Course Objectives

❑ Introduce students to some relevant advanced topics of current interest in academia and industry

❑ Give the students a feel for research topics and what research means

❑ Make students aware of work happening in India

Current Topics

- ❑ Embedded Memory Design
    - ❑ SRAMs (Dr. Rahul Rao, IBM India)
    - ❑ eDRAMs (Dr. Janakriaman, IITM)
    - ❑ Advanced Memories

Learning Objectives for SRAM

❑ Articulate memory hierarchy and the value proposition of SRAMs in the memory chain + utilization in current processors

❑ Explain SRAM building blocks and peripheral operations and memory architecture (with physical arrangement)

❑ Articulate commonly used SRAM cells (6T vs 8T), their advantages and disadvantages

❑ Explain the operation of a non-conventional SRAM cells, and their limitations

❑ Explain commonly used assist methods

❑ Explain how variations impact memory cells

Learning Objectives for EDRAM

❑ Explain the working of a (e)DRAM. What does Embedded mean?

❑ Explain the working of a feedback sense amplifier and modify existing designs to improve performance

❑ Calculate the voltage levels of operation of various components for an eDRAM

❑ Introduce stacked protect devices to reduce voltage stress of the WL driver

Grading

- Assignments - 10%
- Midsem – 30%
- Project - 20%
- End Semester - 40%

Course Schedule

- Friday – 2:00 –5:00
- ESB 207A

Embedded DRAM

Janakiraman V

Assistant Professor

Electrical Department

IIT Madras

Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example
- ❑ Gated Feedback Sense Amplifier (case study)
- ❑ References

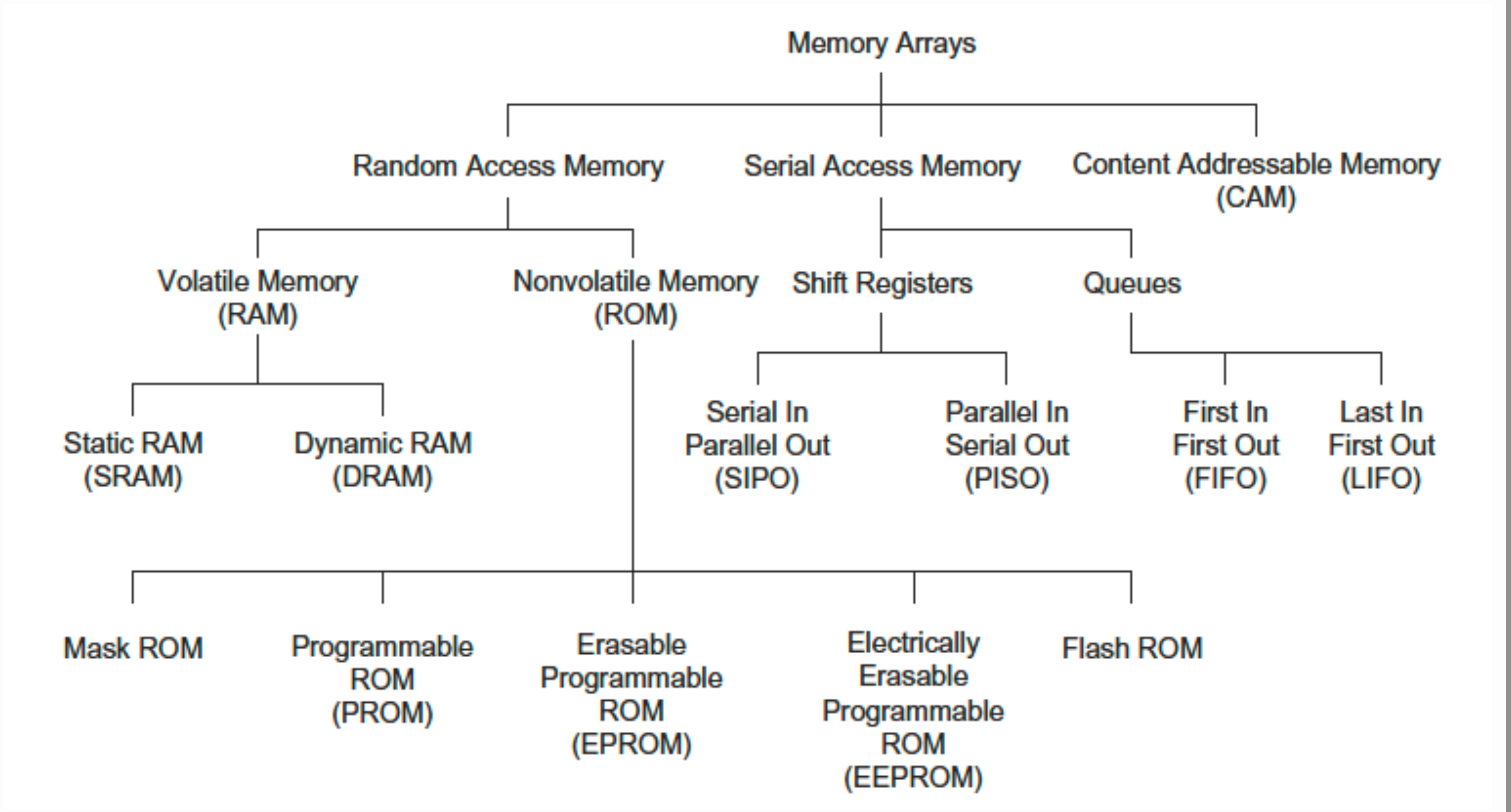
Acknowledgement

- Raviprasad Kuloor (Course slides were prepared by him)
- John Barth, IBM SRDC for most of the slides content
- Madabusi Govindarajan
- Subramanian S. Iyer
- Many Others

Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

Memory Classification revisited



Memory Arrays

Random Access Memory

Serial Access Memory

Content Addressable Memory (CAM)

Volatile Memory (RAM)

Nonvolatile Memory (ROM)

Shift Registers

Queues

Static RAM (SRAM)

Dynamic RAM (DRAM)

Serial In Parallel Out (SIPO)

Parallel In Serial Out (PISO)

First In First Out (FIFO)

Last In First Out (LIFO)

Mask ROM

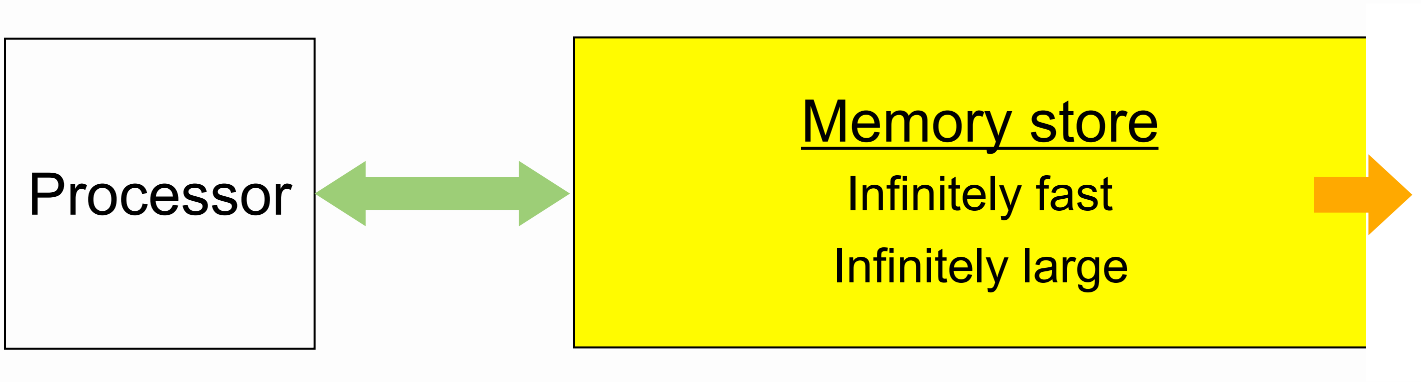
Programmable ROM (PROM)

Erasable Programmable ROM (EPROM)

Electrically Erasable Programmable ROM (EEPROM)

Flash ROM

Motivation for a memory hierarchy – infinite memory



Processor

Memory store

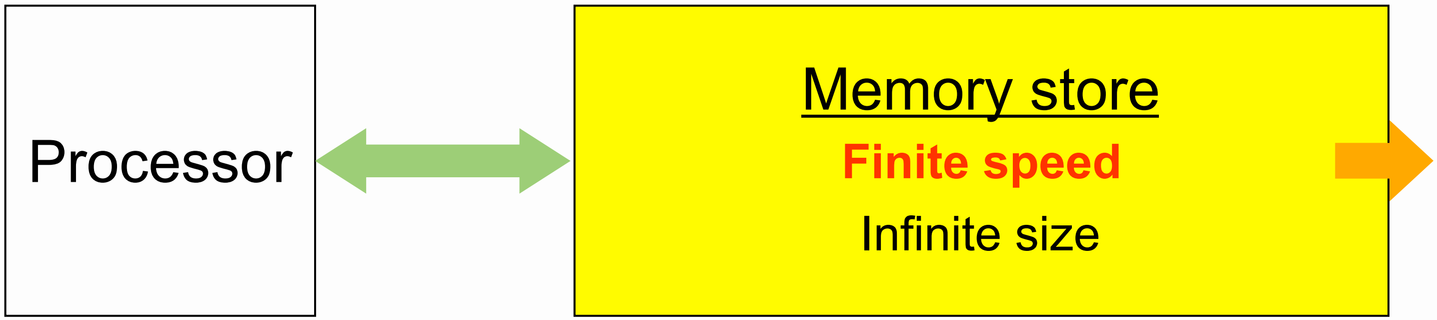
Infinitely fast

Infinitely large

Cycles per Instruction (CPI) = Number of processor clock cycles required per instruction

CPI[∞ cache]

Finite memory speed



Processor

Memory store

Finite speed

Infinite size

$$\mathrm{CPI} = \mathrm{CPI}[\infty\ \mathrm{cache}] + \mathrm{FCP}$$

Finite cache penalty

Locality of reference – spatial and temporal

Temporal

If you access something now you'll need it again soon

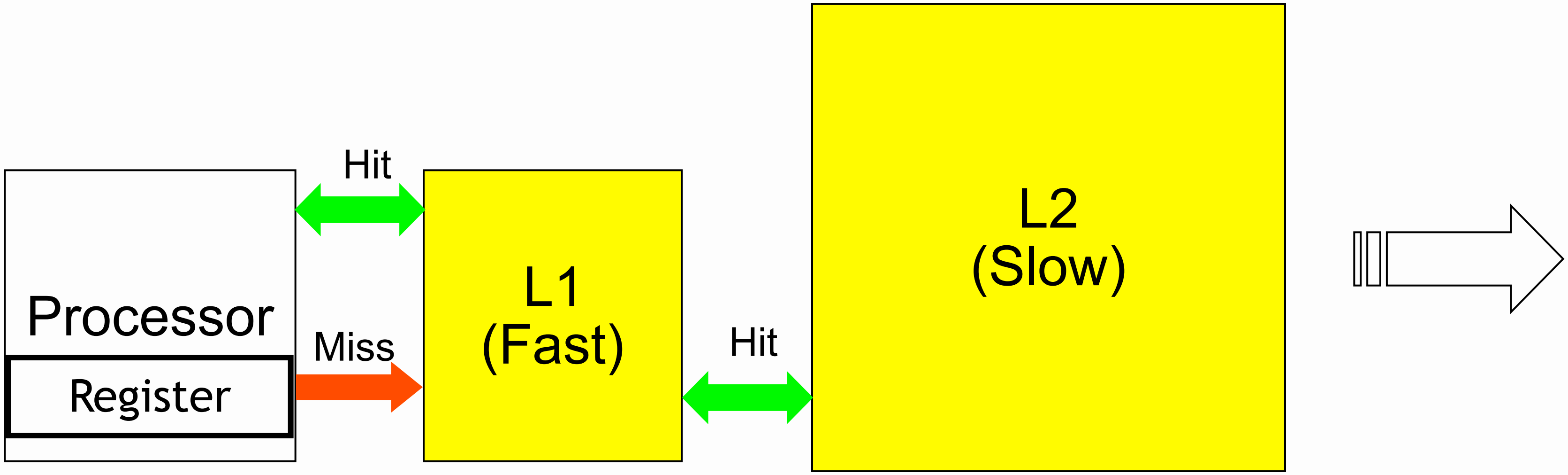
e.g: Loops

Spatial

If you accessed something you'll also need its neighbor

e.g: Arrays

Exploit this to divide memory into hierarchy

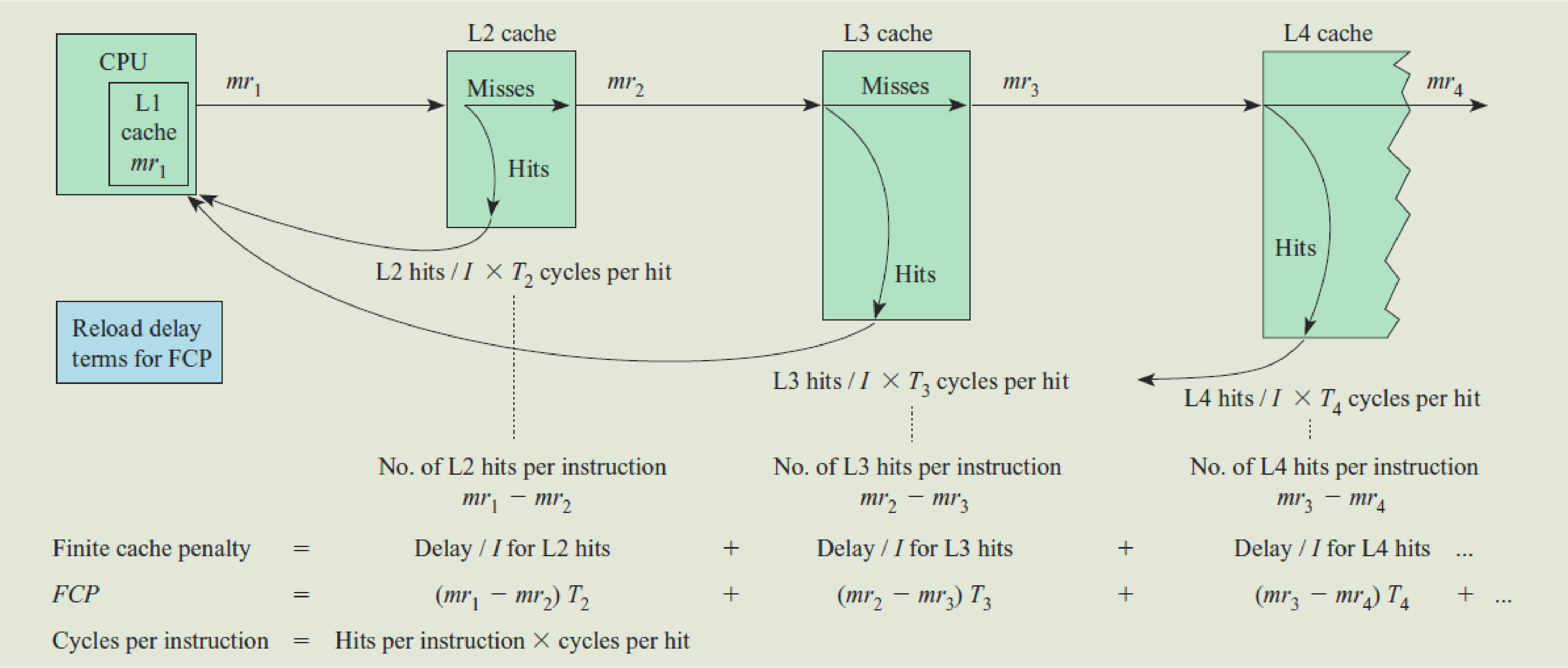


Cache size impacts cycles-per-instruction

Logic-based eDRAM: Origins and rationale for use

R. E. Matick
S. E Schuster

IBM J. RES. & DEV. VOL. 49 NO. 1 JANUARY 2005



CPU

L1 cache mr_1

mr_1

L2 cache

Misses

Hits

mr_2

L3 cache

Misses

Hits

mr_3

L4 cache

Hits

mr_4

L2 hits / I × T_2 cycles per hit

L3 hits / I × T_3 cycles per hit

L4 hits / I × T_4 cycles per hit

Reload delay terms for FCP

No. of L2 hits per instruction $mr_1 - mr_2$

No. of L3 hits per instruction $mr_2 - mr_3$

No. of L4 hits per instruction $mr_3 - mr_4$

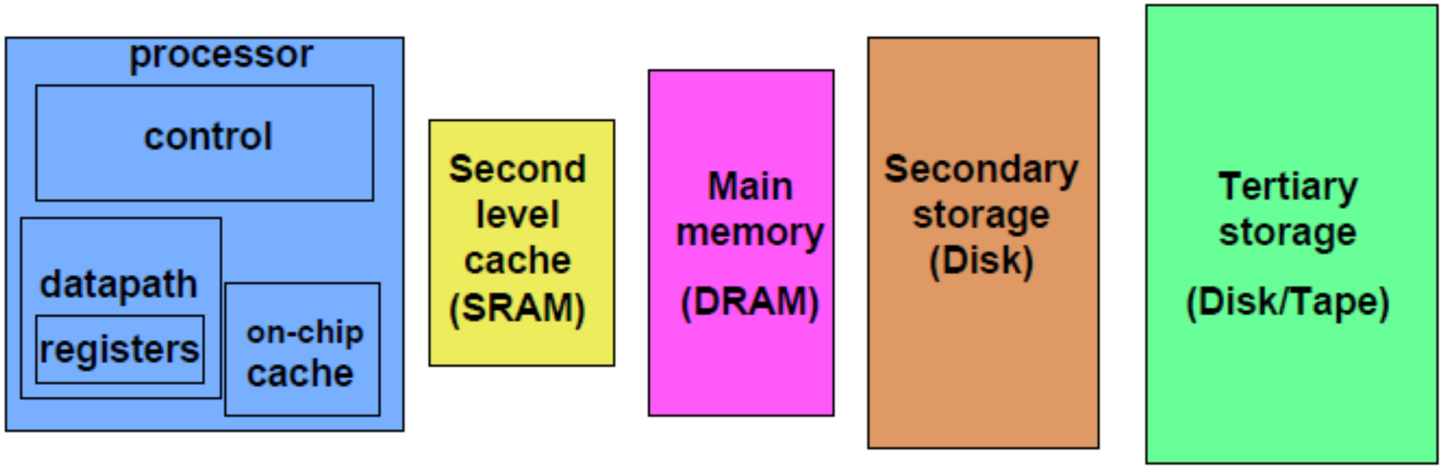
Finite cache penalty = Delay / I for L2 hits + Delay / I for L3 hits + Delay / I for L4 hits ...

$FCP = (mr_1 - mr_2)\,T_2 + (mr_2 - mr_3)\,T_3 + (mr_3 - mr_4)\,T_4 + \ldots$

Cycles per instruction = Hits per instruction × cycles per hit

⟶ Access rate reduces → Slower memory is sufficient

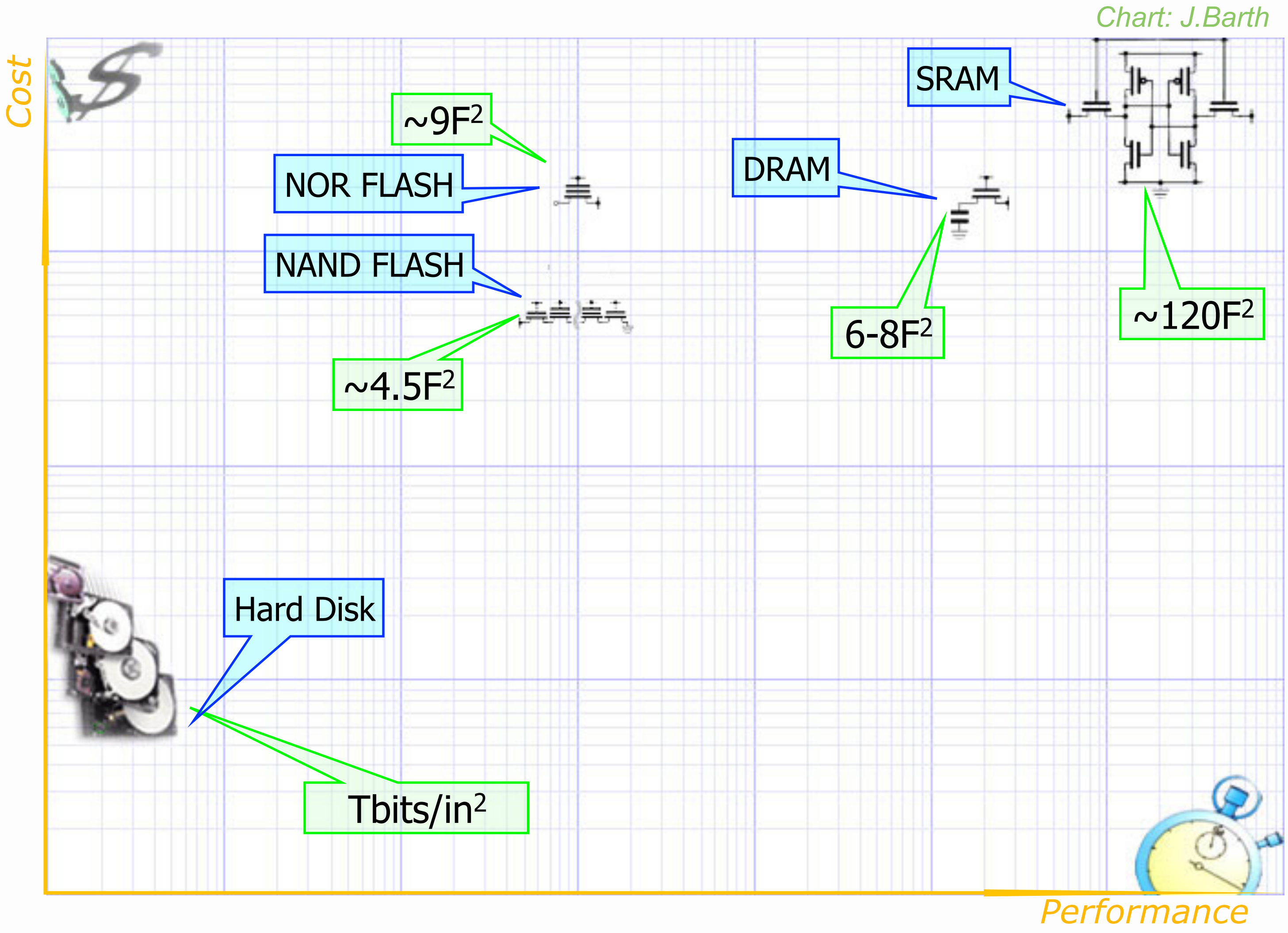
Cache size impacts cycles-per-instruction



Speed	1ns	10ns	100ns	10ms	10sec
Size	B	KB	MB	GB	TB

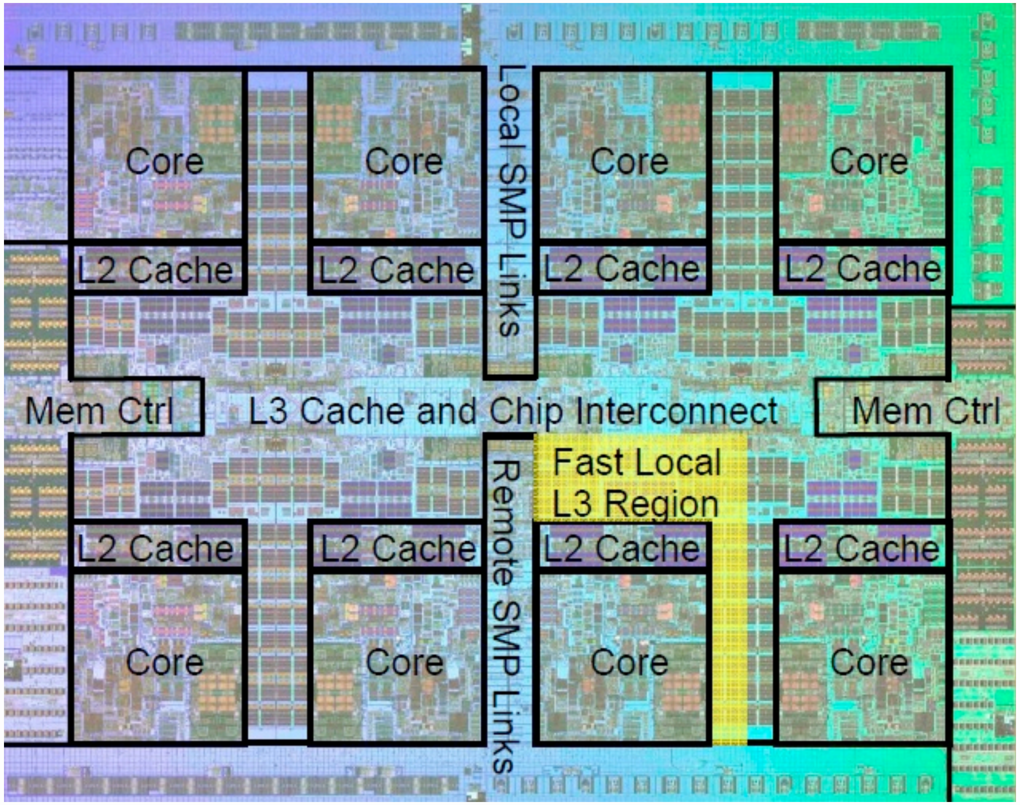
For a 5GHz processor, scale the numbers by 5x

Technology choices for memory hierarchy



eDRAM L3 cache

Power7 processor



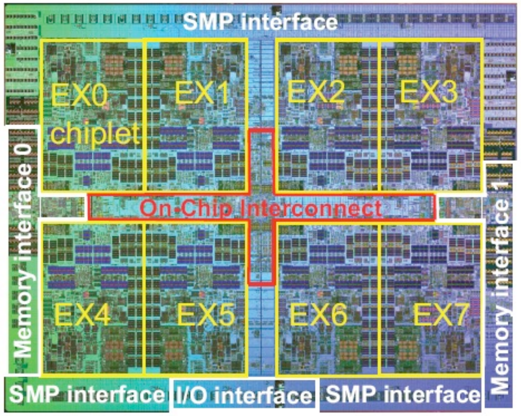
Move L2,L3 Cache inside of the data hungry processor

Higher hit rate → Reduced FCP

JSSCC11

Embedded DRAM Advantages

IBM Power7^{tm}
32MB eDRAM L3

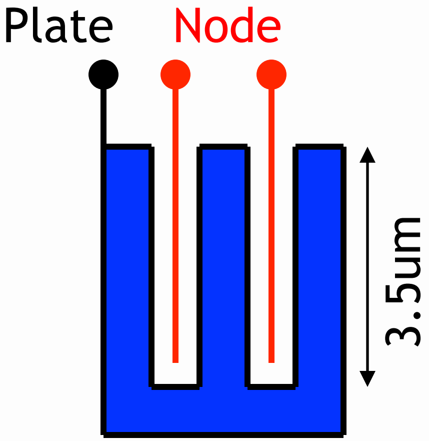


Memory Advantage

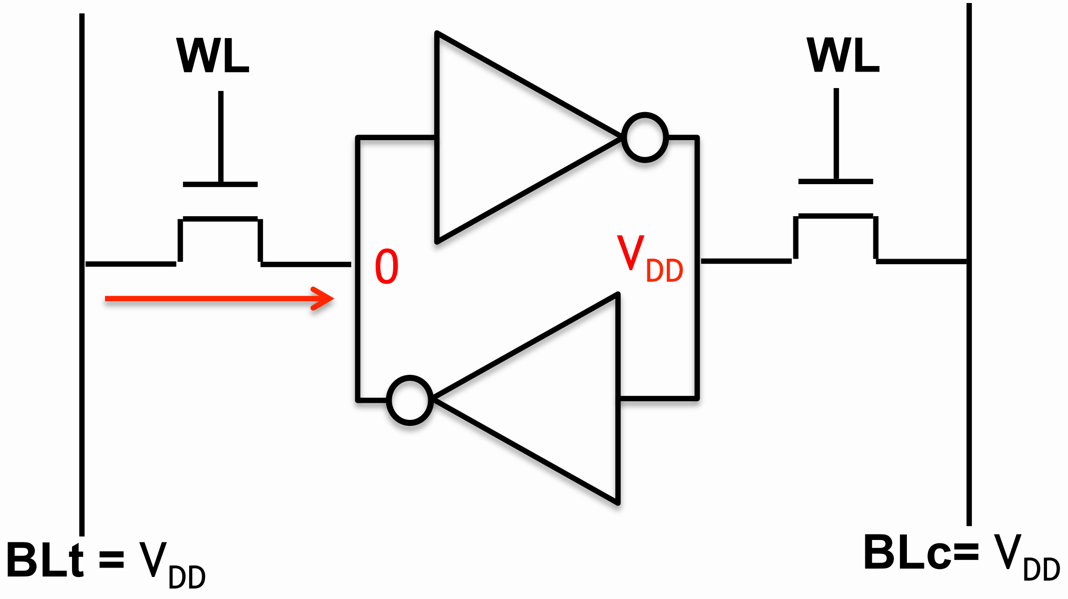
- 2x Cache can provide > 10% Performance
- ~3x Density Advantage over eSRAM
- 1/5x Standby Power Compared to SRAM
- Soft Error Rate 1000x lower than SRAM
- Performance ? DRAM can have lower latency !
- IO Power reduction

Deep Trench Capacitor

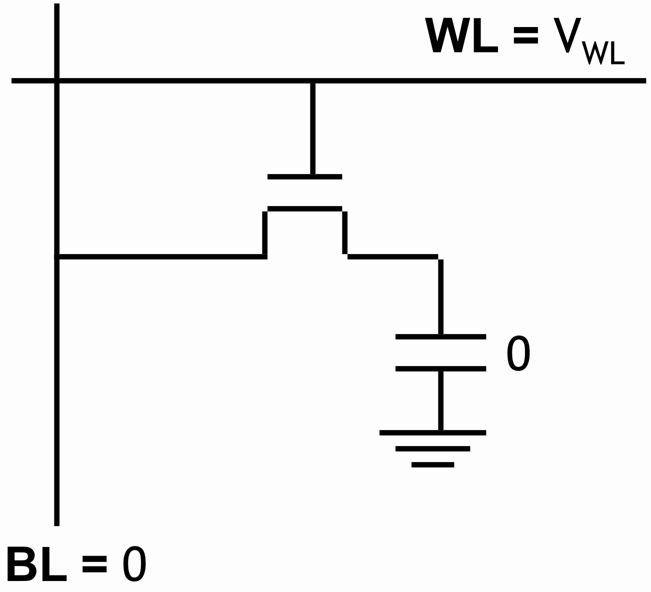
- Low Leakage Decoupling
- 25x more Cap / µm^2 compared to planar
- Noise Reduction = Performance Improvement
- Isolated Plate enables High Density Charge Pump



eDRAM Advantages – Stand By Leakage

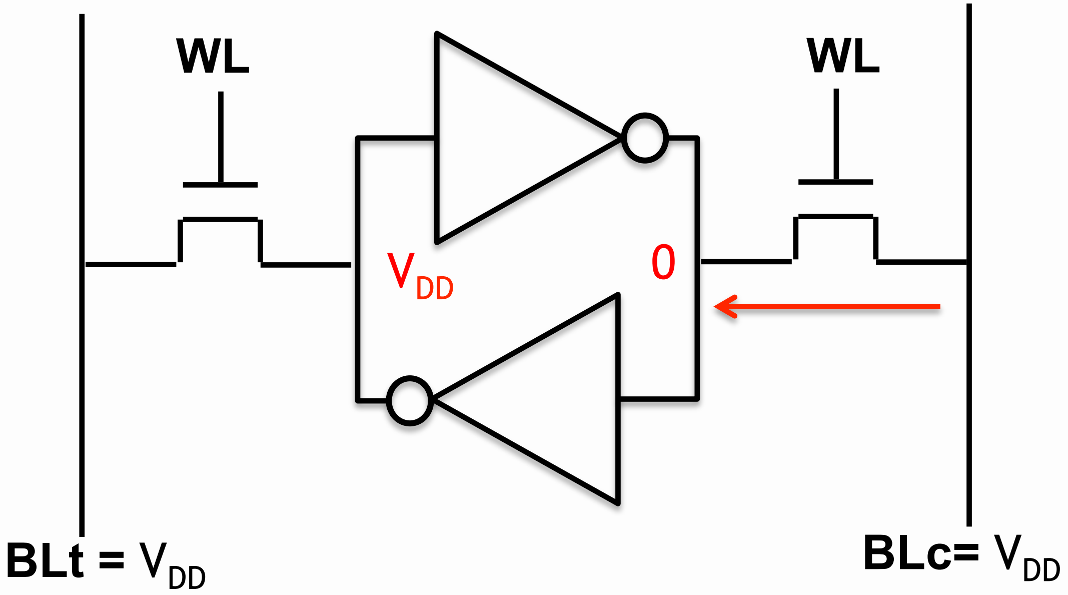


Both inverters leak
True Pass transistor leaks

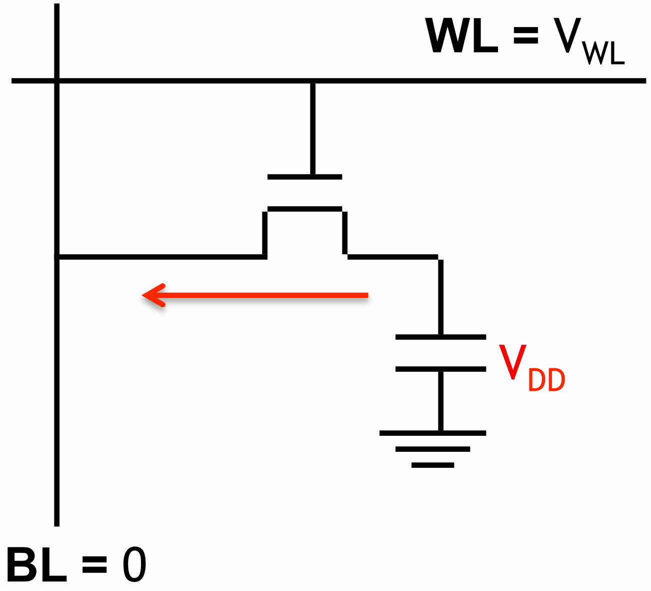


NO Leakage

eDRAM Advantages – Stand By Leakage

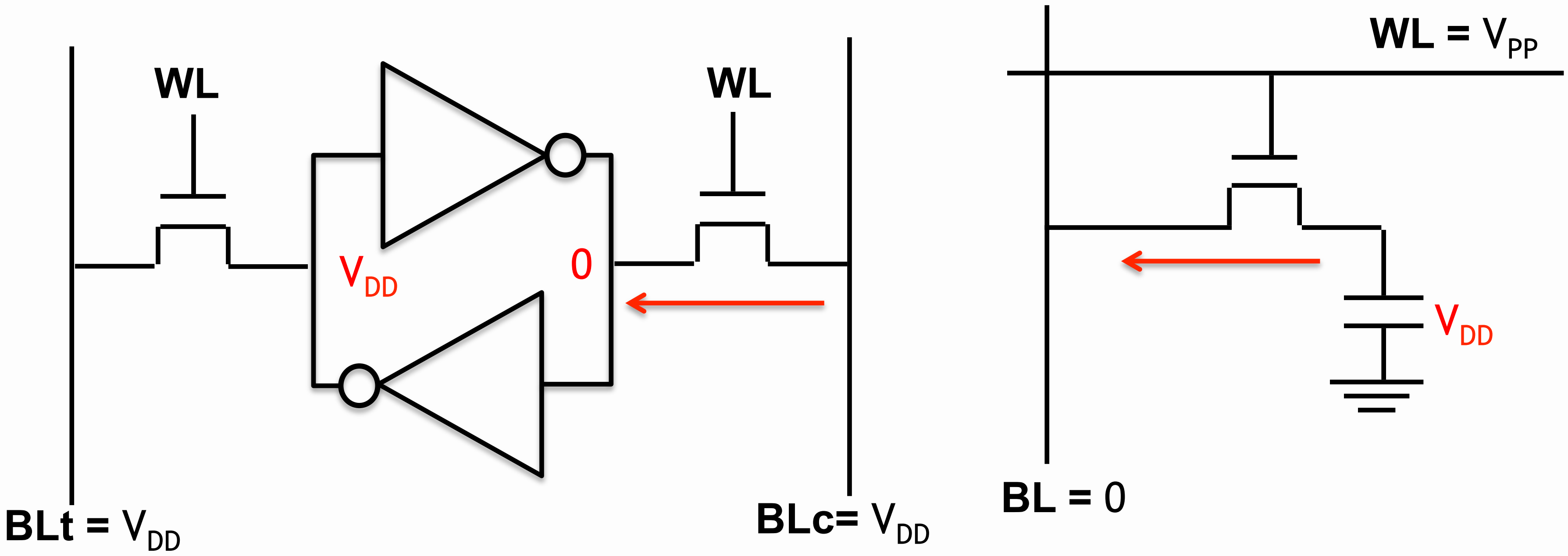


Both inverters leak
Complement Pass transistor leaks



Leakage exists

eDRAM Advantages – Stand By Leakage

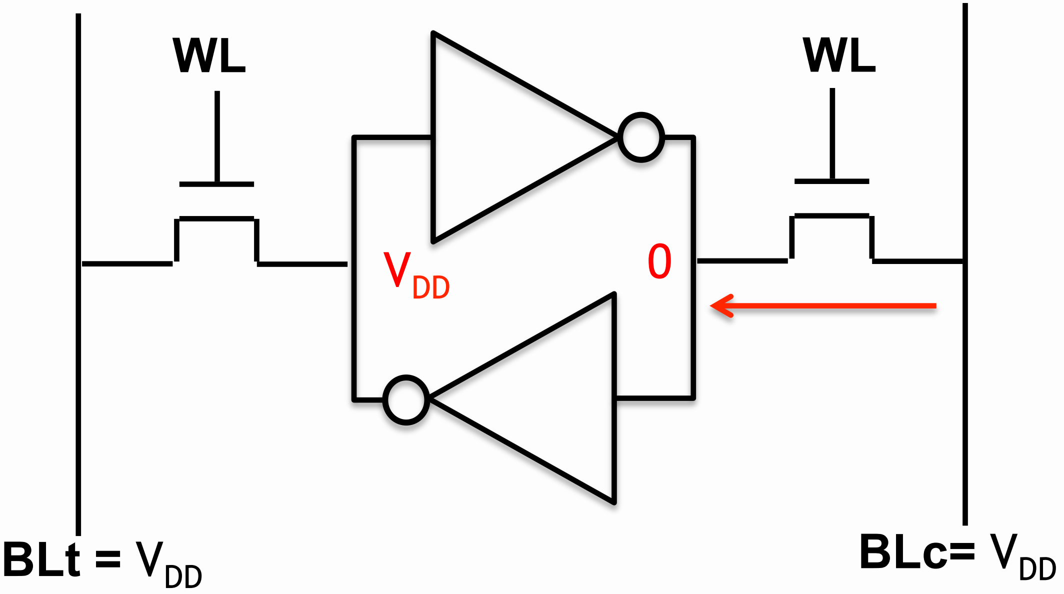


Both inverters leak
Complement Pass transistor leaks

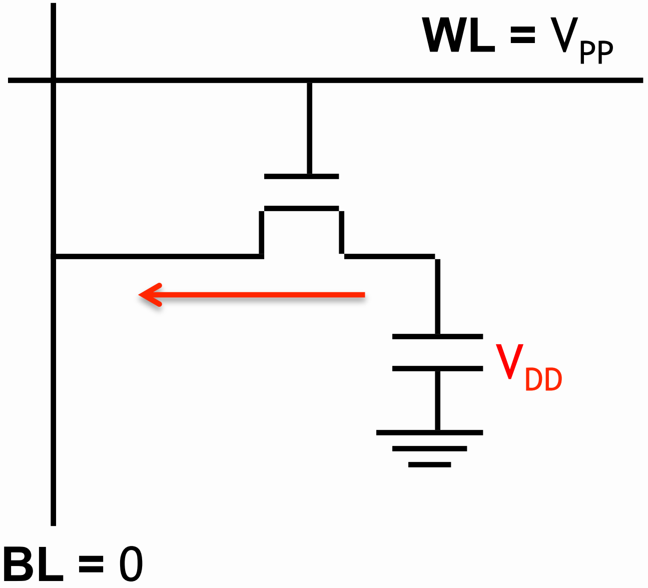
Leakage exists

On average: eDRAMs have 1/5x Standby Power Compared to SRAM

eDRAM Advantages – Performance

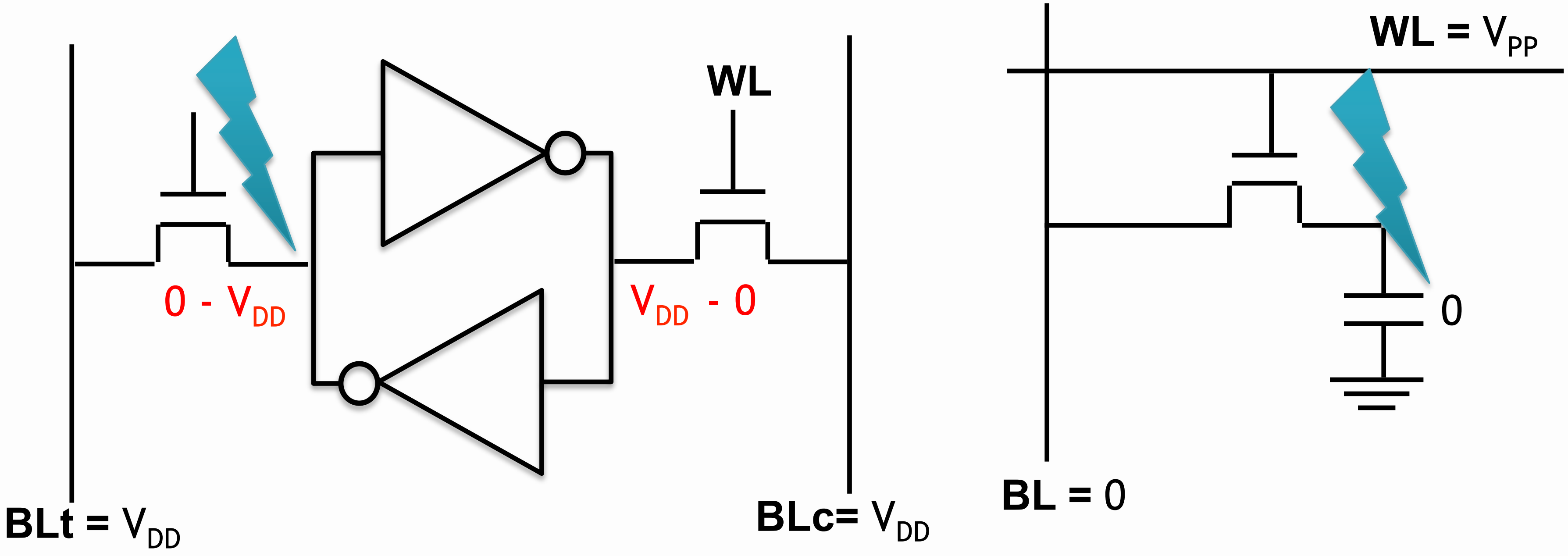


Both inverters leak
Complement Pass transistor leaks



Leakage exists

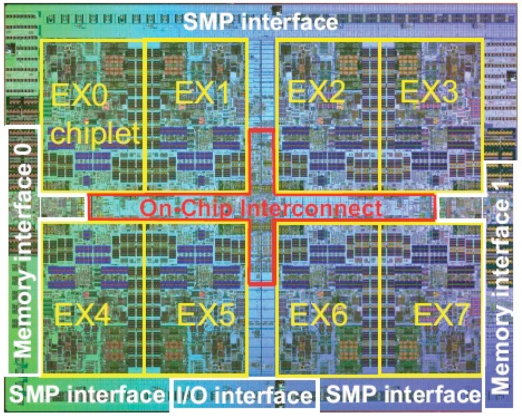
eDRAM Advantages – Soft Error Rate



- Cosmic particles can bombard the cell and cause a bump in the cell voltage
- If voltage bump is large enough SRAM can permanently flip
  - Static cross couple inverters
- Voltage on DRAM capacitor node can also bump
- But will leak away with time –
  - Only those cells which get refreshed in a certain period will flip
- Soft Error Rate 1000x lower than SRAM

