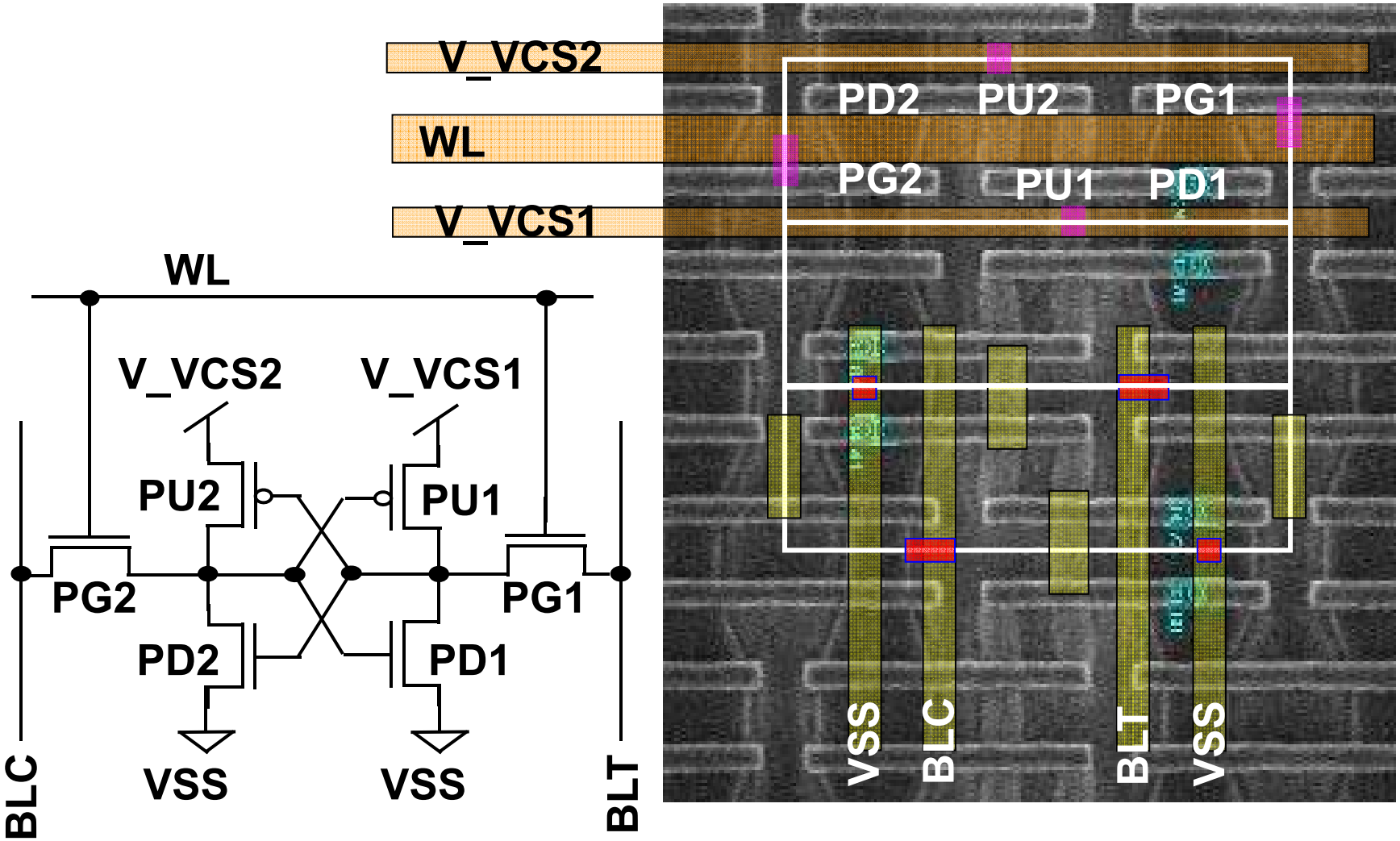

## Fine-Granularity Power-Gating (FGPG)



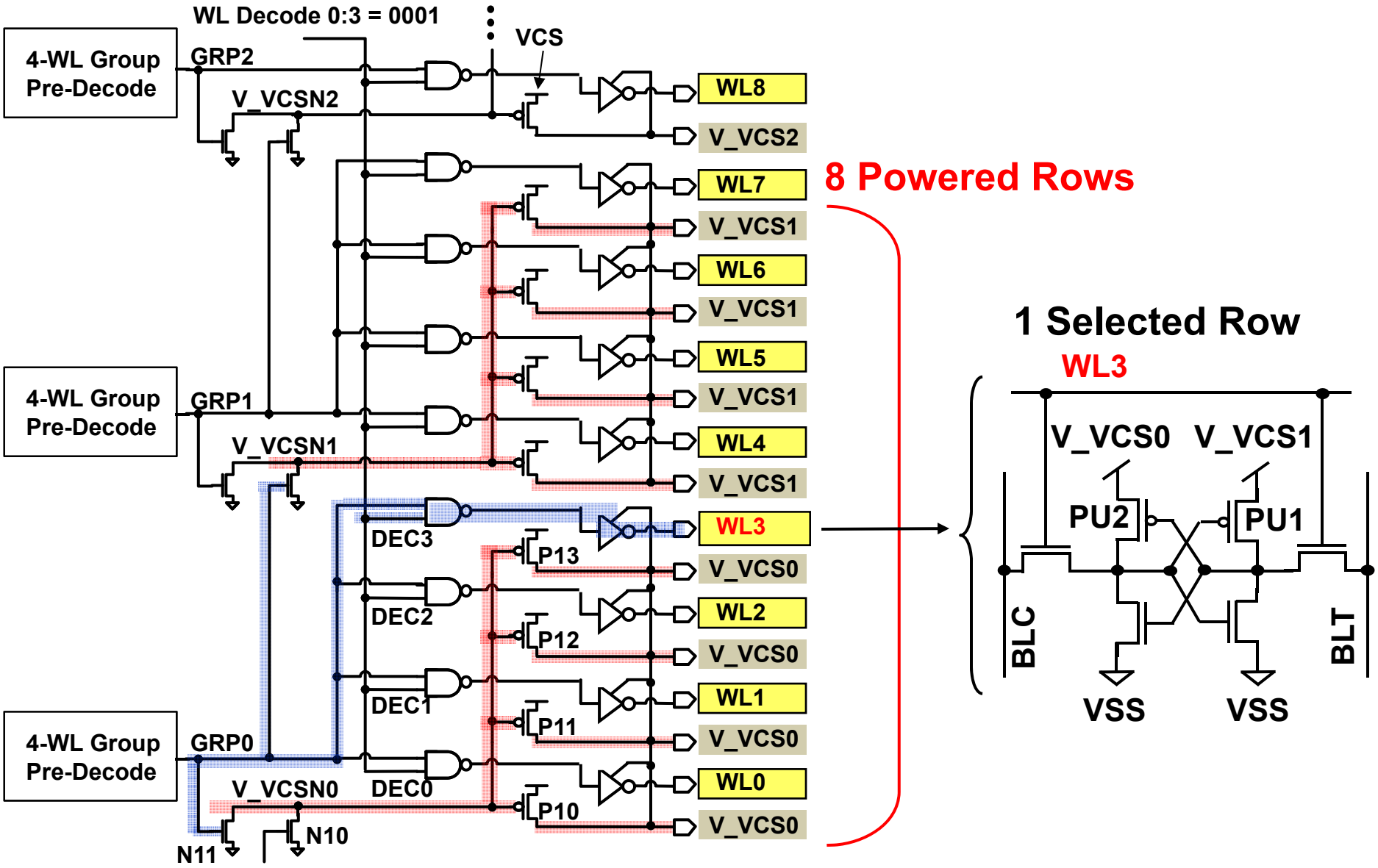




## 2X Improvement in Power Savings

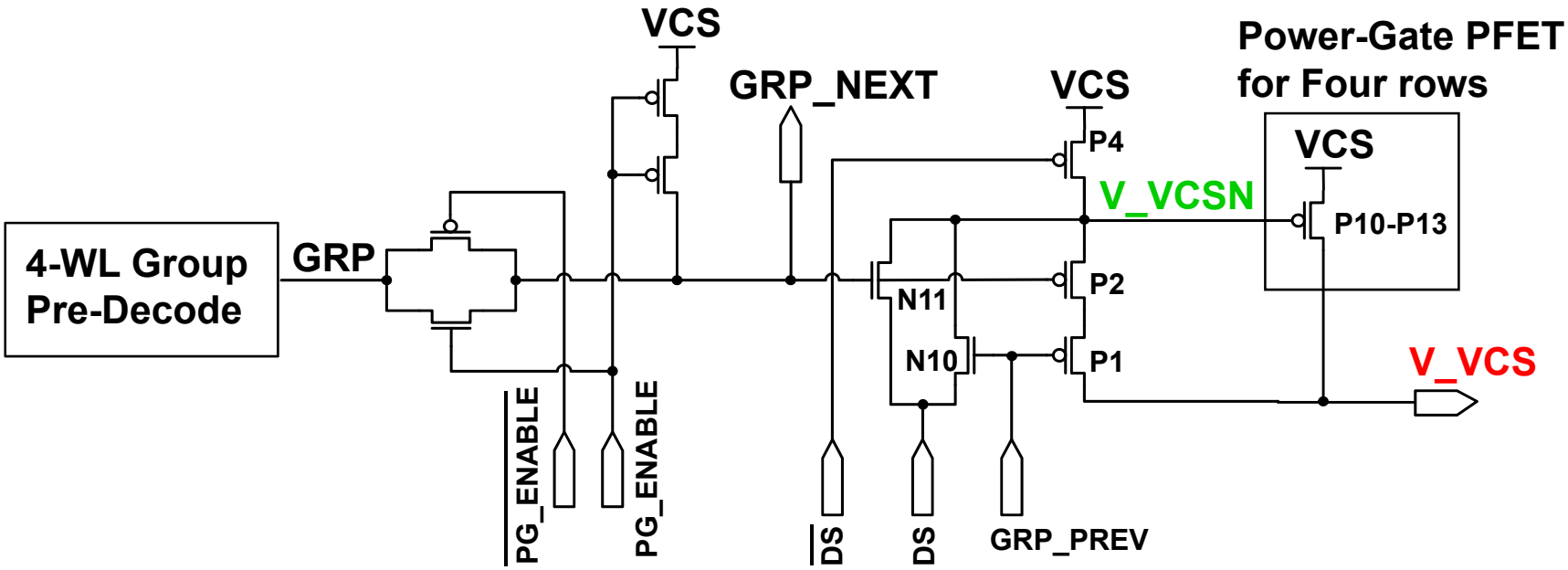
# FGPG Bitcell Implementation

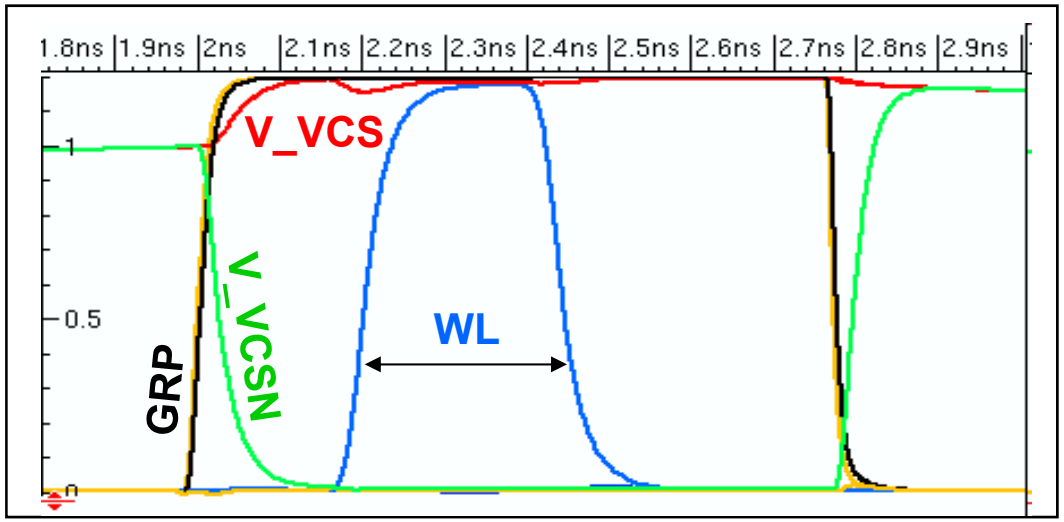


# FGPG Circuit Implementation



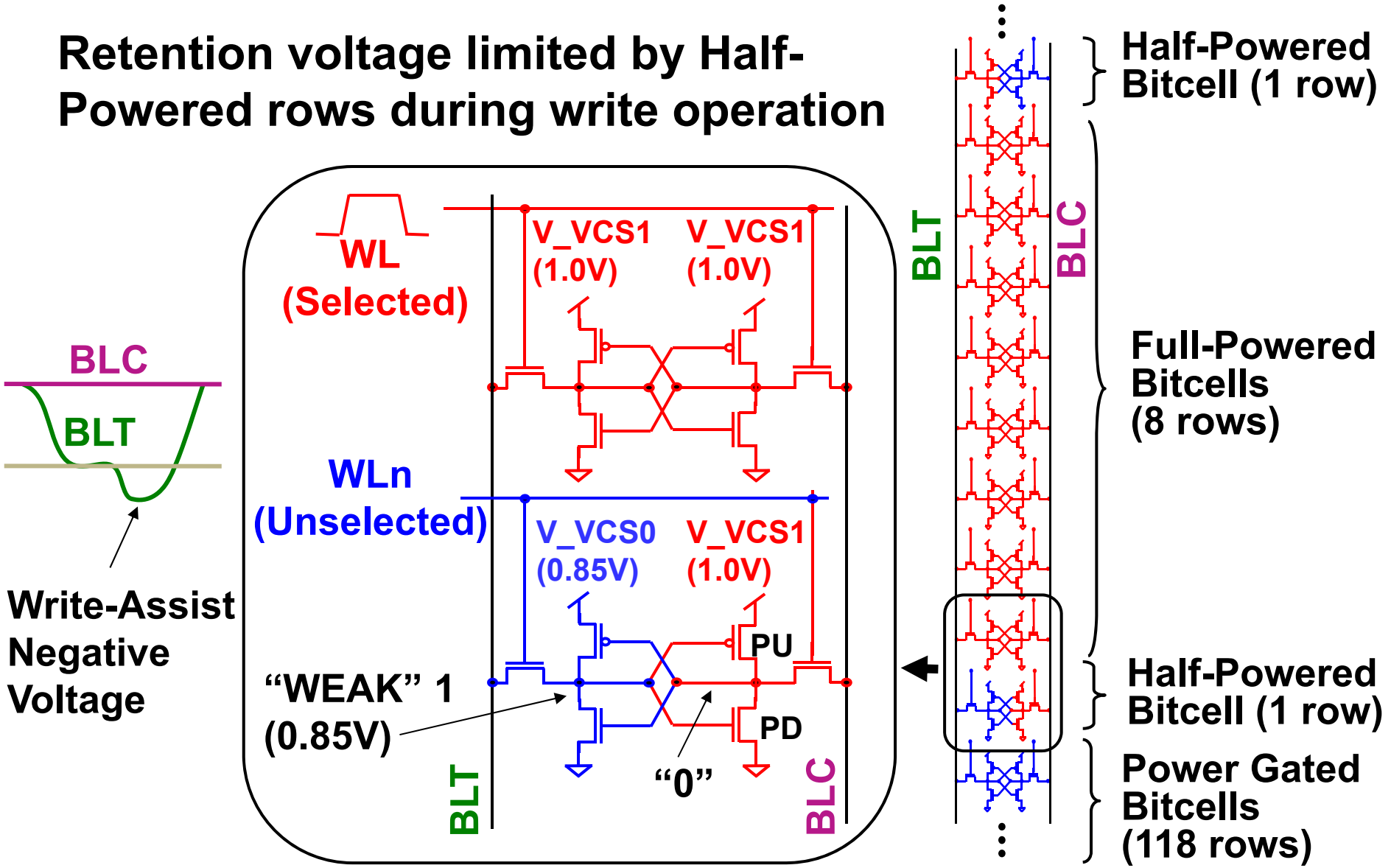