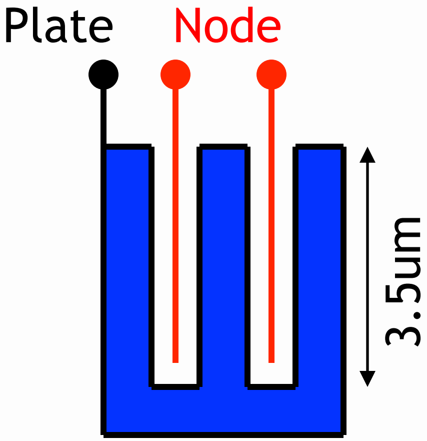
Embedded DRAM Advantages

IBM Power7tm
32MB eDRAM L3



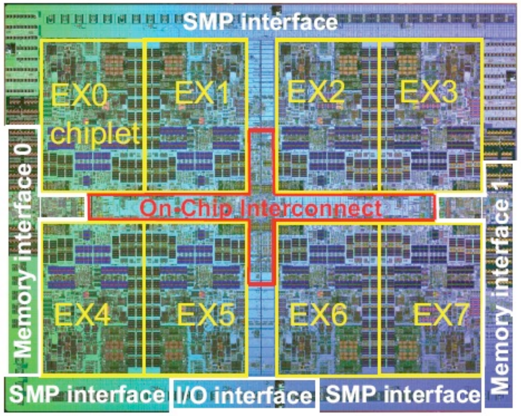
Deep Trench Capacitor

- Low Leakage Decoupling
- 25x more Cap / µm^2 compared to planar
- Noise Reduction = Performance Improvement
- Isolated Plate enables High Density Charge Pump



Embedded DRAM Advantages

IBM Power7tm
32MB eDRAM L3

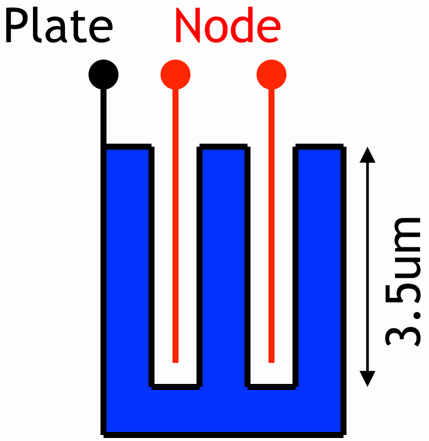


Memory Advantage

- 2x Cache can provide > 10% Performance
- ~3x Density Advantage over eSRAM
- 1/5x Standby Power Compared to SRAM
- Soft Error Rate 1000x lower than SRAM
- Performance ? DRAM can have lower latency !
- IO Power reduction

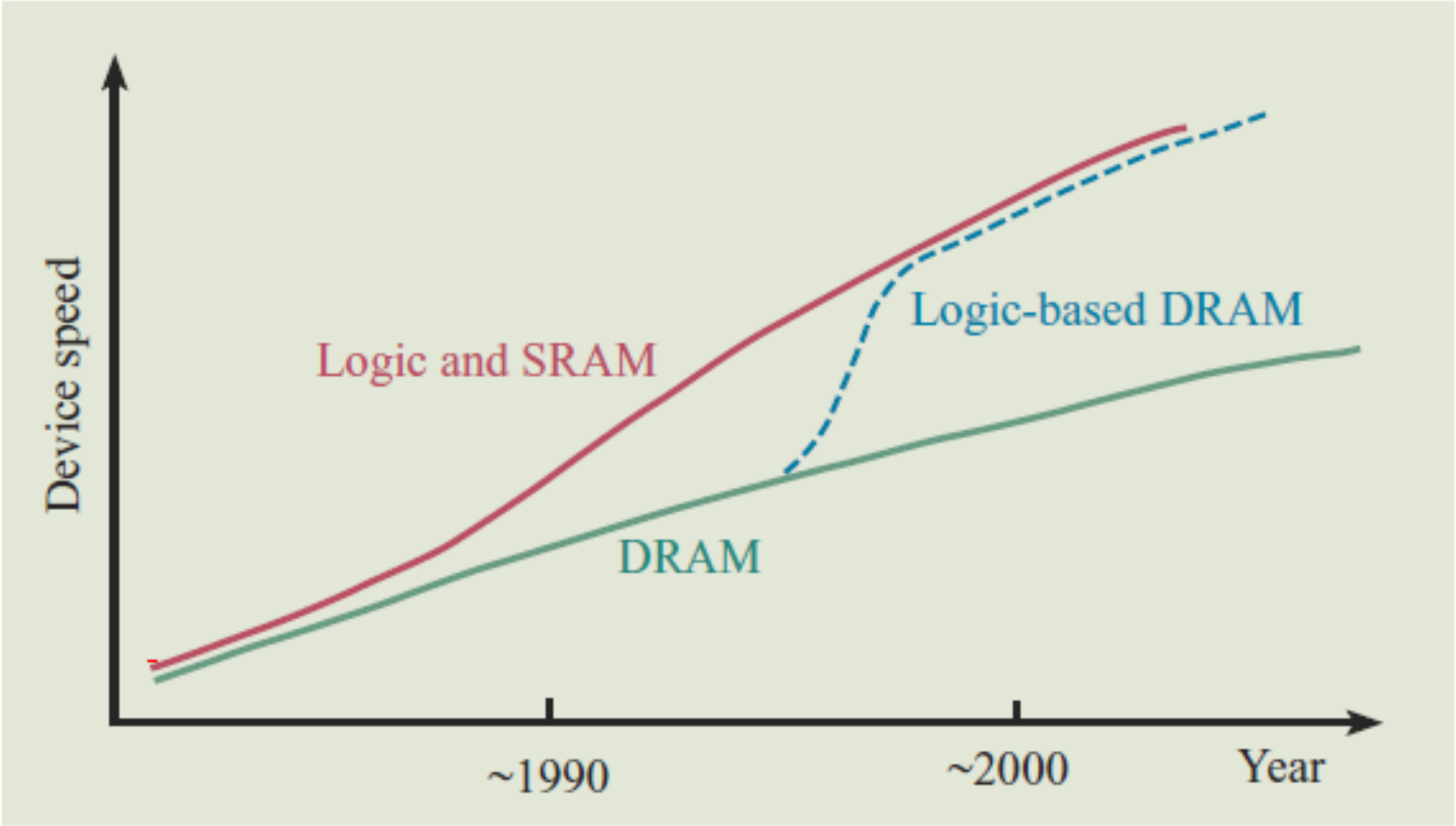
Deep Trench Capacitor

- Low Leakage Decoupling
- 25x more Cap / µm^2 compared to planar
- Noise Reduction = Performance Improvement
- Isolated Plate enables High Density Charge Pump

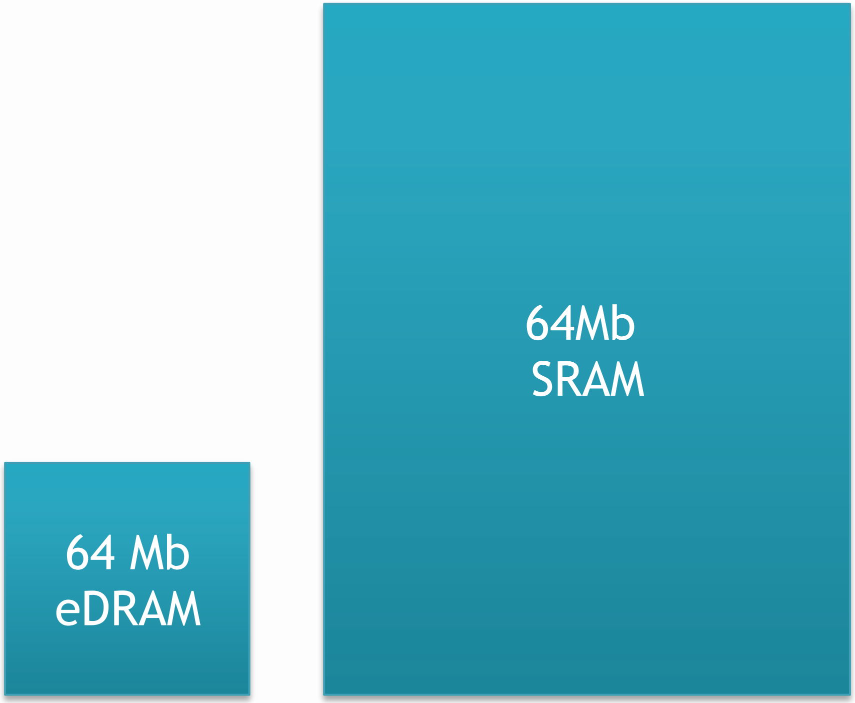


Cache performance – SRAM vs. DRAM

Chart: Matick & Schuster, op. cit.

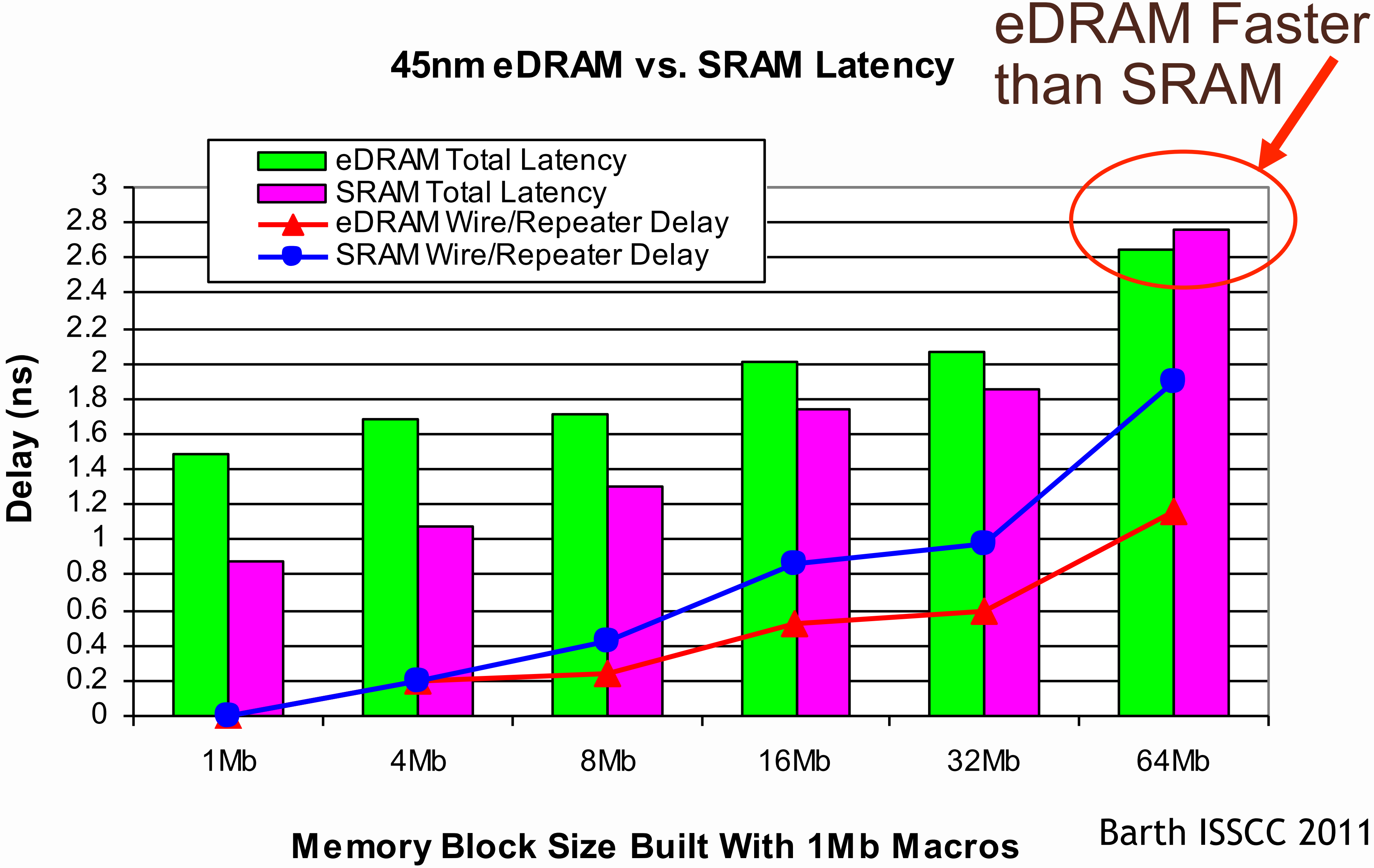


Cache performance – SRAM vs. DRAM



Time to access the farthest word-line determines performance
Access time = Cell access time + time of flight interconnect delay

Embedded DRAM Performance



eDRAM Faster than SRAM

Barth ISSCC 2011

Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM Write Analysis
- ❑ eDRAM Sense-Amplifier Specification
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

Fundamental DRAM Operation

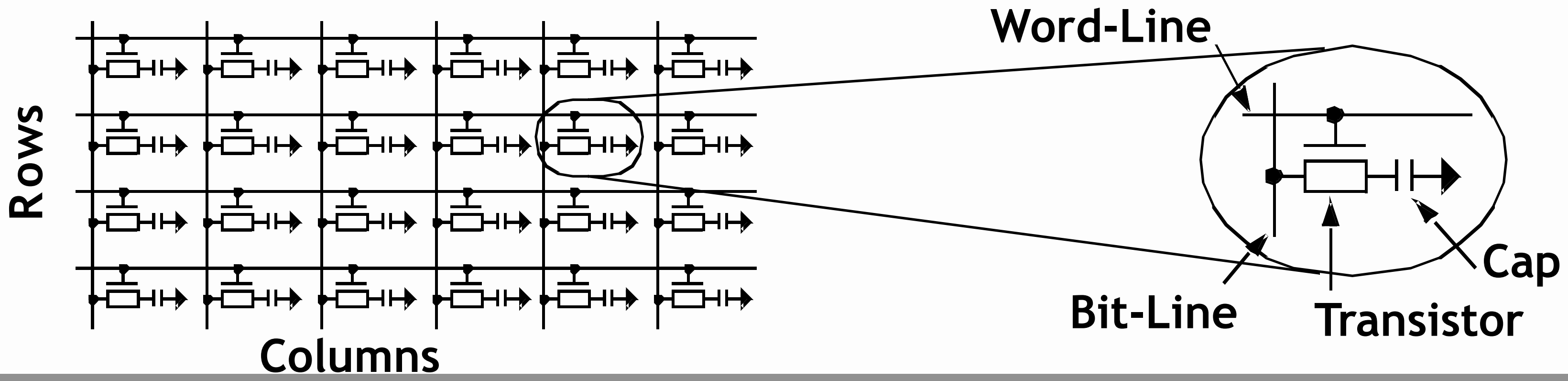
Memory Arrays are composed of Row and Columns

Most DRAMs use 1 Transistor as a switch and
1 Cap as a storage element (Dennard 1967)

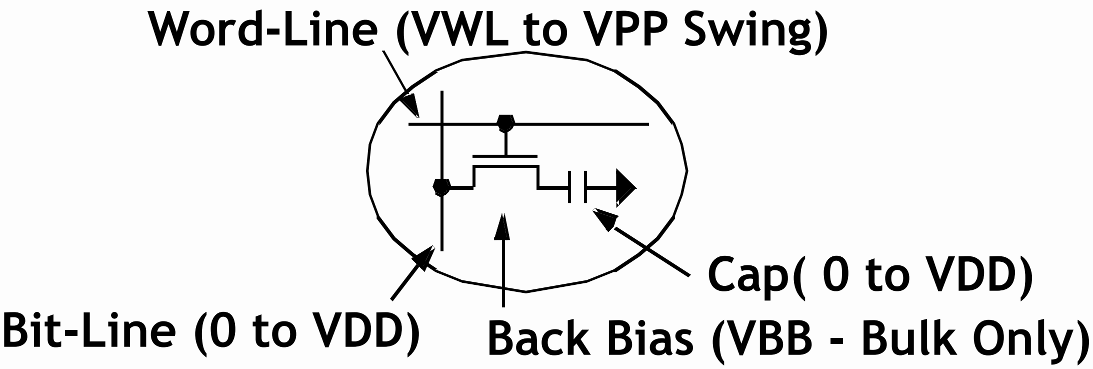
Single Cell Accessed by Decoding One Row / One Column (Matrix)

Row (Word-Line) connects storage Caps to Columns (Bit-Line)

Storage Cap Transfers Charge to Bit-Line, Altering Bit-Line Voltage



1T1C DRAM Cell Terminals

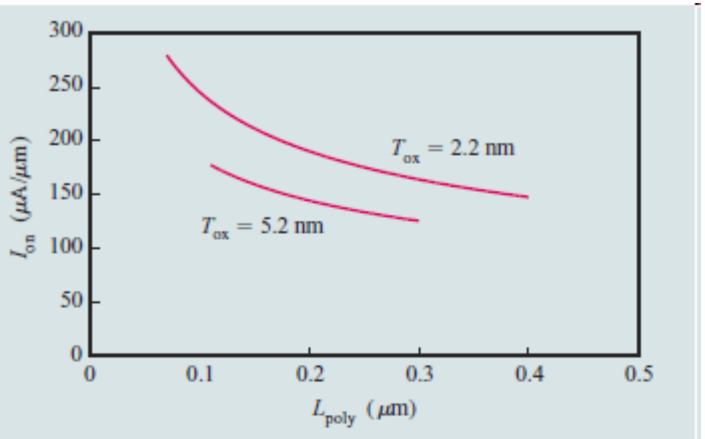


Word-Line (VWL to VPP Swing)

Cap(0 to VDD)

Bit-Line (0 to VDD)

Back Bias (VBB - Bulk Only)



VWL: Word-Line Low Supply, GND or Negative for improved leakage

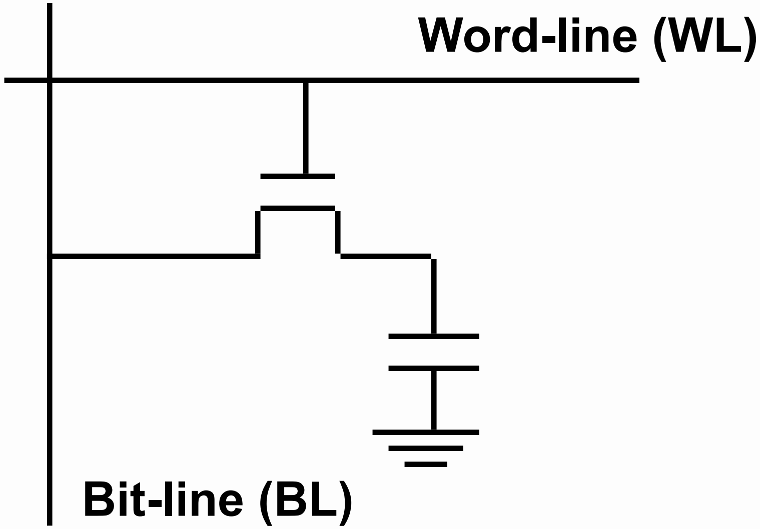
VPP: Word-Line High Supply, 1.8V up to 3.5V depending on Technology
Required to be at least a Vt above VDD to write full VDD

VBB: Back Bias, Typically Negative to improve Leakage
Not practical on SOI

IBM J RES & DEV 2005

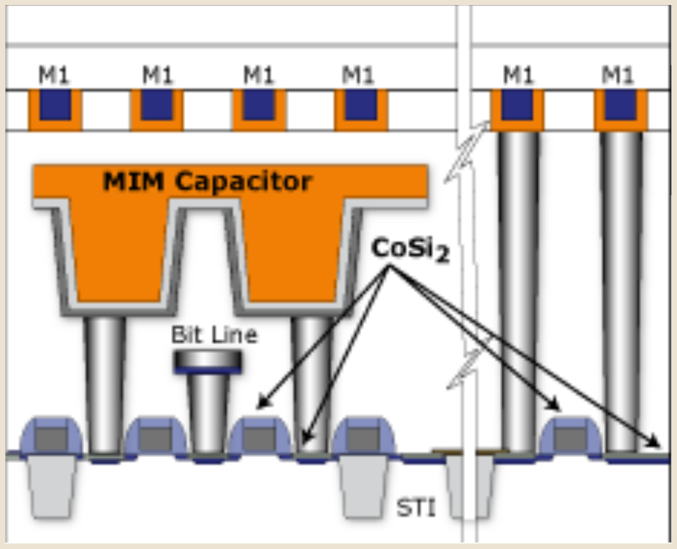
Choice of Access Transistor

- **DRAMs are limited by sub-threshold leakage**
  - I_{off} α $1/t_{OX}$
  - Use thick oxide transistor
    - $t_{OX} \approx$ 3nm in 14nm Technology
    - Thin oxide transistors ($t_{OX} \approx$ 1nm)
  - What should be the width of the device?
    - Density constraints => Unit size
  - Unit size transistor also provides least leakage

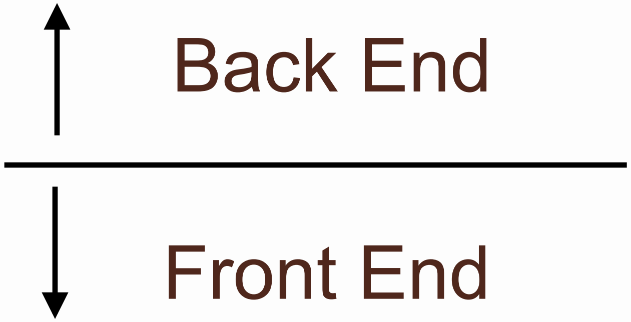


IBM J RES & DEV 2005

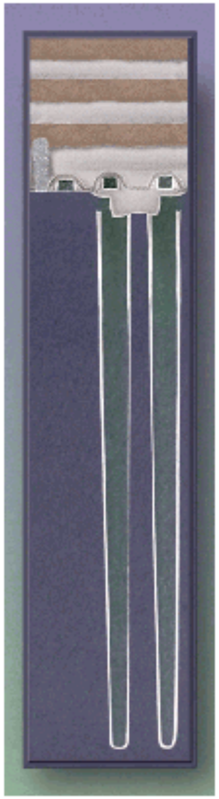
MIM Cap v/s Trench



MIM eDRAM Process



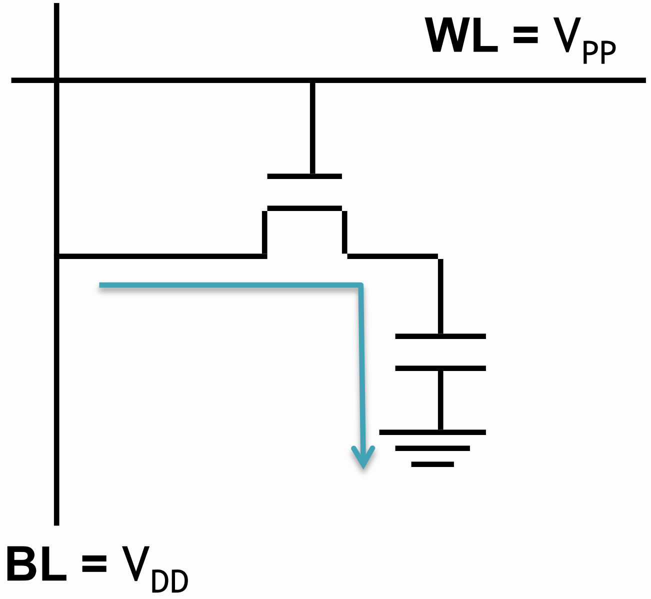
Trench eDRAM Process



- Stack capacitor requires more complex process
- M1 height above gate is increased with stacked capacitor
  - M1 parasitics significantly change when wafer is processed w/o eDRAM
  - Drives unique timings for circuit blocks processed w/ and w/o eDRAM
    - Logic Equivalency is compromised – Trench is Better Choice

Word-line Swing - High

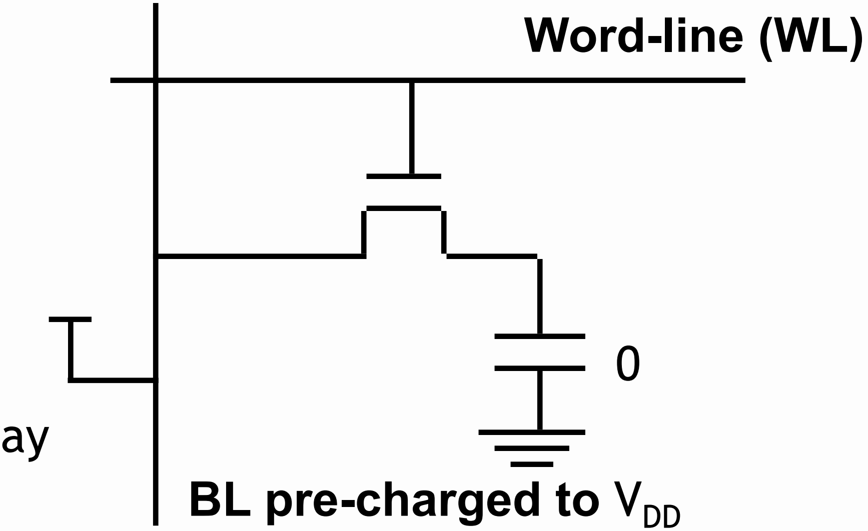
- WL High Voltage – ON State
  - Technology maximum high voltage?
    - Access device is thick oxide
    - Can handle a swing of more than V_{DD}.
  - How high?
    - We wish to write a logic-1 completely
    - Logic-1 = V_{DD}
    - Access device is an NMOS transistor
      - Cannot pass V_{DD} fully if WL= V_{DD}
    - WL high = $V_{PP} \geq V_{DD} + V_{Tn}$
    - What about VTn variability
    - $V_{PP} \geq V_{DD} + V_{Tn} + \Delta V_{Tn}$
- Typical value of $V_{PP} = 0.9 + 0.4 + 0.2 = 1.5V$

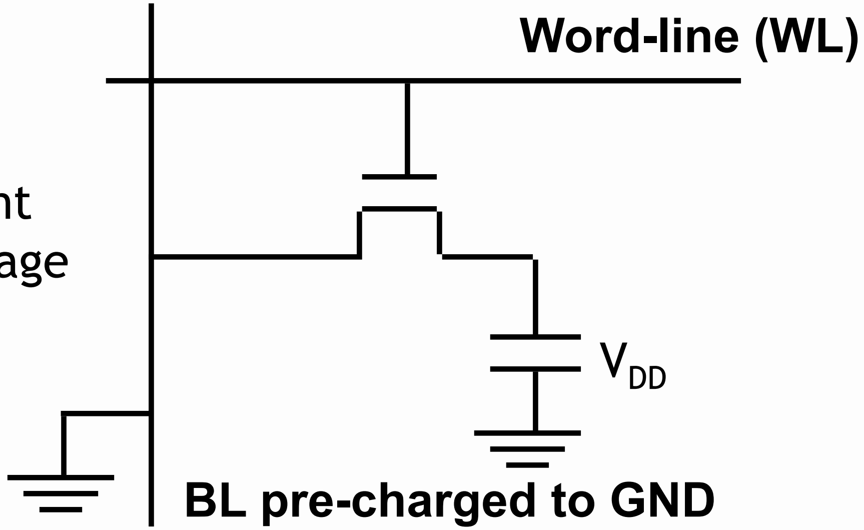


IBM J RES & DEV 2005

Word-line Swing - Low

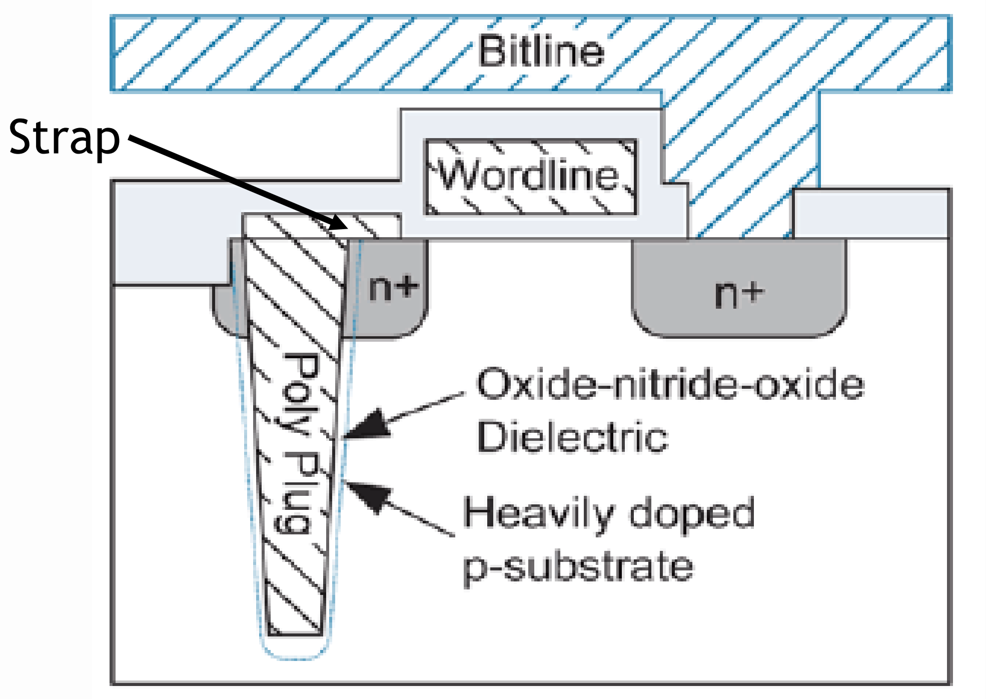
- WL Low Voltage (V_{WL})– OFF State
- BL will be pre-charged to either 0 or V_{DD}
  - BL pre-charged to 0
    - Cell storing a 1 causes leakage
  - BL pre-charged to 1
    - Cell storing a 0 causes leakage
- Need to minimize leakage current either way
- V_{GS} of access device needs to be as low as possible
  - IOFF decreases exponentially with V_{GS}
  - Can we lower the WL down to –V_{DD}?
  - What is the limit?
  - Lower the WL voltage down to the point where GIDL – Gate Induced Drain Leakage sets in
  - Typical value of V_{WL} = -300 mV



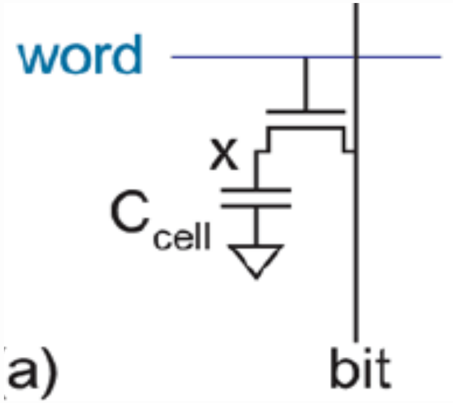


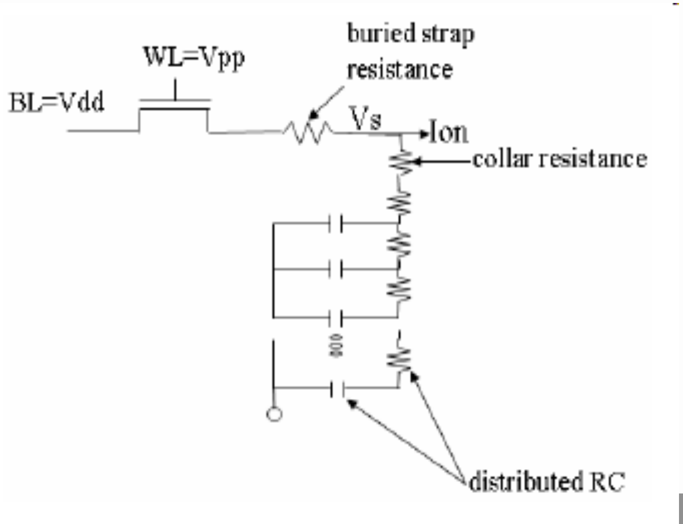
DRAM cell Cross section

- Store their contents as charge on a capacitor rather than in a feedback loop.
- 1T dynamic RAM cell has a transistor and a capacitor

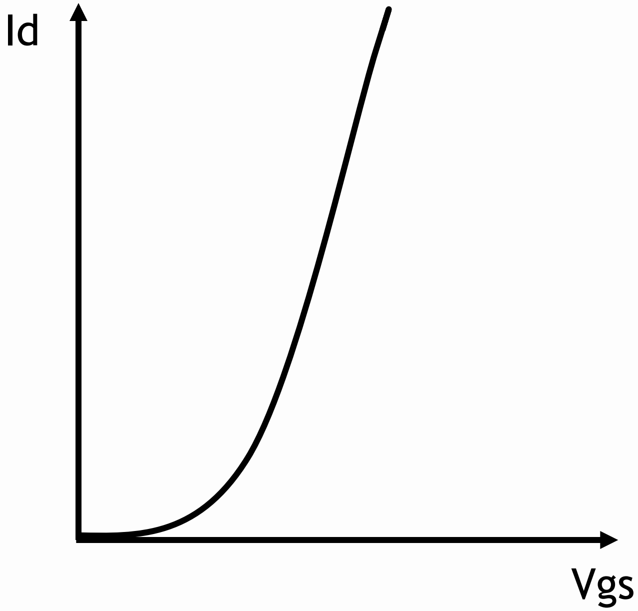


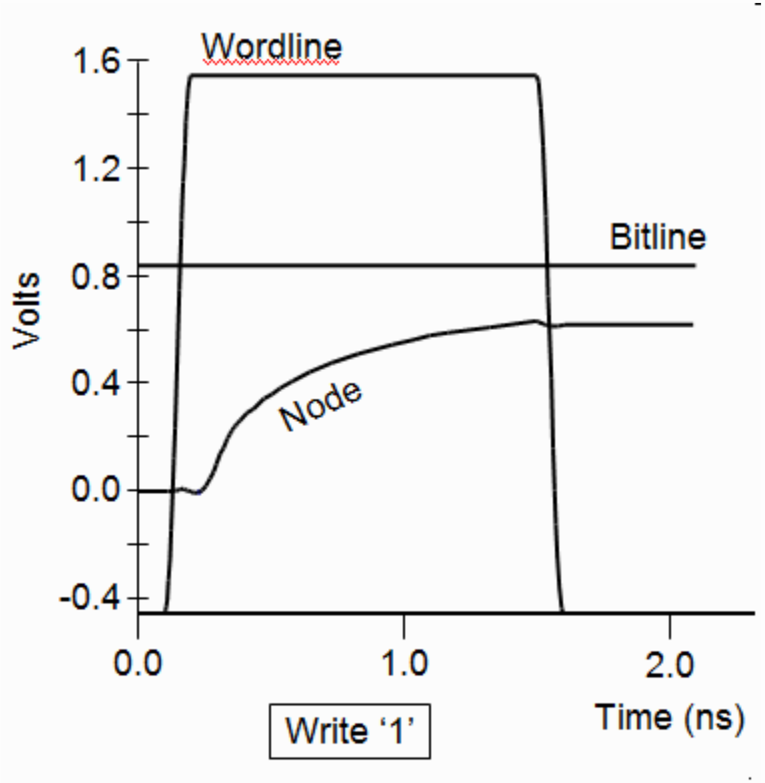
CMOS VLSI design - PEARSON





Storing data ‘1’ in the cell

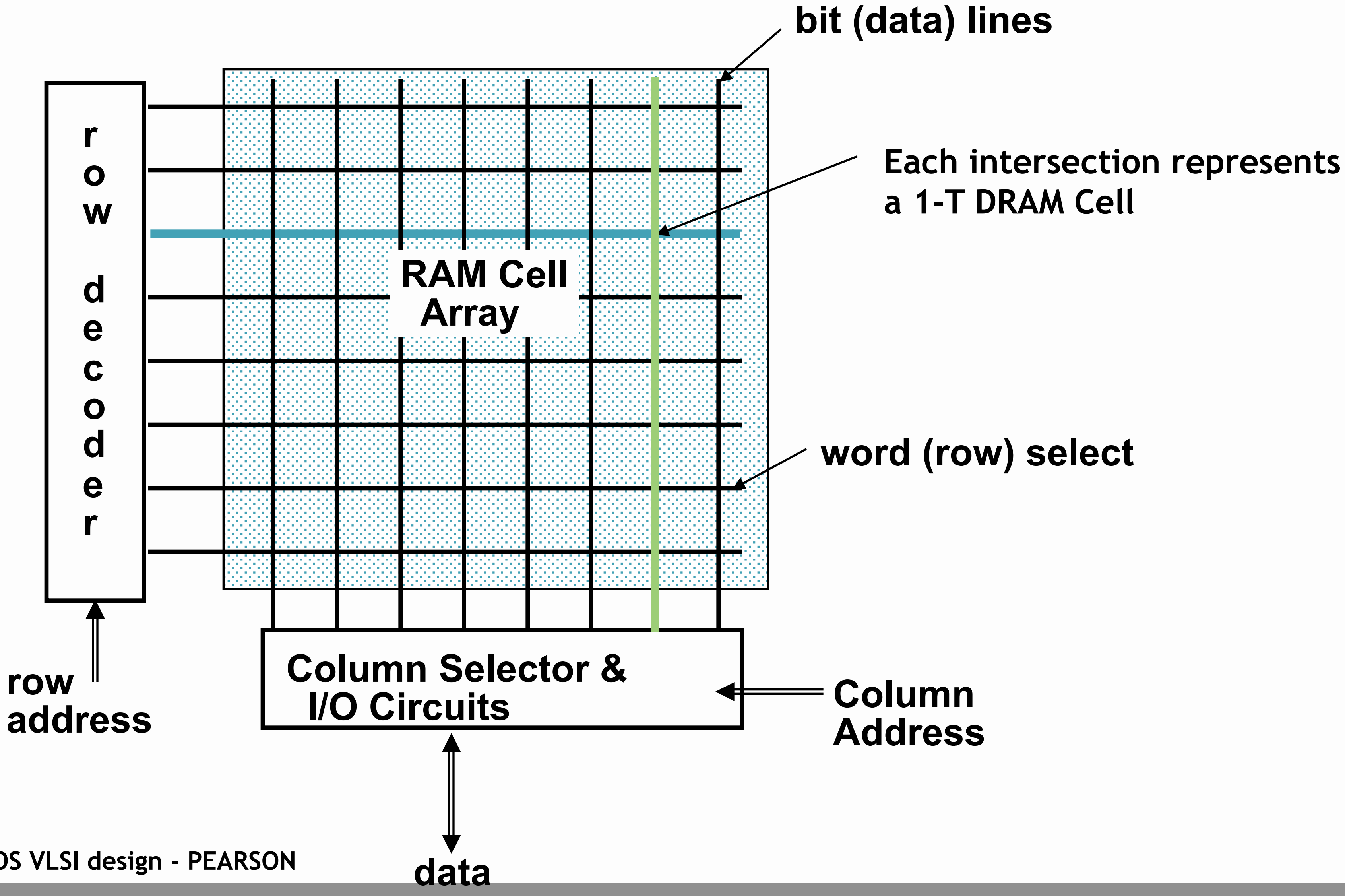




Vgs for pass transistor reduces as bitcell voltage rises, increasing Ron

Why there is a reduction in cell voltage after WL closes? Experiment

Classical DRAM Organization



CMOS VLSI design - PEARSON

DRAM Subarray

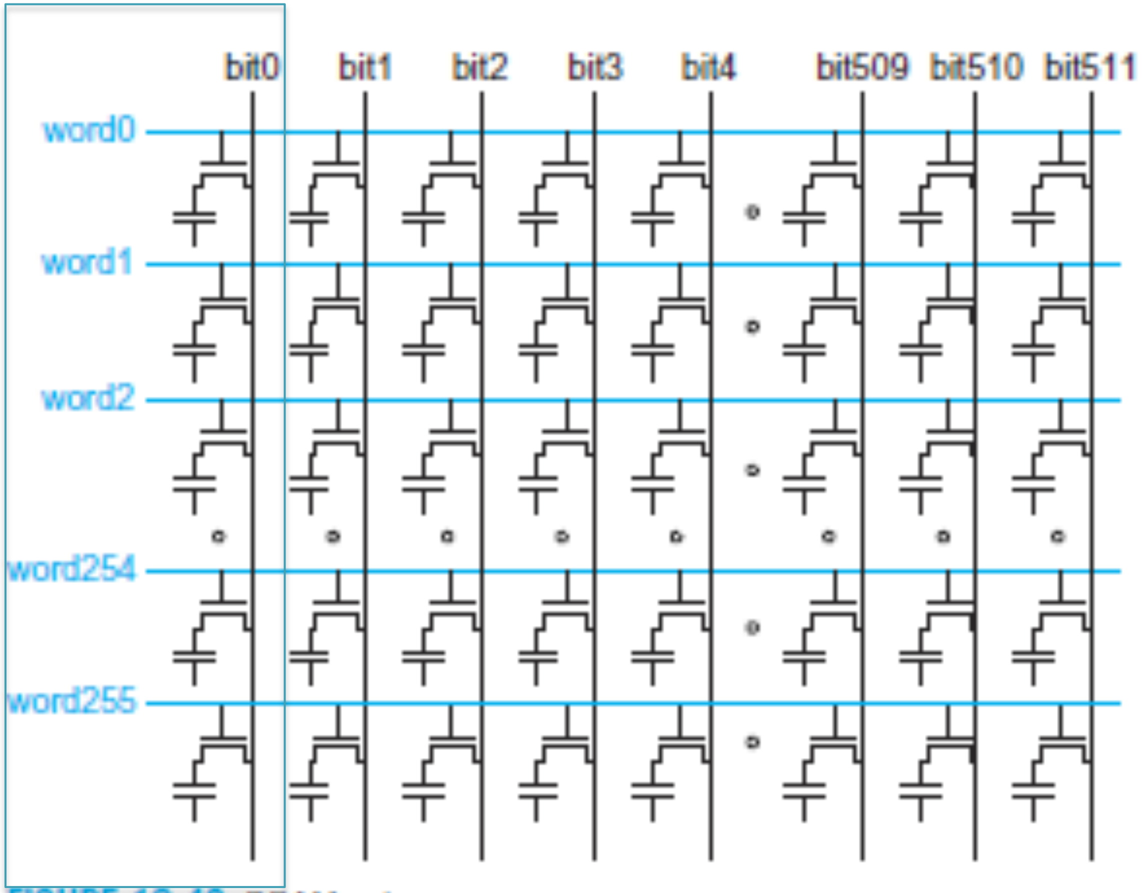
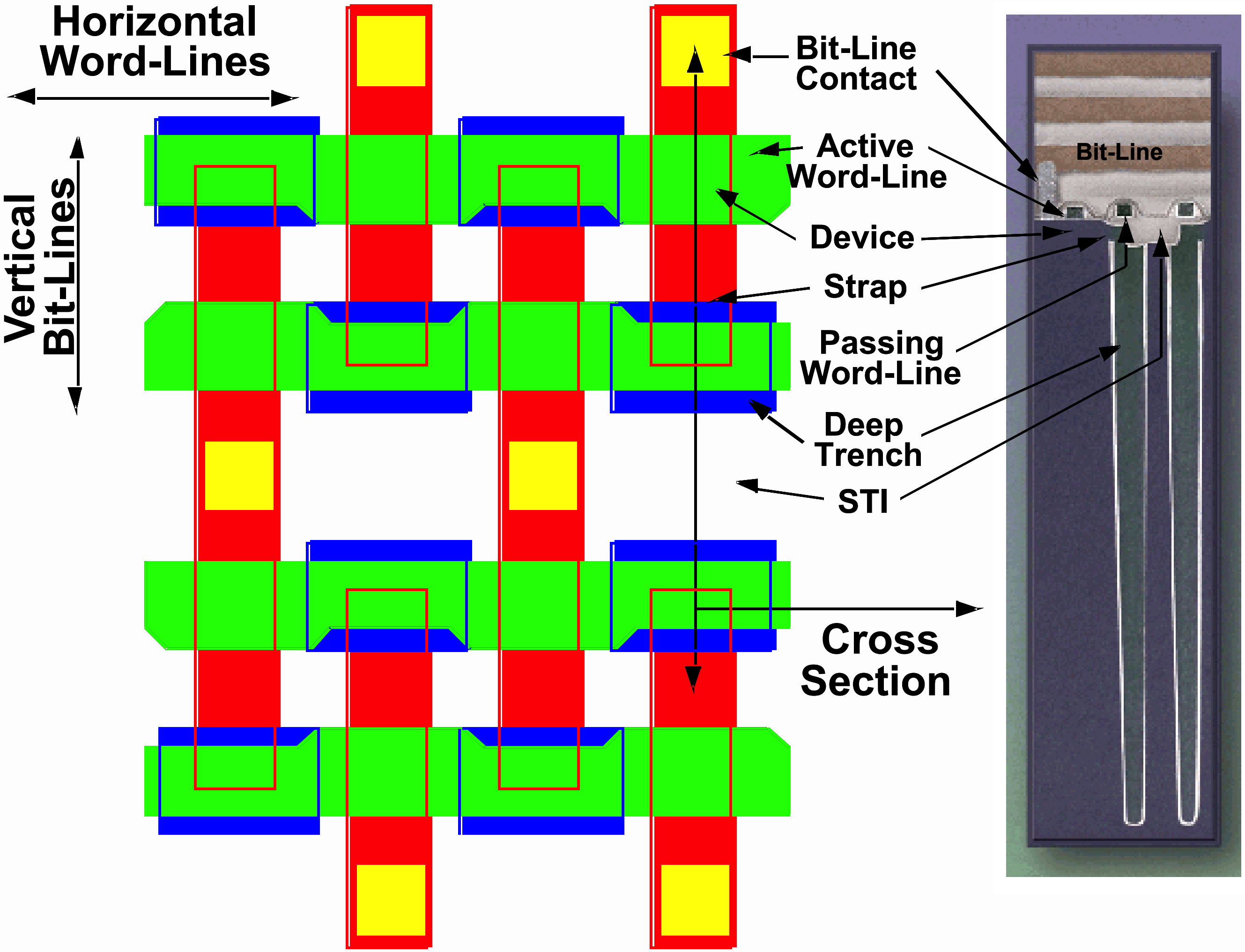
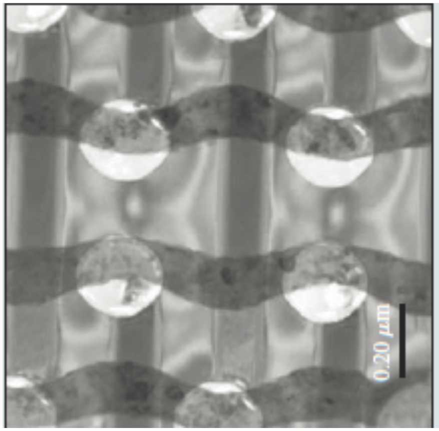


FIGURE 12.43 DRAM subarray

CMOS VLSI design - PEARSON

Trench cell layout and cross-section





Silicon Image

References so far

Barth, J. et al., “A 300MHz Multi-Banked eDRAM Macro Featuring GND Sense, Bit-line Twisting and Direct Reference Cell Write,” ISSCC Dig. Tech. Papers, pp. 156-157, Feb. 2002.