# Power Gating Logic and Simulation



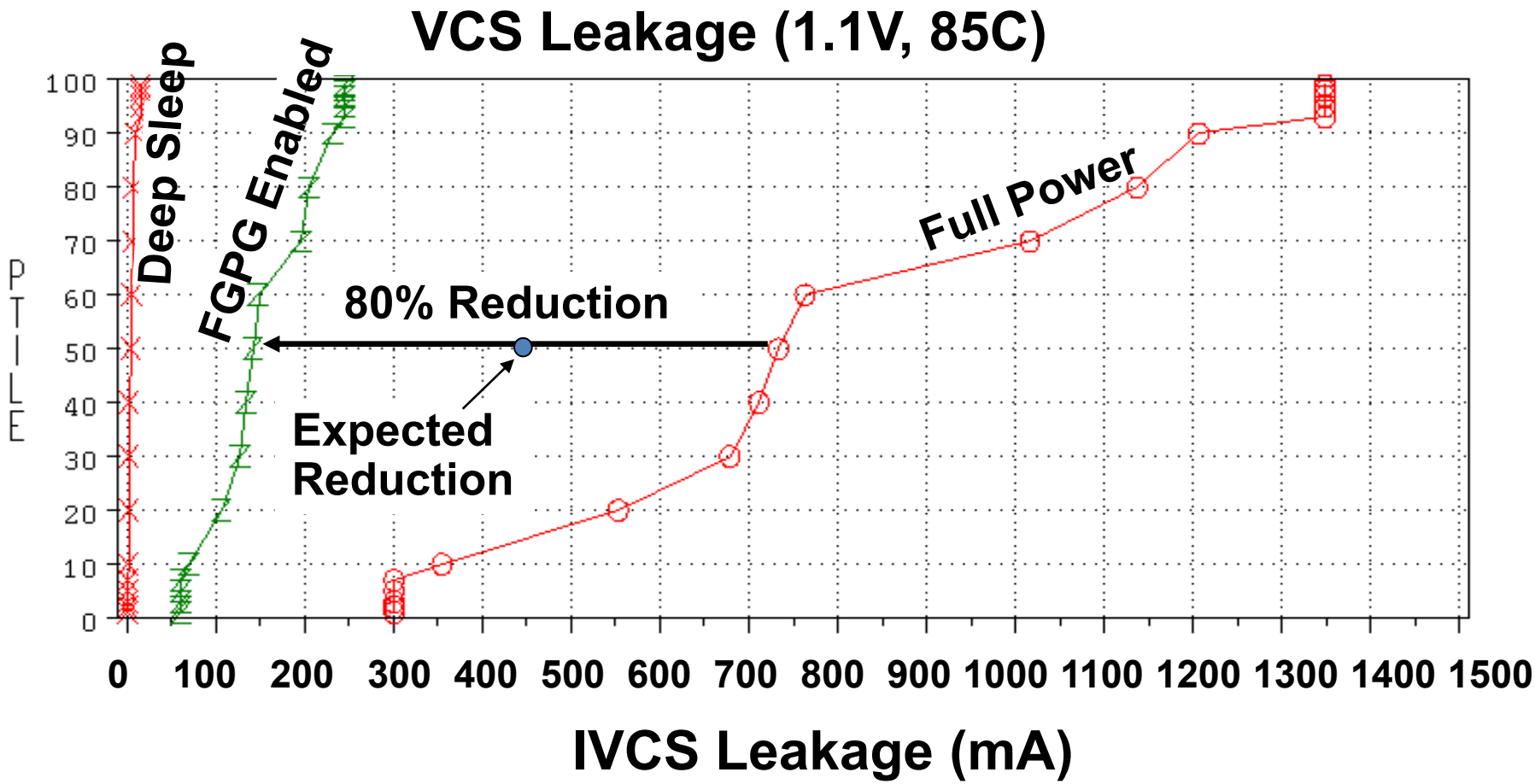



# Half-Powered Rows / Write Conflict

**Retention voltage limited by Half-Powered rows during write operation**



# Power Reduction Measurements

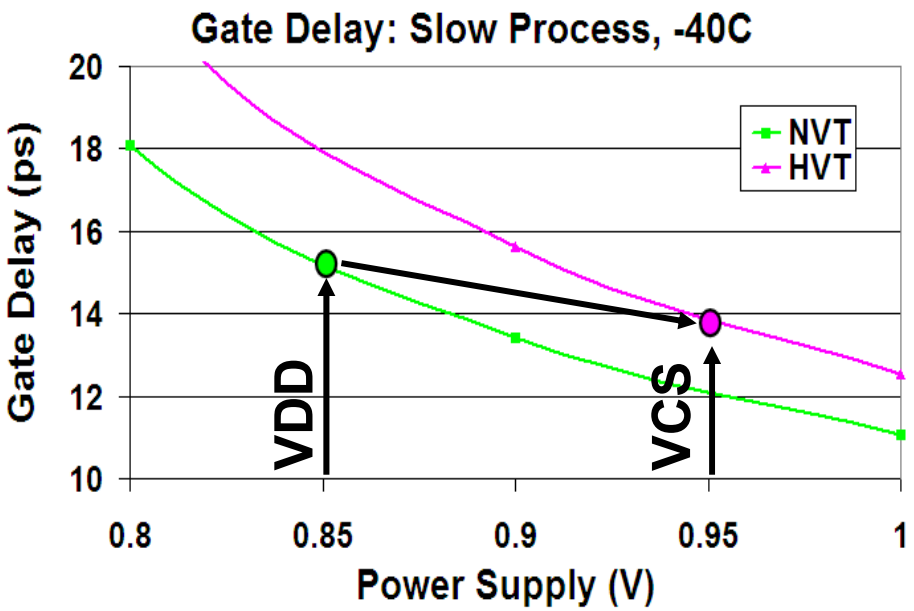


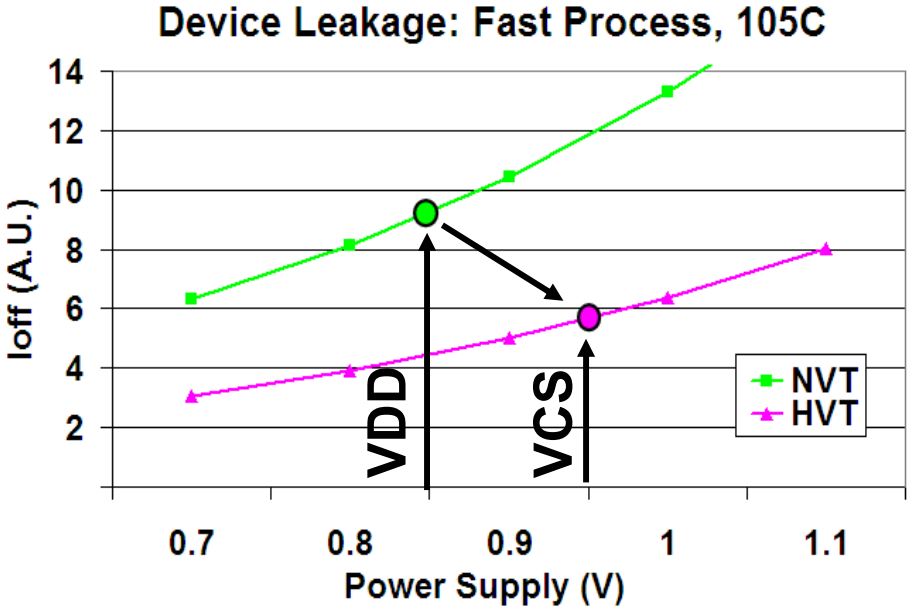

- Full Power: FGPG disabled; array power supply = VCS
- Deep Sleep: Complete power down; data state is lost

# Outline

- Introduction
- Fine-Granularity Power Gating
- **Low-Energy Power-Supply Partition**
- Measurement Results
- Conclusion

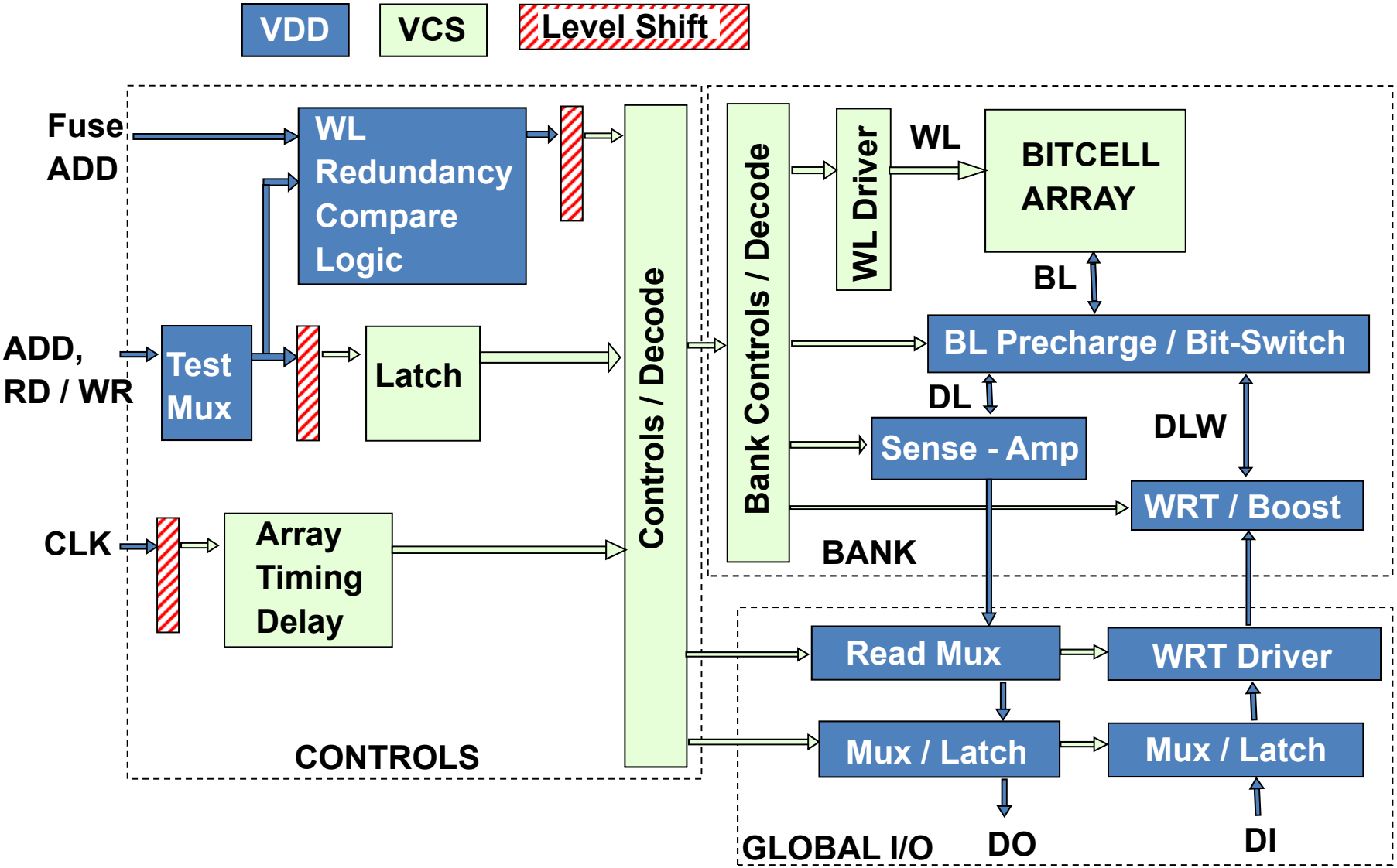
# Optimum Vt Usage in Dual Supply Design




- Low-activity, narrow paths (clocking, logic, decode)
  - High-Vt devices at higher supply (VCS) improve performance and reduce leakage
- Wide I/O data busses, high activity paths
  - Normal Vt devices at lower supply (VDD) reduce impact to active power

# SRAM Macro Power Supply Partition



# Outline

- Introduction
- Fine-Granularity Power Gating
- Low-Energy Power-Supply Partition
- **Measurement Results**
- **Conclusion**

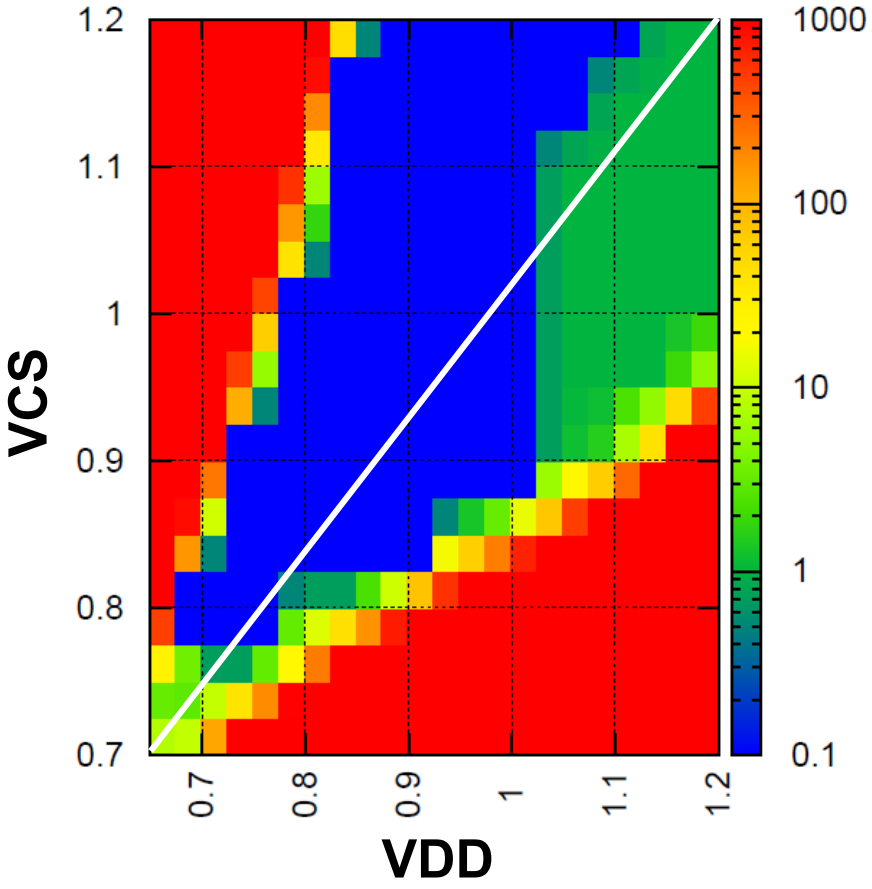


© 2013 IEEE

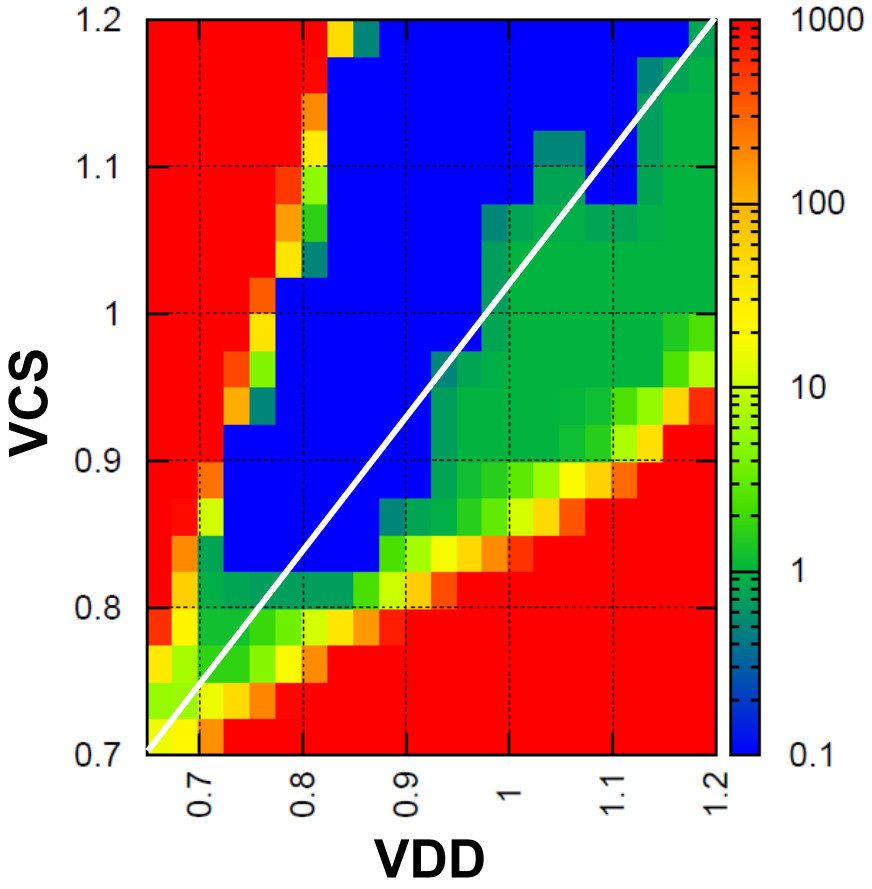
# Silicon Results

## 8Mb Fails @ 25C

### FGPG Disabled




### FGPG Enabled



# Conclusion

- 64Mb Test Chip is fabricated in 22nm HKMG
- Implemented power reduction techniques
  - Fine-Granularity Power Gating
  - Low-Energy Power-Supply Partition
  - Small Area Overhead: 1.5%
- Power savings include
  - 37% Leakage savings at the macro level
  - 10% Total power savings at the chip level

# 7GHz L1 Cache SRAMs for the 32nm zEnterprise™ EC12 Processor

J.Davis[1], P.Bunce[1], D. Henderson[1],
Y. Chan[1], U.Srinivasan[1], D. Rodko[1],
P. Patel[1], T. Knips[1], T. Werner[2]

[1] IBM Poughkeepsie, NY
[2] IBM Böblingen, Germany

# zEnterprise™ EC12 L1 Cache

- Large Split L1 cache
  - 96KB data cache
  - 64KB instruction cache
- Fast Cycle Time - 5.5 GHz clock
  - SRAM must be able to perform a read and a write operation every 164 ps (after tester guardband)
- Fast access times to support multi-cycle system paths
  - Requires aggressive use of dynamic circuits