Barth, J. et. al., “A 500MHz Multi-Banked Compilable DRAM Macro with Direct Write and Programmable Pipeline,” ISSCC Dig. Tech. Papers, pp. 204-205, Feb. 2004.

Barth, J. et al., “A 500MHz Random Cycle 1.5ns-Latency, SOI Embedded DRAM Macro Featuring a 3T Micro Sense Amplifier,” ISSCC Dig. Tech. Papers, pp. 486-487, Feb. 2007.

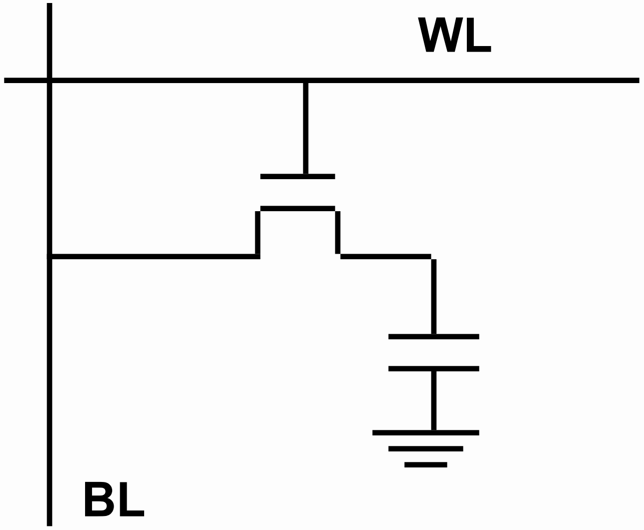
Barth, J. et al., “A 45nm SOI Embedded DRAM Macro for POWER7TM 32MB On-Chip L3 Cache,” ISSCC Dig. Tech. Papers, pp. 342-3, Feb. 2010.

Butt,N., et al., “A 0.039um2 High Performance eDRAM Cell based on 32nm High-K/Metal SOI Technology,” IEDM pp. 27.5.1-2, Dec 2010.

Bright, A. et al., “Creating the BlueGene/L Supercomputer from Low-Power SoC ASICs,” ISSCC Dig. Tech. Papers, pp. 188-189, Feb. 2005.

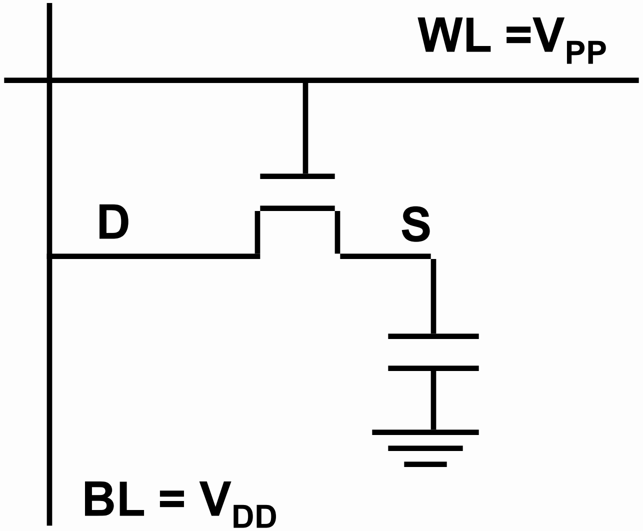
DRAM Operations

- Write
- Read
- Refresh



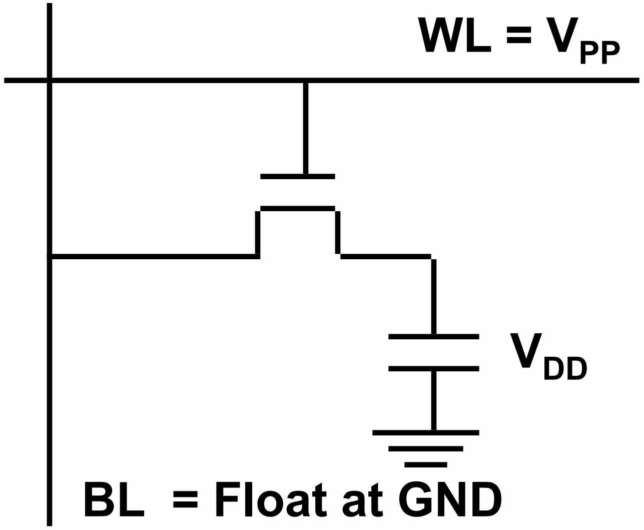
DRAM Read, Write and Refresh

- Write:
  - 1. Drive bit line
  - 2. Select row

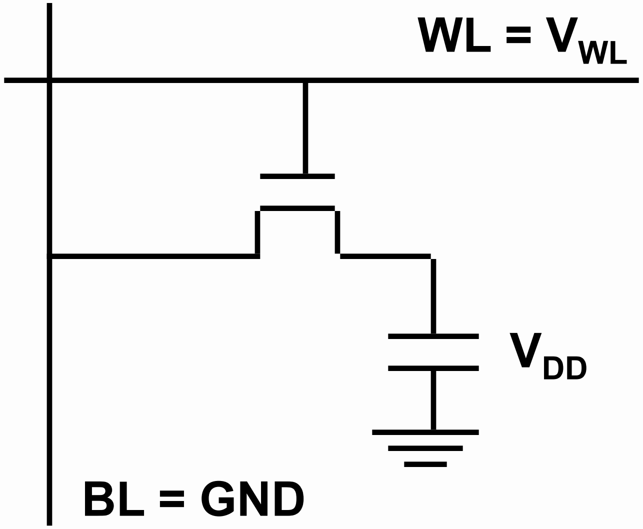


DRAM Read, Write and Refresh

- Read:
  - 1. Pre-charge bit line
  - 2. Select row – Turn ON WL
  - 3. Cell and bit line share charges
    - Signal developed on bitline
  - 4. Sense the data
  - 5. **Write back: restore the value**

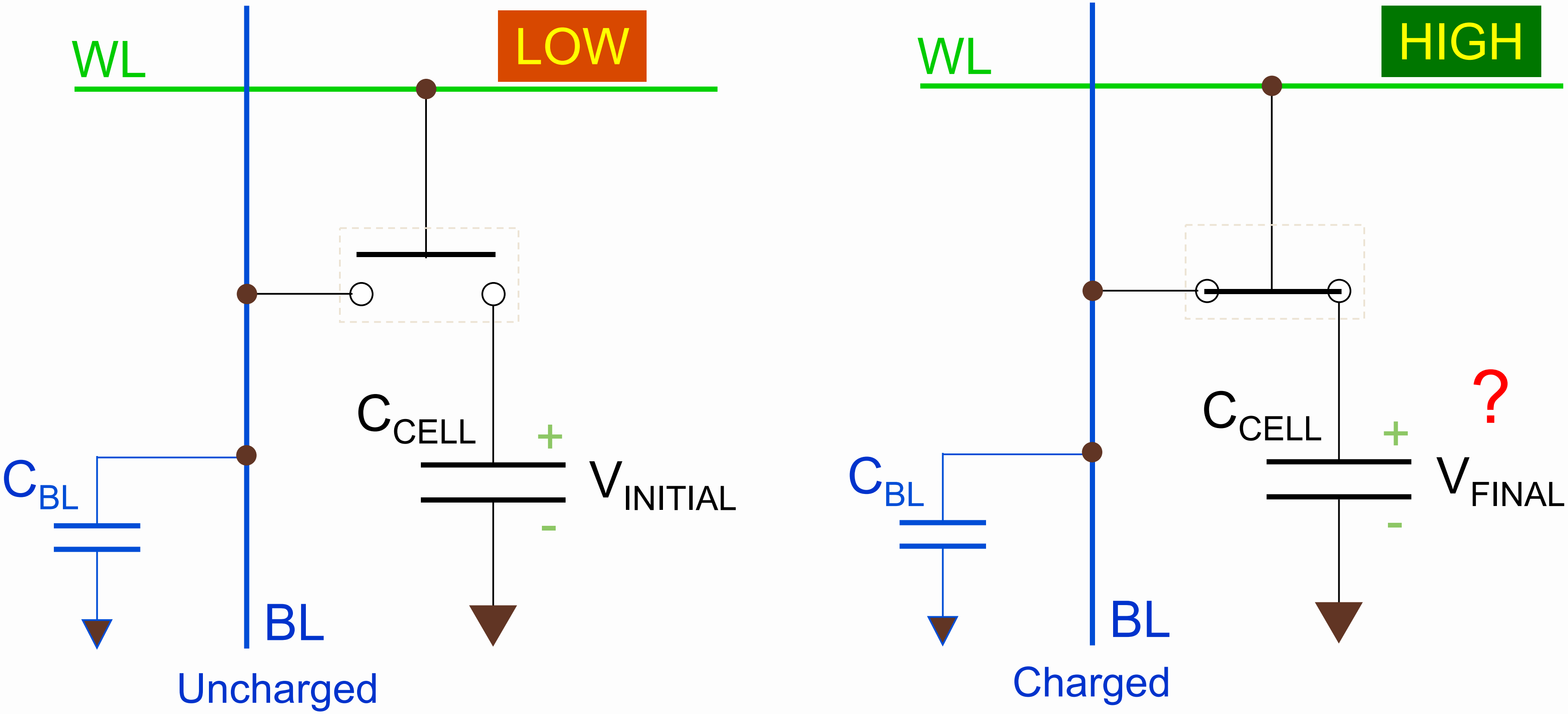


DRAM Read, Write and Refresh



- Refresh
  - 1. Just do a dummy read to every cell → auto write-back

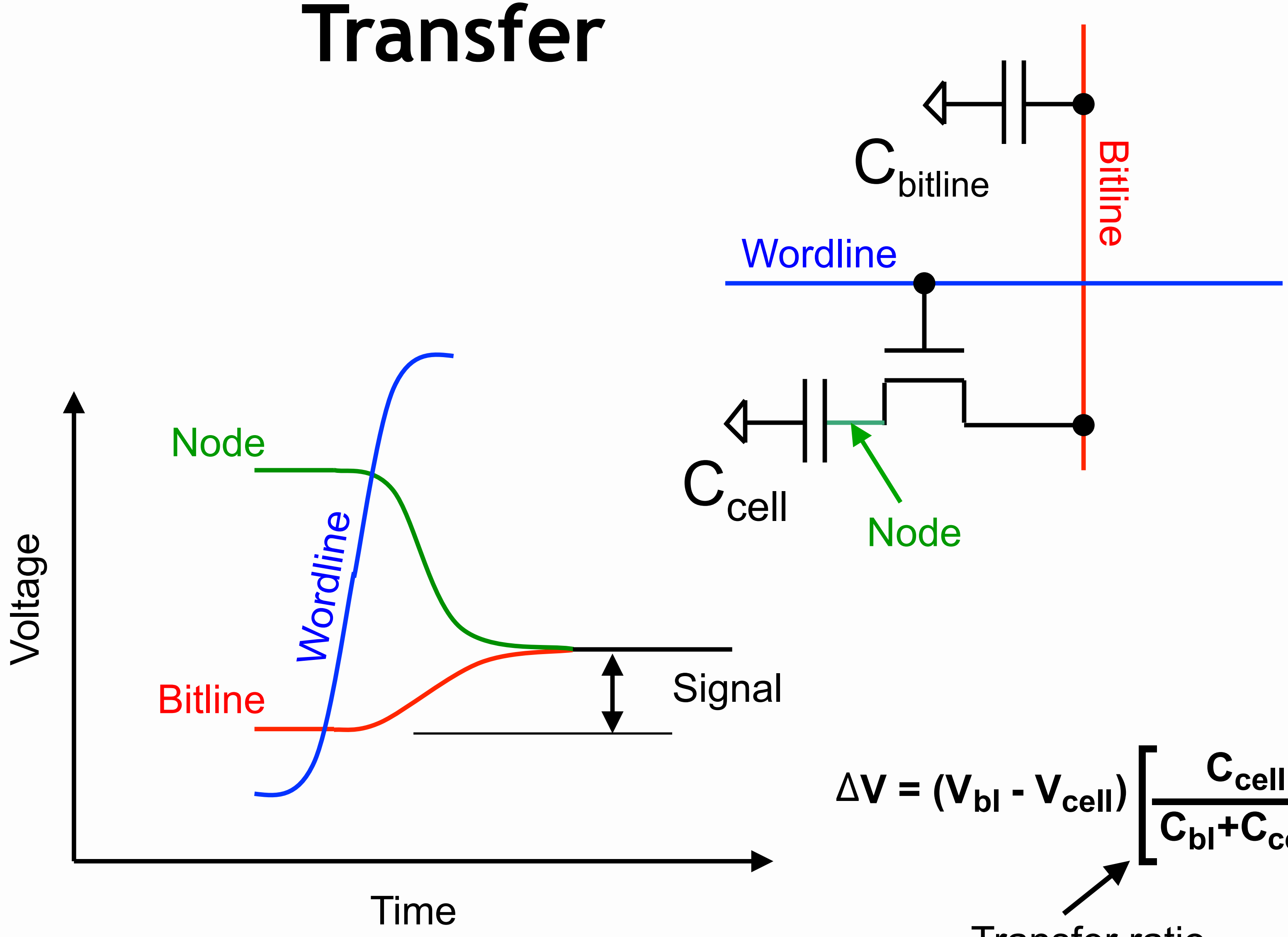
Read - Cell transfer ratio



$$C_{CELL} \times V_{INITIAL} = (C_{CELL} + C_{BL}) \times V_{FINAL}$$

$$\text{Transfer ratio} = C_{CELL} / (C_{CELL} + C_{BL})$$

Cell Charge Transfer

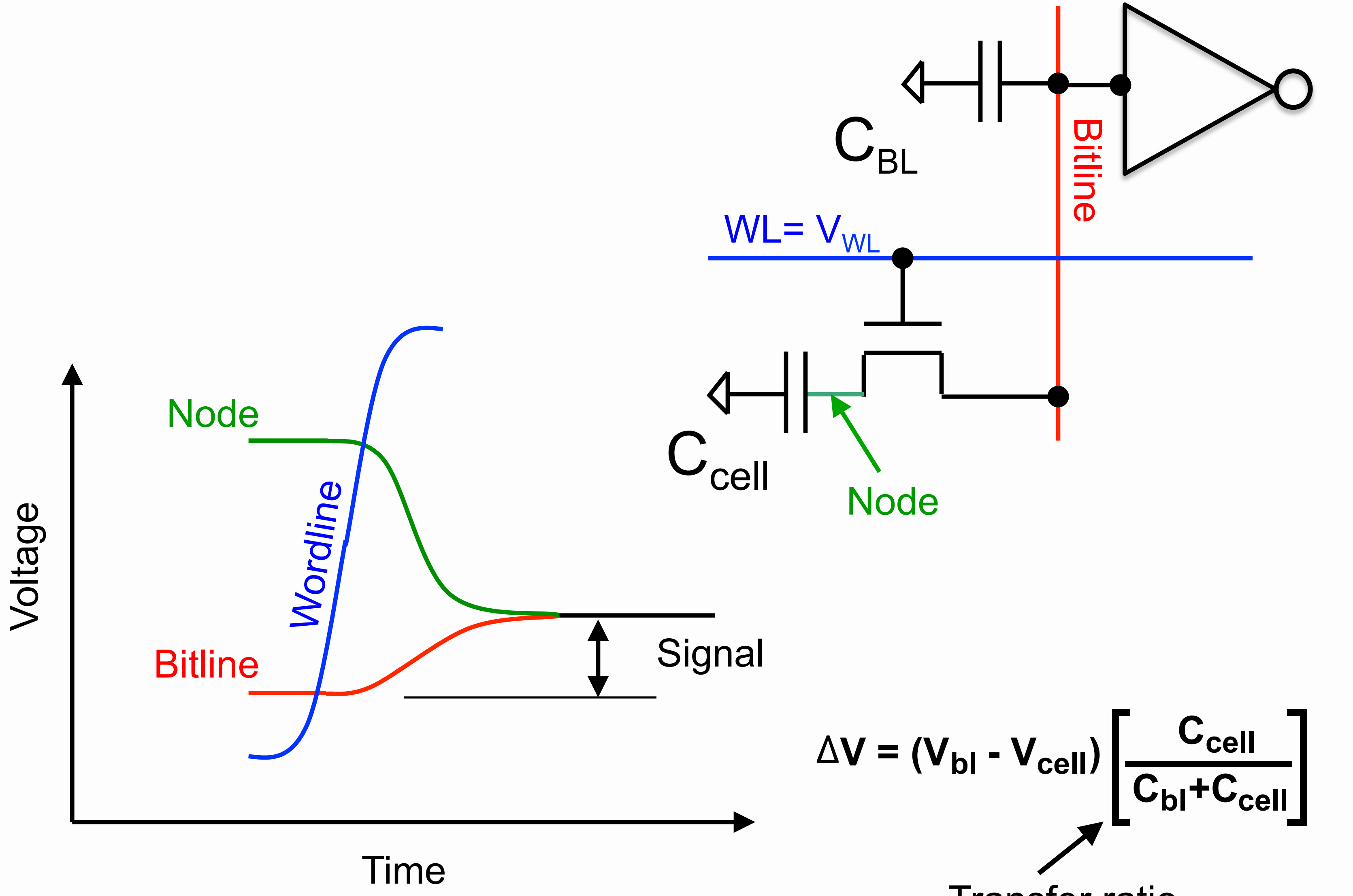


$$\Delta V = (V_{bl} - V_{cell}) \left[\frac{C_{cell}}{C_{bl} + C_{cell}} \right]$$

Transfer ratio

Sensing

Signal ΔV > V_M (Trip Point)

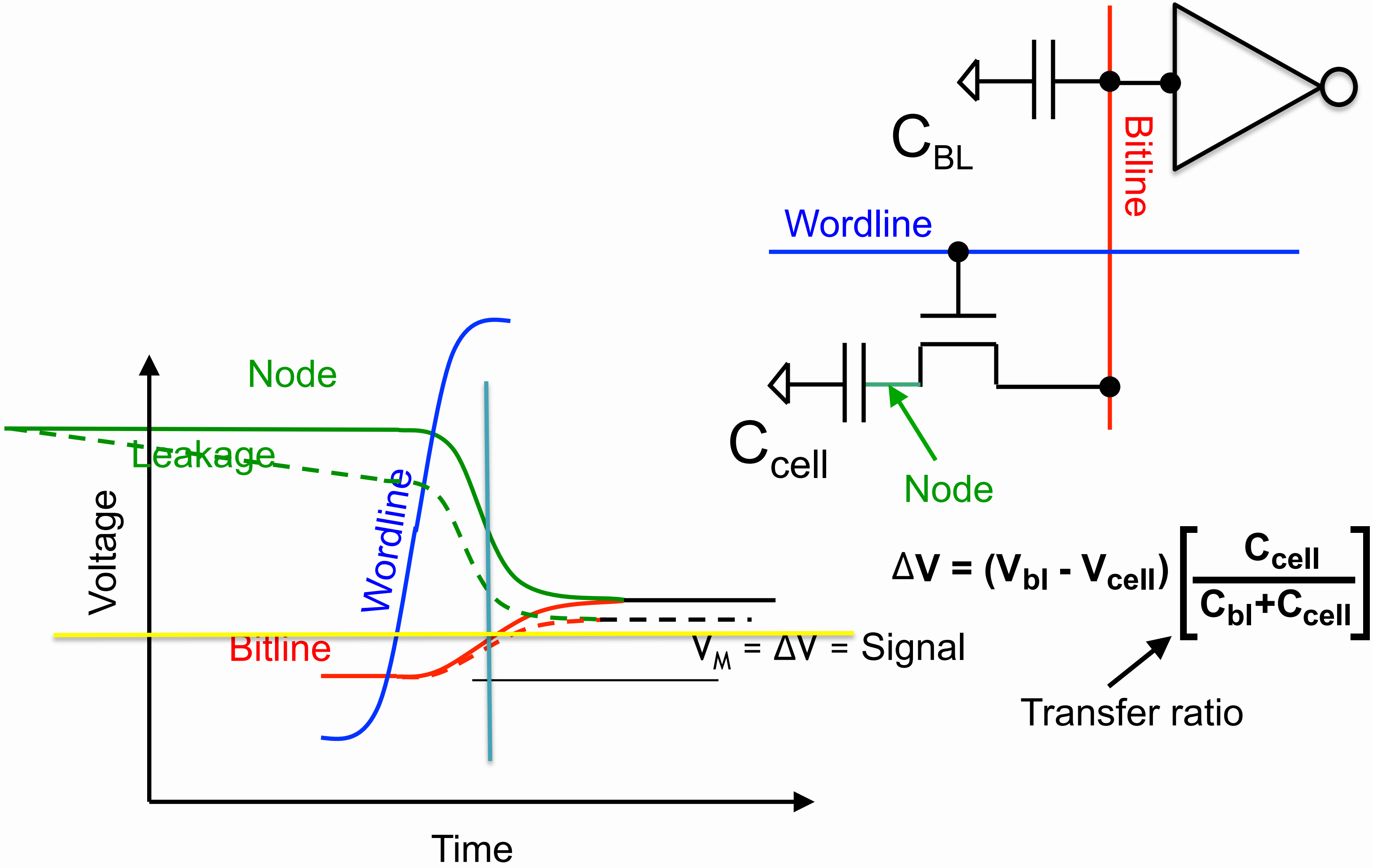


$$\Delta V = (V_{bl} - V_{cell})\left[\frac{C_{cell}}{C_{bl}+C_{cell}}\right]$$

Transfer ratio

Retention

Signal ΔV > V_M (Trip Point)



$$\Delta \mathbf{V} = (\mathbf{V}_{bl} - \mathbf{V}_{cell}) \left[\frac{\mathbf{C}_{cell}}{\mathbf{C}_{bl} + \mathbf{C}_{cell}} \right]$$

Transfer Ratio and Signal

ΔBit-Line Voltage Calculated from Initial Conditions and Capacitances:

$$\Delta V = V_{bl} - V_f = V_{bl} - \frac{Q}{C} = V_{bl} - \left[\frac{C_{bl}*V_{bl}+C_{cell}*V_{cell}}{C_{bl}+C_{cell}}\right]$$

$$\Delta V = (V_{bl} - V_{cell})\left[\frac{C_{cell}}{C_{bl}+C_{cell}}\right]$$

Transfer Ratio (typically 0.2)

ΔBit-Line Voltage is Amplified with Cross Couple “Sense Amp”

Sense Amp Compares Bit-Line Voltage with a Reference

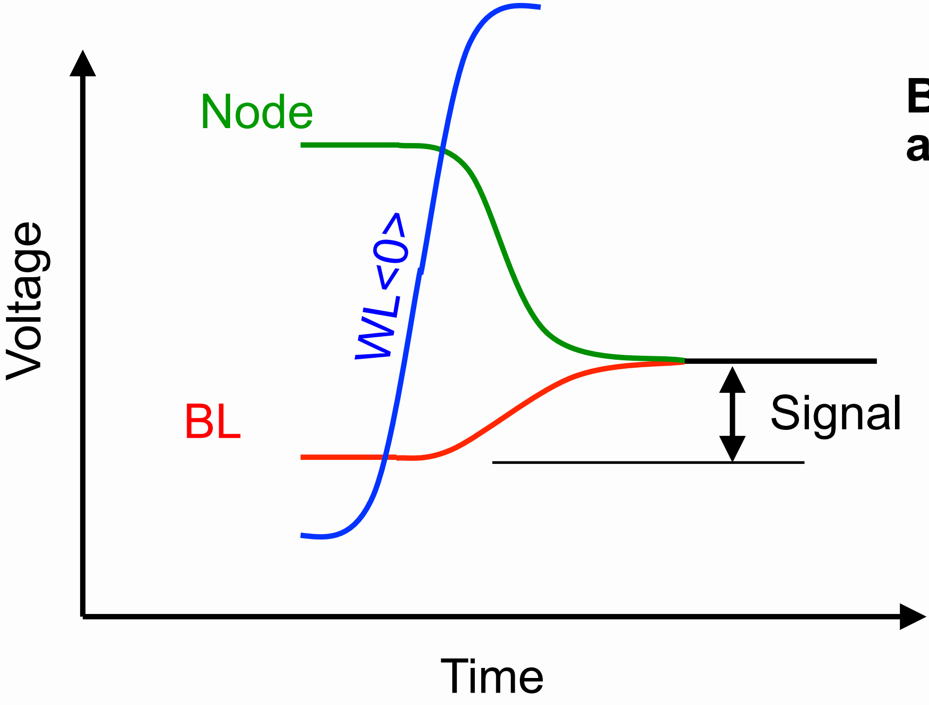
Bit-Line Voltage - Reference = Signal

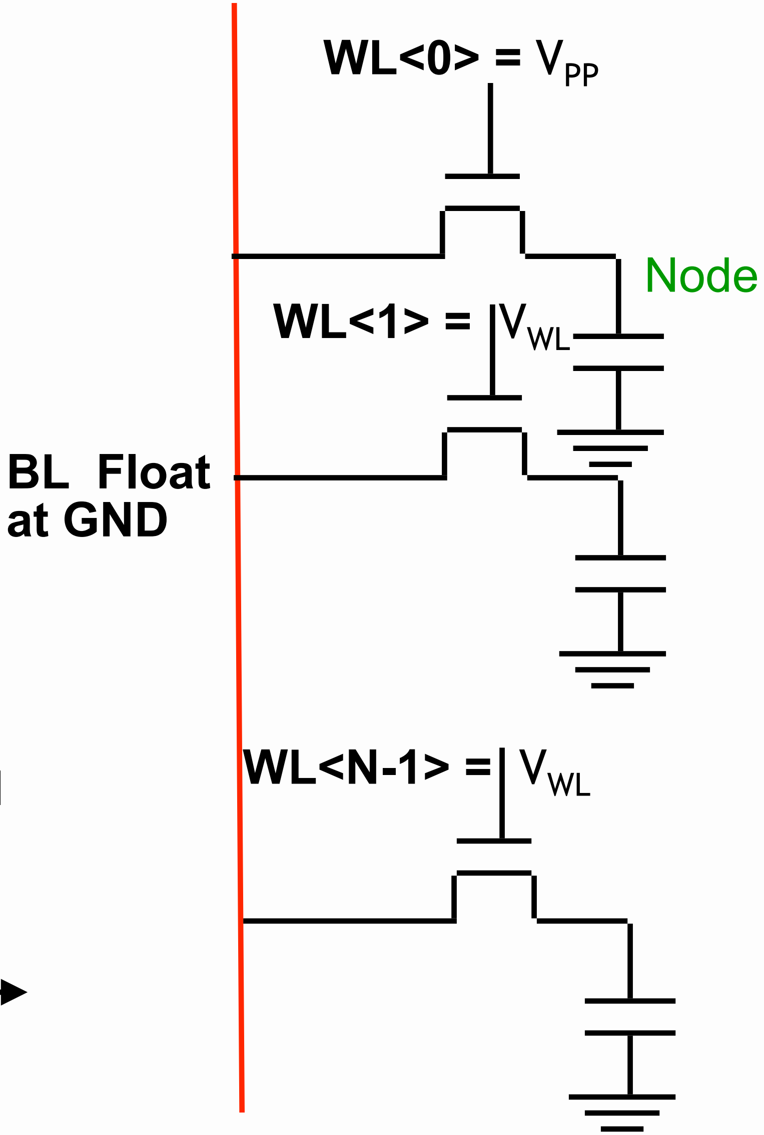
Pos Signal Amplifies to Logical ‘1’, Neg Signal Amplifies to Logical ‘0’

Signal: # WLs on a BL

$C_{BL} = N\,C_{DIFF} + NC_{M1}$

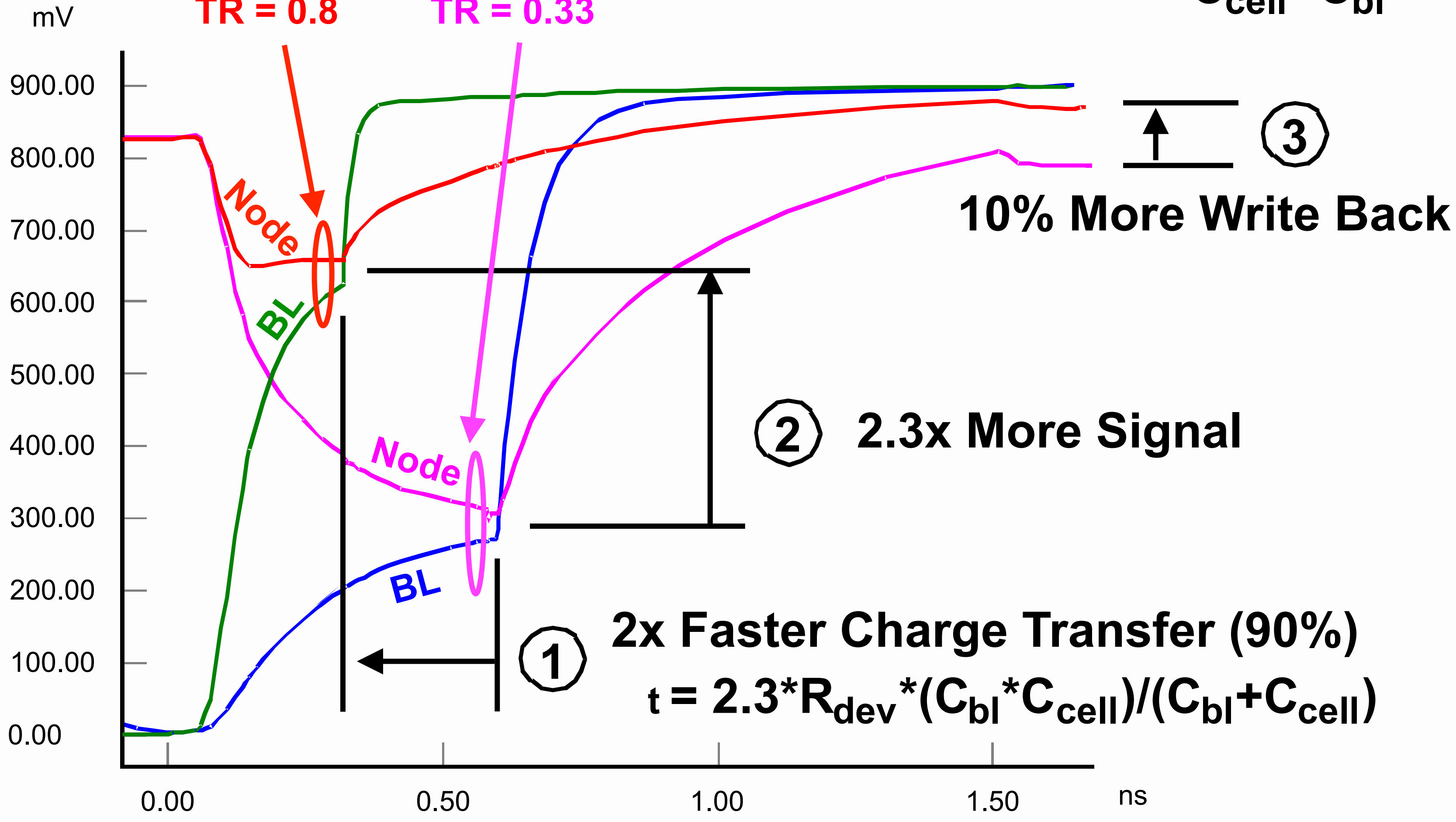
$TR = C_{Cell} / (C_{BL} + C_{Cell})$





Bits per Bit-Line v/s Transfer Ratio

TR = Transfer Ratio = $\frac{C_{cell}}{C_{cell}+C_{bl}}$



32 Bits/BL
TR = 0.8

128 Bits/BL
TR = 0.33

3 10% More Write Back

2 2.3x More Signal

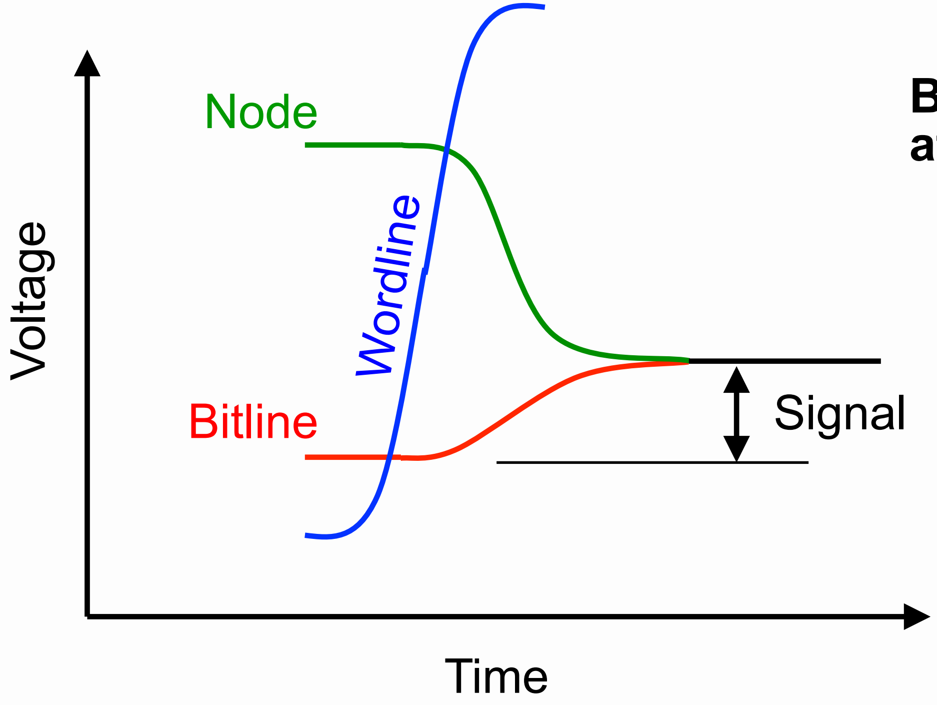
1 2x Faster Charge Transfer (90%)

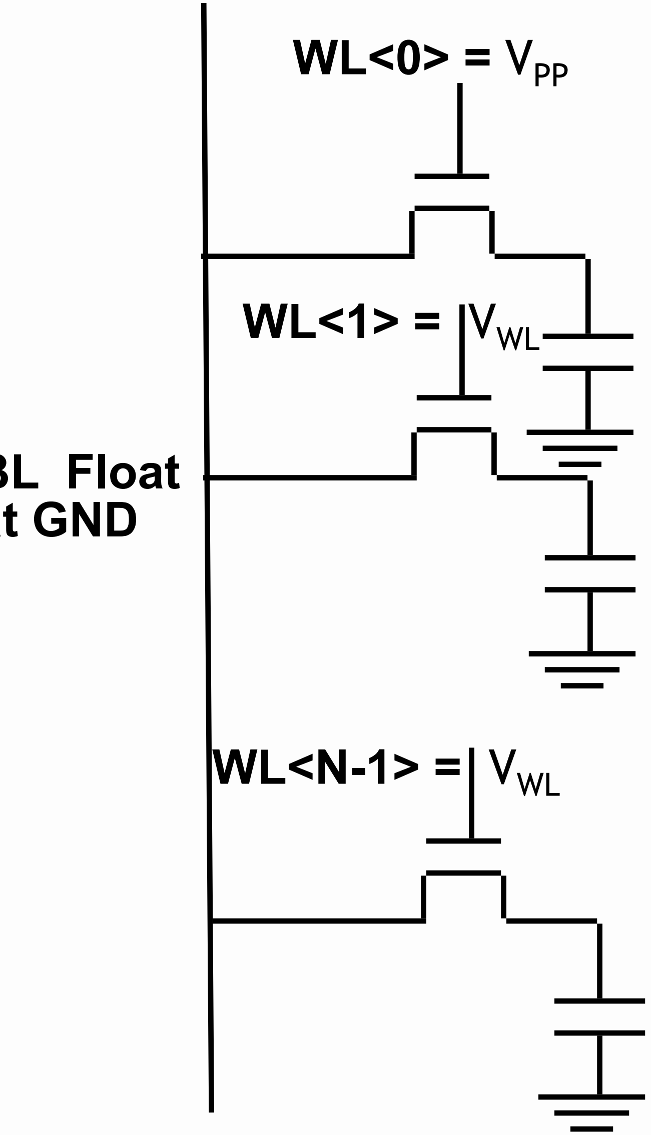
$t = 2.3*R_{dev}*(C_{bl}*C_{cell})/(C_{bl}+C_{cell})$

JSSC08

Signal: # WLs on a BL

Short BLs are Mandatory in DRAMs





Segmentation

Array Segmentation Refers to WL / BL Count per Sub-Array

Longer Word-Line is Slower but more Area efficient (Less Decode/Drivers)

Longer Bit-Line (more Word-Lines per Bit-Line)

Less Signal (Higher Bit-Line Capacitance = Lower Transfer Ratio)
More Power (Bit-Line CV is Significant Component of DRAM Power)
Slower Performance (Higher Bit-Line Capacitance = Slower Sense Amp)
More Area Efficient (Fewer Sense Amps)

Number of Word-Lines Activated determines Refresh Interval and Power

All Cells on Active Word-Line are Refreshed
All Word-Lines must be Refreshed before Cell Retention Expires
64ms Cell Retention / 8K Word Lines = 7.8us between refresh cycles
Activating 2 Word-Lines at a time = 15.6us, 2x Bit-Line CV Power

Choice of SA

Depending on signal developed SA architecture is chosen

Direct sensing

- Requires large signal development
- An inverter can be used for sensing
- Micro sense amp (uSA) is another option

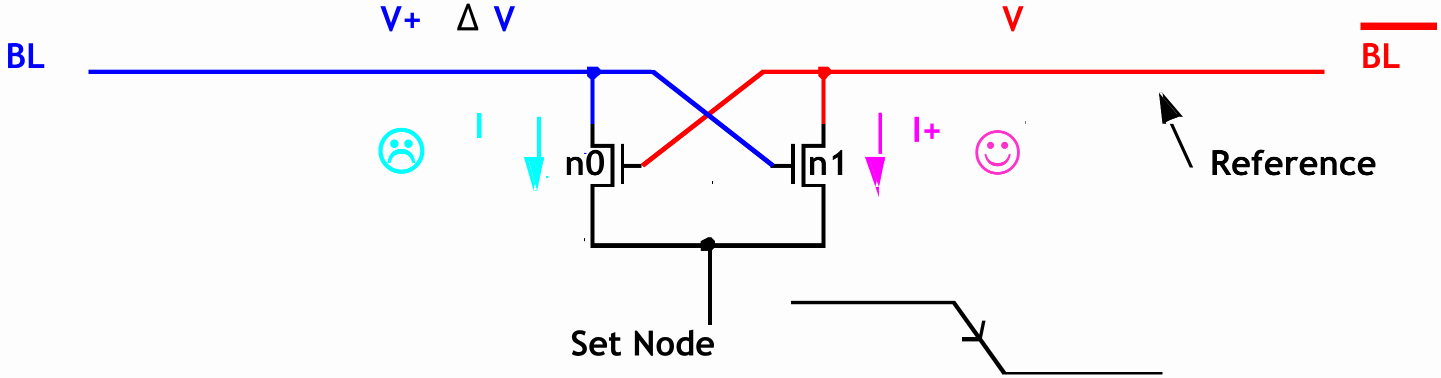
Differential sense amp

- Can sense low signal developed

This is choice between area, speed/performance

Sensing → Signal Amplification

Differential Voltage Amplified by Cross Couple Pair



When Set Node < (V+ΔV) - V_{tn1}, I+ will start to flow (On-Side Conduction)

When Set Node < (V) - V_{tn0}, I will start to flow (Off-Side Conduction)

Off-Side Conduction Modulated by Set Speed and Amount of Signal

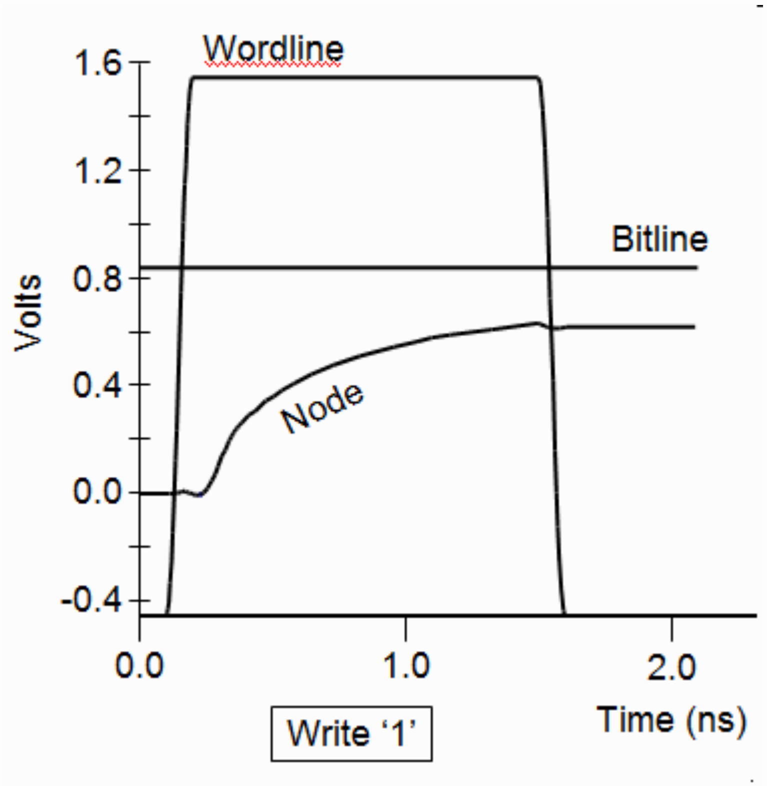
Complimentary X-Couple Pairs provide Full CMOS Levels on Bit-Line

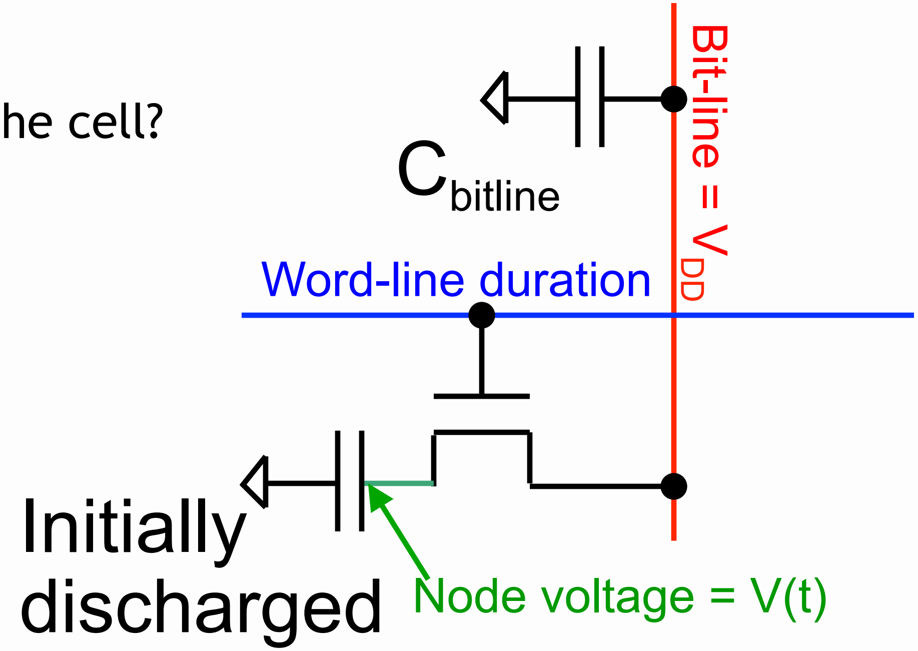
Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM Write Analysis
- ❑ eDRAM Sense-Amplifier Specification
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

Write-Margin

How much signal can we write into the cell?