# Key L1 Cache Specifications

| | Data Cache | Instruction Cache |
|---|---|---|
| Technology | 32nm SOI | 32nm SOI |
| Memory Cell | 6T / 0.291 μm$^2$ | 6T / 0.291 μm$^2$ |
| Organization: Read<br>. Write | 512x36bx6wx2p<br>256x72bx6w | 512x 72bx4wx1p<br>256x144bx4w |
| SRAMs per cache | 8 | 4 |
| Fine Grain Banking | Yes | No |
| Write Bypass | No | Yes |
| Size | 367μm x 291μm | 419μm x 275μm |
| Access Time[1] (simulated) | 235ps (Nominal) | 219ps (Nominal) |
| Power[2] (simulated) | 49.6 mW | 120.5 mW |

[1] 85°C VDD = 0.95V; VCS (CELL SUPPY) = 1.05V

[2] 65°C VDD = 1.08V; VCS = 1.18V ; 6.1 GHz

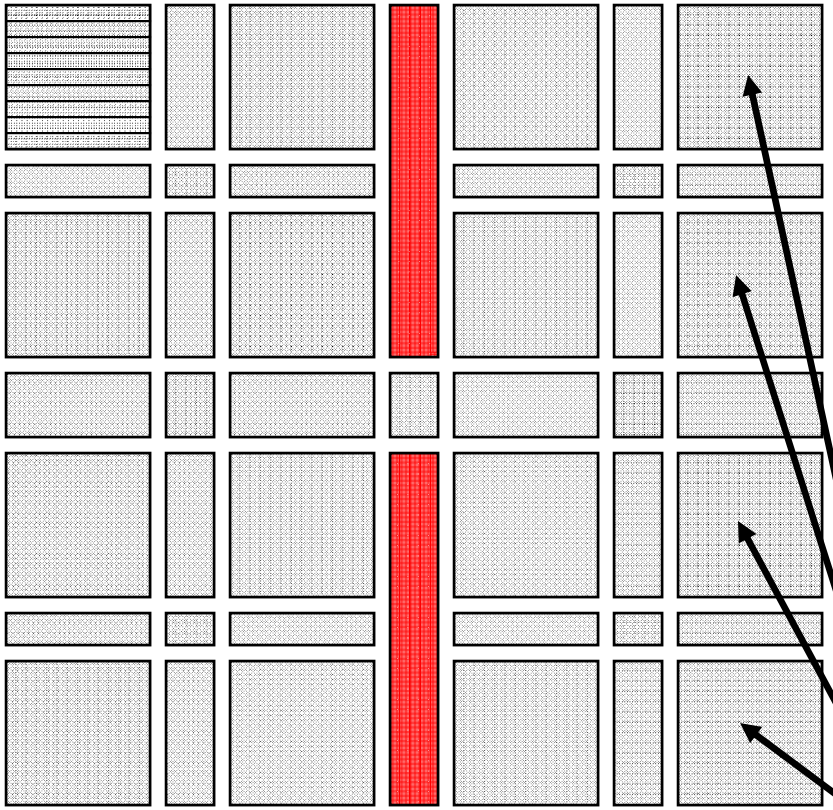
# Design Challenges

- Application specifications drive aggressive use of dynamic circuitry
- Dynamic circuits provide low latency
  - Can be optimized for a specific value
  - Non-optimized value is the starting value
- Dynamic circuits have two phases
  - Evaluation phase - needs to be long enough to “write” dynamic nodes
  - Restore phase - needs to be long enough to generate the starting value for the next cycle

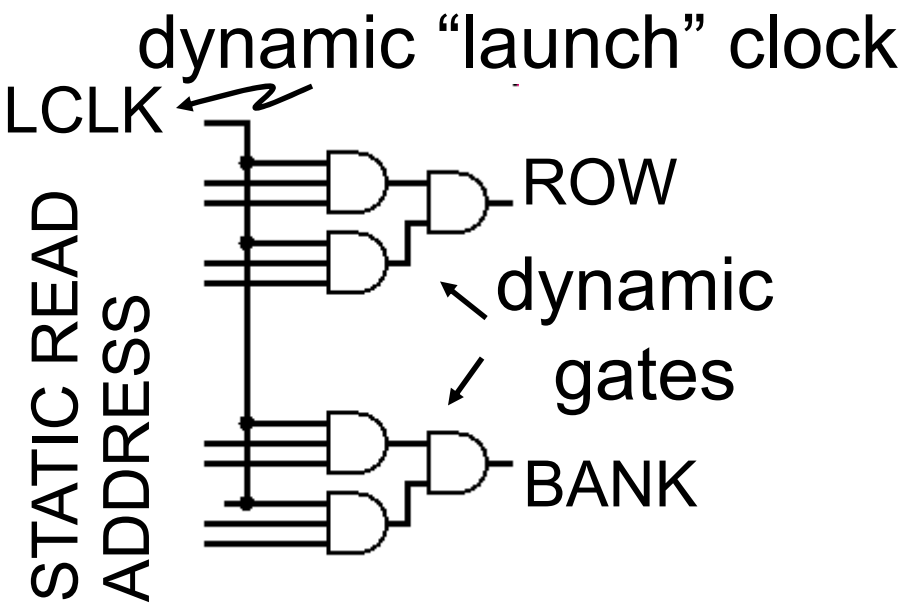
# Design Challenges (cont)

- Shrinking device geometries
  - Increased individual device variability → less "tracking"
- Greater impact to dynamic circuits vs static circuits
  - Static has decreased yield at a given frequency
  - Dynamic
    - decreased yield at **any** frequency (minimum fixed evaluation time needed to write dynamic nodes)
    - evaluation/restore phase contention (minimum underlap to avoid power burn)
    - "Traditional" solutions (wider pulses, larger separation between phases) → limited maximum frequency

# SRAM Architecture

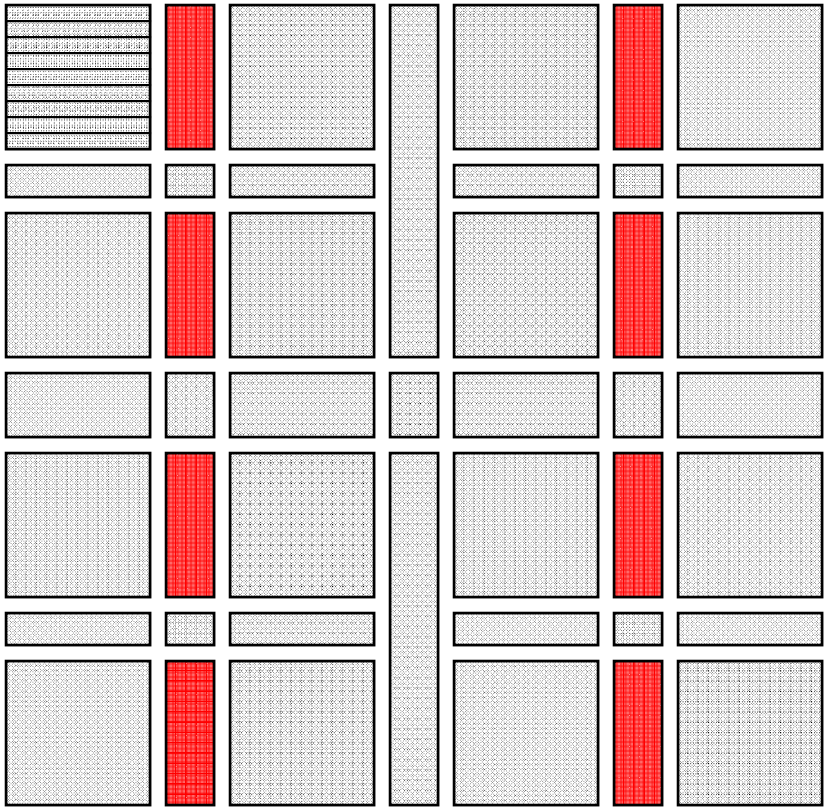


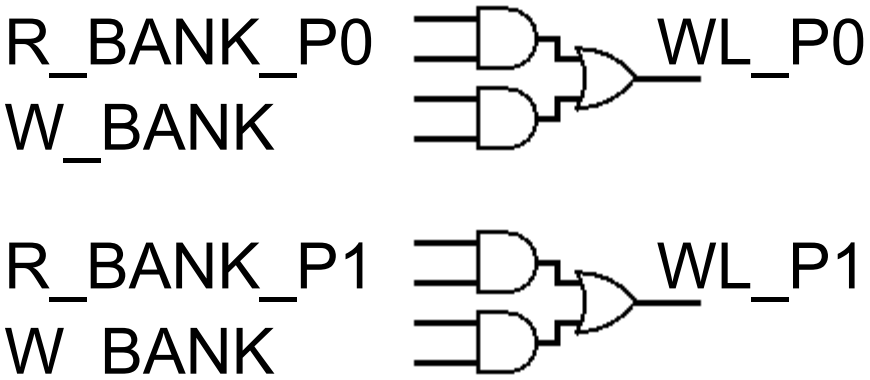
- First and Second Level Dynamic Decoders




- 32 Bank Select signals
- 16 Word in Bank signals
  (16 wordline group = Bank)
- Each block is 8 banks tall
- 4x8 = 32 Banks
- 32x16 = 512 wordlines

# SRAM Architecture (cont)

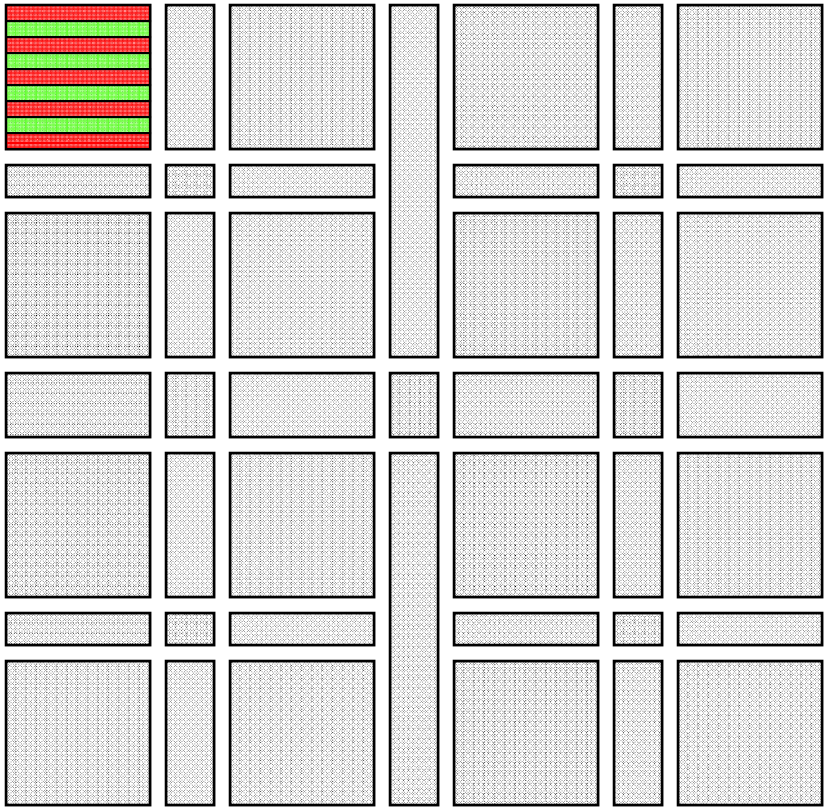




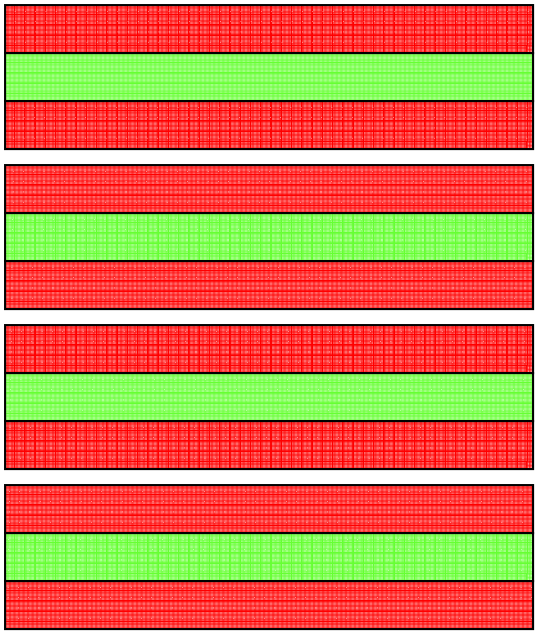

- MUXES READ/WRITE ADDRESSES
- 1 WL PER READ PORT
- 2 WL PER ROW
  - P0 = Port 0
  - P1 = Port 1

- Final word decoders
- Dual read ports

# SRAM Architecture (cont)



- SRAM Cell Core

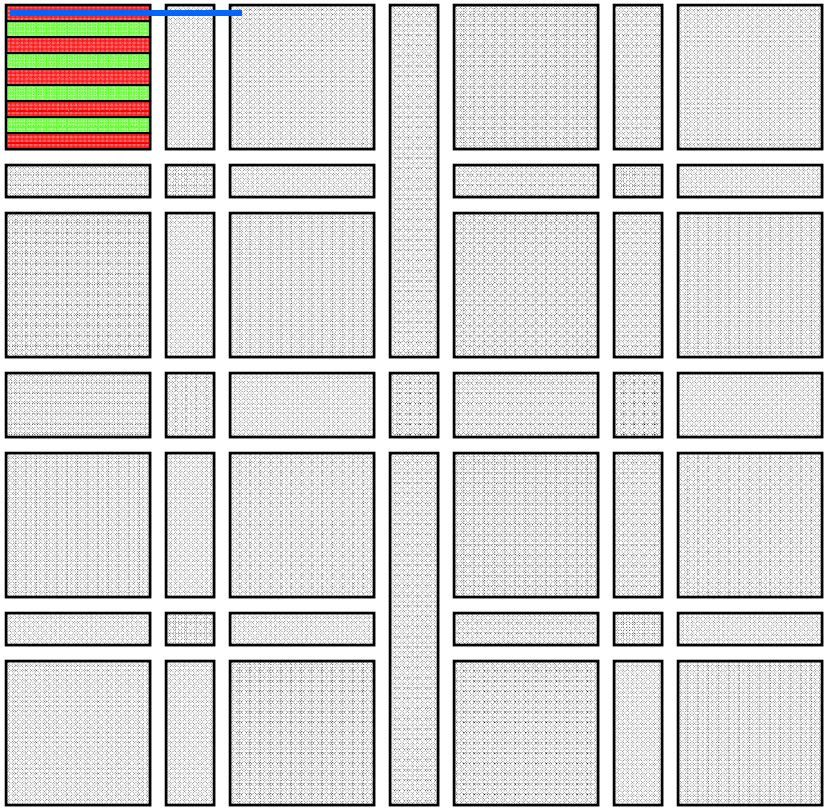


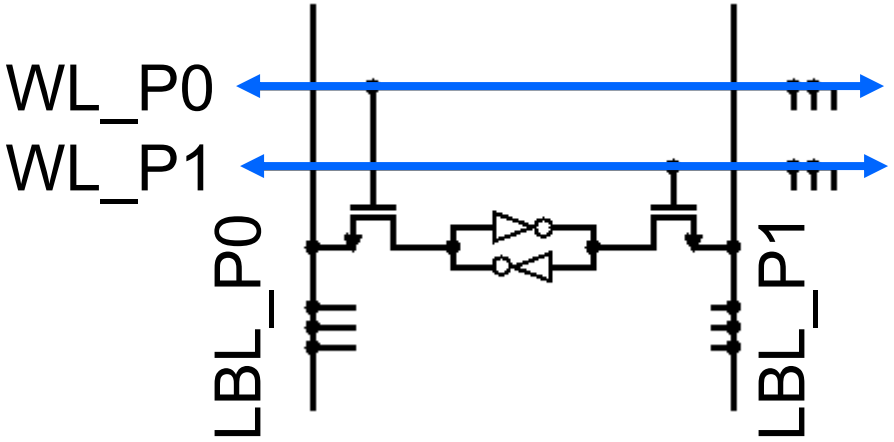
- RED = SRAM CELLS
  - 16 Rows Per Bank
- GREEN = “LOCAL EVAL”
- 8 Banks per Cell Core