$C_{bitline}$

Bit-line = V_{DD}

Word-line duration

Initially discharged

Node voltage = V(t)

Write-Margin

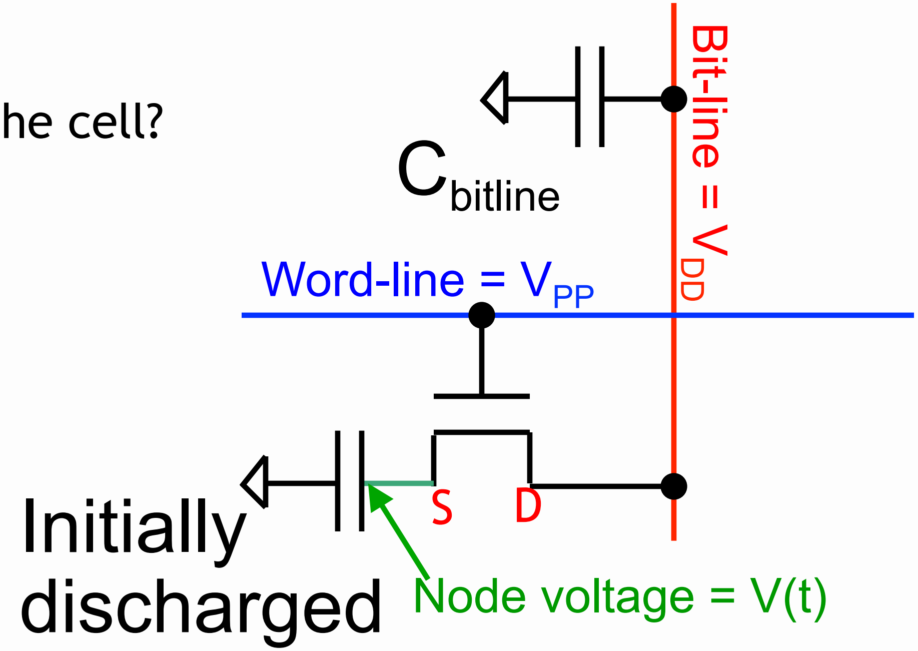
How much signal can we write into the cell?

Assume:

- Bit-line is charged to V_{DD}
- Word-line rises instantly to V_{PP}

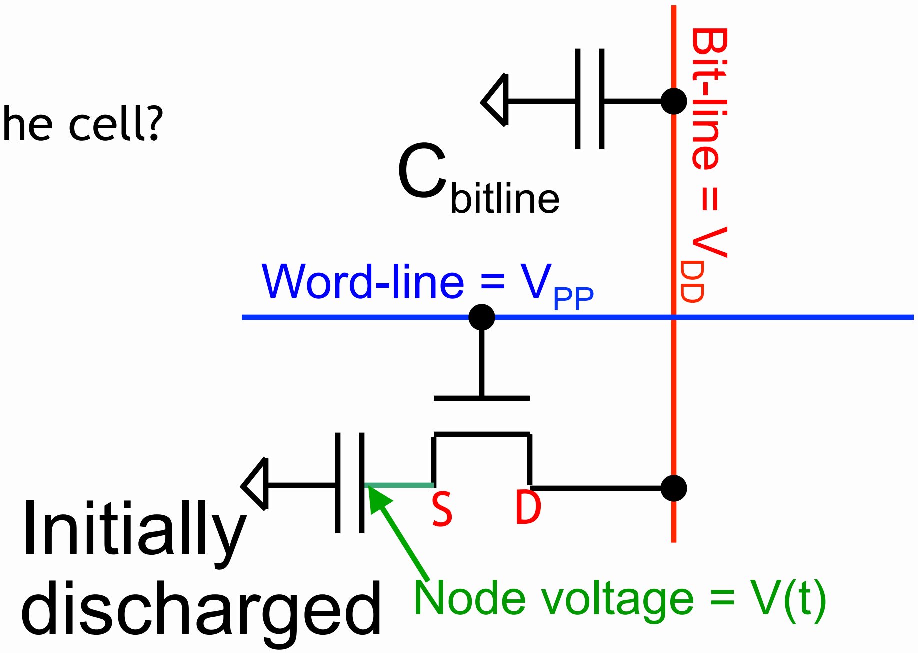
$V_{GS} = V_{PP} – V(t)$

$V_{DS} = V_{DD} – V(t)$



Write-Margin

How much signal can we write into the cell?



Assume:

- Bit-line is charged to V_{DD}
- Word-line rises instantly to V_{PP}

$V_{GS} = V_{PP} – V(t)$

$V_{DS} = V_{DD} – V(t)$

Access transistor is a thick oxide device => $V_{DSAT} >> V_{PP}$

$V_{DS} = V_{DD} – V(t) < V_{GS} - V_{Tn} = V_{PP} – V(t) - V_{Tn}$

$V_{PP} \geq V_{DD} + V_{Tn} + \Delta V_{Tn}$

Access device is in the linear region throughout

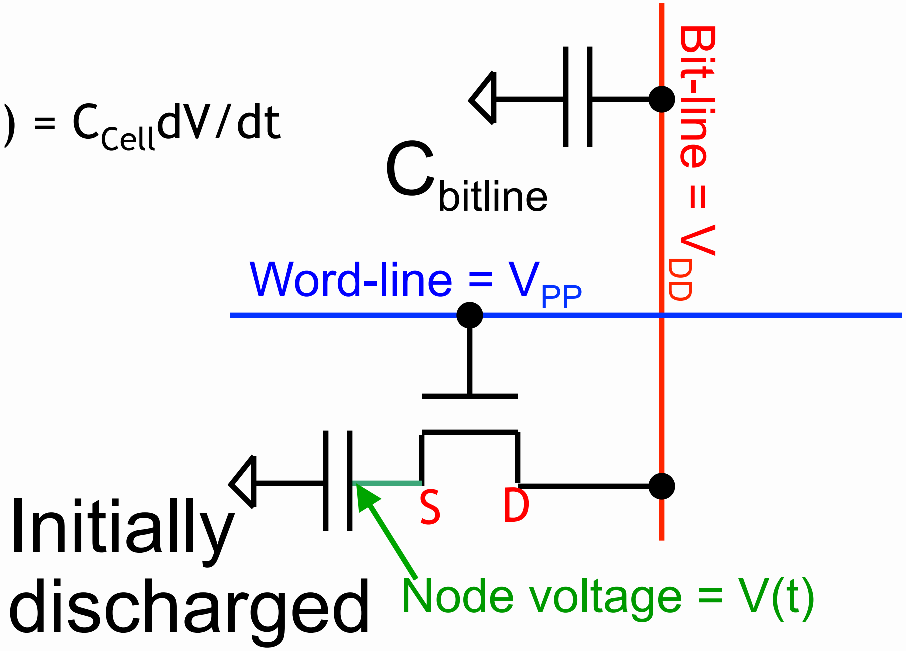
Write-Margin

$I_{DS} = \mu_n C_{OX}\,(W/L) V_{DS}\,(V_{GS} - V_{Tn} - V_{DS}/2) = C_{Cell}\,dV/dt$

$V_{GS} = V_{PP} - V(t)$

$V_{DS} = V_{DD} - V(t)$

Assume you wish to write a voltage V_f into the cell in a time T_{WRITE}



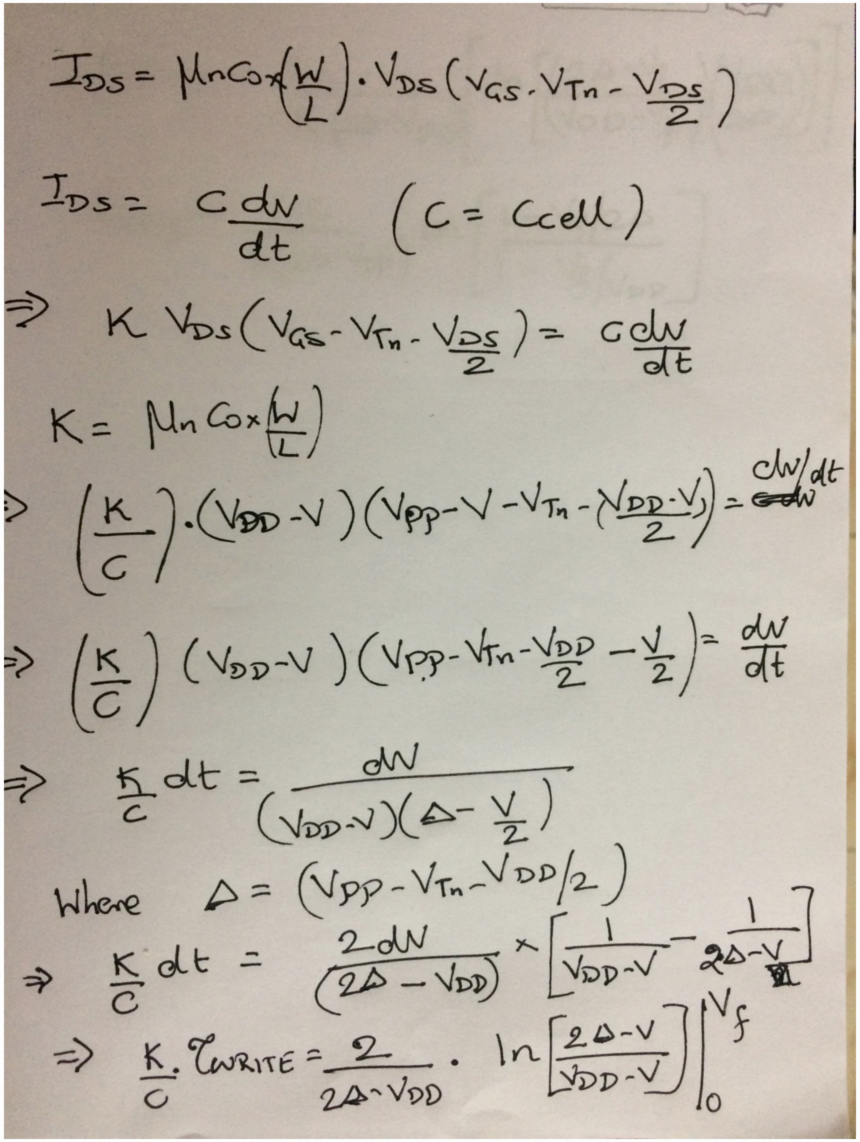
Solving you will get:

$T_{WRITE} = 2\,T_0 \ln[(1-V_f/2\Delta)/(1-V_f/V_{DD})]$

Where

$T_0 = \{C_{cell}/[\mu_n C_{OX}\,(W/L)(2\Delta - V_{DD})]\}$

$\Delta = V_{PP} - V_{Tn} - V_{DD}/2$



$$I_{DS} = \mu_n C_{ox}\left(\frac{W}{L}\right) \cdot V_{DS}\left(V_{GS} - V_{Tn} - \frac{V_{DS}}{2}\right)$$

$$I_{DS} = C\frac{dV}{dt} \qquad (C = C_{cell})$$

$$\Rightarrow K\,V_{DS}\left(V_{GS} - V_{Tn} - \frac{V_{DS}}{2}\right) = C\frac{dV}{dt}$$

$$K = \mu_n C_{ox}\left(\frac{W}{L}\right)$$

$$\Rightarrow \left(\frac{K}{C}\right) \cdot (V_{DD} - V)\left(V_{PP} - V - V_{Tn} - \left(\frac{V_{DD} - V}{2}\right)\right) = \overset{dV/dt}{\cancel{cdV}}$$

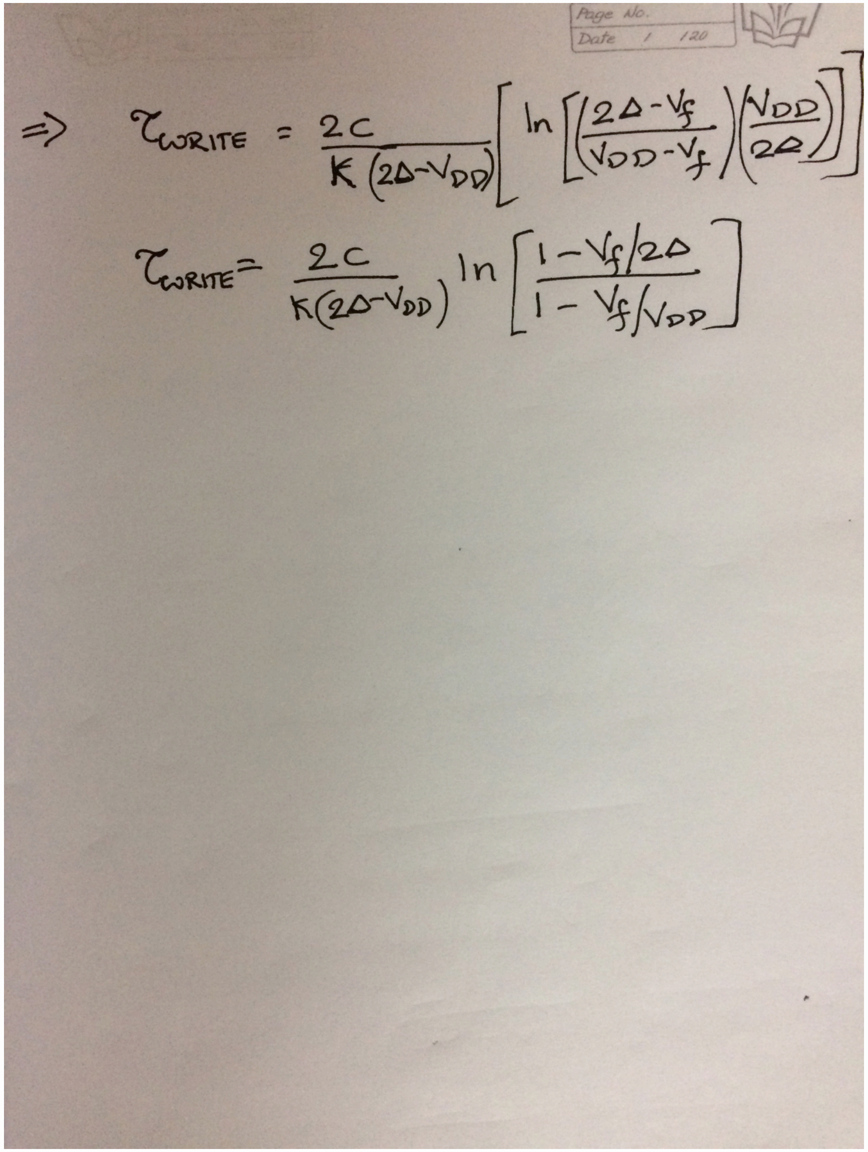
$$\Rightarrow \left(\frac{K}{C}\right)(V_{DD} - V)\left(V_{PP} - V_{Tn} - \frac{V_{DD}}{2} - \frac{V}{2}\right) = \frac{dV}{dt}$$

$$\Rightarrow \frac{K}{C}dt = \frac{dV}{(V_{DD} - V)\left(\Delta - \frac{V}{2}\right)}$$

Where $\Delta = (V_{PP} - V_{Tn} - V_{DD}/2)$

$$\Rightarrow \frac{K}{C}dt = \frac{2\,dV}{(2\Delta - V_{DD})} \times \left[\frac{1}{V_{DD} - V} - \frac{1}{2\Delta - V}\right]$$

$$\Rightarrow \frac{K}{C} \cdot \tau_{WRITE} = \frac{2}{2\Delta - V_{DD}} \cdot \ln\left[\frac{2\Delta - V}{V_{DD} - V}\right]\Bigg|_0^{V_f}$$



Page No.
Date / /20

$$\Rightarrow \tau_{WRITE} = \frac{2C}{K(2\Delta - V_{DD})}\left[\ln\left[\left(\frac{2\Delta - V_f}{V_{DD} - V_f}\right)\left(\frac{V_{DD}}{2\Delta}\right)\right]\right]$$

$$\tau_{WRITE} = \frac{2C}{K(2\Delta - V_{DD})}\ln\left[\frac{1 - V_f/2\Delta}{1 - V_f/V_{DD}}\right]$$

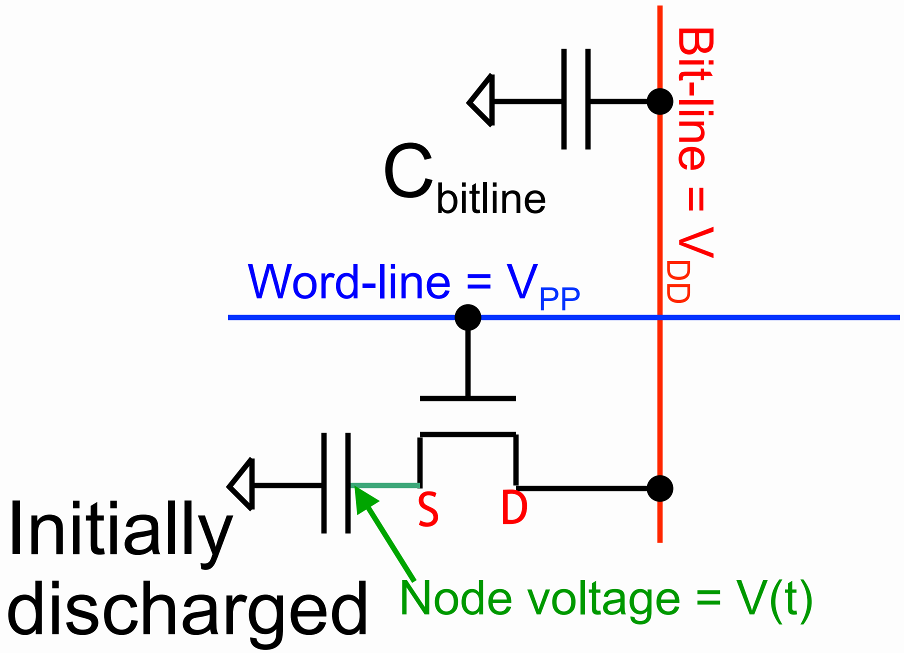
Write-Margin – Analysis

$T_{WRITE} = 2\ T_0 \ln[(1-V_f/2\Delta)/(1-V_f/V_{DD})]$

Where

$T_0 = \{C_{cell}/[\mu_n C_{OX}\ (W/L)(2\Delta - V_{DD})]\}$

$\Delta = V_{PP} - V_{Tn} - V_{DS}/2$



1. What happens when VPP is increased?
2. What happens when V_f → V_{DD}

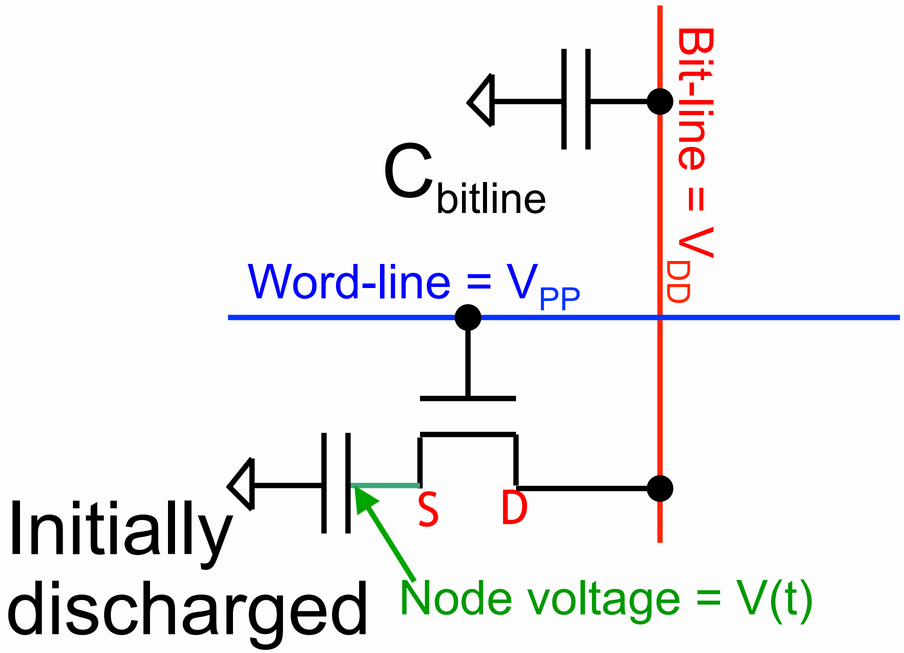
Write-Margin – Analysis

$T_{WRITE} = 2\ T_0 \ln[(1-V_f/2\Delta)/(1-V_f/V_{DD})]$

Where

$T_0 = \{C_{cell}/[\mu_n C_{OX}\ (W/L)(2\Delta - V_{DD})]\}$

$\Delta = V_{PP} – V_{Tn} – V_{DD}/2$



1. What happens when VPP is increased?
   - Δ increases and T_0 decreases
   - $\ln[(1-V_f/2\Delta)/(1-V_f/V_{DD})]$ increases logarithmically
   - T_{WRITE} effectively decreases.
   - However WL power goes up as $(V_{PP} - V_{WL})^2$

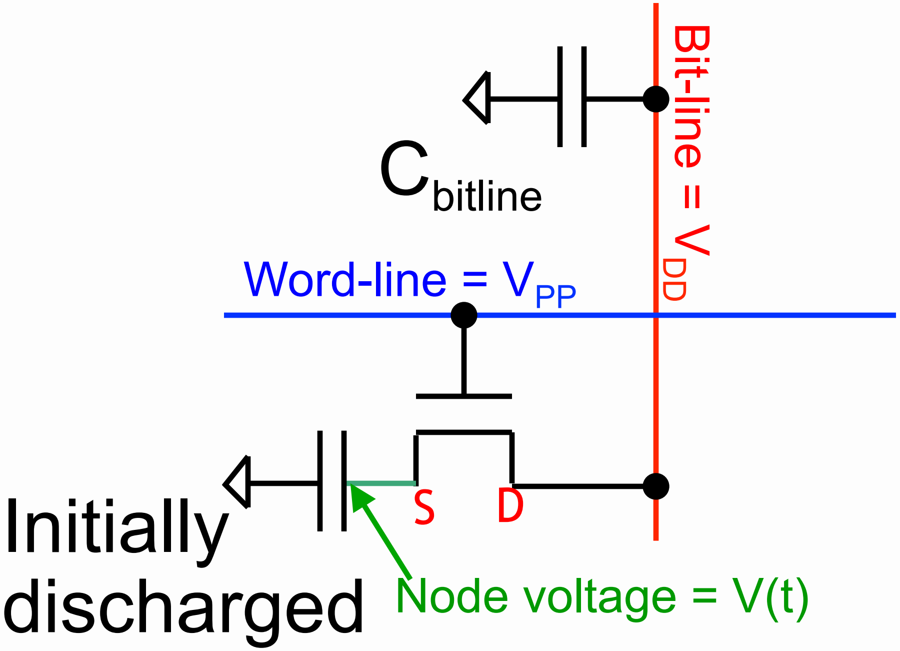
Write-Margin – Analysis

$T_{WRITE} = 2\ T_0 \ln[(1\text{-}V_f/2\Delta)/(1\text{-}V_f/V_{DD})]$

Where

$T_0 = \{C_{cell}/[\mu_n C_{OX}\ (W/L)(2\Delta - V_{DD})]\}$

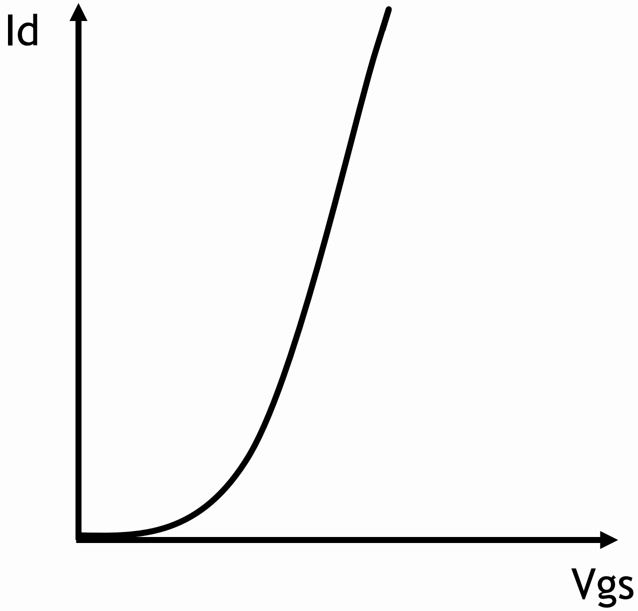
$\Delta = V_{PP} - V_{Tn} - V_{DS}/2$

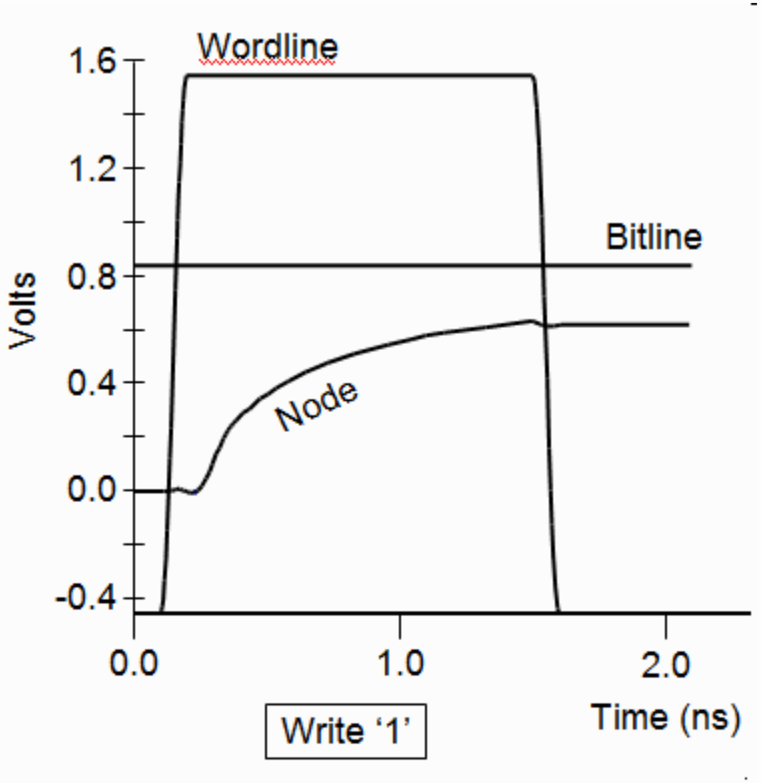


1. What happens when V_f ➔ V_{DD}
   - $(1\text{-}V_f/V_{DD})$ approaches 0
   - $\ln[(1\text{-}V_f/2\Delta)/(1\text{-}V_f/V_{DD})]$ shoots to ∞

You cannot write a FULL V_{DD} into the cell in finite time

Storing data ‘1’ in the cell

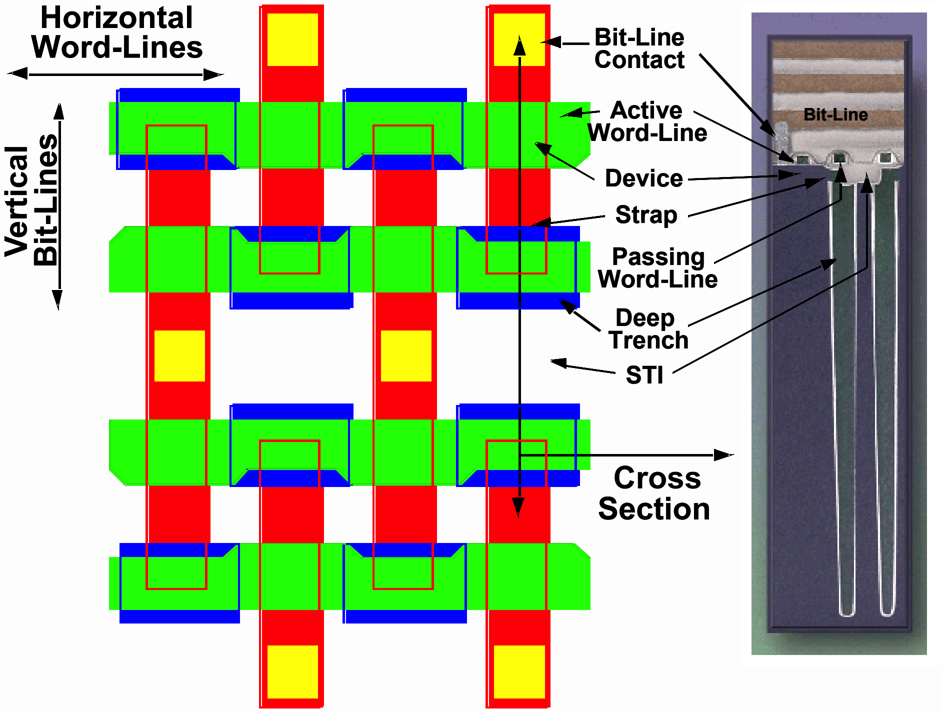


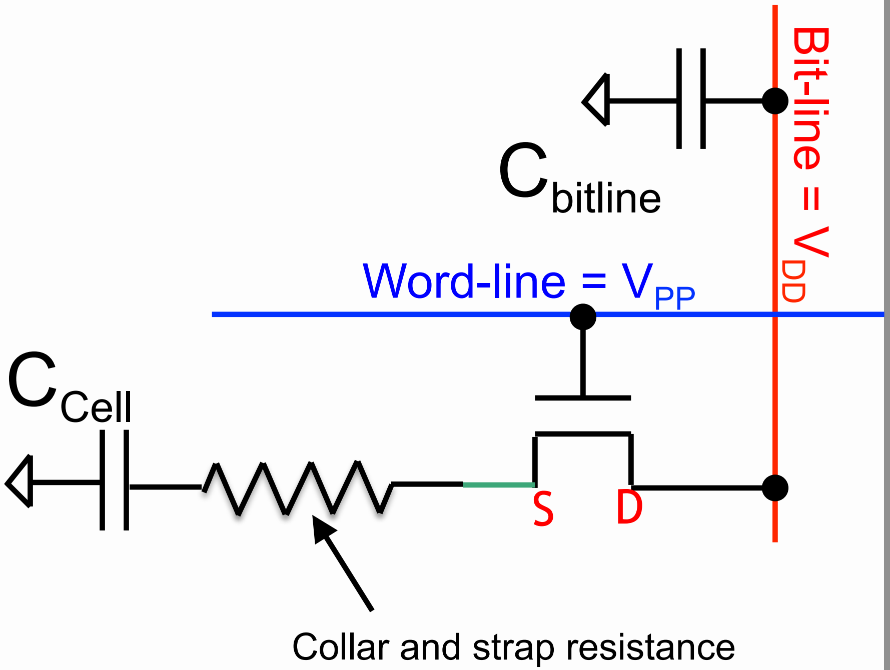


Vgs for pass transistor reduces as bitcell voltage rises, increasing Ron

Why there is a reduction in cell voltage after WL closes? Experiment

Reality





The node takes longer to charge than what we calculated analytically

Conclusions do not change:

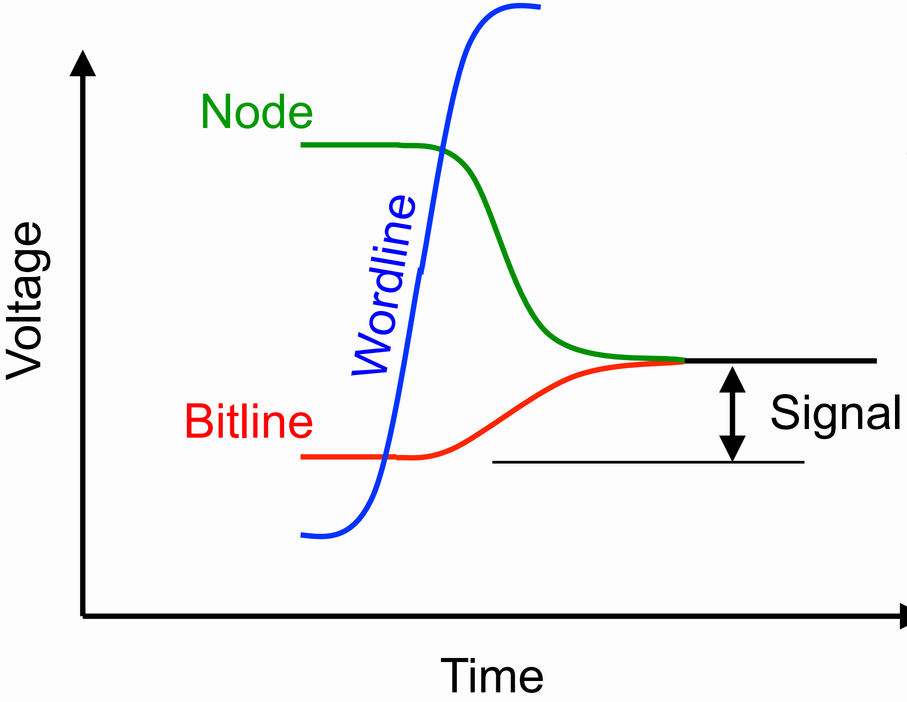
- Increasing V_{PP} decreases the write time
- Cannot write a full V_{DD} in finite time

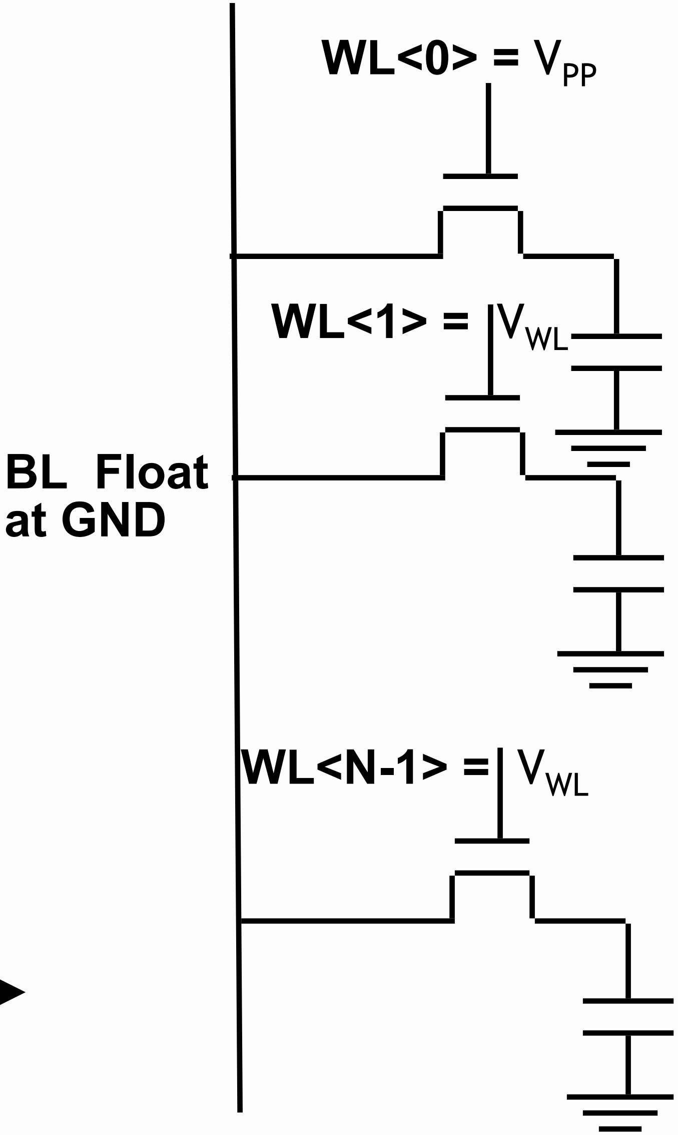
Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM Write Analysis
- ❑ eDRAM Sense-Amplifier Specification
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

Signal: # WLs on a BL

Short BLs are Mandatory in DRAMs



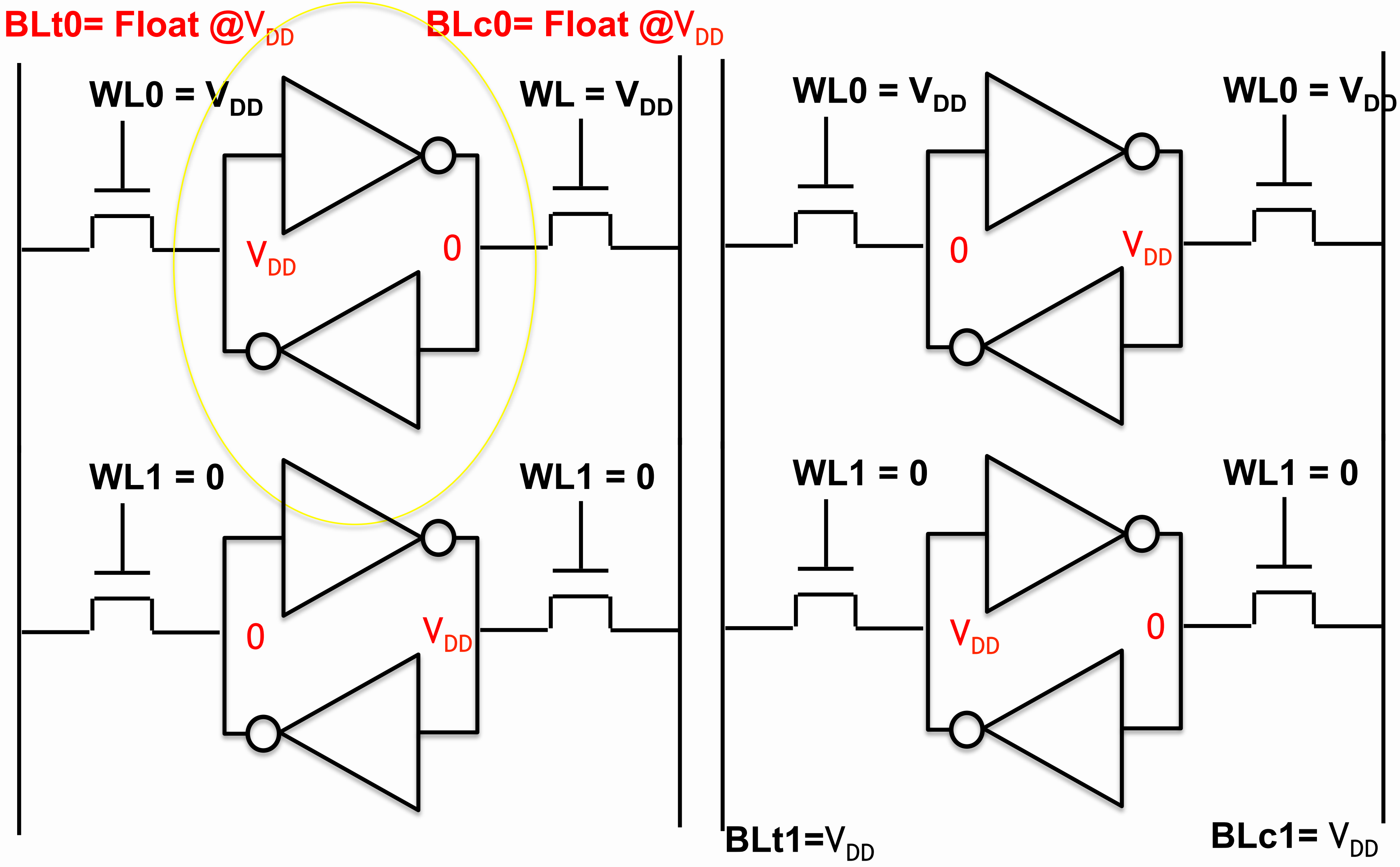


SRAM vs DRAM

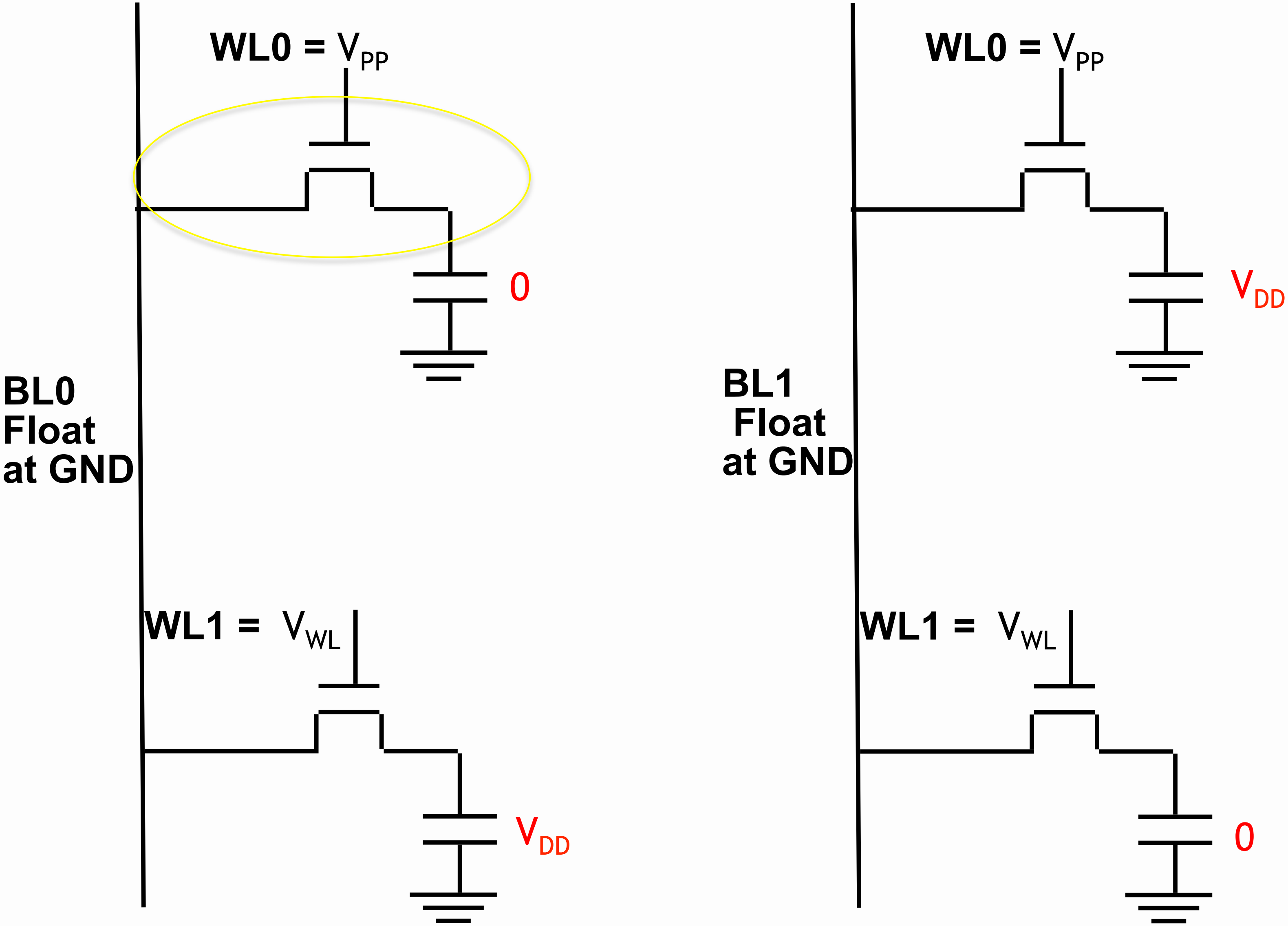
Assume a 1Mb memory with 512 WLs and 2048 BLs and 8 Columns. i.e. 256 DLs

	SRAM	DRAM
#WLs per BL	512	32
# BLs per DL	8	8
Sharing Sense Amps across columns	Yes	Possible?
Effective number of cells connected to a SA	4096	?
# Sense Amps	256	?