© 2013 IEEE

# SRAM Architecture (cont)

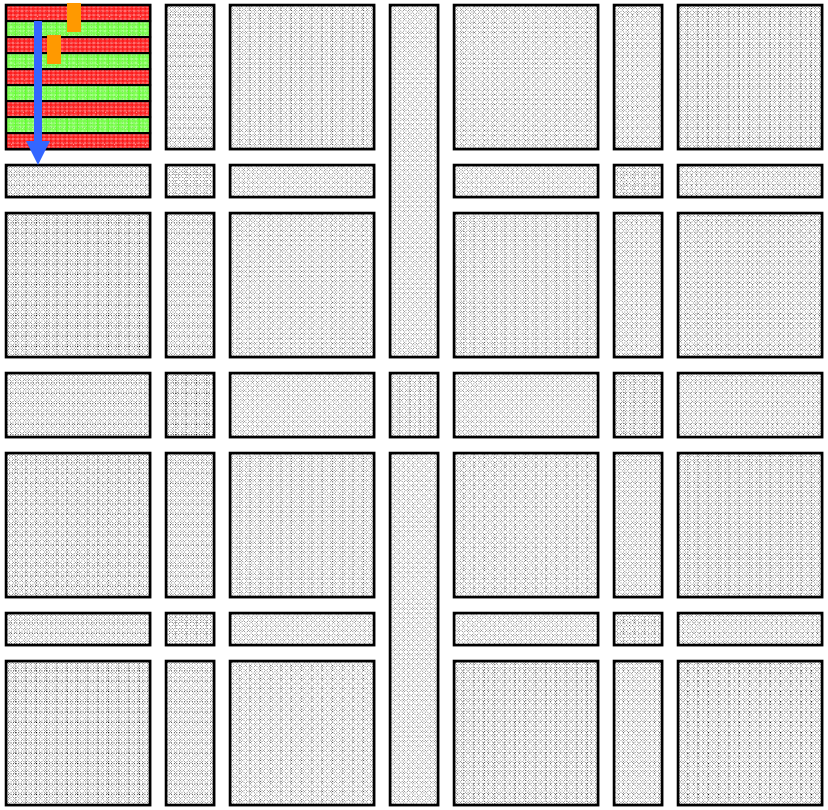


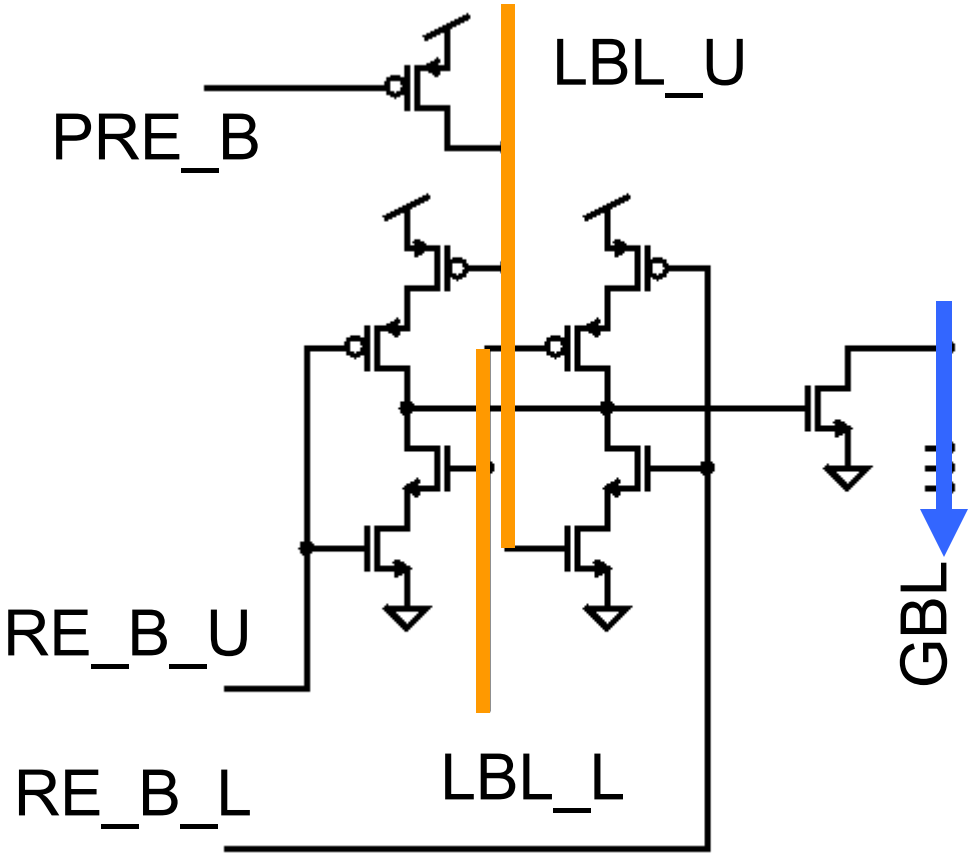



- Conventional 6T Cell
- 16 Cells per Local Bit Line

- SRAM Cells
- Local Bit Lines

# SRAM Architecture (cont)

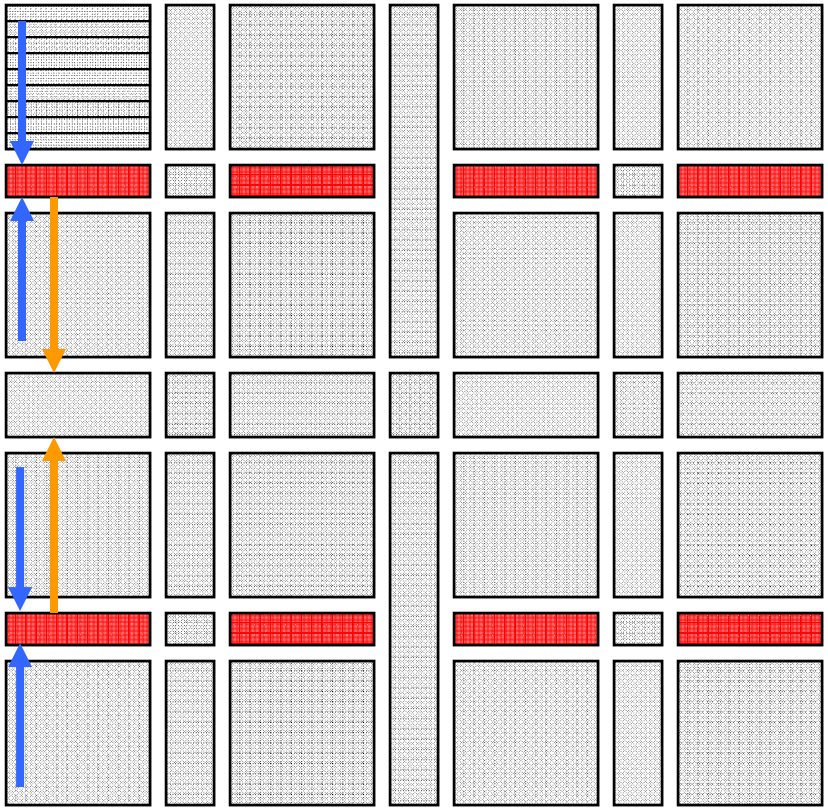


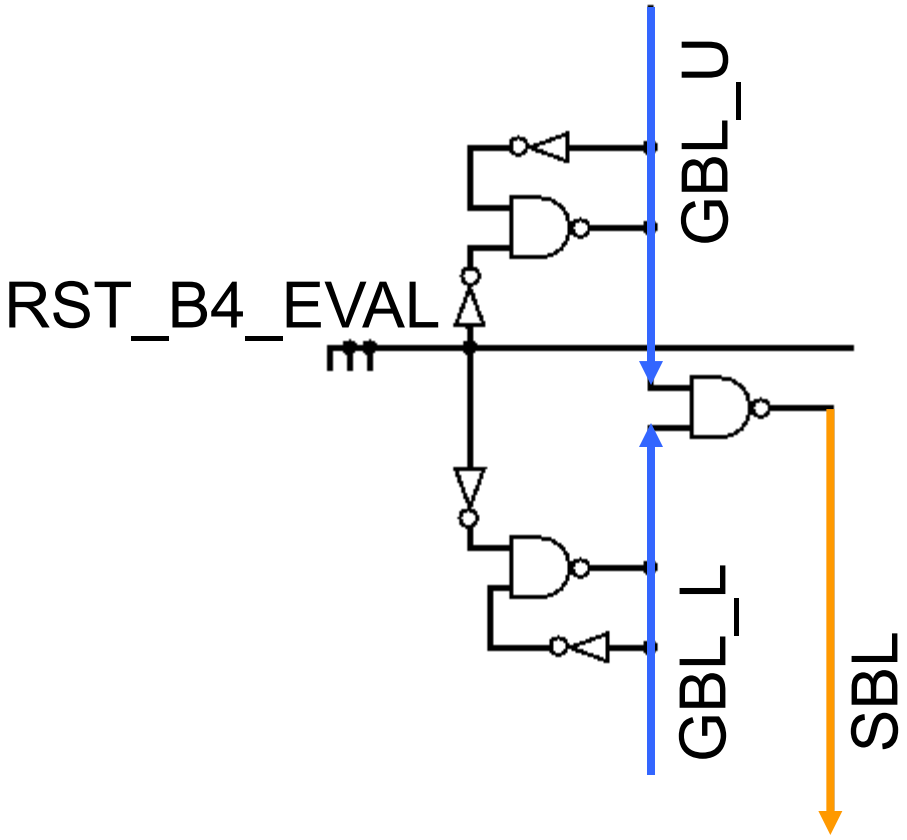



- Local Bit Lines (LBLs)
- Read enables (RE_B)
- Global Bit Lines (GBLs)

- Combine 2 Local Bitlines
- 4 LOCAL_EVALs per GBL

# SRAM Architecture (cont)

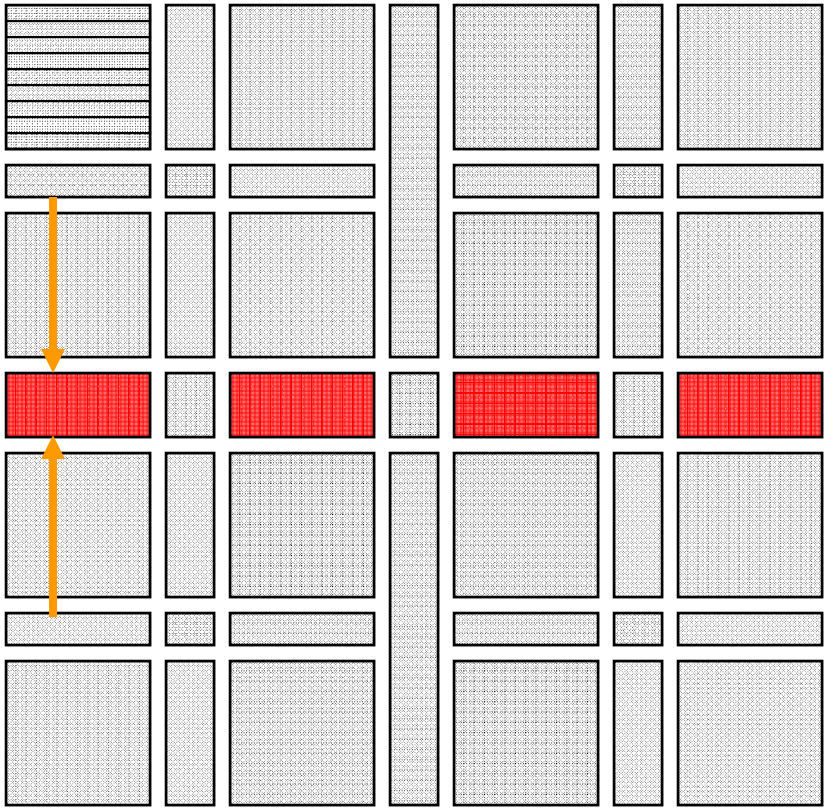


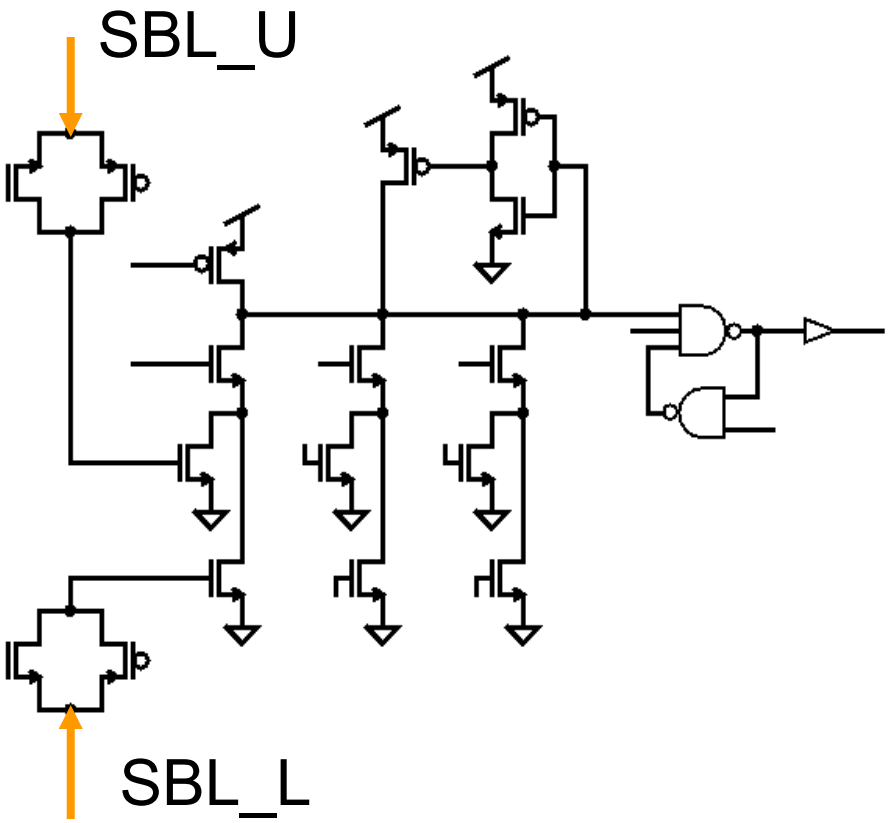



- Global Bit Line (GBL) Latches
- Solar Bit Lines (SBL)

- Combine 2 Global Bitlines
- 4 full keeper “latches” per column

# SRAM Architecture (cont)

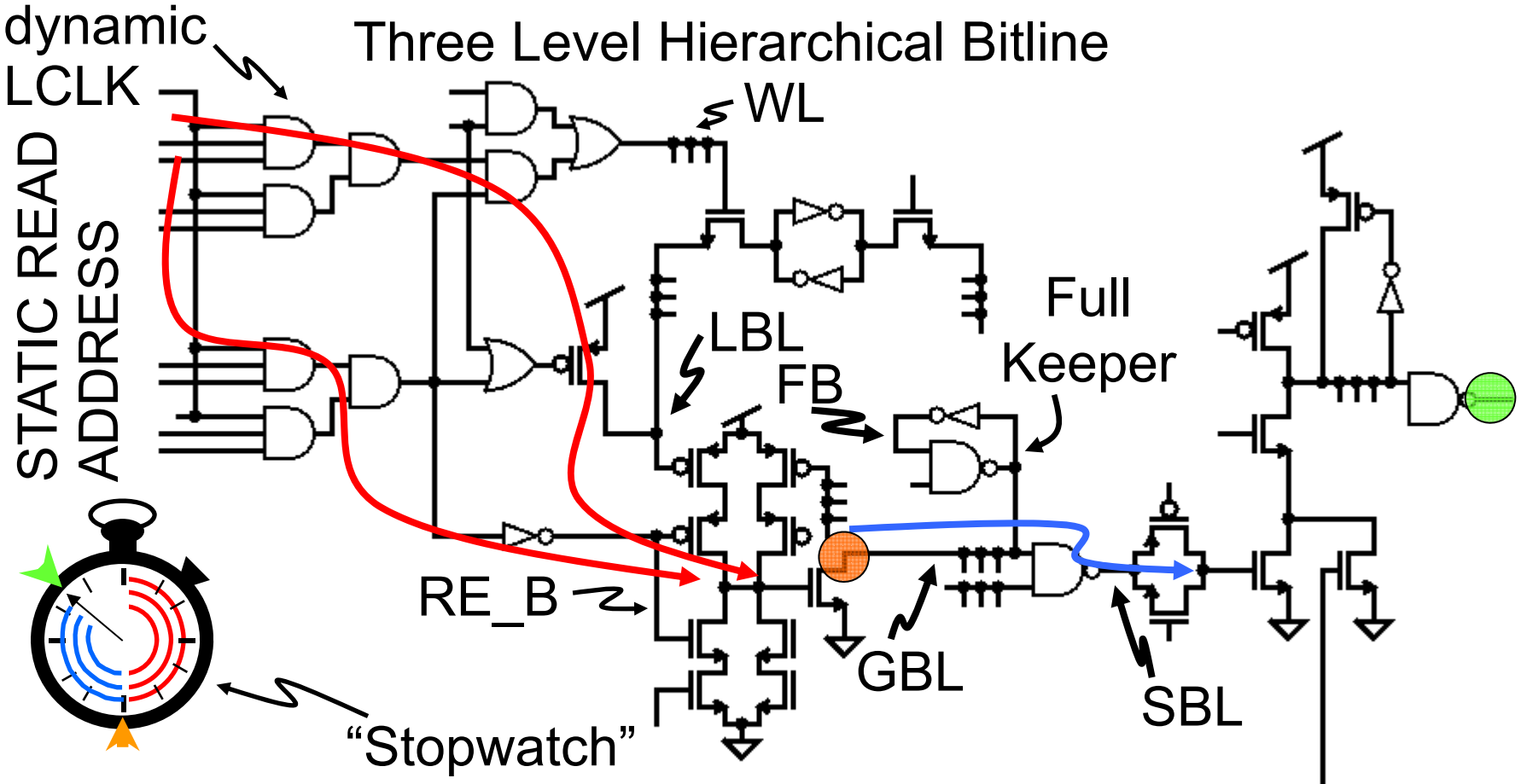





- Way Mux
- Output Driver

- Combine 2 Solar Bitlines

# Full Read Path




(path breakdown by propagation delay type)

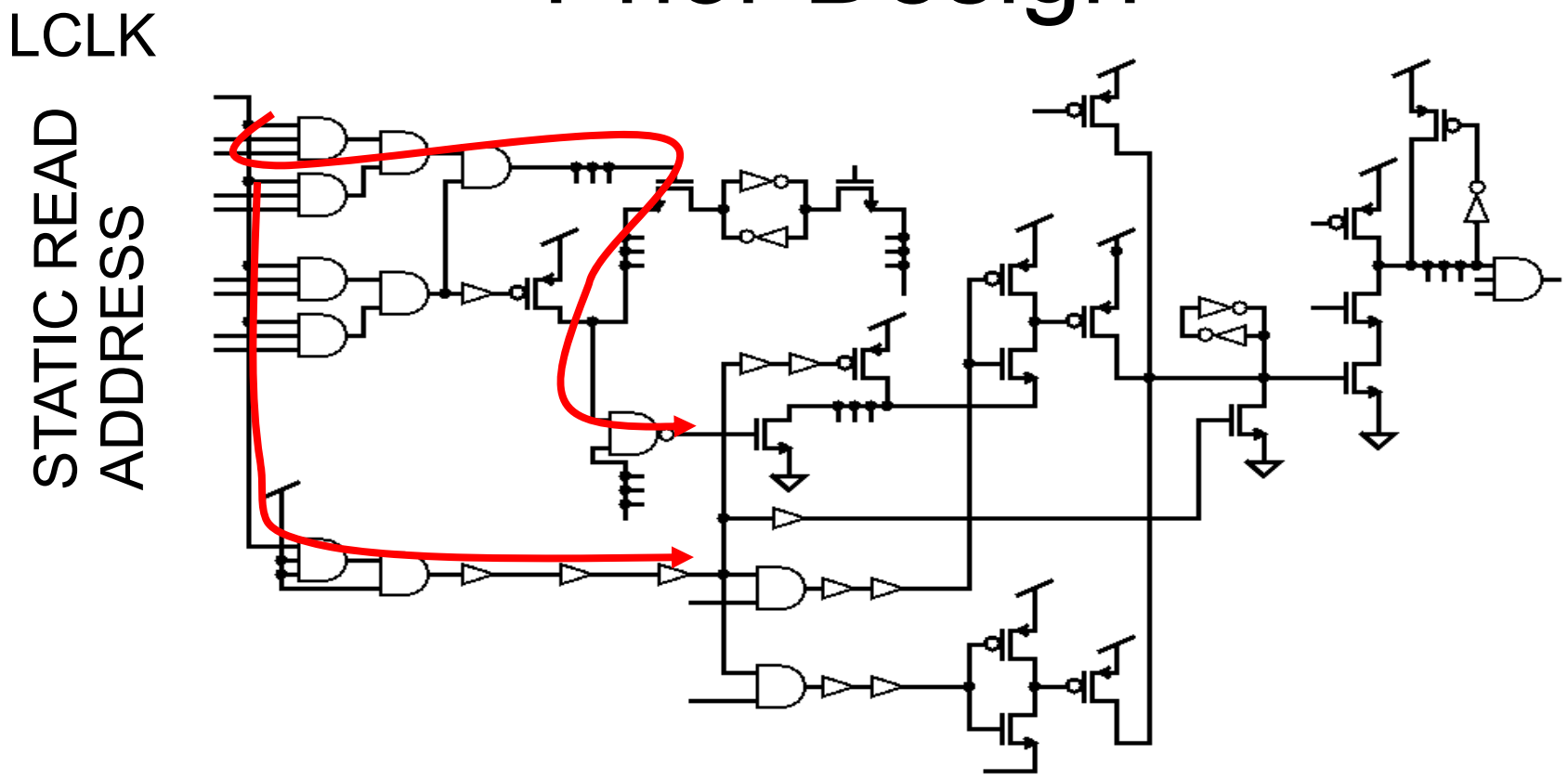
dynamic signal propagation

static signal propagation

ORANGE = latch point

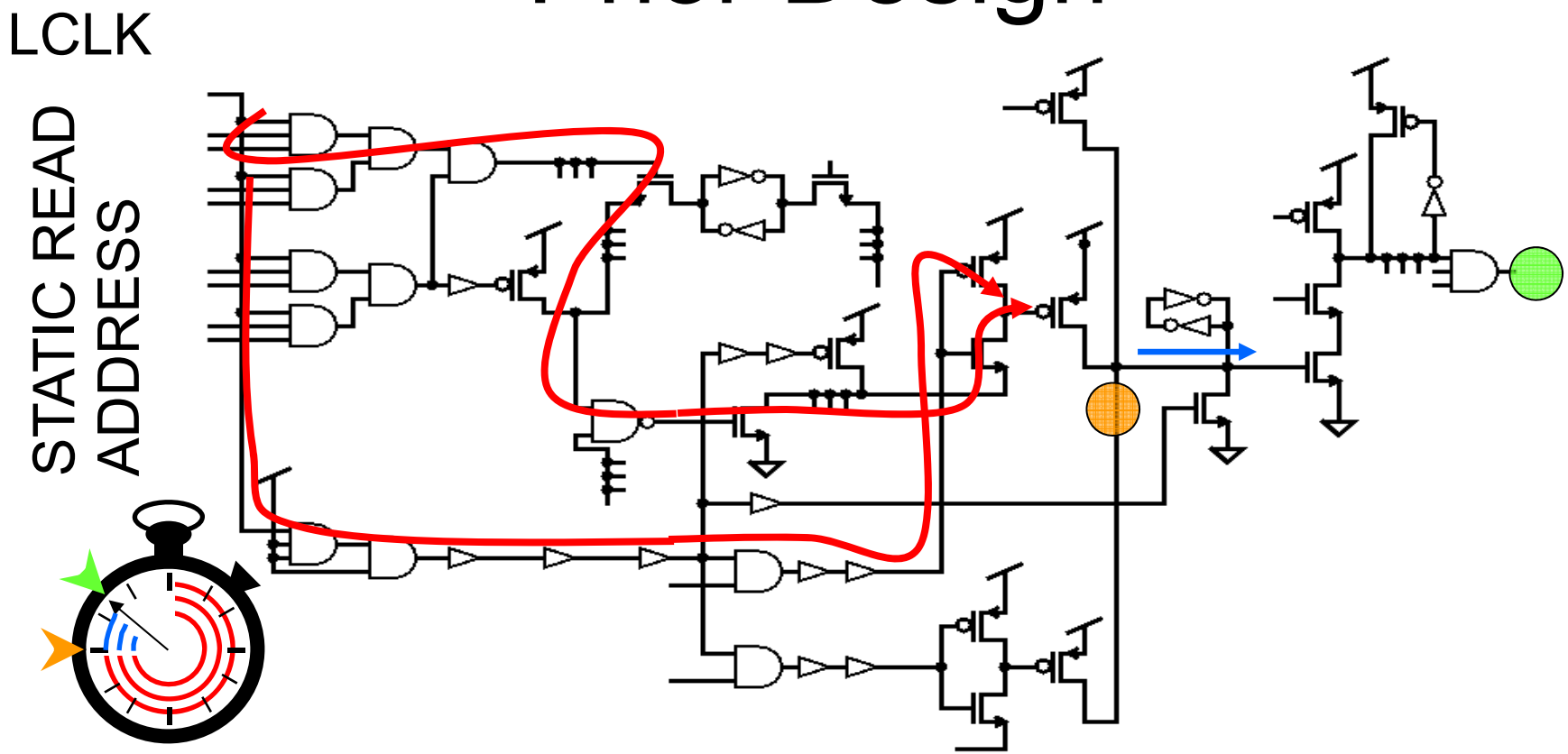
GREEN = latched output

# Prior Design



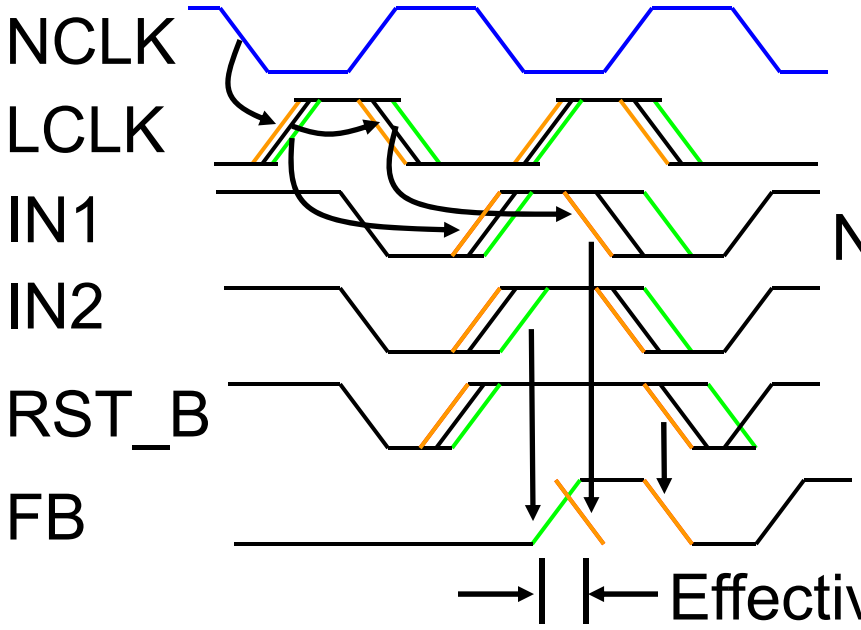

- More “efficient” (1 latch per column vs 4)
- Similar stage count / latency for each topology

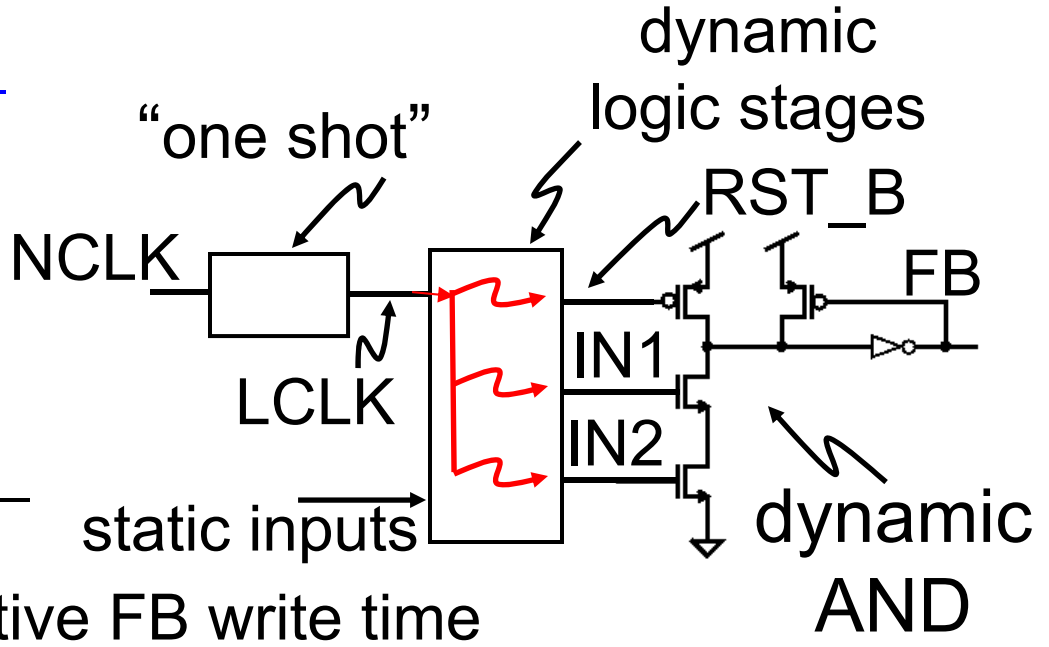
# Prior Design




- More “efficient” (1 latch per column vs 4)
- Similar stage count / latency for each topology
- Longer (in terms of propagation time) dynamic paths