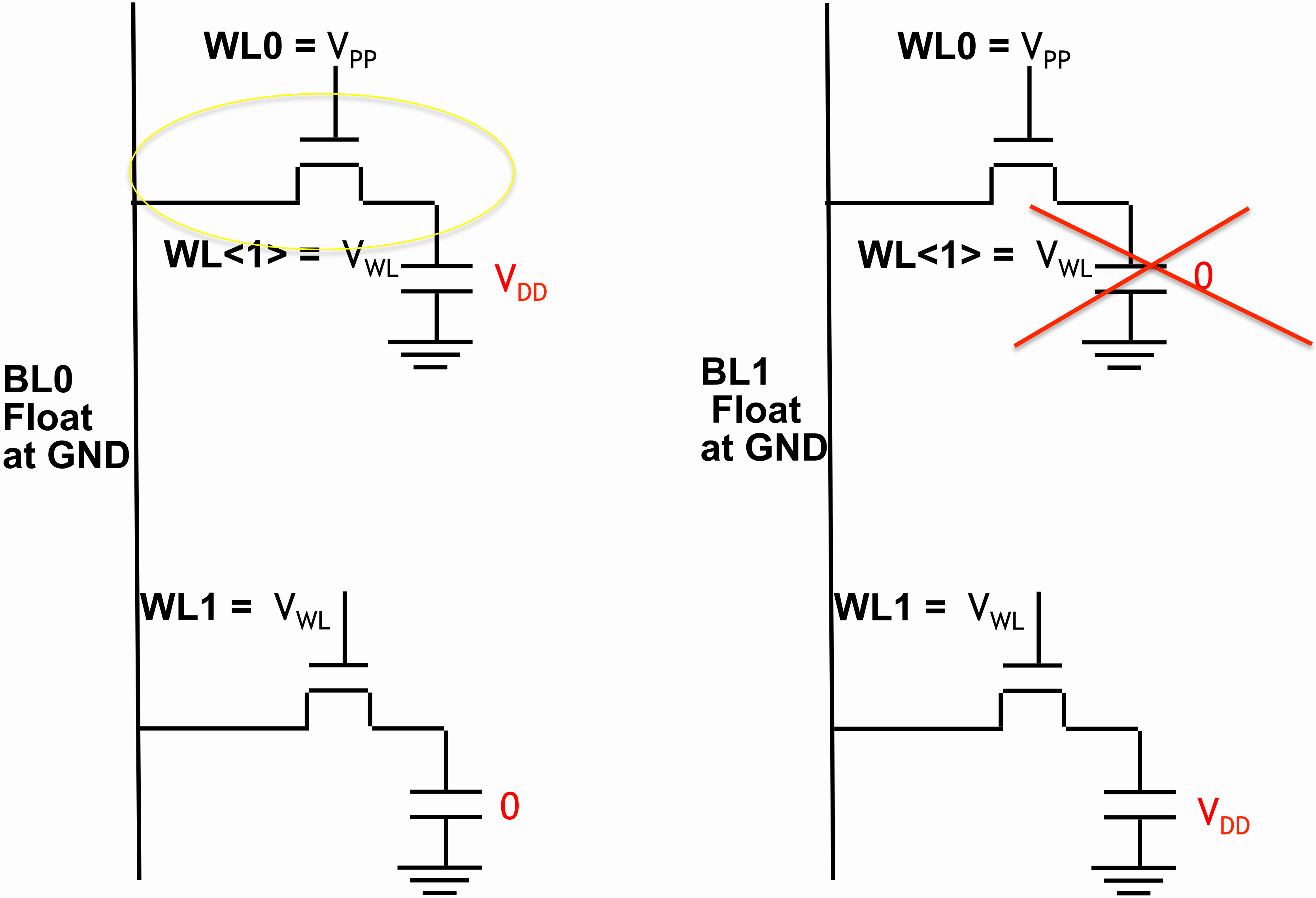
Half Select Condition - SRAM



Half Select Condition - eDRAM



Half Select Condition - eDRAM



SRAM vs DRAM

Assume a 1Mb memory with 512 WLs and 2048 BLs and 8 Columns. i.e. 256 DLs

	SRAM	DRAM
#WLs per BL	512	32
# BLs per DL	8	8
Sharing Sense Amps across columns	Yes	No
Effective number of cells connected to a SA	4096	32
# Sense Amps	256	

Extremely small SA

- Number of transistors
- Sizes of transistors

Topics

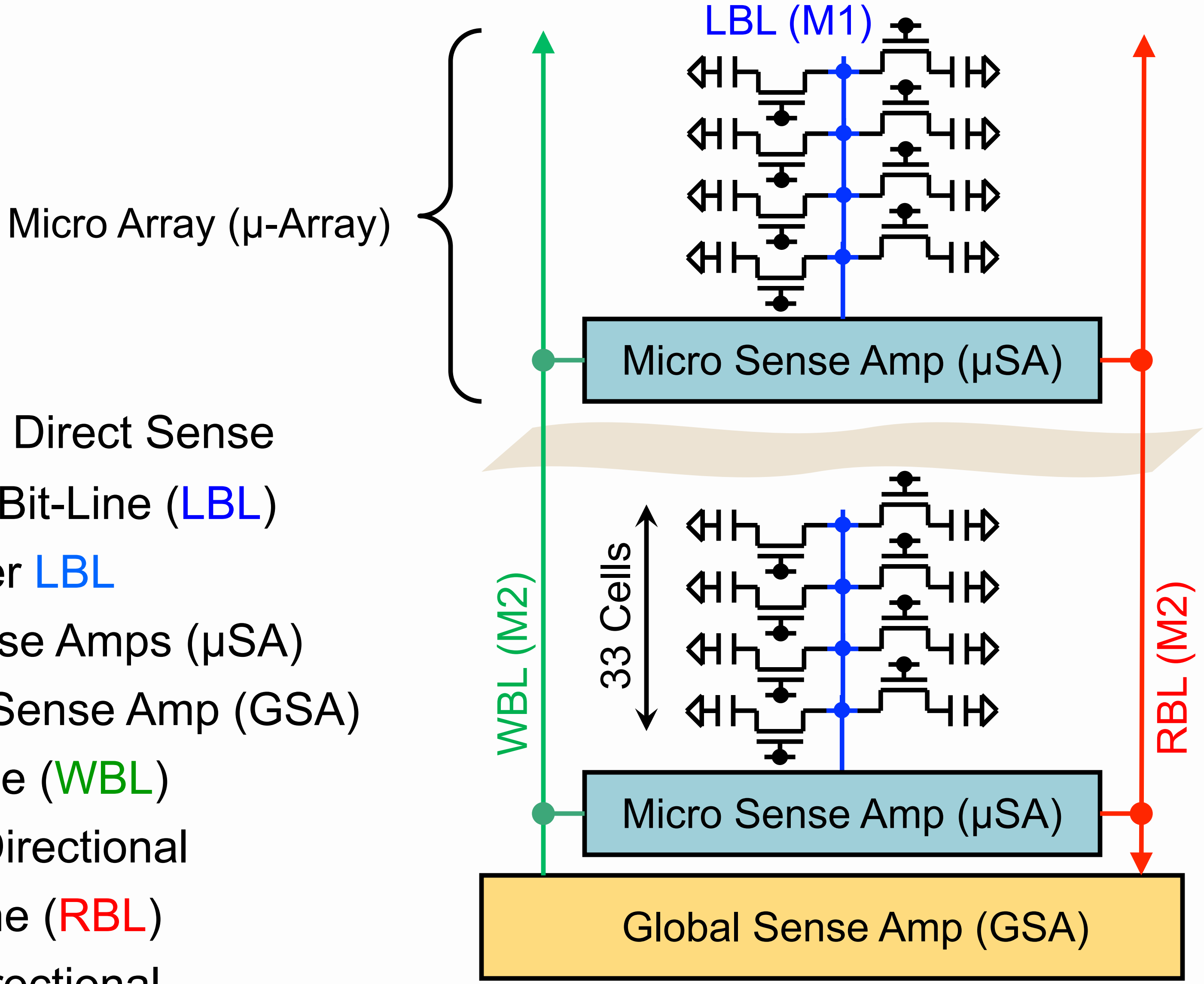
- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM Write Analysis
- ❑ eDRAM Sense-Amplifier Specification
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

DRAM Operation Details (Case Study)

IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 43, NO. 1, JANUARY 2008

A 500 MHz Random Cycle, 1.5 ns Latency, SOI Embedded DRAM Macro Featuring a Three-Transistor Micro Sense Amplifier (John Barth/IBM)

Micro Sense Architecture



- Hierarchical Direct Sense
- Short Local Bit-Line (LBL)

  - 33 Cells per LBL
- 8 Micro Sense Amps (μSA)

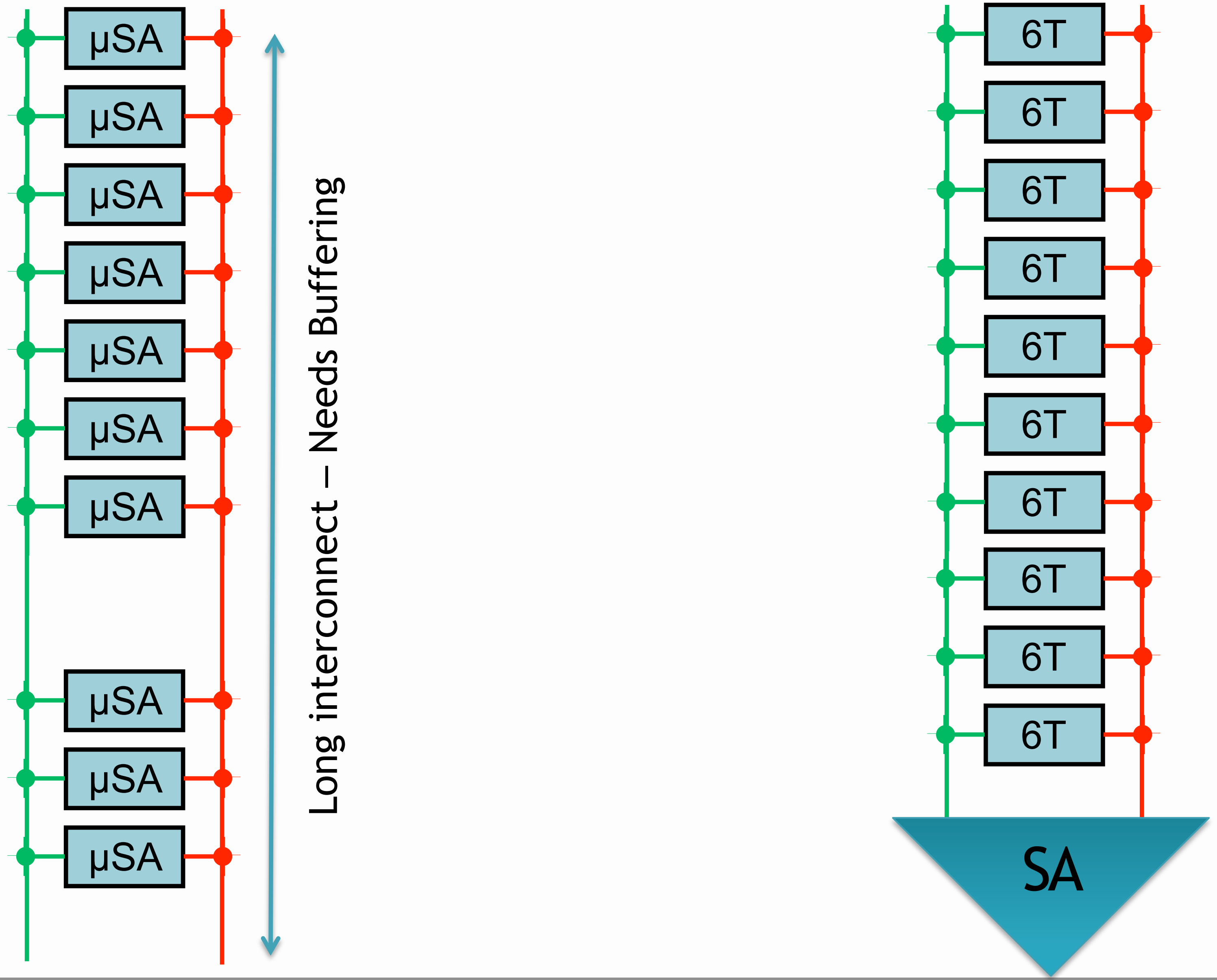
  per Global Sense Amp (GSA)
- Write Bit-Line (WBL)

  Uni-Directional
- Read Bit-Line (RBL)

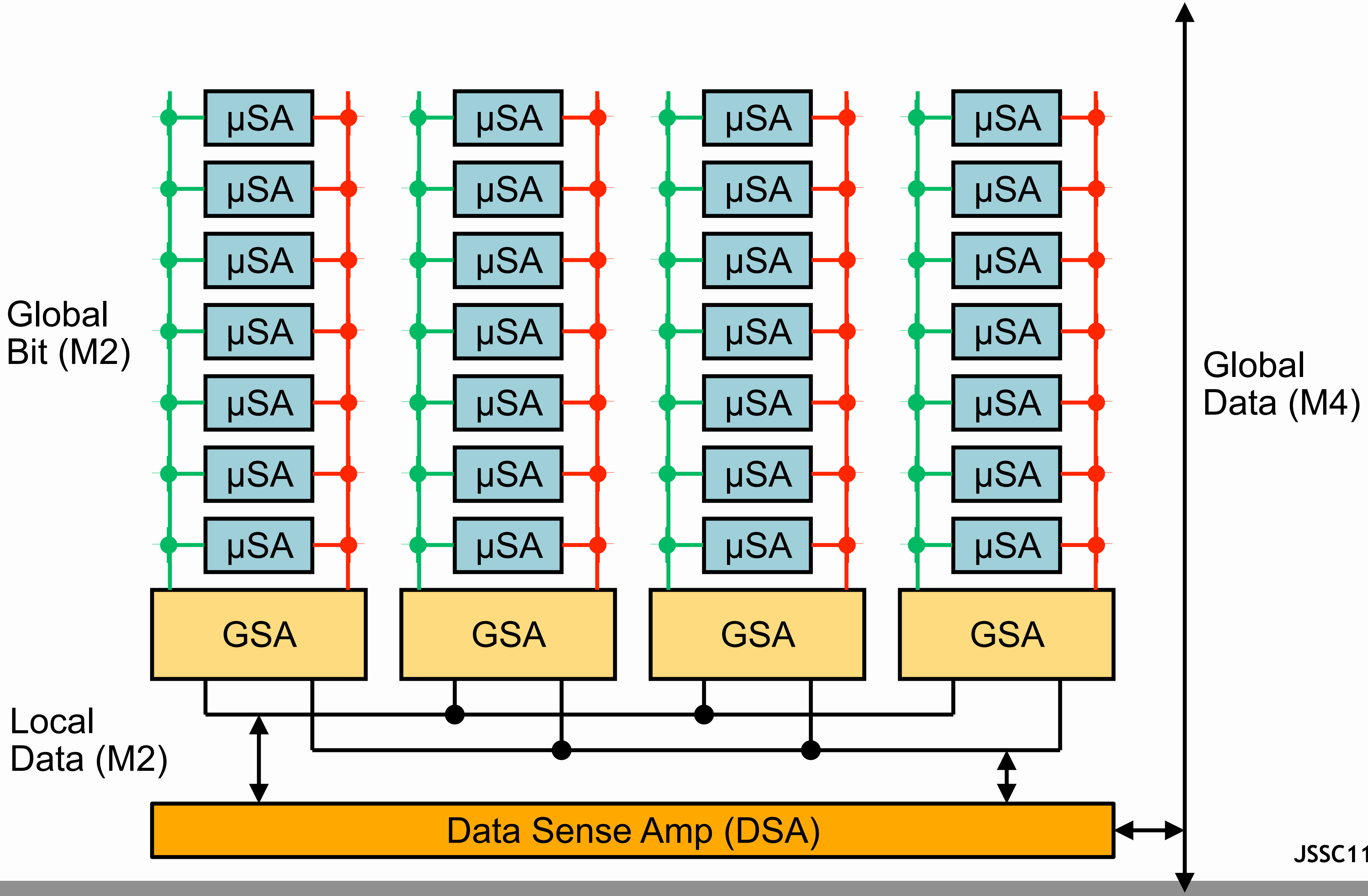
  Bi-Directional

JSSC11

Sense Hierarchy - Motivation

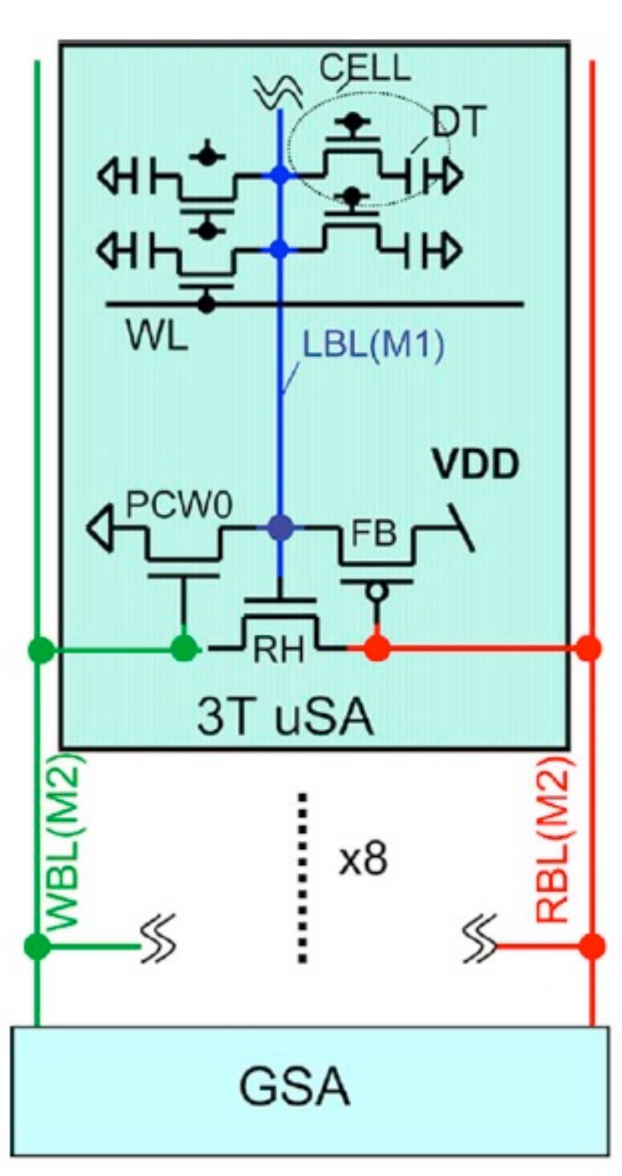


Micro Sense Hierarchy - Three levels



JSSC11

3T uSA operation



Pre-charge

WL is low. WBL and RBL both pre-charged to HIGH.
Next GSA drives WBL low. **LBL floats to GND level**

Read “0”

LBL remains LOW. RBL is HIGH. Sensed as a “0”

Read “1”

LBL is HIGH. Turns on RH, pulls RBL LOW.
+ feedback as pFET FB turns ON. Sensed as a “1”

Write “1”

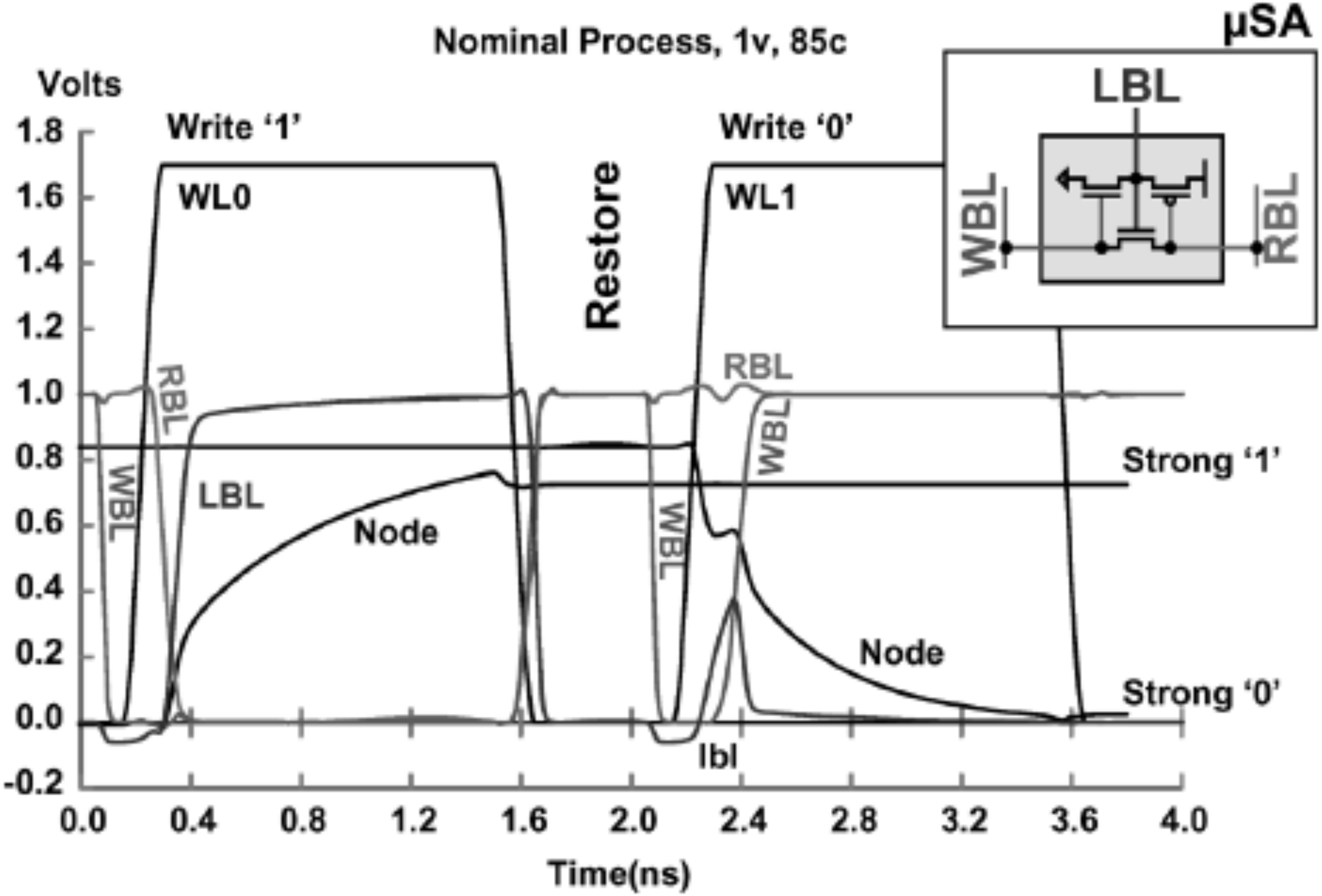
GSA pulls RBL to GND. FB pFET turns ON
Happens while WL rises (direct write)

Write “0”

WBL is HIGH, PCW0 ON. Clamps LBL to GND
As WL activates.

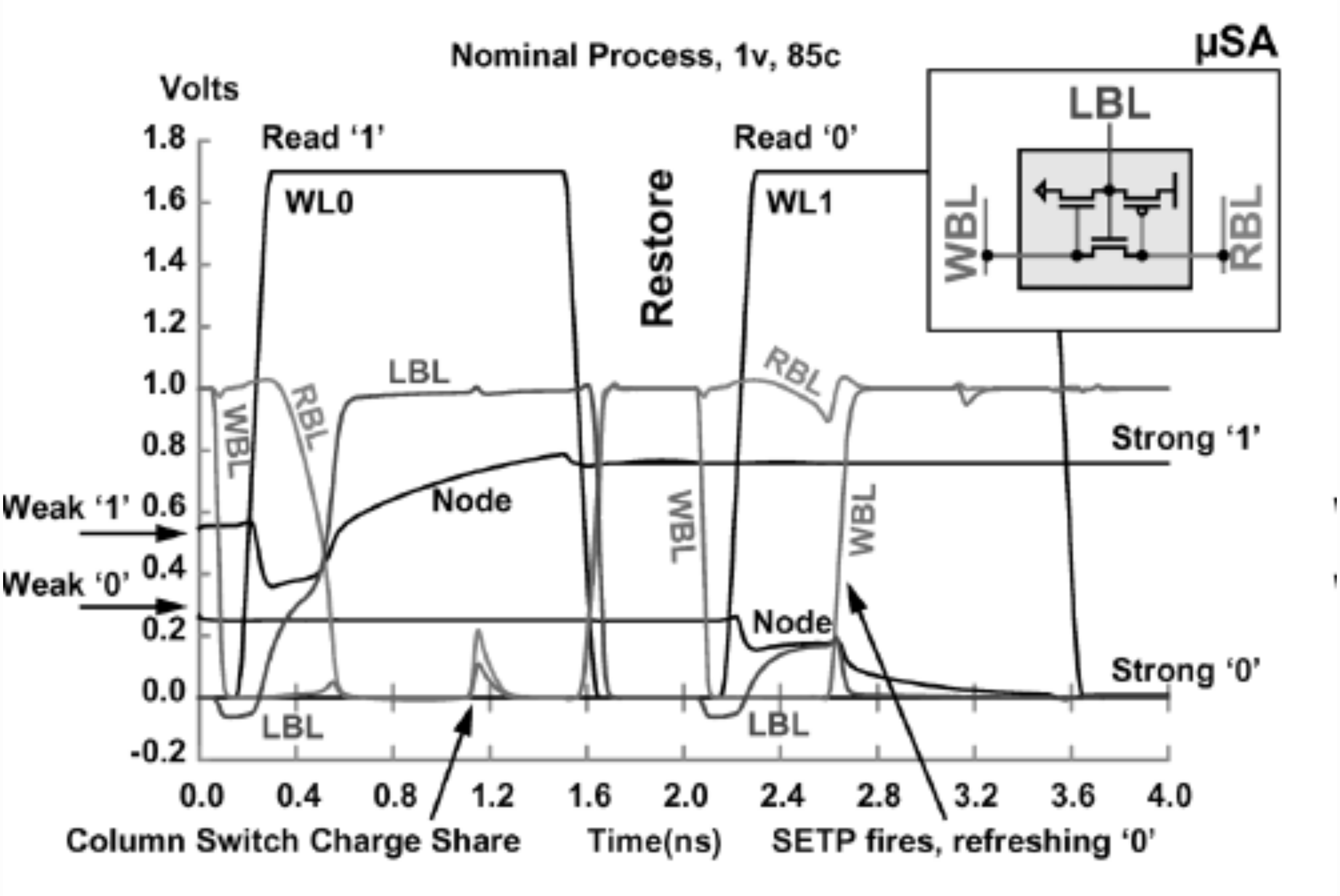
JSSC11

Simulations - Write



JSSC08

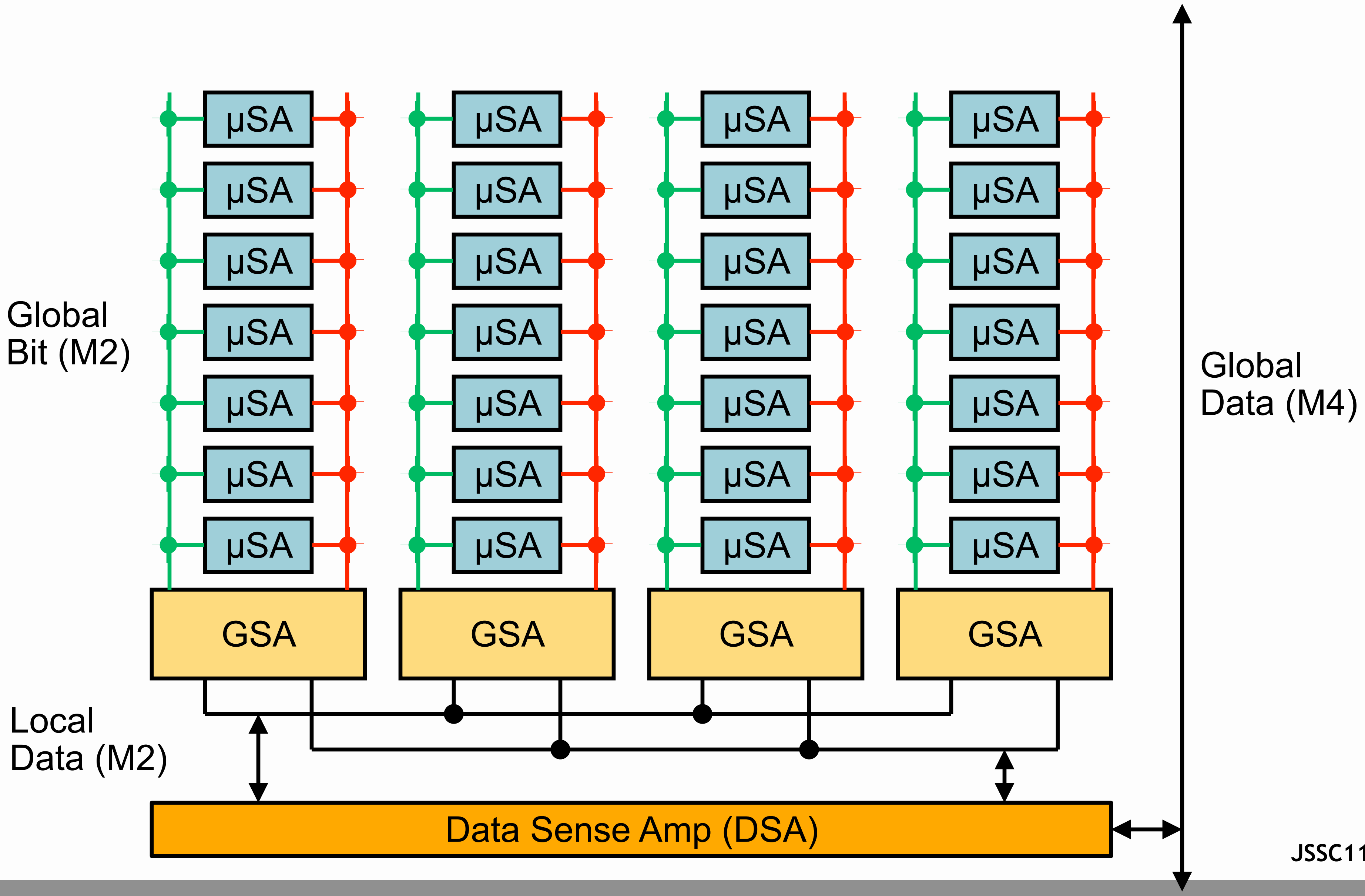
Simulations - Read



(b)

JSSC08

Micro Sense Hierarchy - Three levels

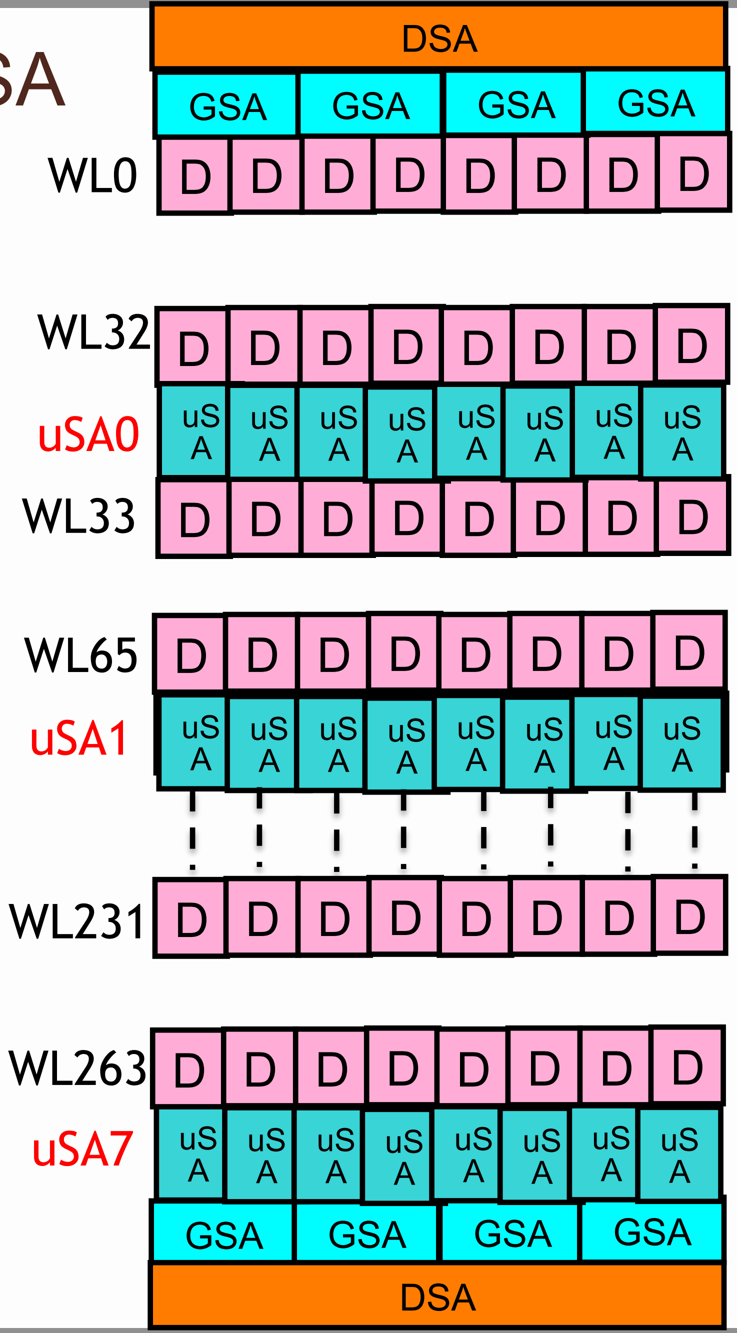


JSSC11

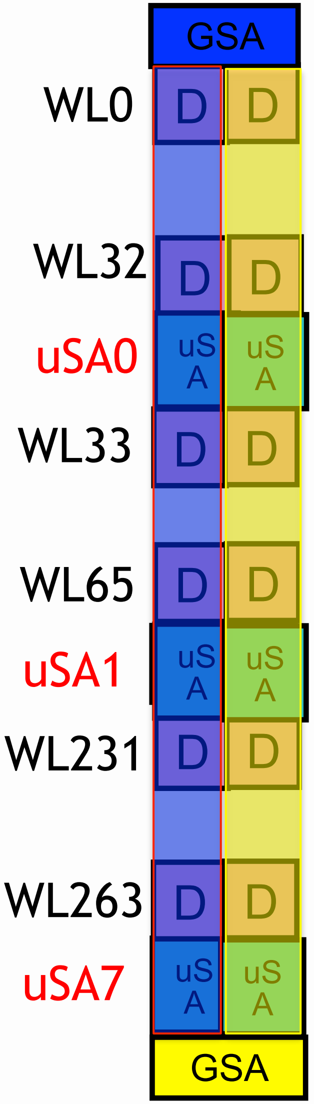
Layout Floor plan of Array+SA

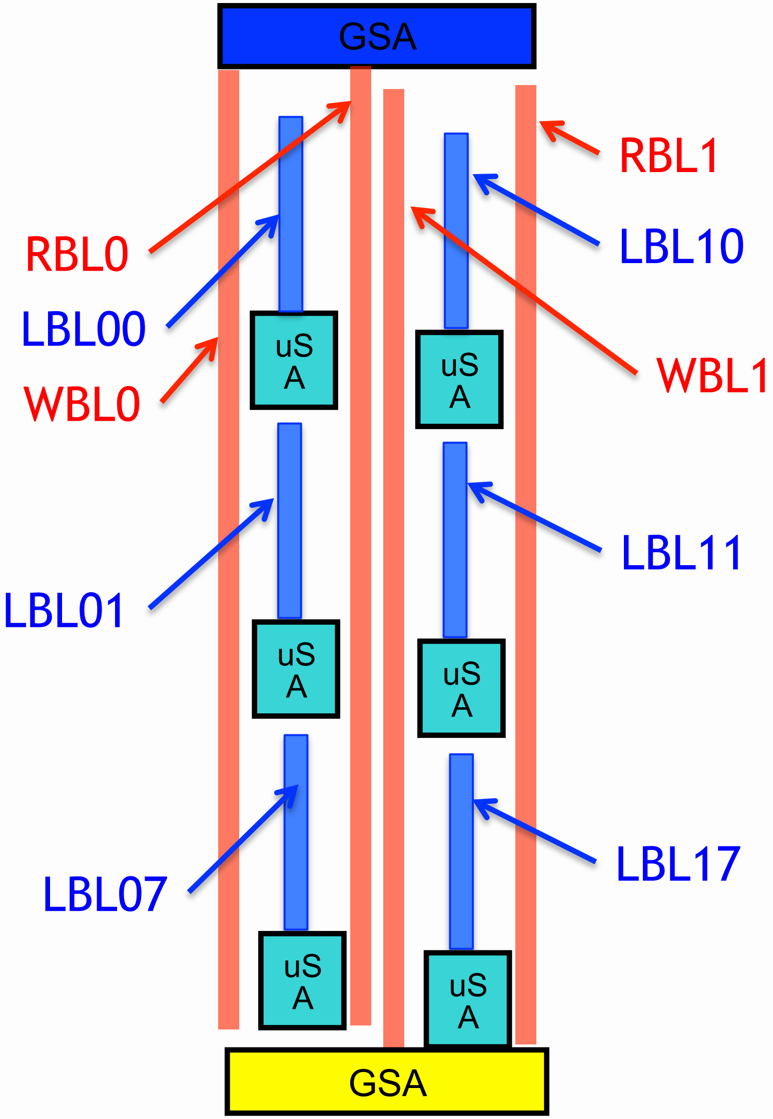
GSA Should fit into the bitcell width or n*bitcell width