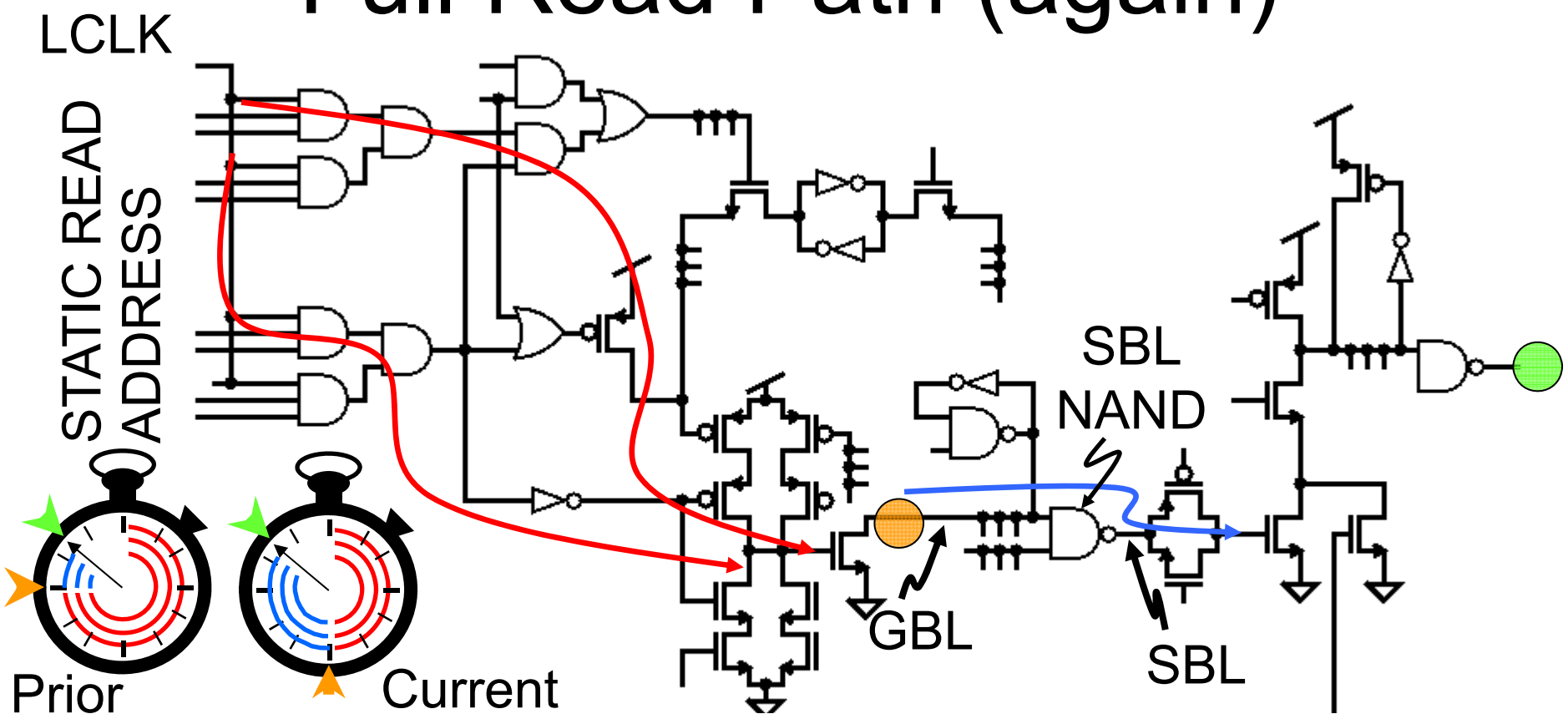
# Challenges of Long Dynamic Paths




- Edge variations are a function of individual device variations in path delays from the common reference point - rising LCLK
- Long delays in paths increase edge variation / misalignment
  - Reduce the time FB (feedback) is actively being written
  - Create evaluate / restore contention (overlap)
  - Can compensate with wider pulses & initial under lapping but; minCycleTime~evaluationTime+underlap+restoreTime

# Full Read Path (again)

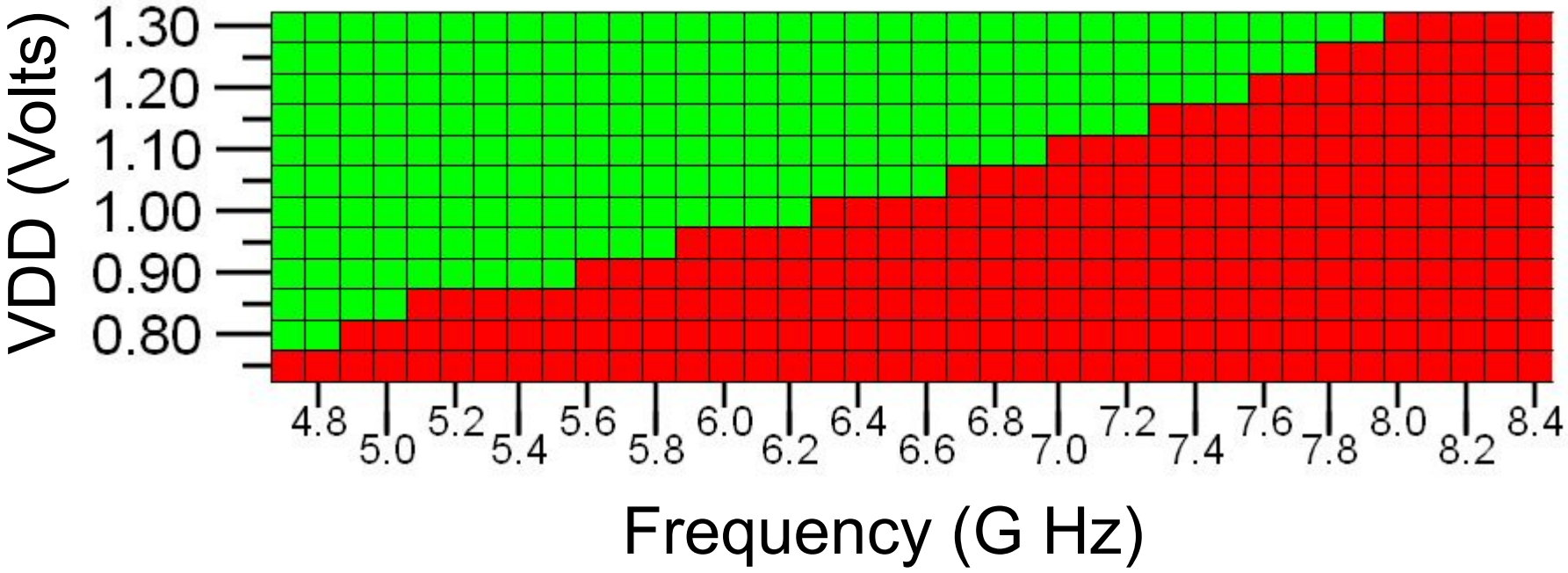



- Shorter dynamic paths without sacrificing latency
  - More complex (local read enables and more latches)
  - Better pulse width control and alignment
  - Shorter evaluate and restore times – higher frequency

# Data Cache Shmoo

GREEN = PASS

RED = FAIL




Sample running at 7.9 GHz

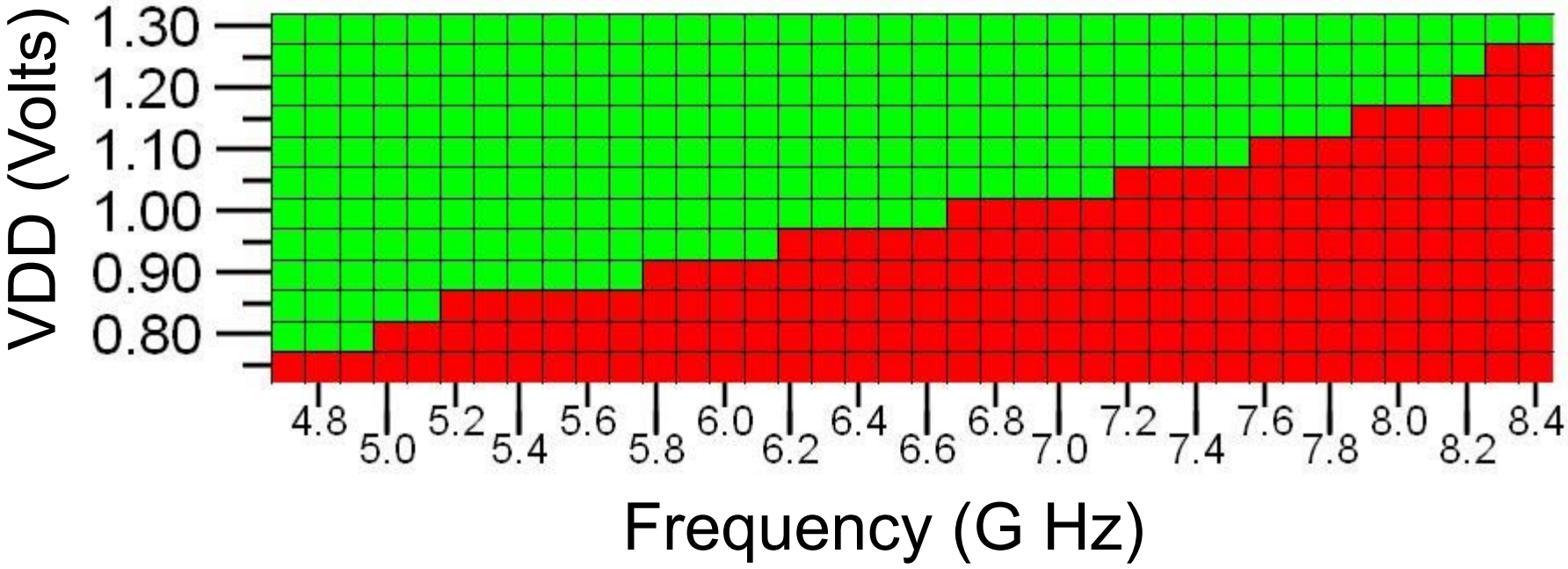


© 2013 IEEE

# Instruction Cache Shmoo

GREEN = PASS

RED = FAIL




Sample running over 8 GHz !

# Conclusion

- Dynamic circuits can deliver fast access
- Device and path variability can drive up dynamic evaluation and restore times which can limit frequency
- Careful attention to dynamic signal length and alignment can result in higher frequencies
- Demonstrated by these 32nm SRAMs running in excess of 7G Hz

# Acknowledgements


- D. Wendel, J. Warnock, B. Huott, J. Vora, M. Canada and the IBM Server and Technology development teams

# Conclusion

- Dynamic circuits can deliver fast access
- Device and path variability can drive up dynamic evaluation and restore times which can limit frequency
- Careful attention to dynamic signal length and alignment can result in higher frequencies
- Demonstrated by these 32nm SRAMs running in excess of 7G Hz