Thus, distributed GSA on two sides of bitcell array

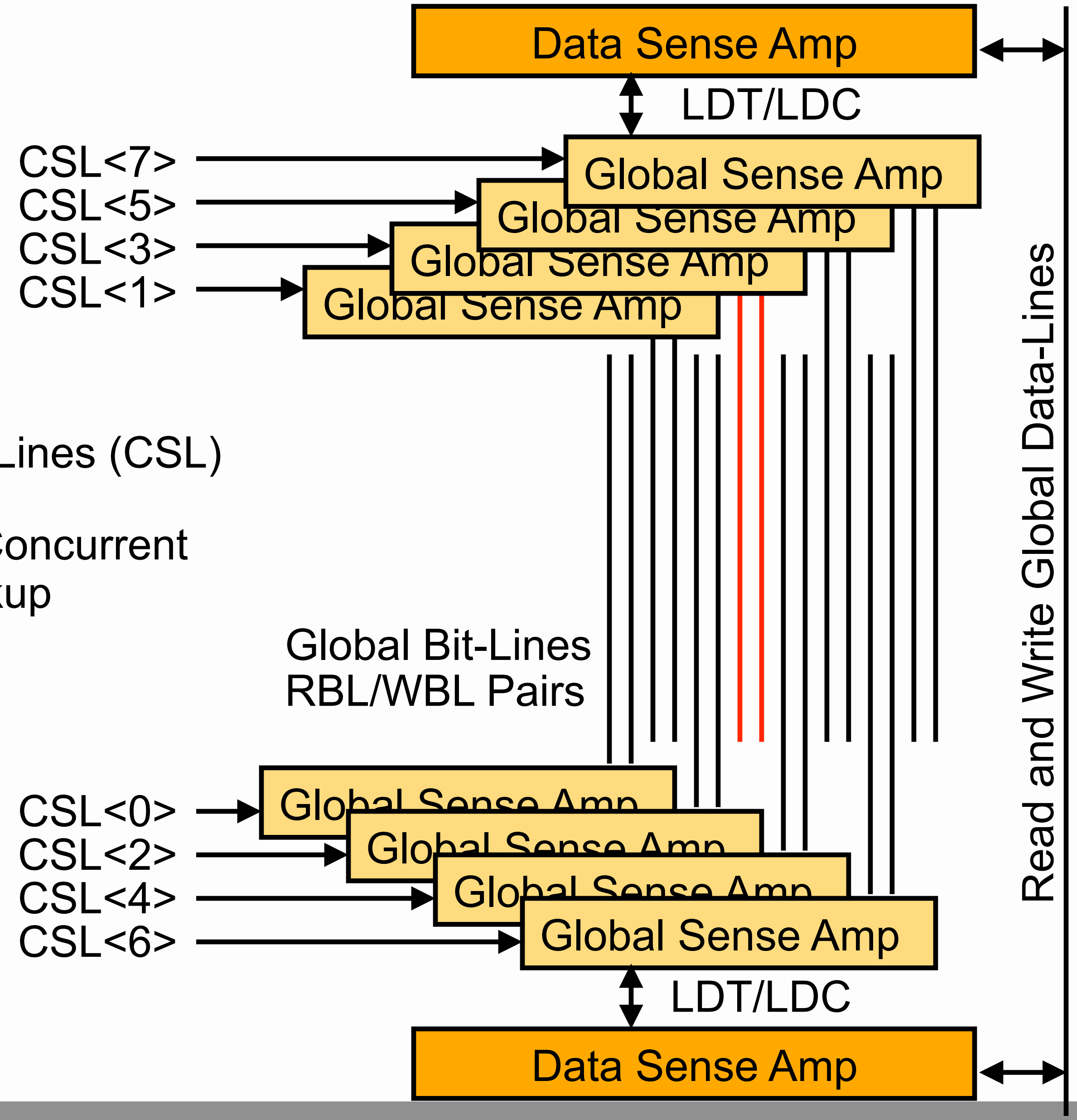


Layout Floor plan of Array+SA



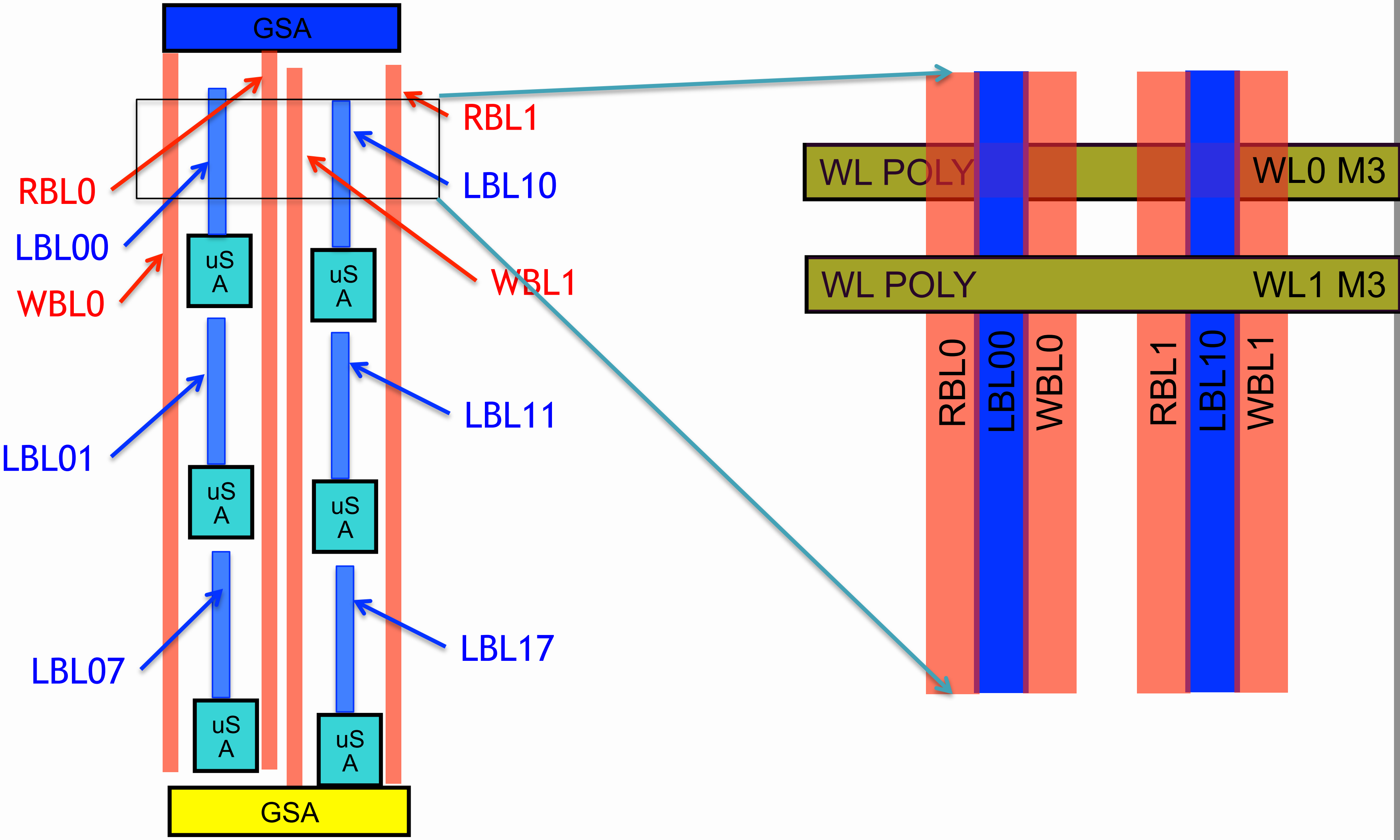


Column Interleave

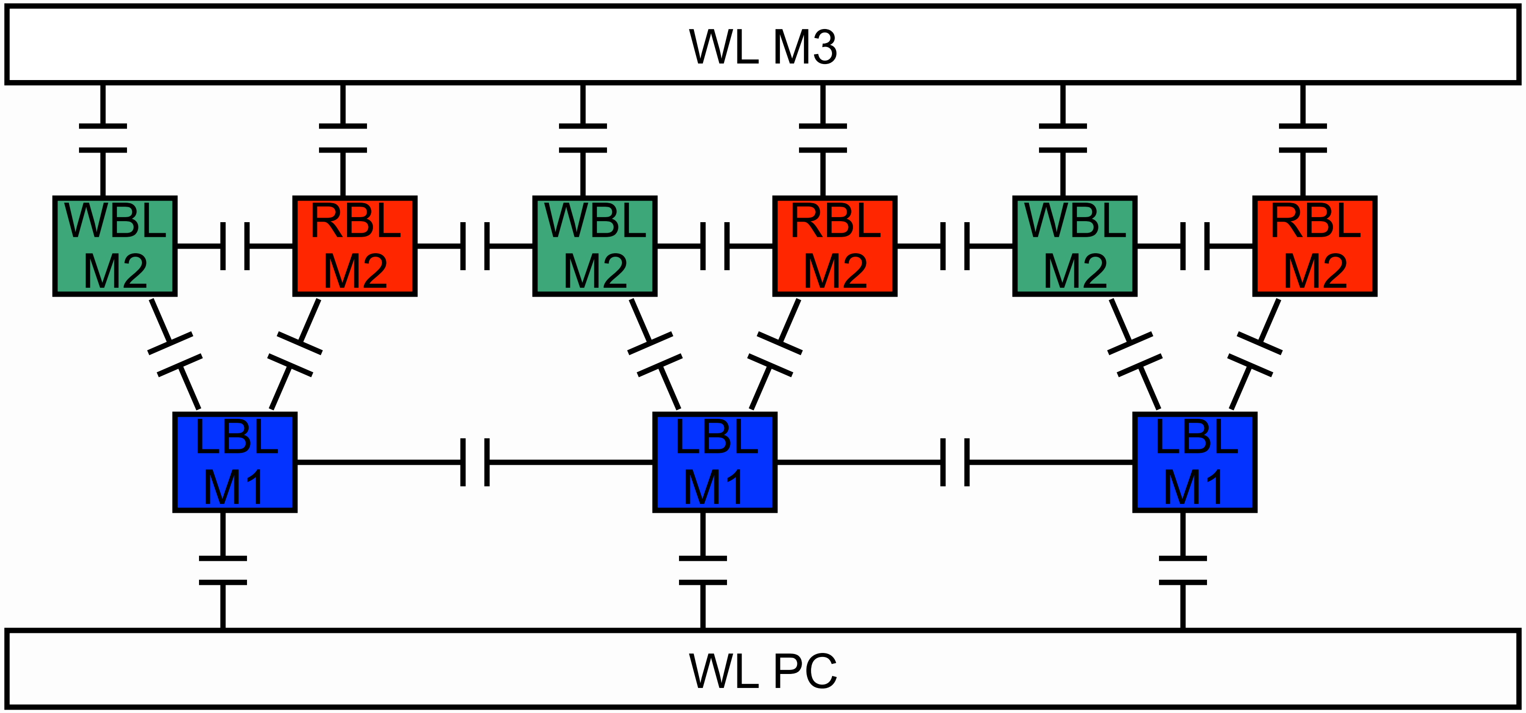


- 1 of 8 Column Select Lines (CSL)
- Fire Early for Write
- Fire Late to Support Concurrent Cache Directory Lookup

LAYOUT of array

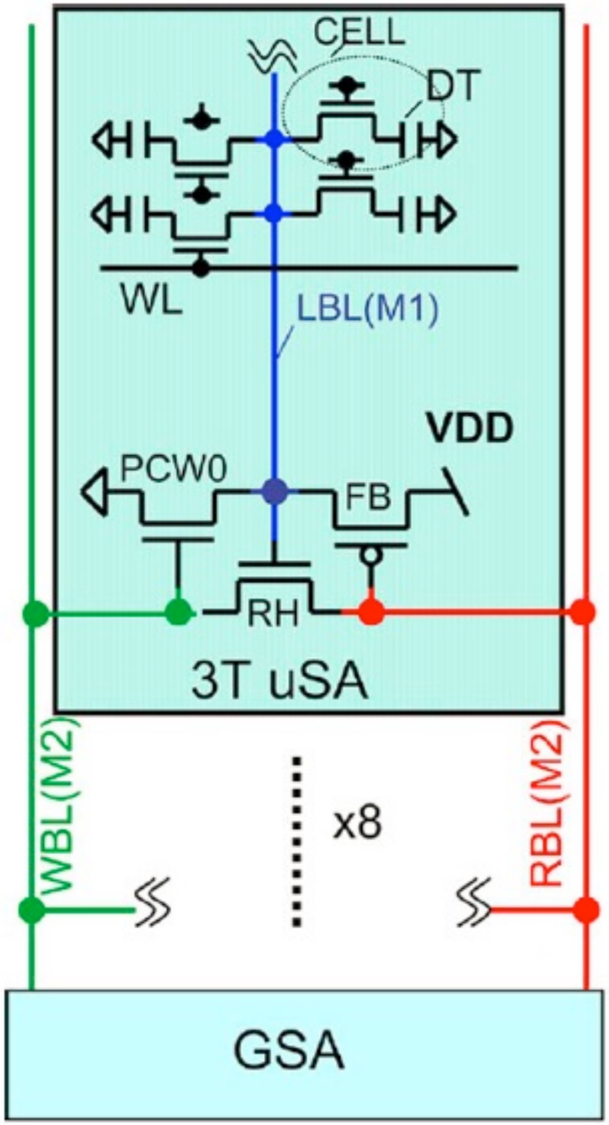


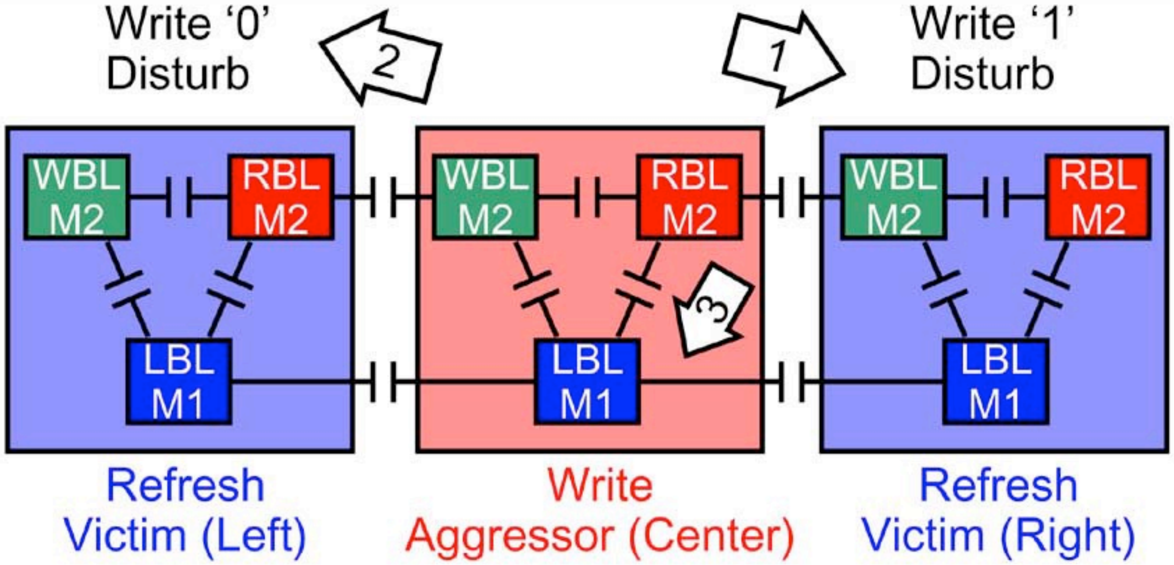
Micro Sense Local Bit-line Cross Section



Single Ended Sense – Twist not effective
Line to Line Coupling must be managed

Micro Sense Coupling Mechanisms

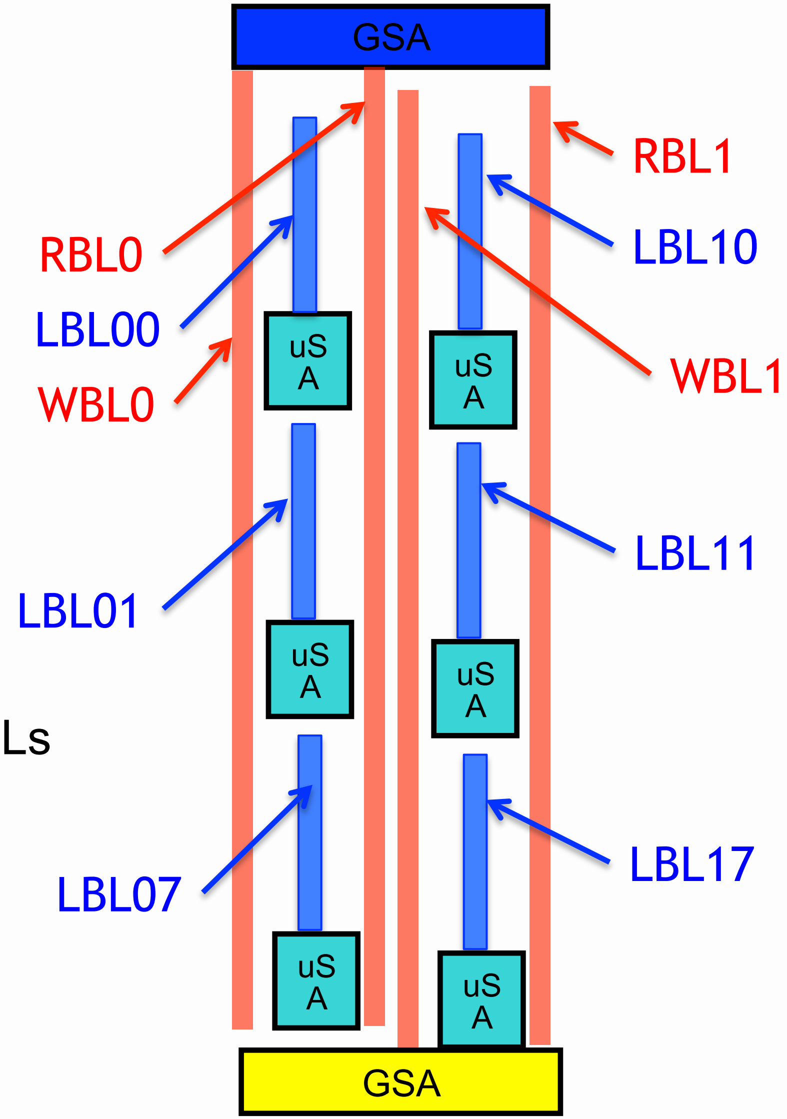




1. Write '1' Couples WBL below Ground
   Increasing RH leakage during Refresh '0'
2. Write '0' Couples RBL above VDD
   Delaying Feedback during Refresh '1'
3. Read '1' Couples Half-Selected LBL Below GND
   Increasing Array Device Sub-VT Leakage

JSSC11

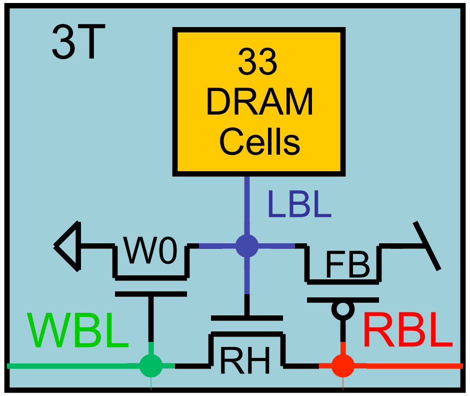
Half Selected LBL



- Accessing one of WL0-32
- Cell connected is on LBL00
- LBL01-07 – Half Selected LBLs

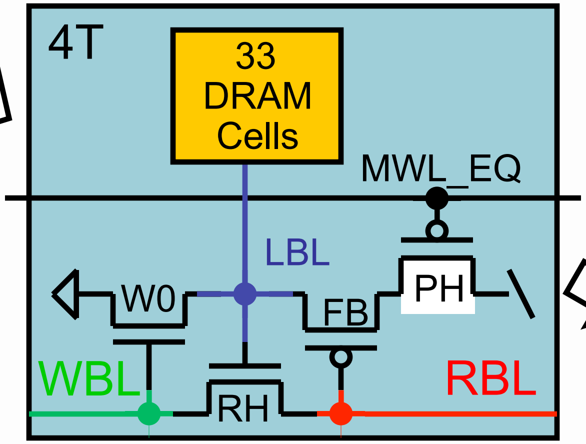
Micro Sense Evolution

1. Write Zero (W0)
2. Read Head (RH)
3. Feed-Back (FB)



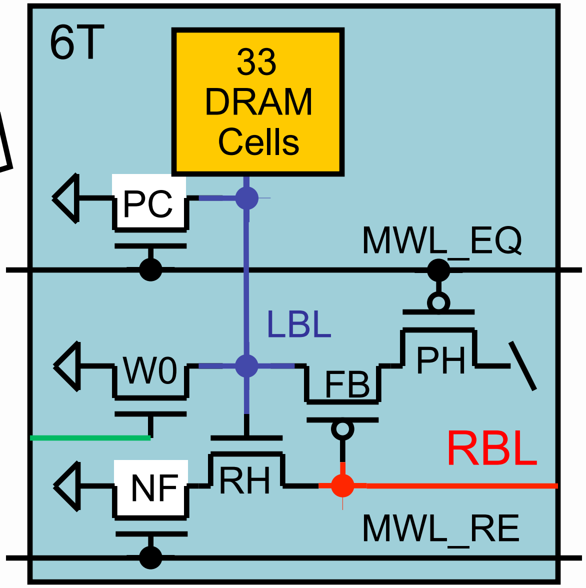
Barth, ISSCC'07

4. PFET Header (PH)
   - LBL Power Gate
   - LBL Leakage



Klim, VLSI'07

5. Pre-Charge (PC)
   - WBL Power (Write '0' Only)
6. NFET Footer (NF)
   - RBL Leakage
   - Decompose Pre-Charge and Read Enable (MWL_RE)



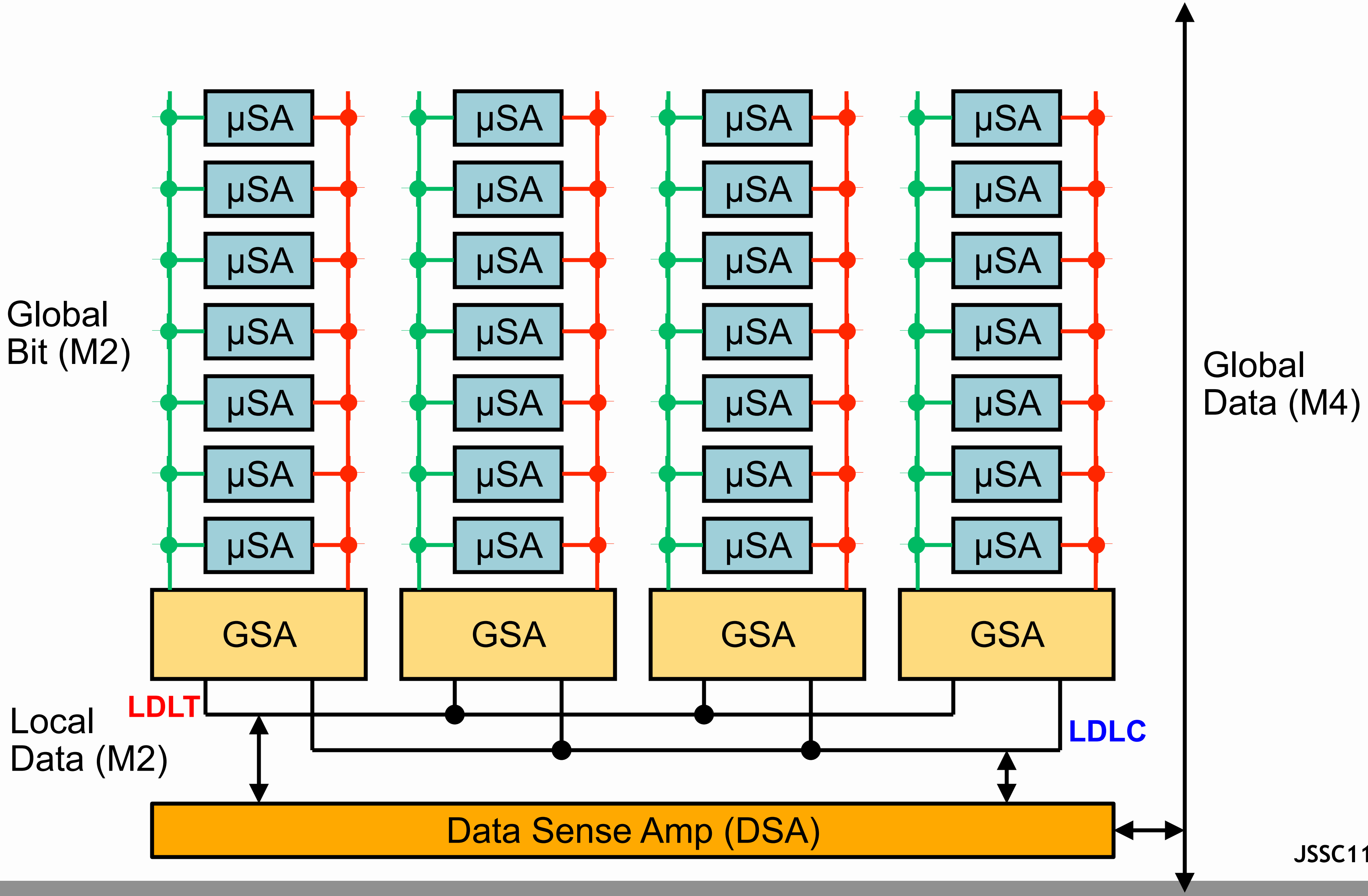
JSSC11

Power Reduction

Power Reduction
Traded for Transistor Count

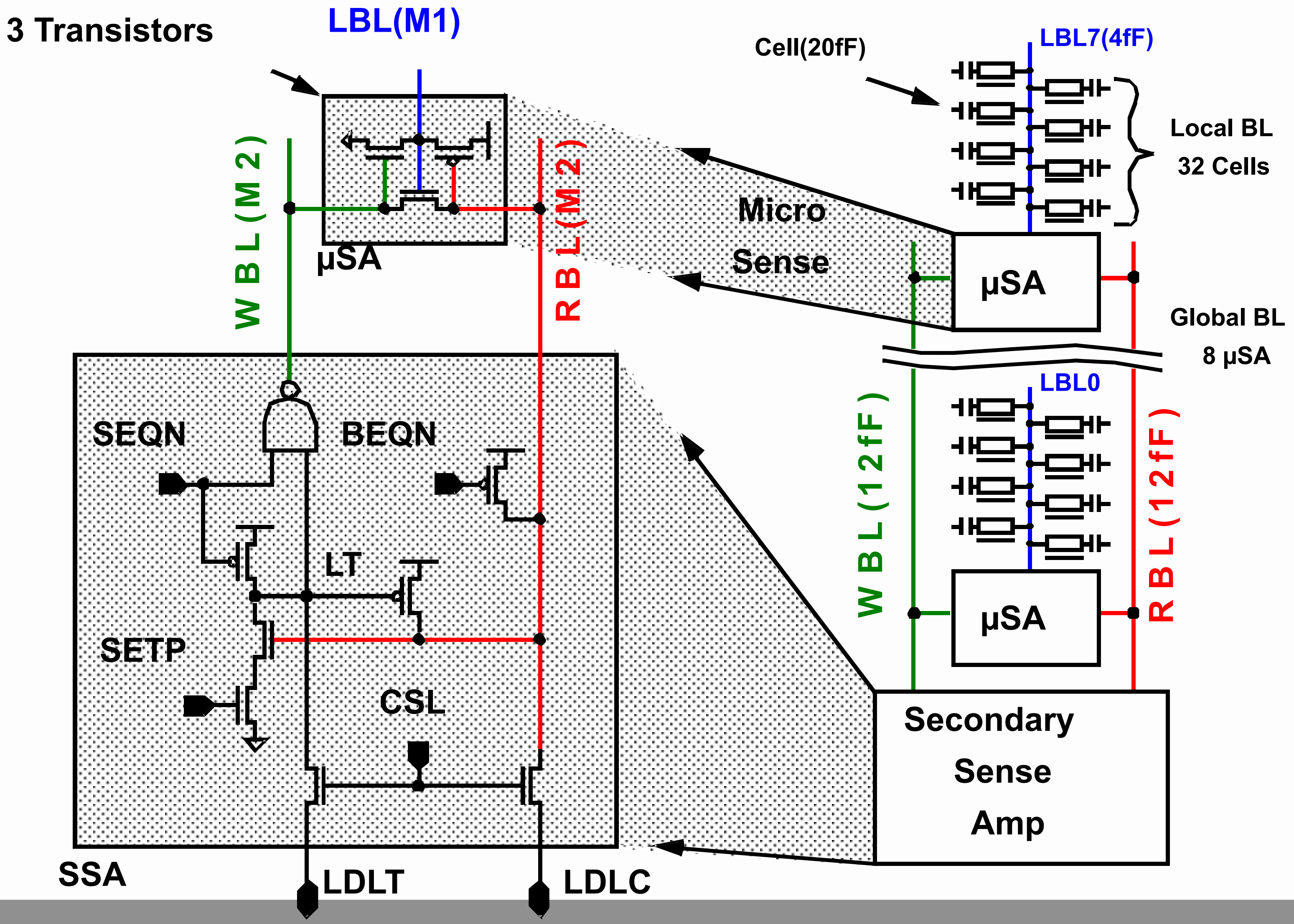
Increased Transistor Count

Micro Sense Hierarchy - Three levels



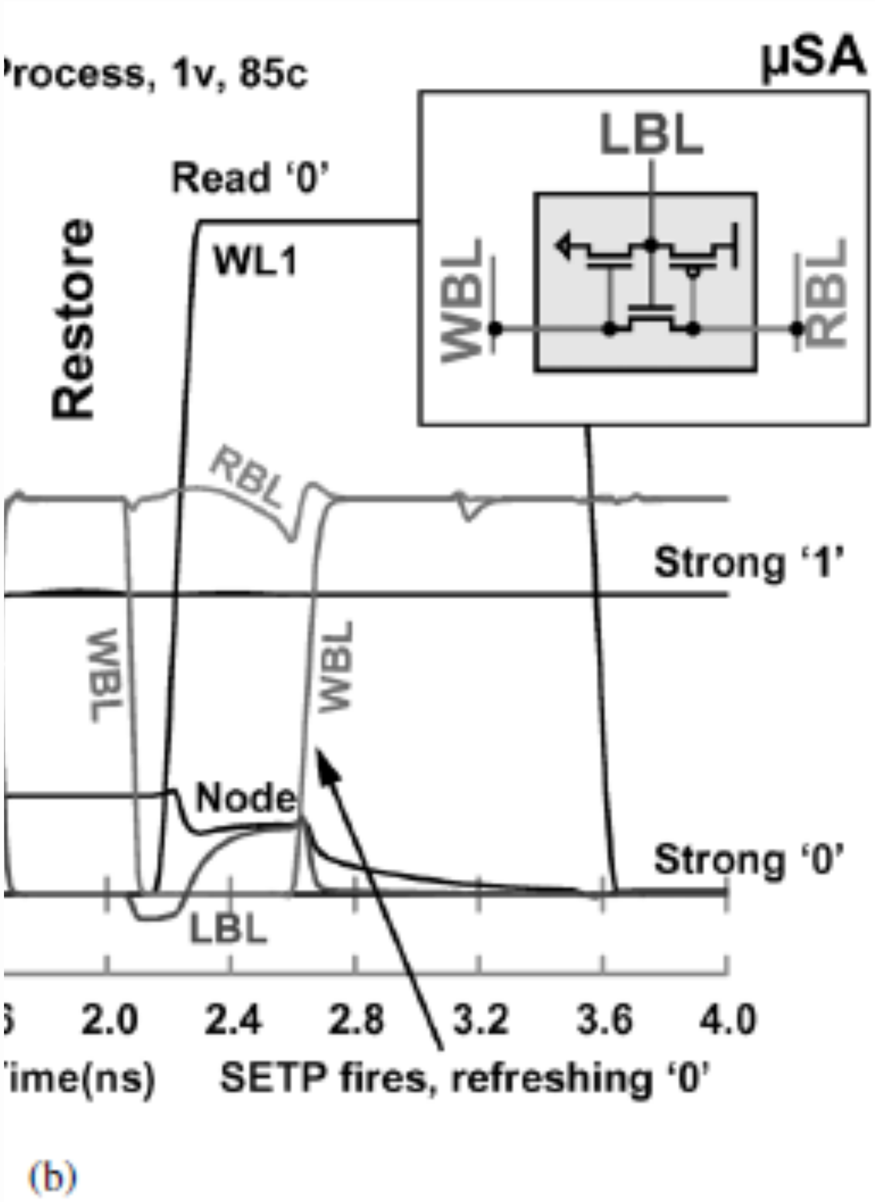
JSSC11

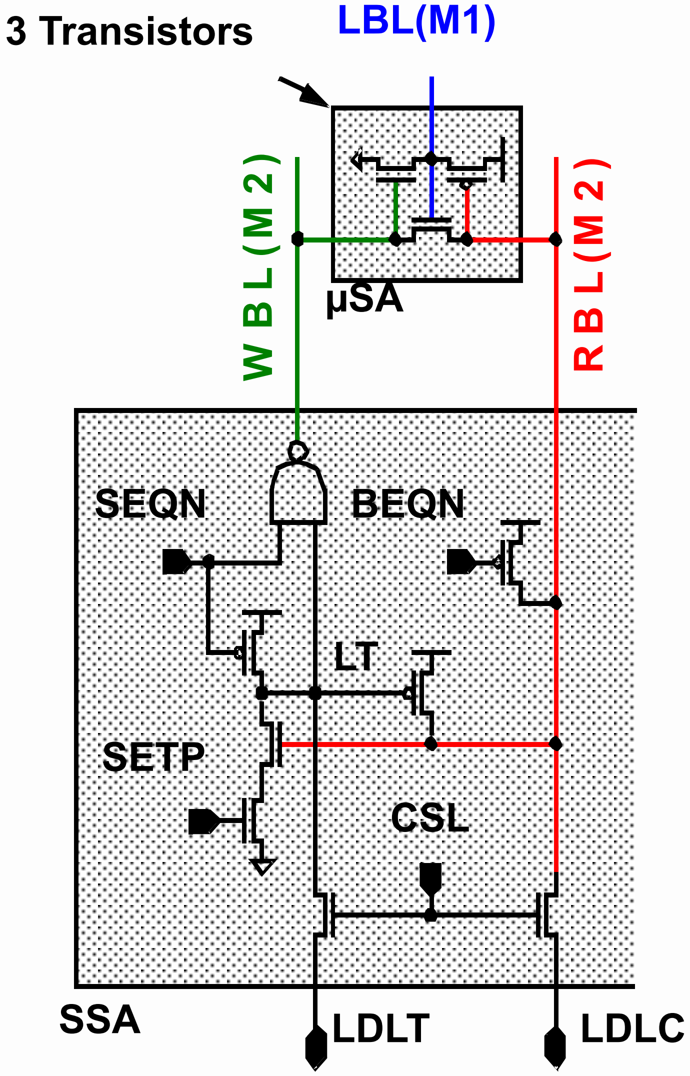
Micro Sense Architecture (µSA)



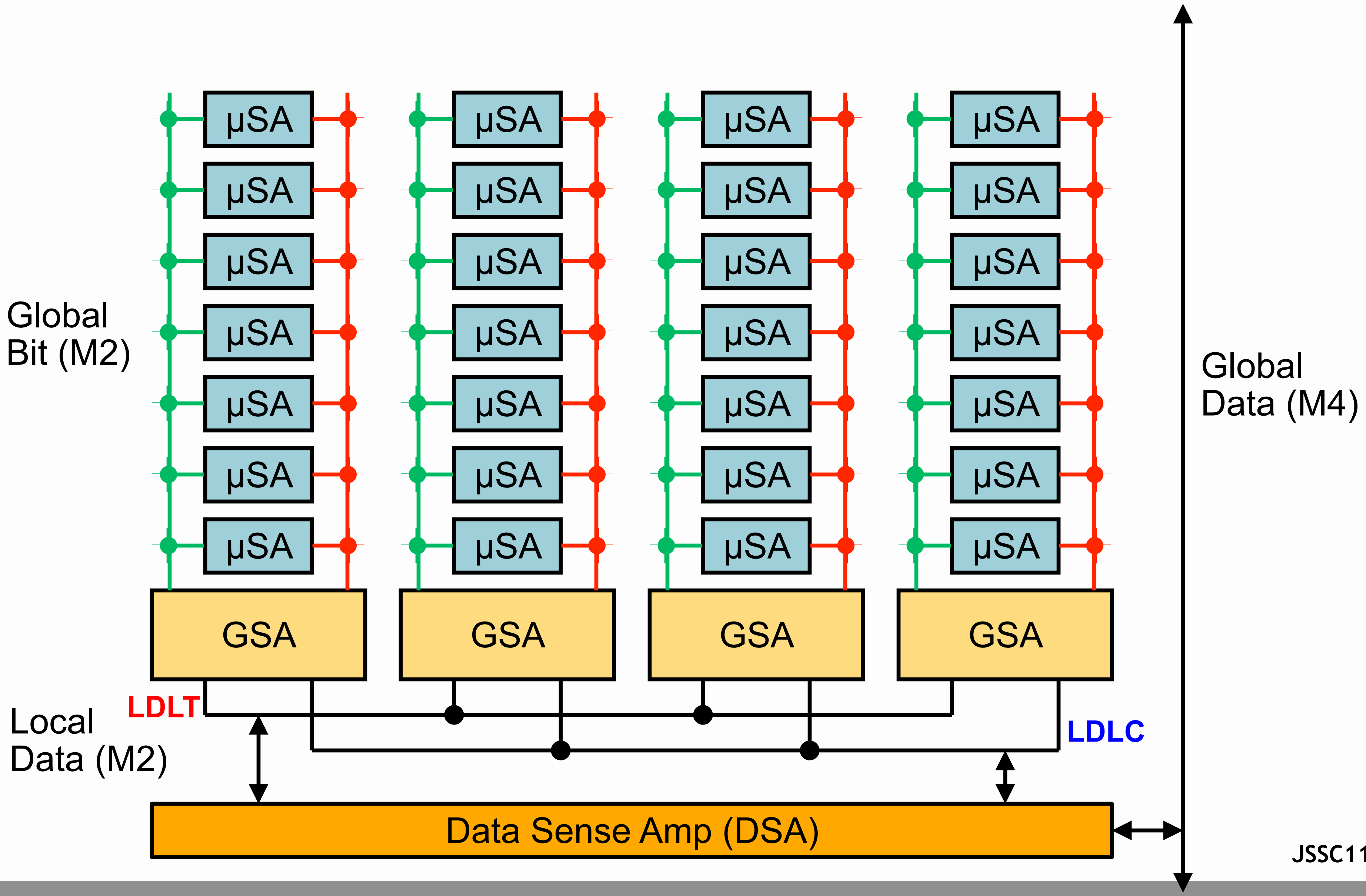
JSSC08

Micro Sense Architecture (μSA)



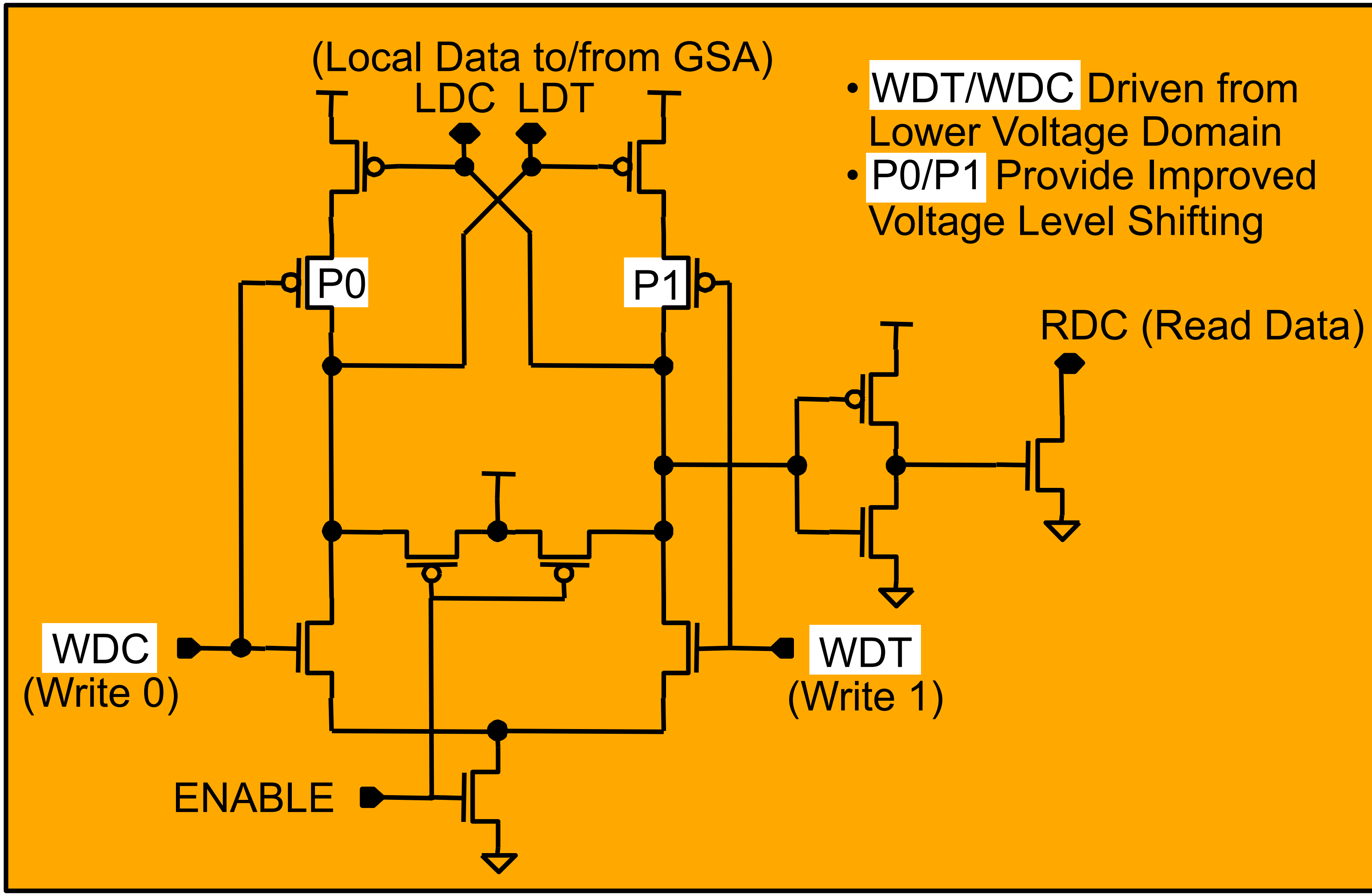


Micro Sense Hierarchy - Three levels



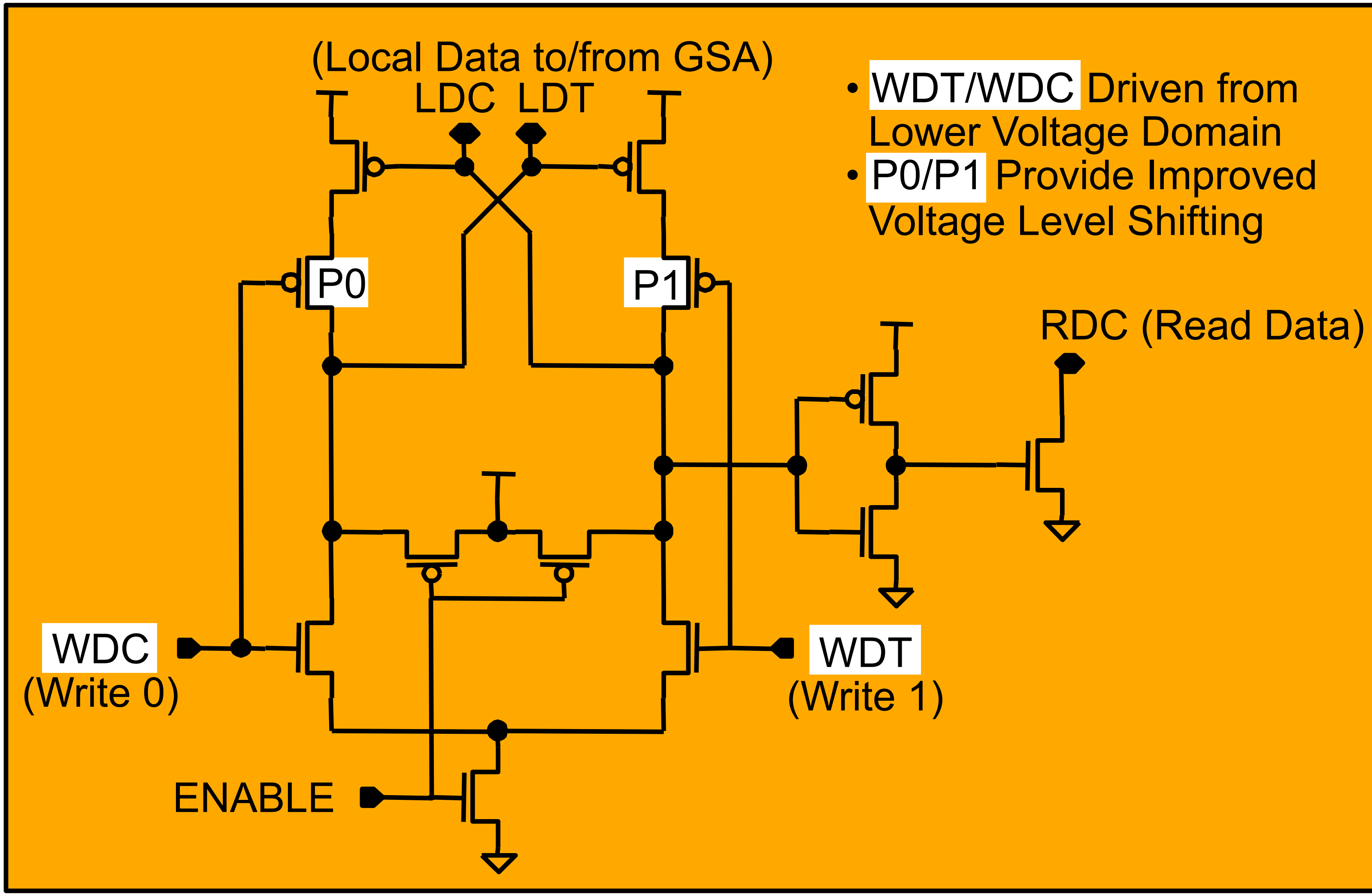
JSSC11

Data Sense Amp (DSA)



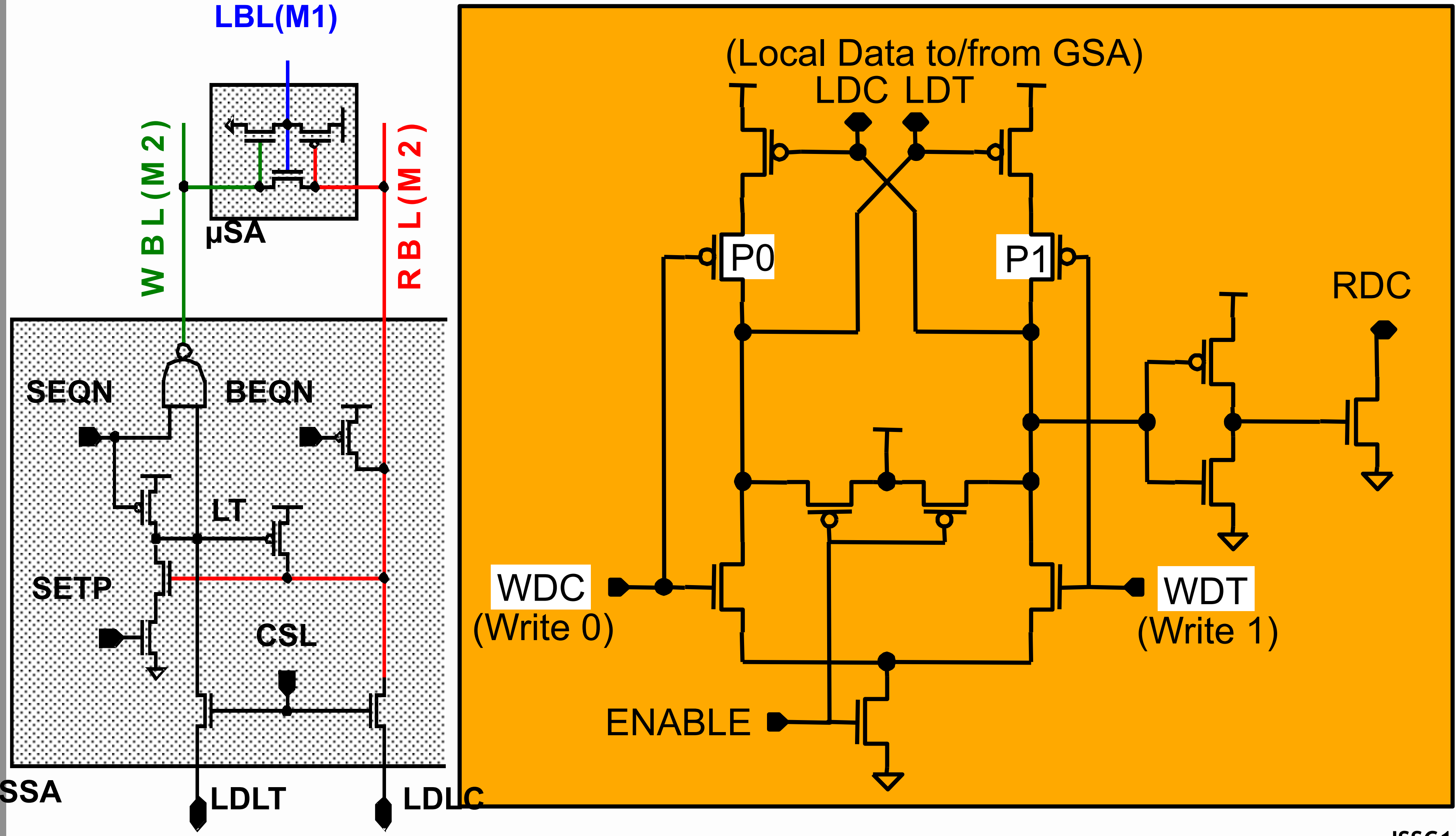
- WDT/WDC Driven from Lower Voltage Domain
- P0/P1 Provide Improved Voltage Level Shifting

Data Sense Amp (DSA) – Write



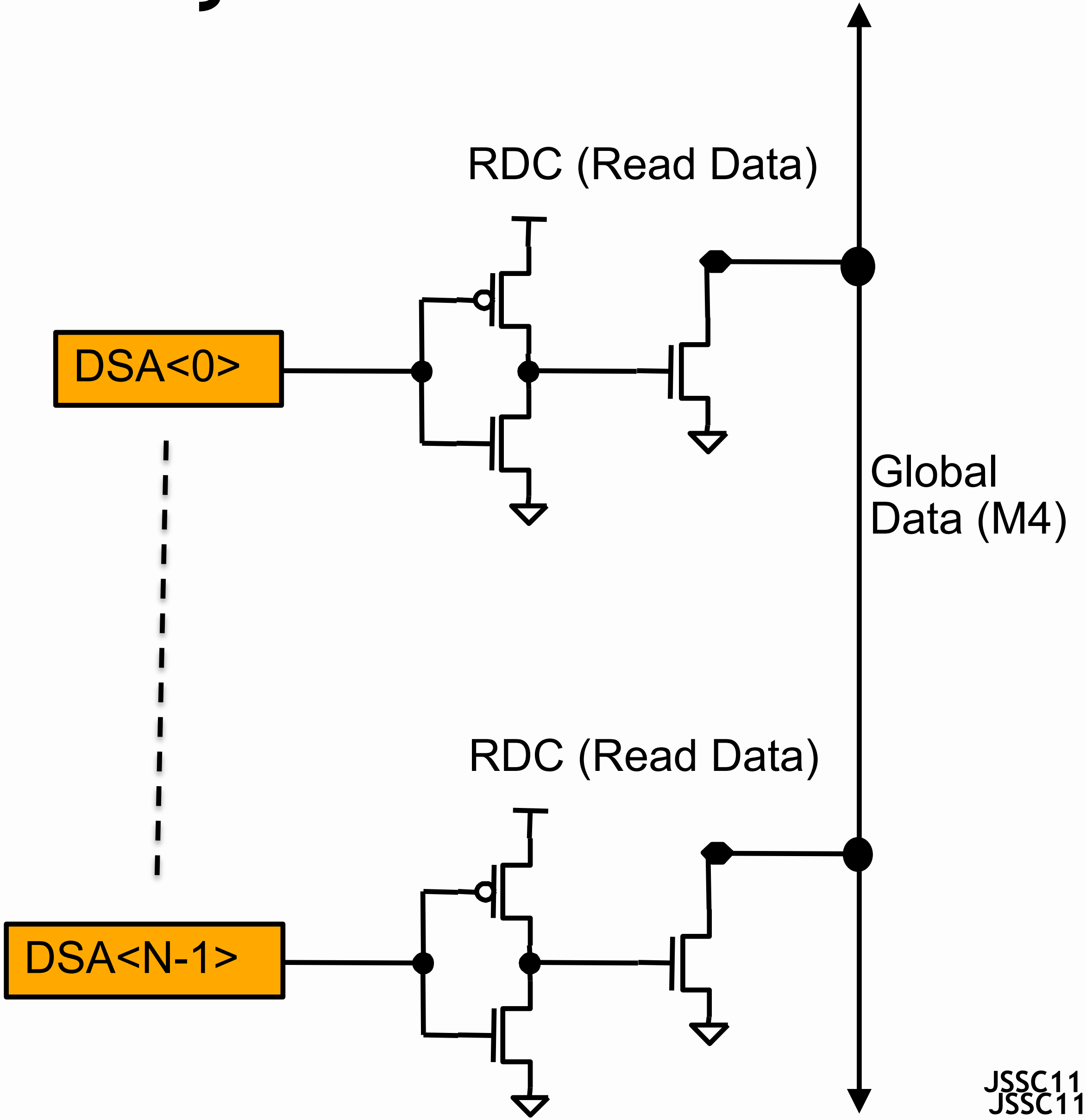
- WDT/WDC Driven from Lower Voltage Domain
- P0/P1 Provide Improved Voltage Level Shifting

Data Sense Amp (DSA) – Read



JSSC11

Combining DSA's- Dynamic NOR Gate



JSSC11

JSSC11

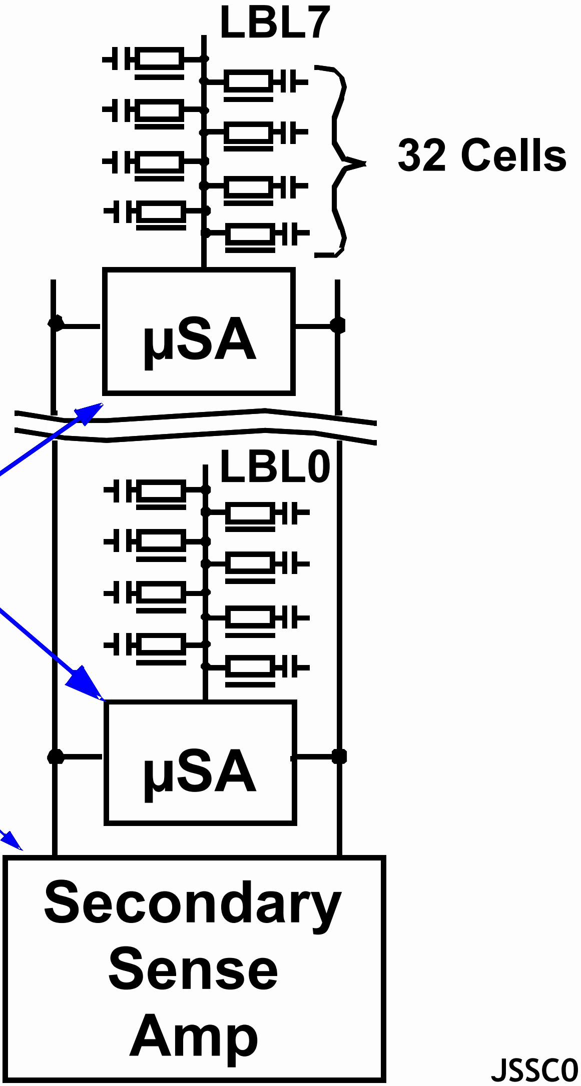
Micro Sense Advantage

Fast Performance of Short Bit-Line

Area Overhead of 4x Longer Bit-Line

Bits/BL	256	128	32
Sense Amp	10%	20%	19%
Reference Cells	2.3%	4%	-
Twist Region	2%	2.6%	-
Second Sense Amp	-	-	8%
Total	14.3%	26.6%	27%

Same Overhead

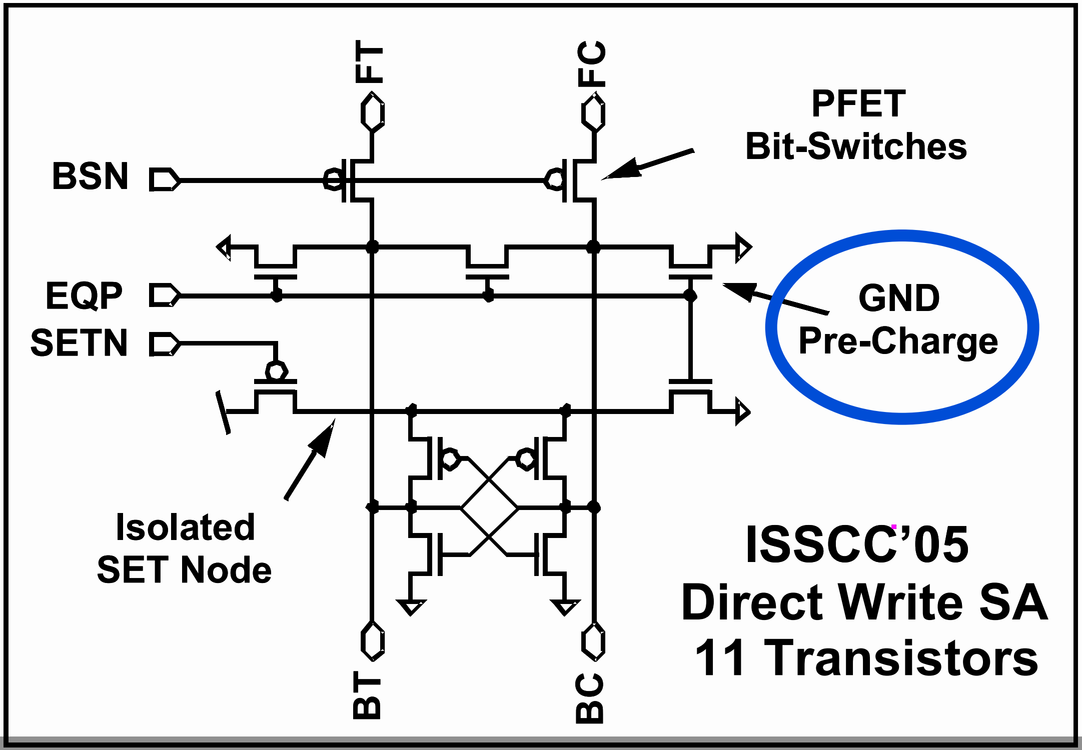


JSSC08

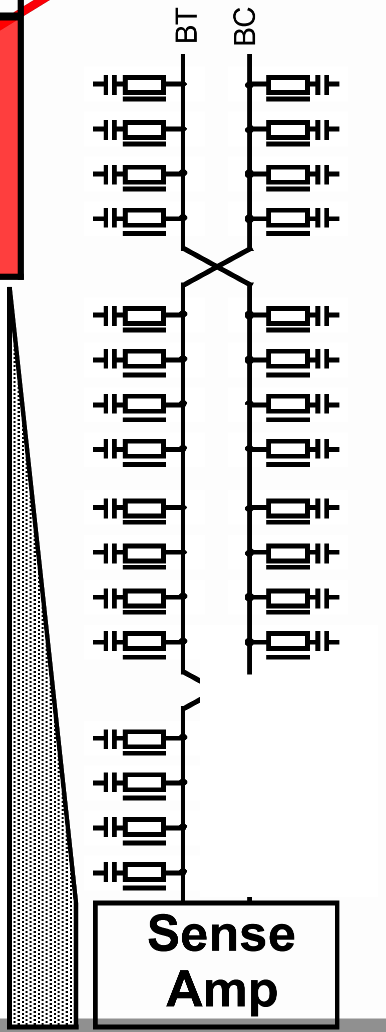
Bit-Line area overhead

Bits/BL	256	128	32
Sense Amp	10%	20%	> 80%
Reference Cells	2.3%	4%	
Twist Region	2%	2.6%	

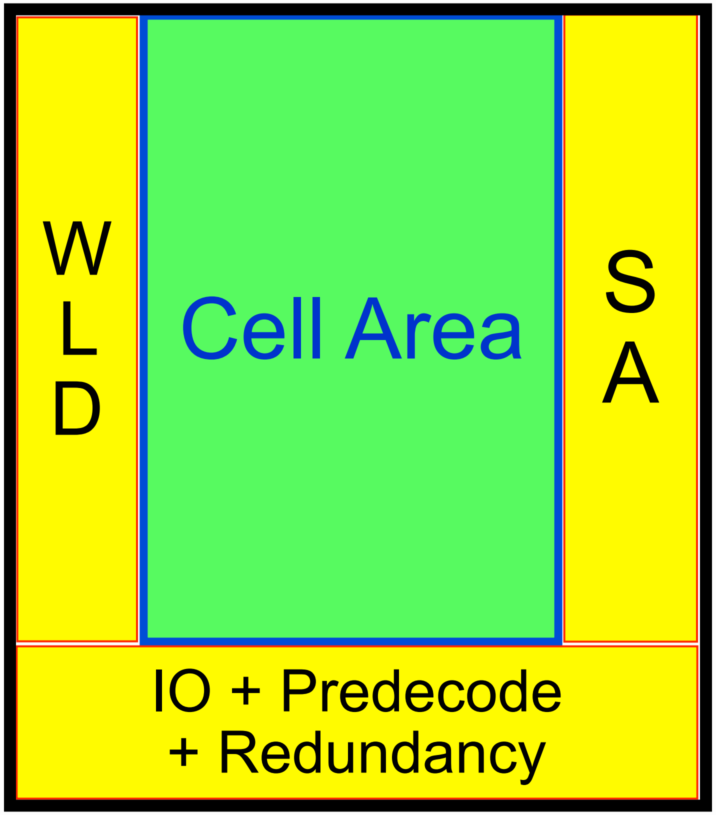
Unacceptable



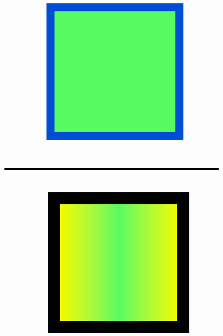
ISSCC'05
Direct Write SA
11 Transistors



Array utilization



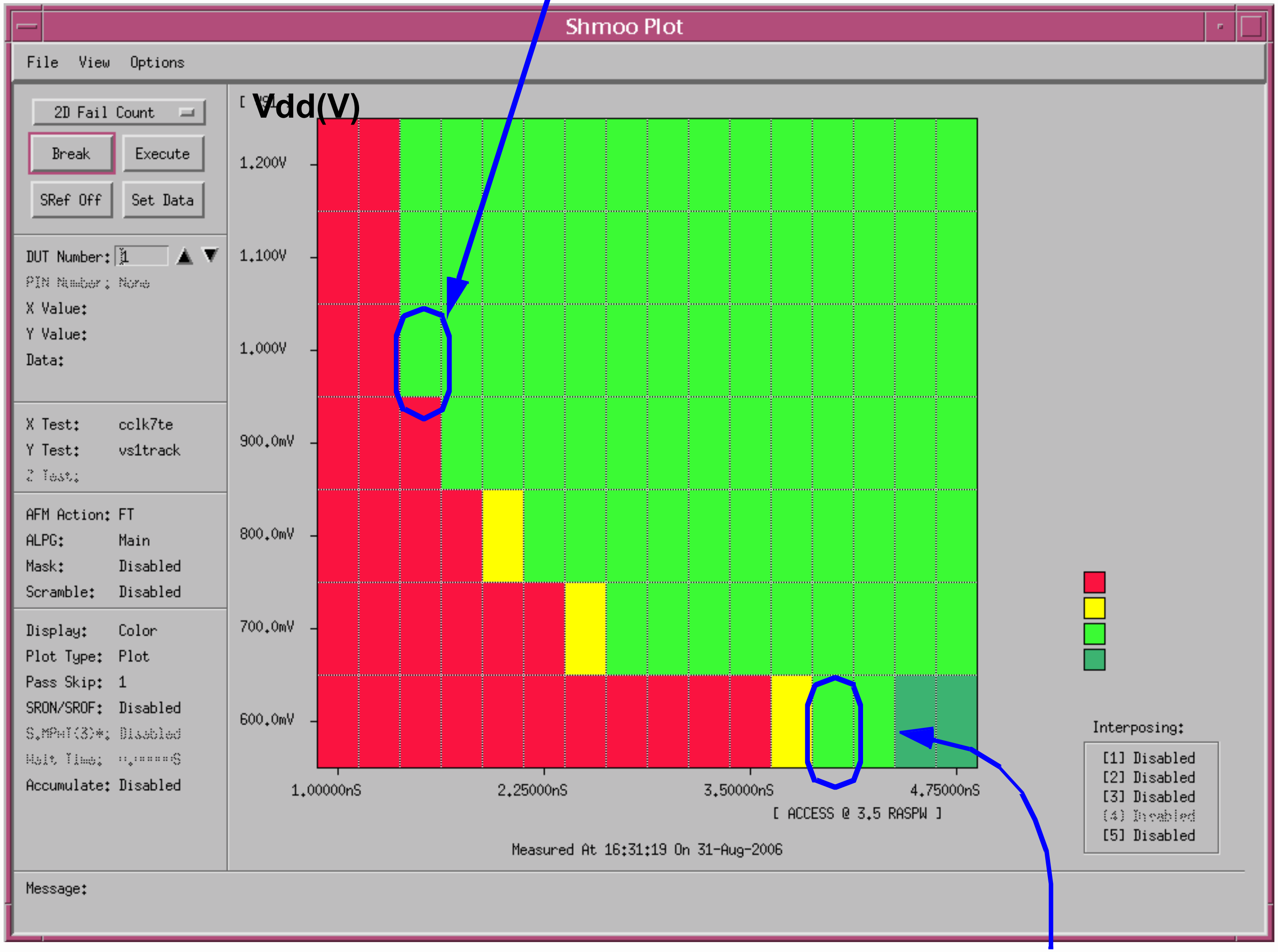
Utilization =



Mbits/mm^2

Access Shmoo

1.5ns Access @1V 85C



DUT Number: 1

PIN Number: None

X Value:

Y Value:

Data:

X Test: cclk7te

Y Test: vs1track

Z Test:

AFM Action: FT

ALPG: Main

Mask: Disabled

Scramble: Disabled

Display: Color

Plot Type: Plot

Pass Skip: 1

SRON/SROF: Disabled

S.MPHT(3)*: Disabled

Wait Time: [illegible]

Accumulate: Disabled

Measured At 16:31:19 On 31-Aug-2006

Interposing:

[1] Disabled
[2] Disabled
[3] Disabled
[4] Disabled
[5] Disabled

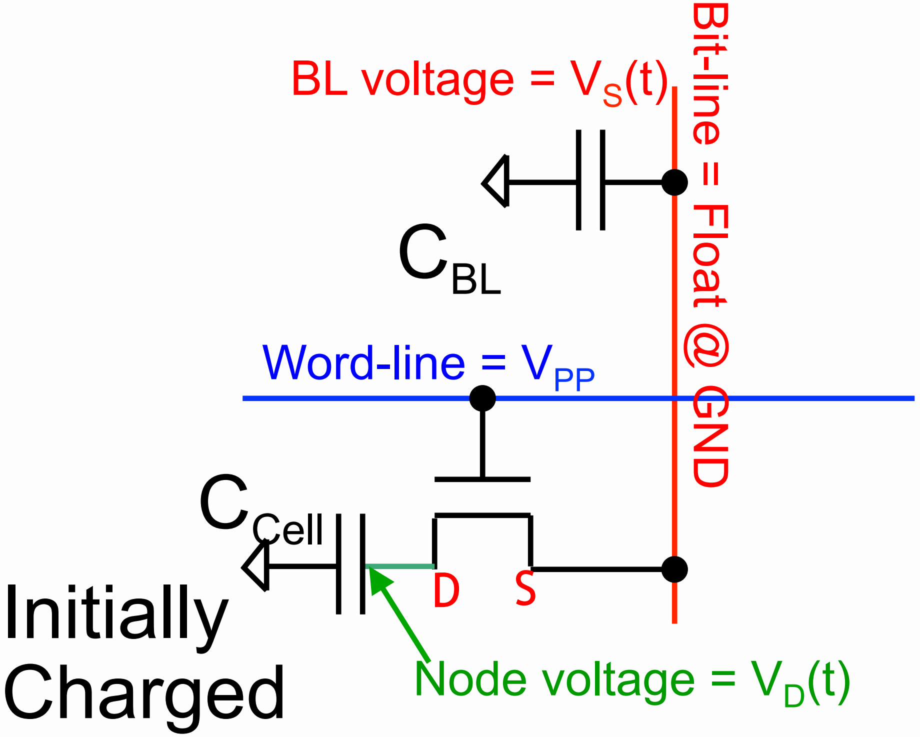
Message:

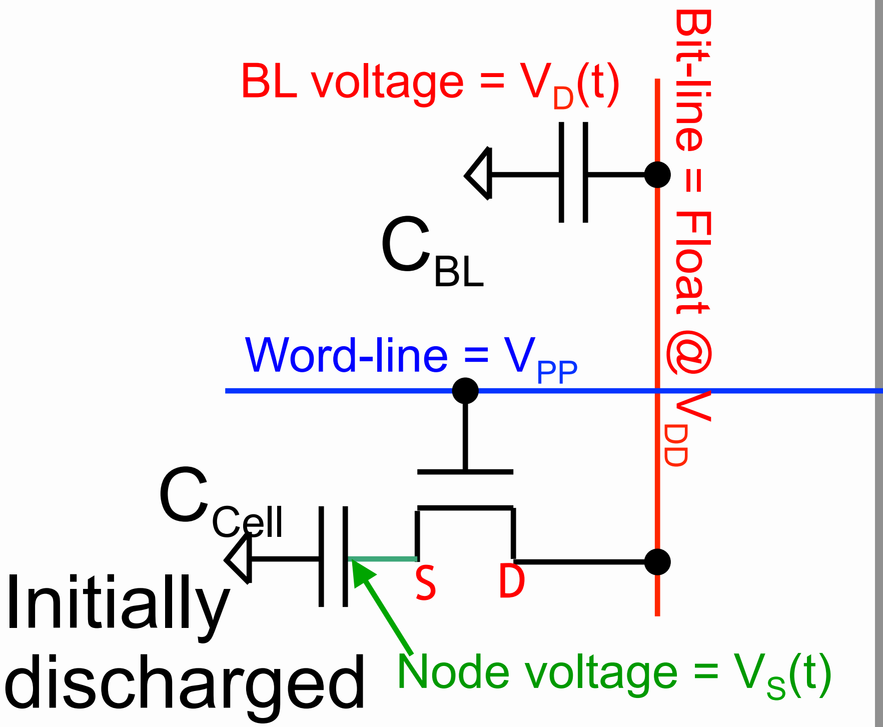
4ns Access @600mV

Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM Write Analysis
- ❑ eDRAM Sense-Amplifier Specification
- ❑ eDRAM operational details (case study)
- ❑ eDRAM Read Analysis
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM

Read Time Calculation





BL – Pre-discharged
Read-1

BL – Pre-charged
Read-0

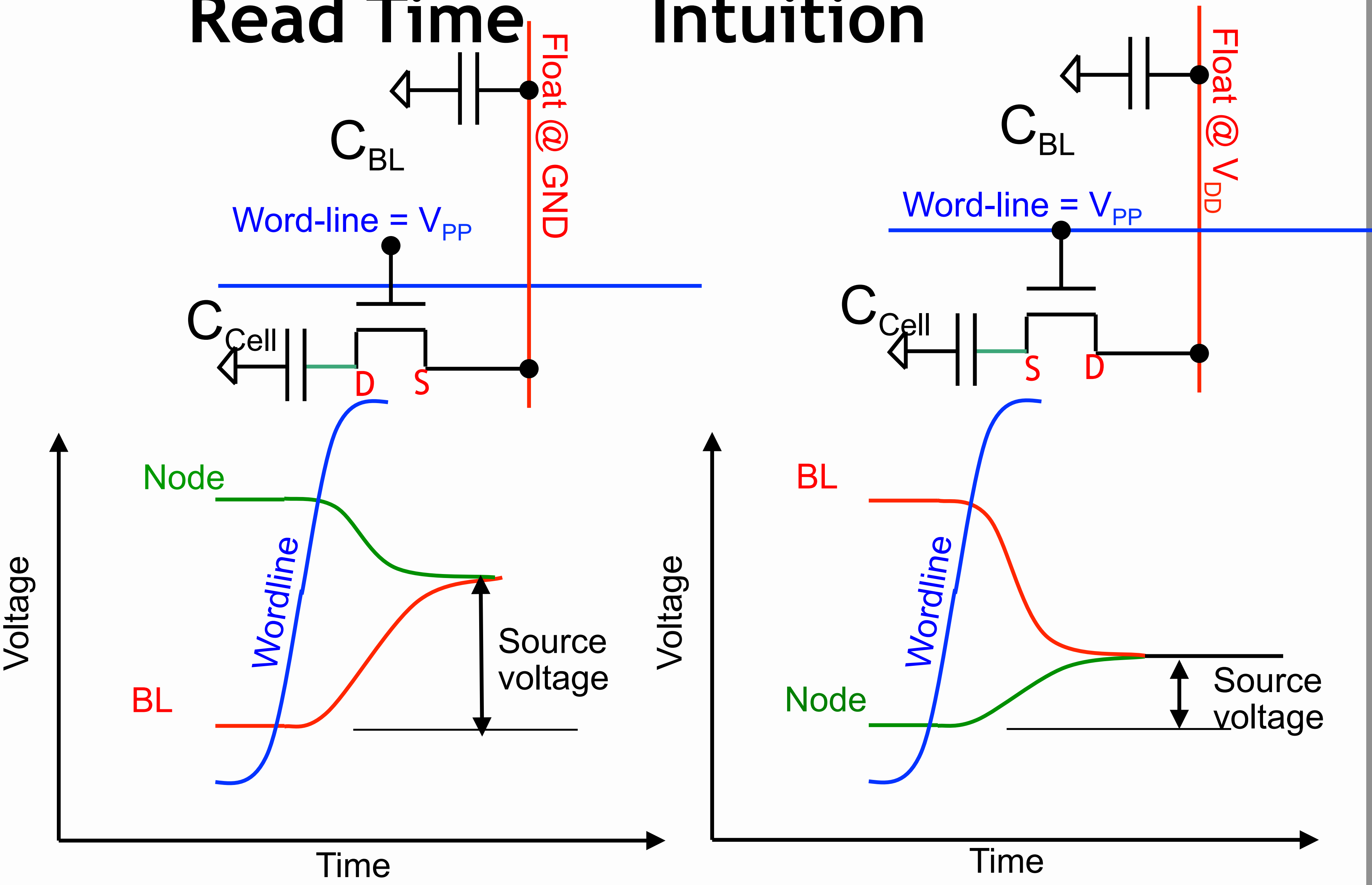
Read time : Let us define it as time required for BL to reach $V_{DD}/2$

Read Time

	BL pre-discharged	BL pre-charged
Source	BL	Node
Drain	Node	BL
Initial Charge	$C_{Cell}V_{DD}$	$C_{BL}V_{DD}$
Charge	$C_{Cell}V_D + C_{BL}V_S = C_{Cell}V_{DD}$	$C_{Cell}V_S + C_{BL}V_D = C_{BL}V_{DD}$
Charging Equation	$I_{DS}[\frac{1}{C_{Cell}} + \frac{1}{C_{BL}}] = -\frac{dV_{DS}}{dt}$	
Current	$I_{DS} = \mu_n C_{OX}\frac{W}{L}V_{DS}(V_{PP} - V_{Tn} - V_S - V_{DS}/2)$	
V_S	$\alpha(V_{DD} - V_{DS})$	$(1-\alpha)(V_{DD} - V_{DS})$
α	$\frac{C_{Cell}}{C_{Cell}+C_{BL}}$	
I_{DS}	$KV_{DS}(\Delta_1 + (\alpha - 0.5)V_{DS})$	$KV_{DS}(\Delta_2 - (\alpha - 0.5)V_{DS})$
K	$\frac{C_{EFF}}{\mu_n C_{OX}(W/L)}$; $C_{EFF} = \frac{C_{Cell}C_{BL}}{C_{Cell}+C_{BL}}$	
Δ	$\Delta_1 = V_{PP} - V_{Tn} - \alpha V_{DD}$	$\Delta_2 = V_{PP} - V_{Tn} - (1-\alpha)V_{DD}$

	BL pre-discharged	**BL pre-charged**
Read Threshold	$V_S = V_{DD}/2$	$V_D = V_{DD}/2$
V_{DS-TH} (V_{FINAL})	$\frac{V_{DD}}{2}(1 - \frac{C_{BL}}{C_{Cell}})$	
T_{WRITE}	$-K \int_{V_{DD}}^{V_{FINAL}} \frac{dV_{DS}}{V_{DS}(\Delta_1+(\alpha-0.5)V_{DS})}$	$-K \int_{V_{DD}}^{V_{FINAL}} \frac{dV_{DS}}{V_{DS}(\Delta_2-(\alpha-0.5)V_{DS})}$
T_{WRITE}	$\frac{K}{\Delta_1} ln(\frac{V_{DD}(\Delta_1+(\alpha-0.5)V_{FINAL})}{V_{FINAL}(\Delta_1+(\alpha-0.5)V_{DD})})$	$\frac{K}{\Delta_2} ln(\frac{V_{DD}(\Delta_2+(0.5-\alpha)V_{FINAL})}{V_{FINAL}(\Delta_2+(0.5-\alpha)V_{DD})})$

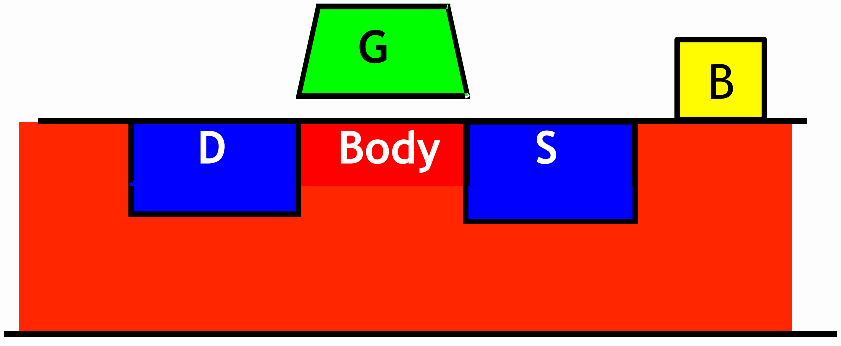
Read Time Intuition

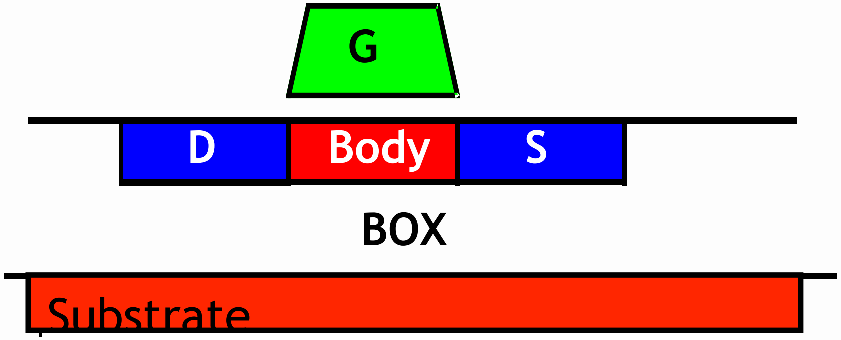


Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM Write Analysis
- ❑ eDRAM Sense-Amplifier Specification
- ❑ eDRAM operational details (case study)
- ❑ eDRAM Read Analysis
- ❑ SOI Technology
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM

Bulk vs SOI Technology





Body contact can be used to fix the body potential in bulk technology

SOI Technology

- Floating body is a problem
- History effect
- SOI provides better sub-threshold slope
- Higher performance with lower leakage

Floating Body Effects

Body potential modulated by coupling and leakage

Better source follower vs. bulk during write back (body coupling)

  Improved write ‘1’ cell voltage

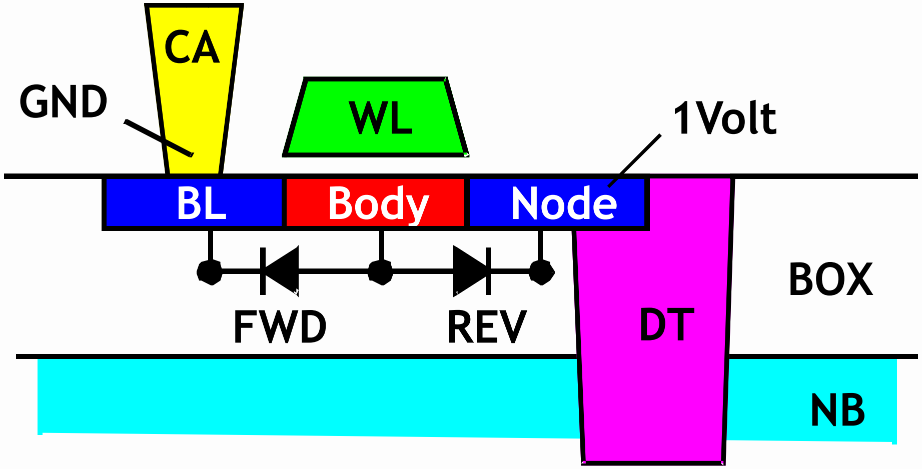
Degraded I_{off}/ Retention if body floats high (body leakage)

  GND pre-charge keeps body low

  Eliminate long periods with BL high (limit page mode)

$\text{ILeak}_{FWD} > \text{ILeak}_{REV}$

When BL = GND
Body ~ GND



JSSC08

Floating Body Effects

Body potential modulated by coupling and leakage

Better source follower vs. bulk during write back (body coupling)

- Improved write ‘1’ cell voltage

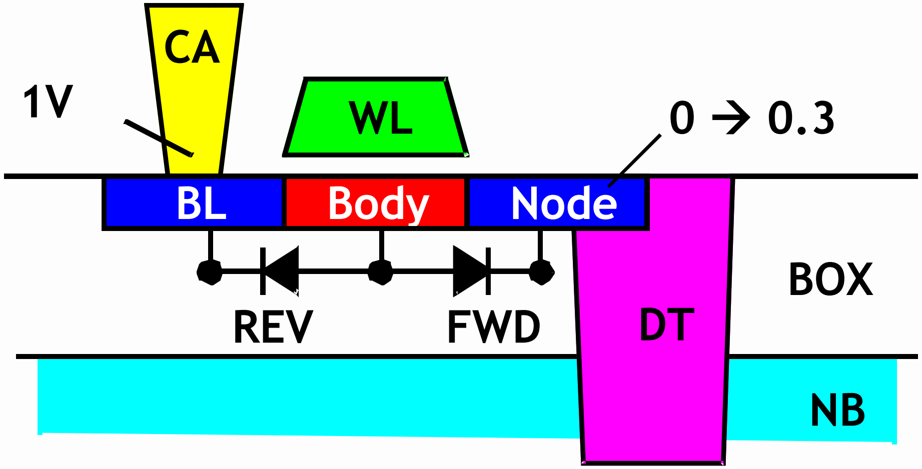
Degraded I_{off}/ Retention if body floats high (body leakage)

- GND pre-charge keeps body low
- Eliminate long periods with BL high (limit page mode)

$ILeak_{FWD} > ILeak_{REV}$

When BL = GND

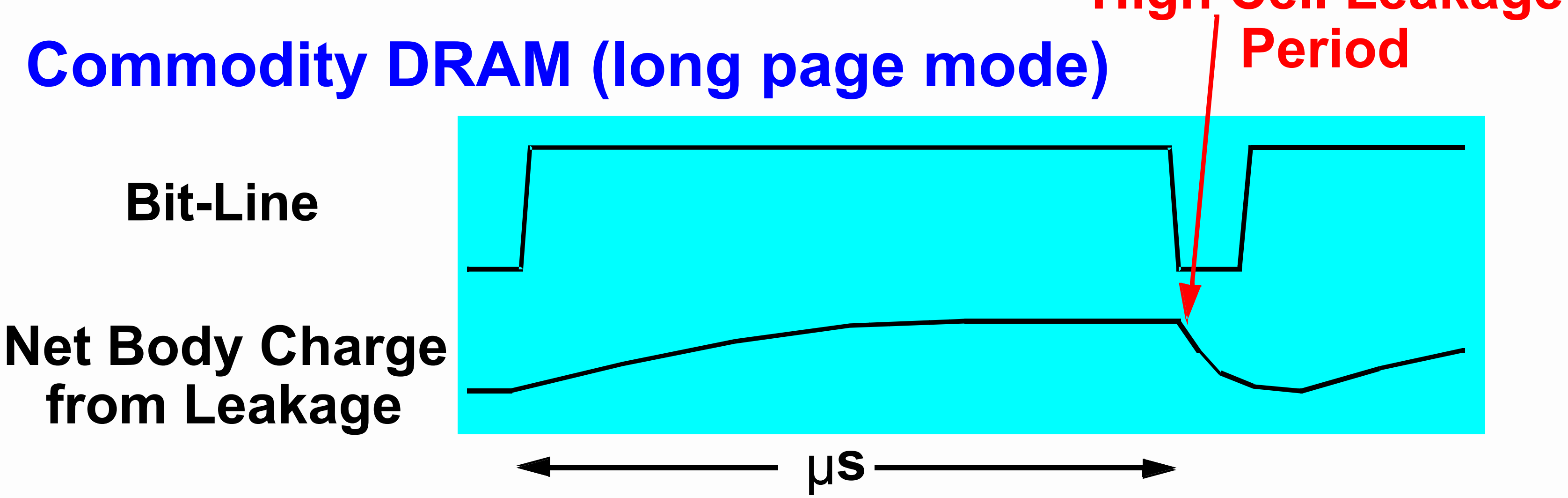
Body ~ GND



JSSC08

Array Body Charging

Commodity DRAM (long page mode)



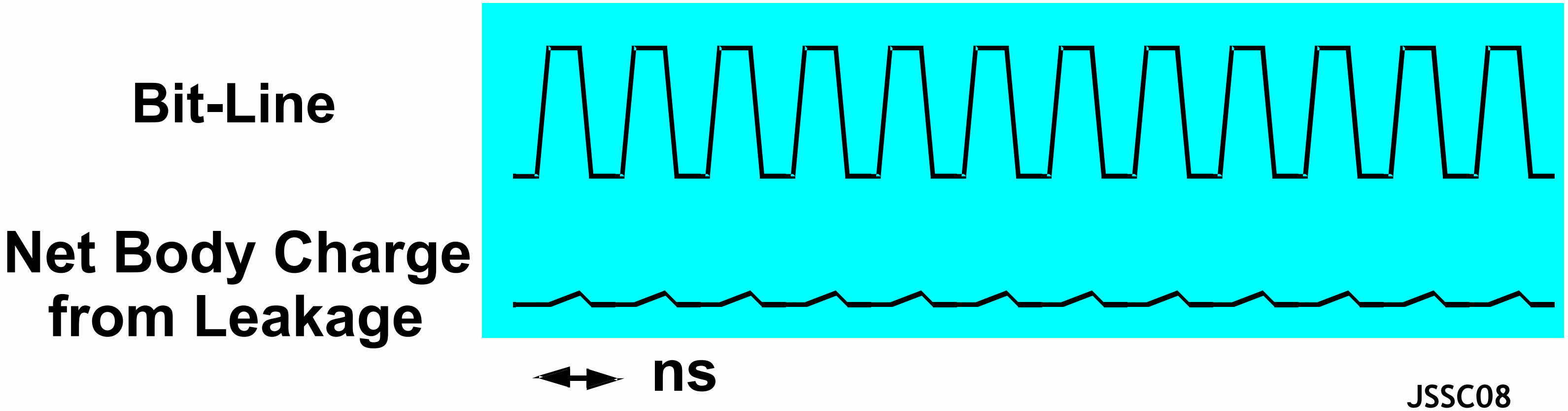
High Cell Leakage Period

Bit-Line

Net Body Charge from Leakage

µs

embedded DRAM (limited page mode)



Bit-Line

Net Body Charge from Leakage

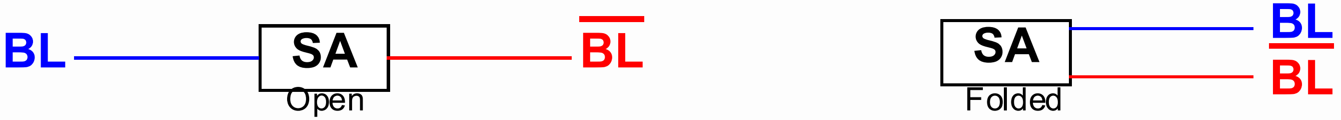
ns

JSSC08

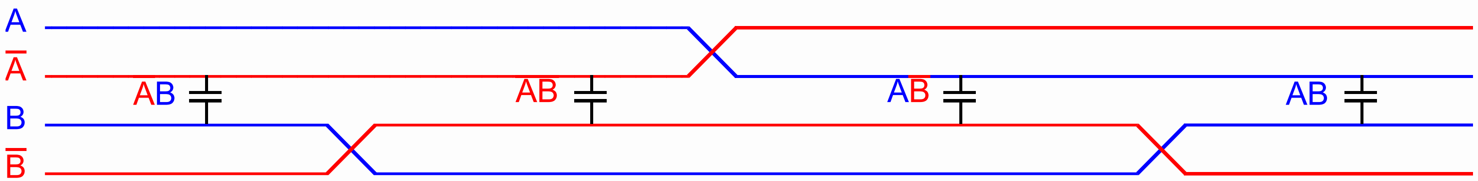
Noise

Coupling and Local Process Variation effectively degrades signal

External Noise (Wire or Sx) Reduced to Common Mode by Folding



Line to Line Coupling Limited by Bit-Line Twisting



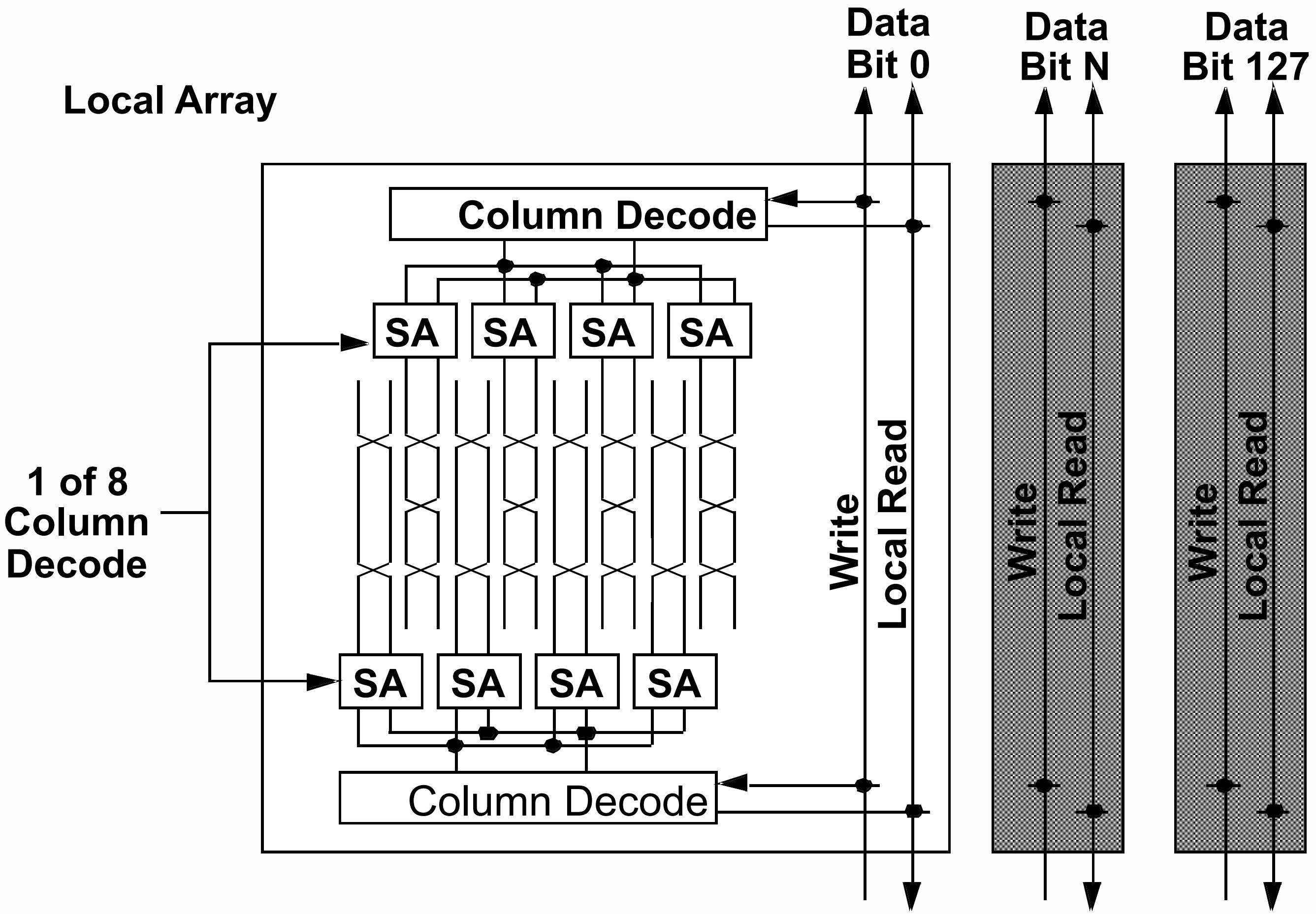
A Couples Equally into B and $\overline{B}$

V_t and DL Mis-Match Limited by Longer Channel Length

Overlay Mis-Alignment Limited by Identical Orientation

Capacitive Mis-Match Limited by careful Physical Design (Symmetry)

Interleaved Sense Amp w/ Bit-Line Twist



CMOS VLSI design - PEARSON

Open and Folded Bitline Schematic

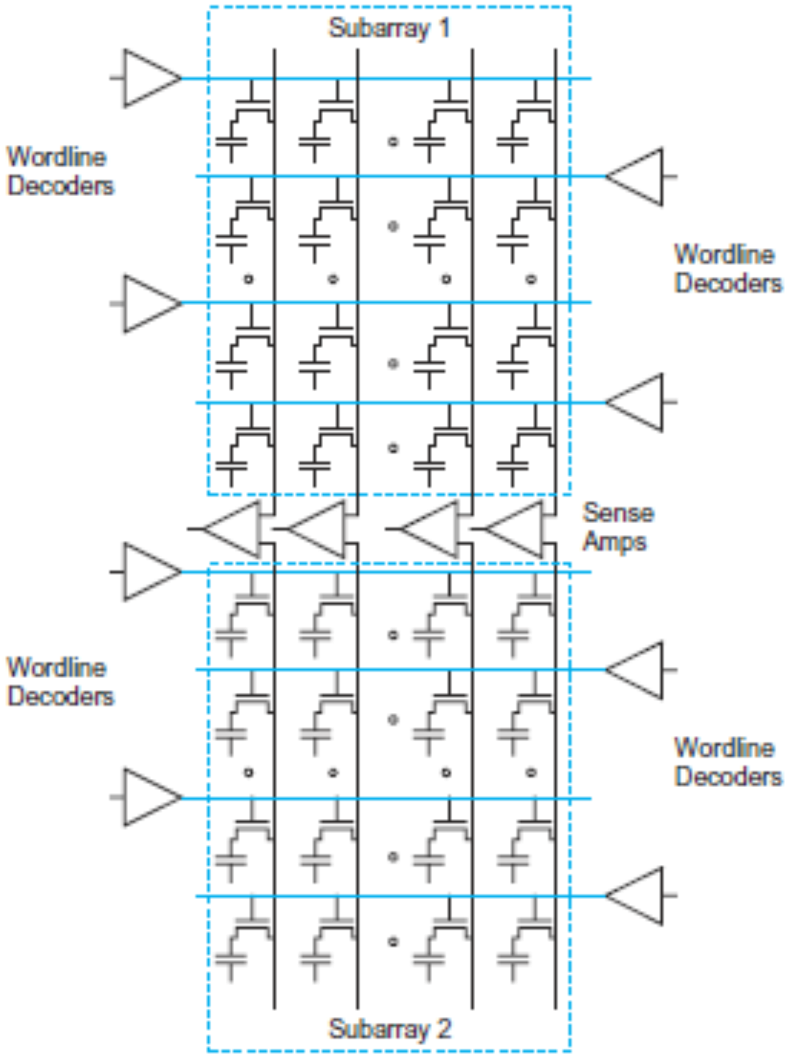


FIGURE 12.44 Open bitlines

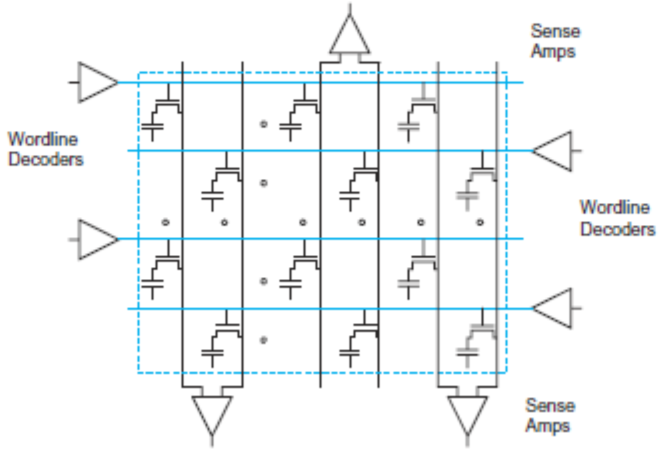


FIGURE 12.45 Folded bitlines

CMOS VLSI design - PEARSON

Folded Bitline Layout

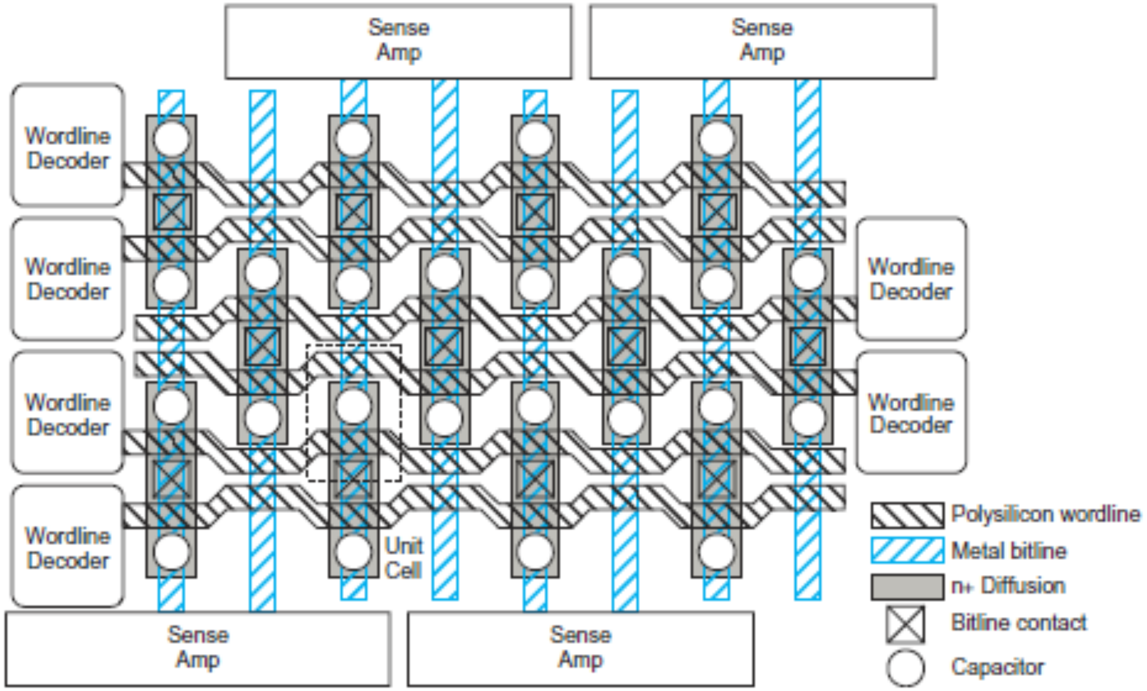


FIGURE 12.46 Layout of folded bitline subarray

CMOS VLSI design - PEARSON

Topics

□ Introduction to memory

□ DRAM basics and bitcell array

□ eDRAM operational details (case study)

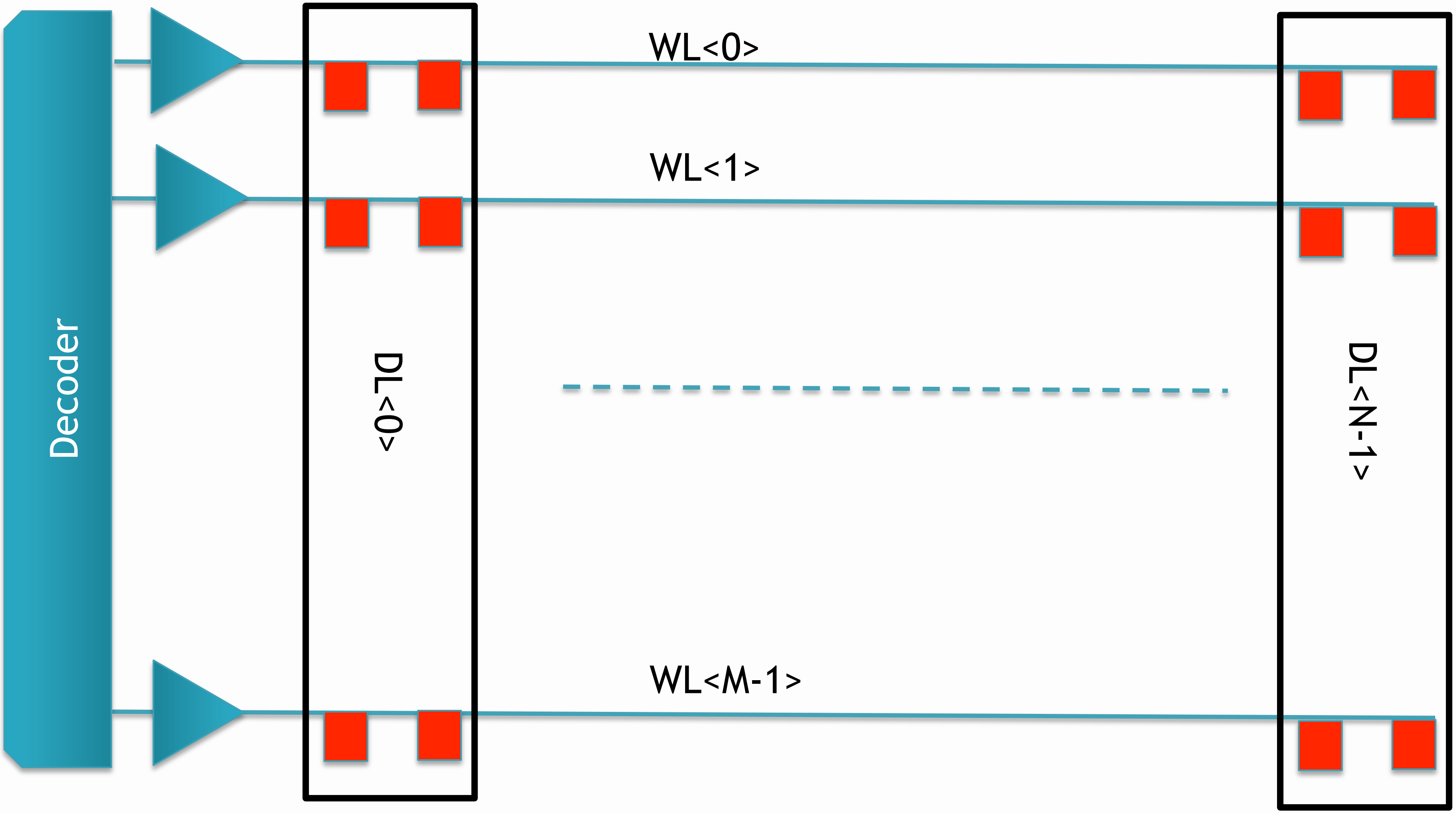
□ Noise concerns

□ Wordline driver (WLDRV) and level translators (LT)

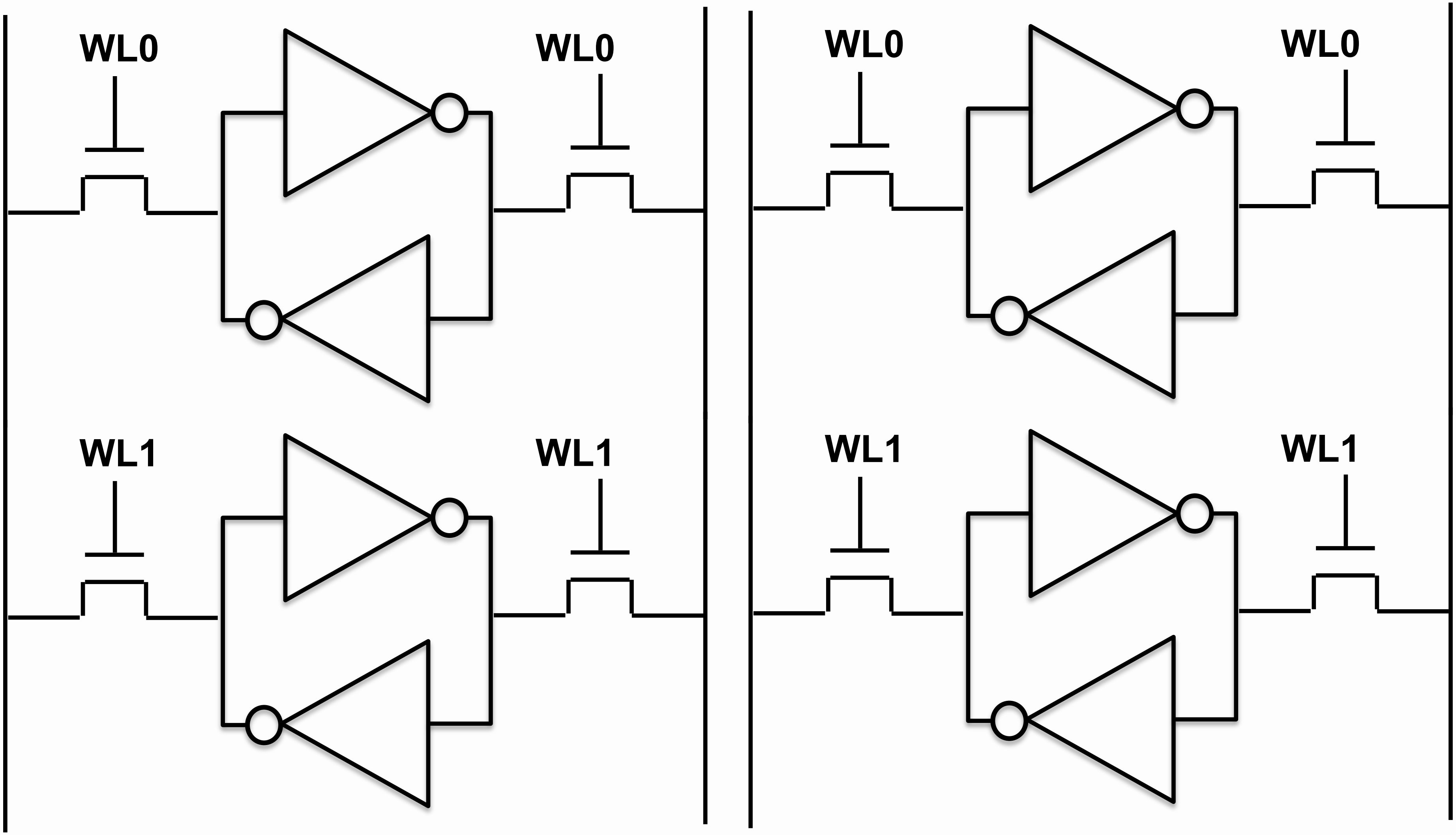
□ Challenges in eDRAM

□ Understanding Timing diagram – An example

SRAM – Word-line Driver

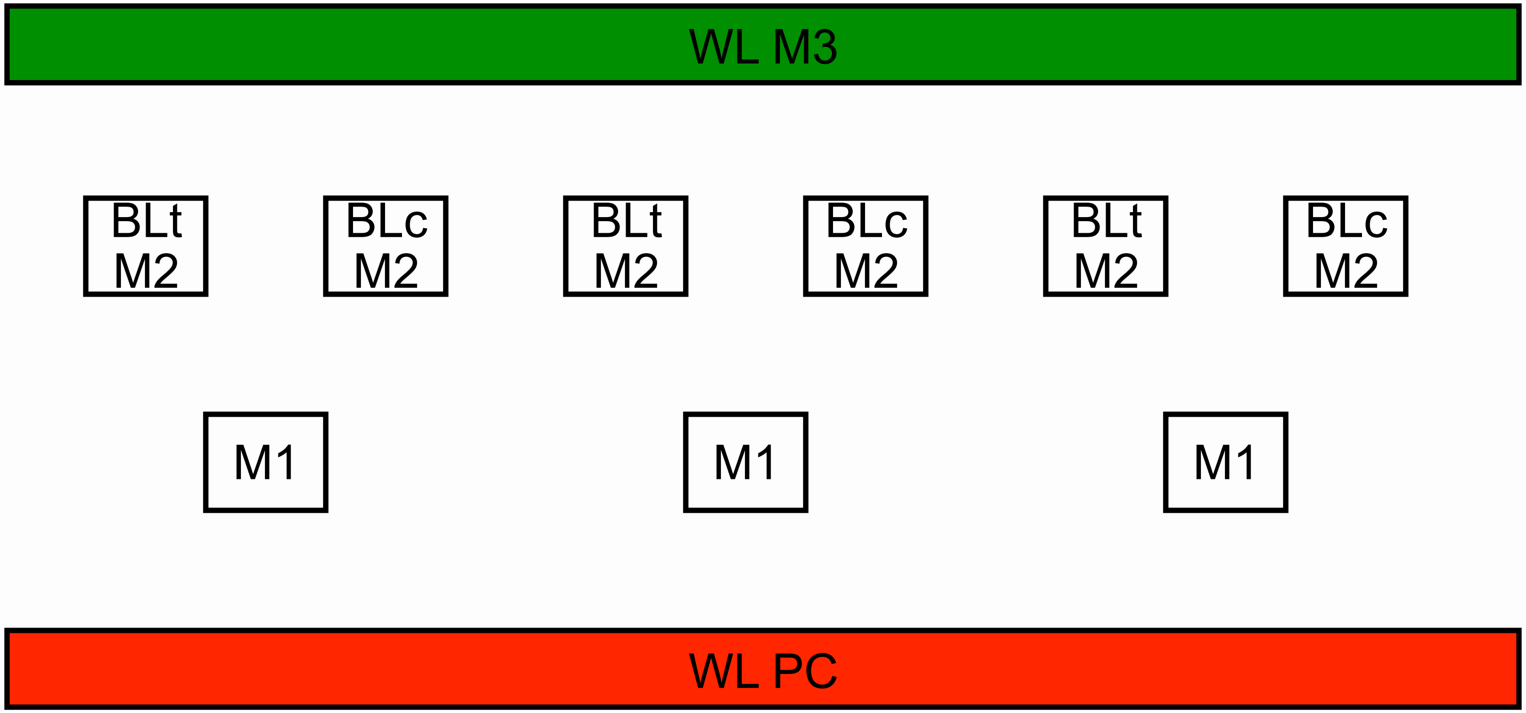


WL Capacitance Load



Large load capacitance = #DL x #Columns x Access Transistor Capacitance

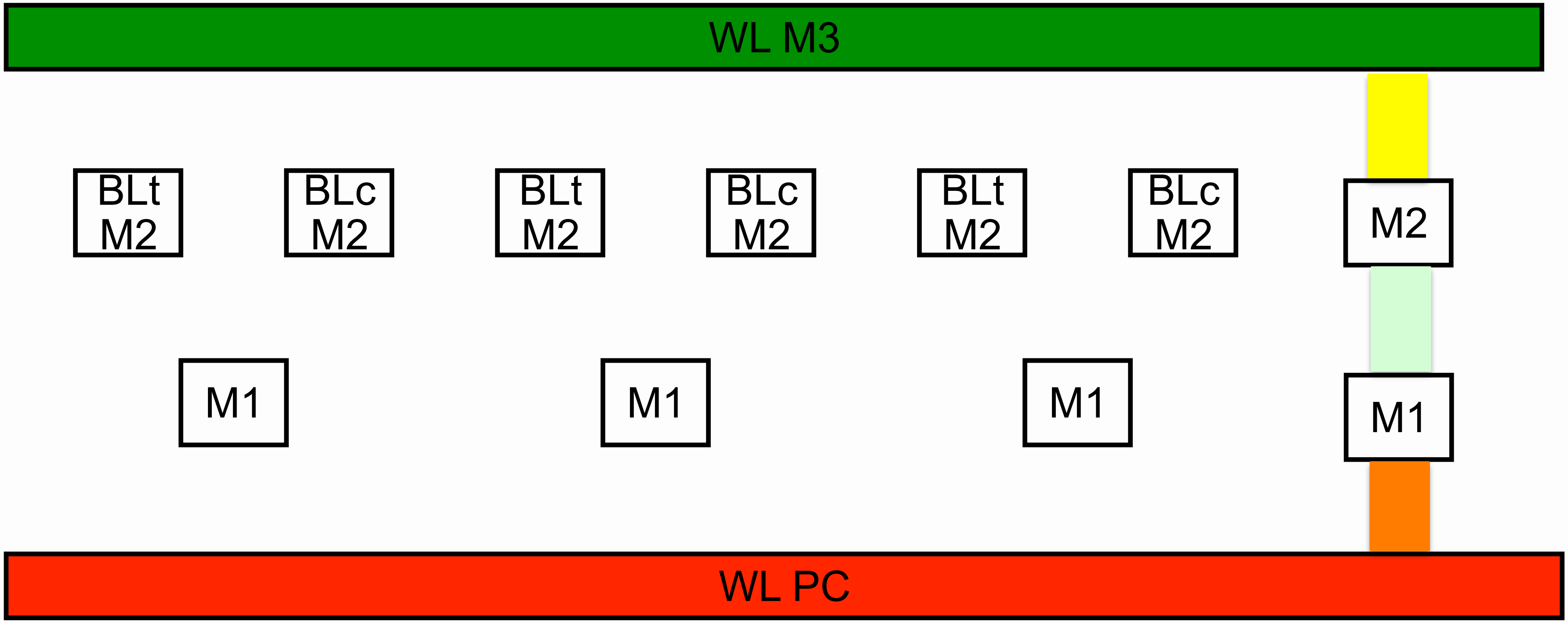
WL – Metal Capacitance Load



RC delay due to Metal 3 + Poly

Length = #DL x #Columns x width of each cell?

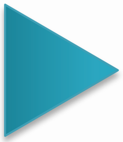
WL – Metal Capacitance Load

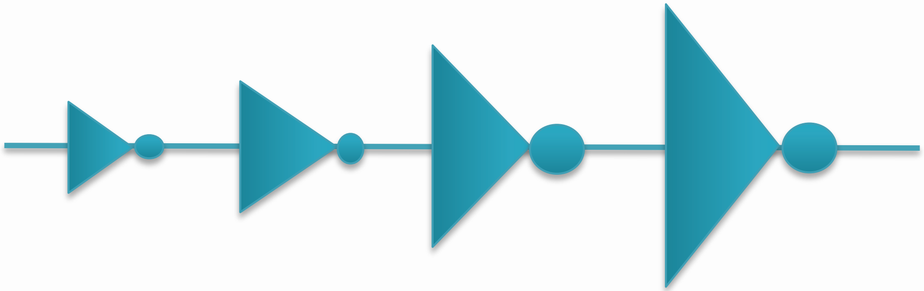


Stitch cell – Connect M3 to PC

Length > #DL x #Columns x width of each cell?

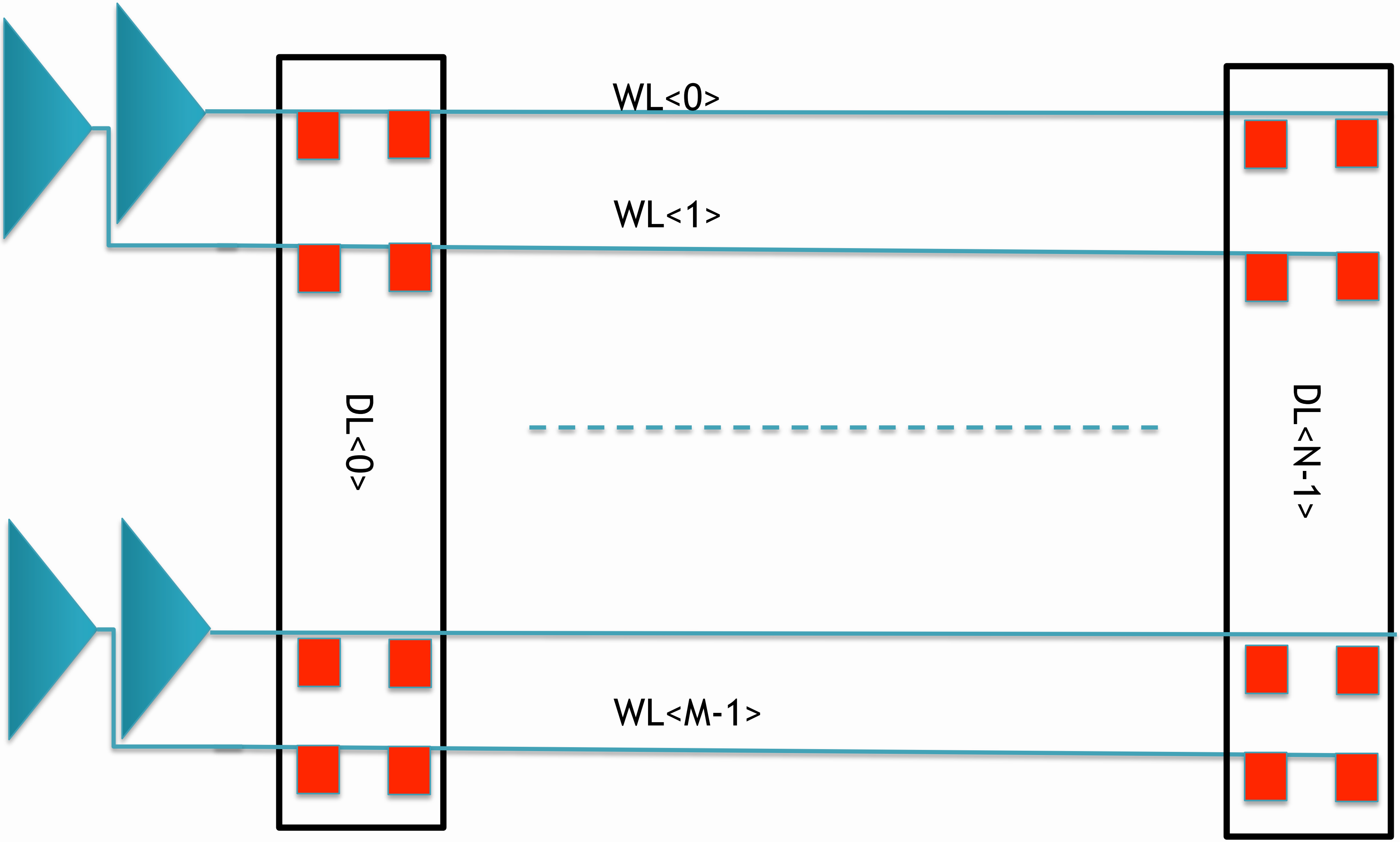
Word Line Driver



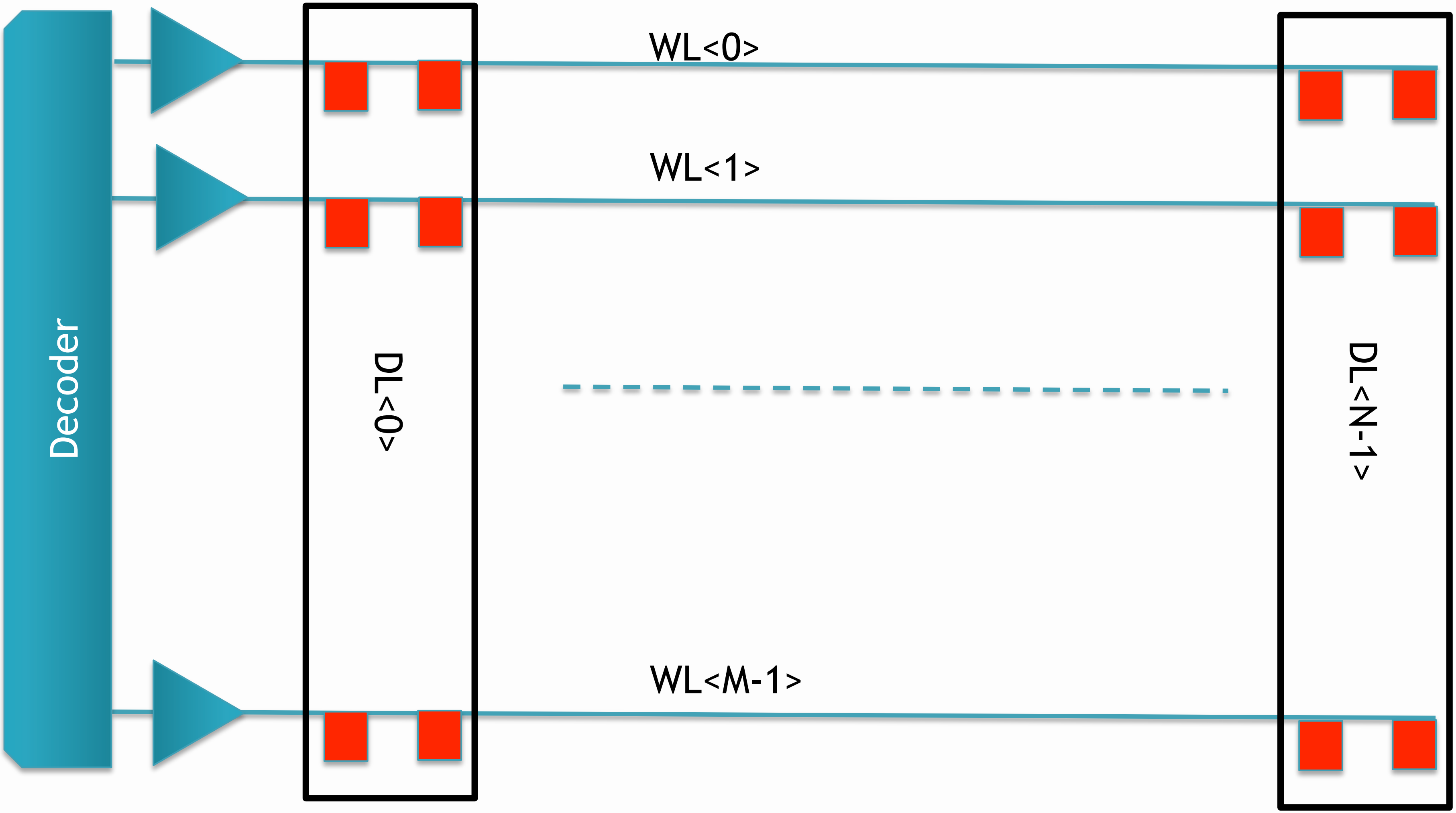


- Need to drive a large load capacitance
- For SRAMs – It is nothing but a buffer
- Need to accommodate within y direction of a cell

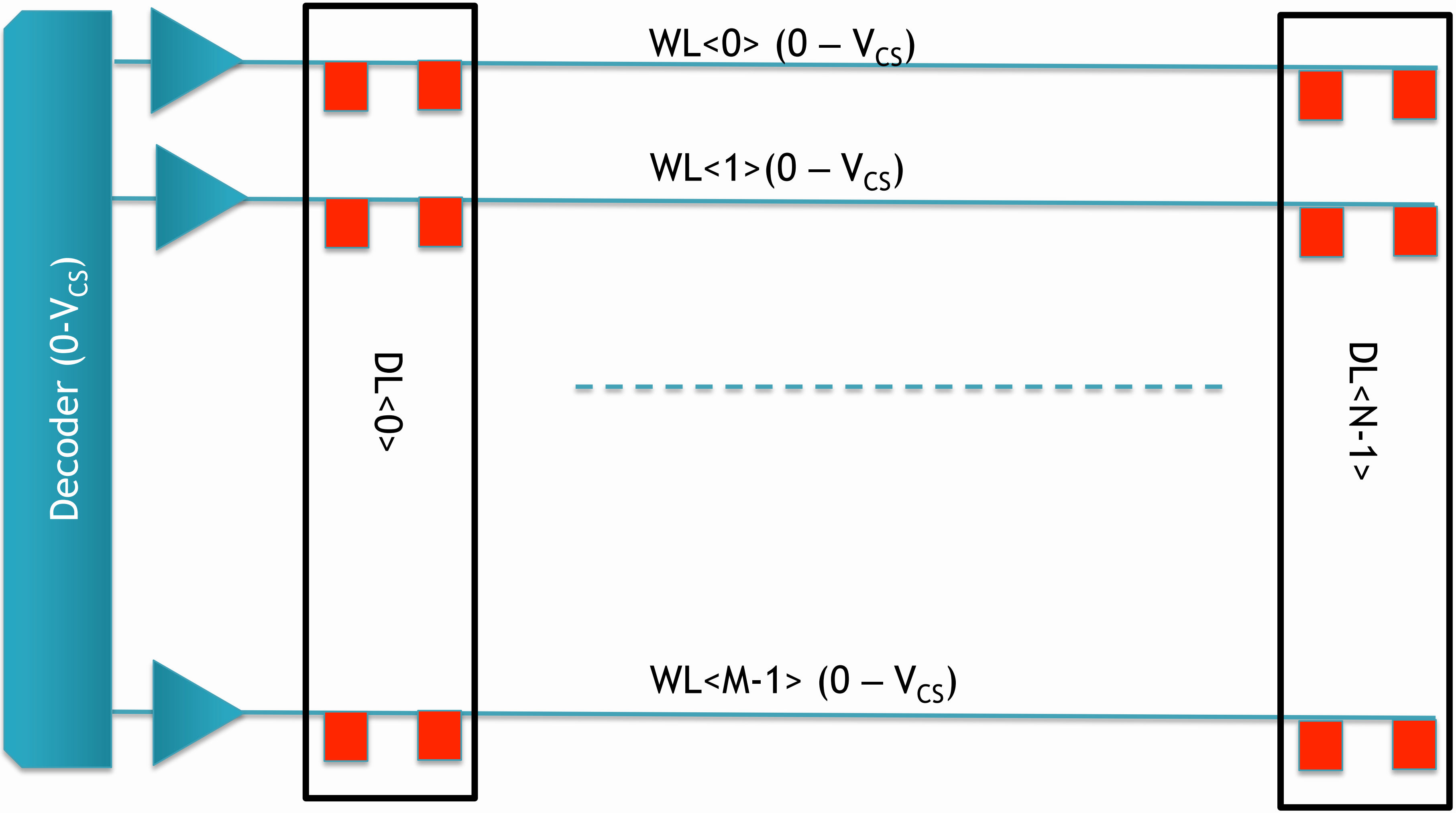
SRAM - WLDRV



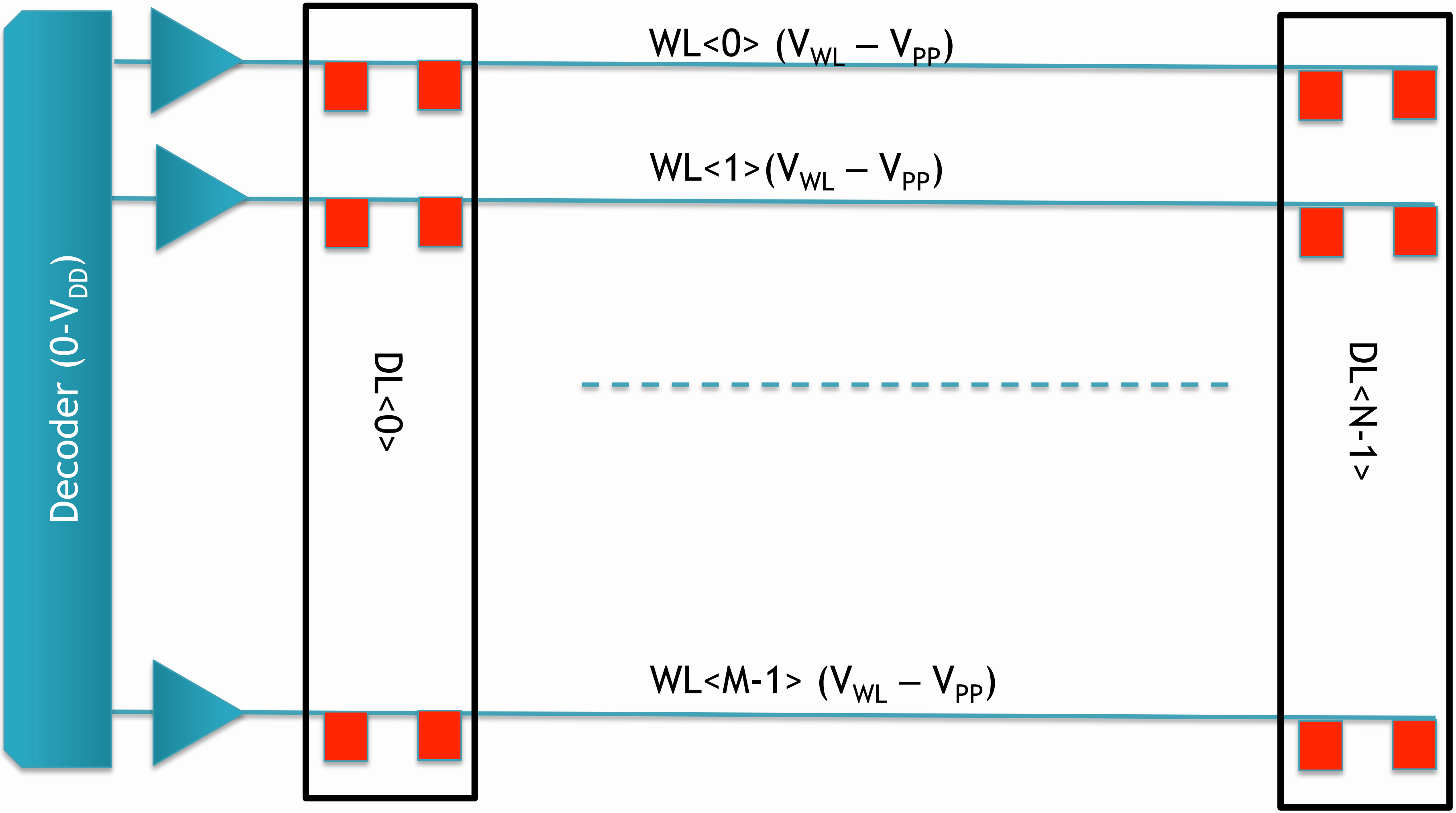
SRAM – Word-line Driver



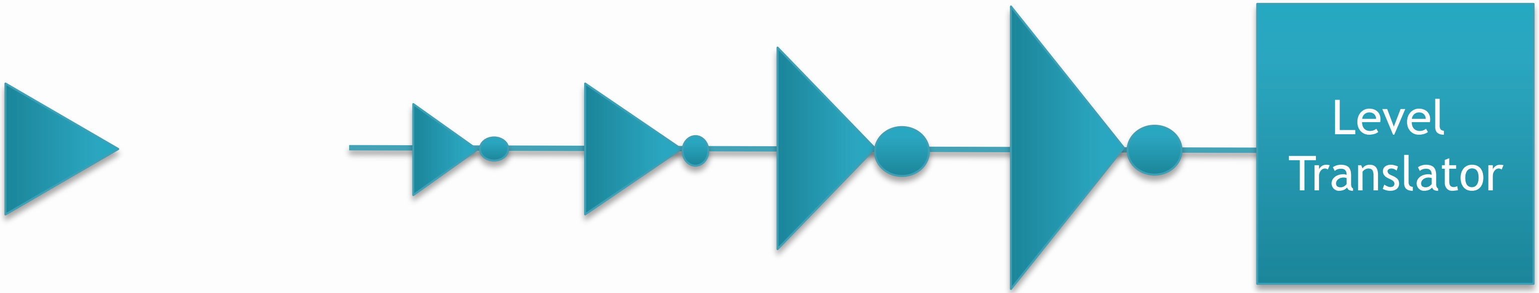
SRAM – Word-line Driver



eDRAM – Word-line Driver

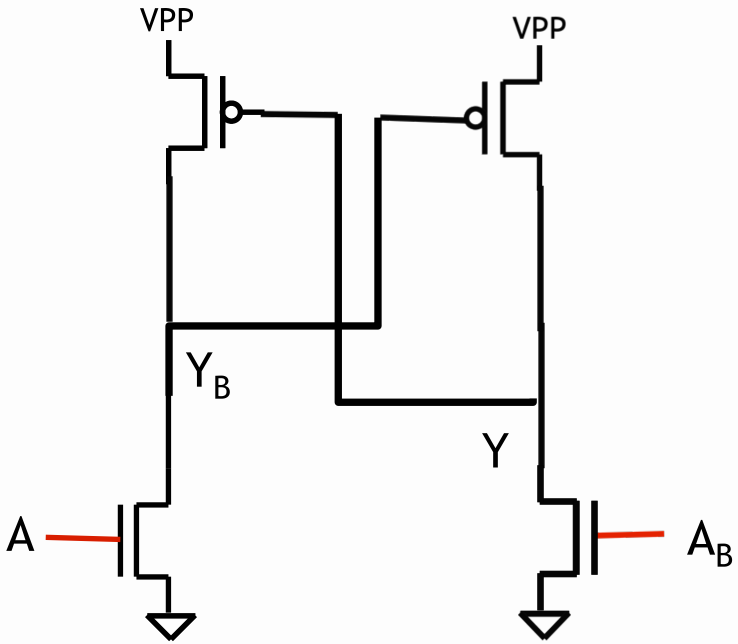

Decoder (0-V_{DD})
WL<0> (V_{WL} – V_{PP})
WL<1>(V_{WL} – V_{PP})
WL<M-1> (V_{WL} – V_{PP})
DL<0>
DL<N-1>

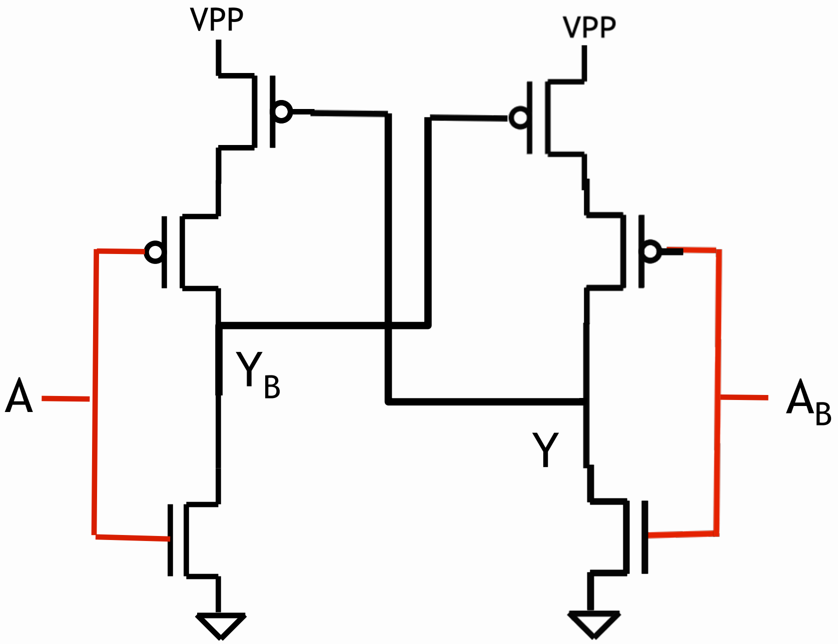
eDRAM - Word Line Driver



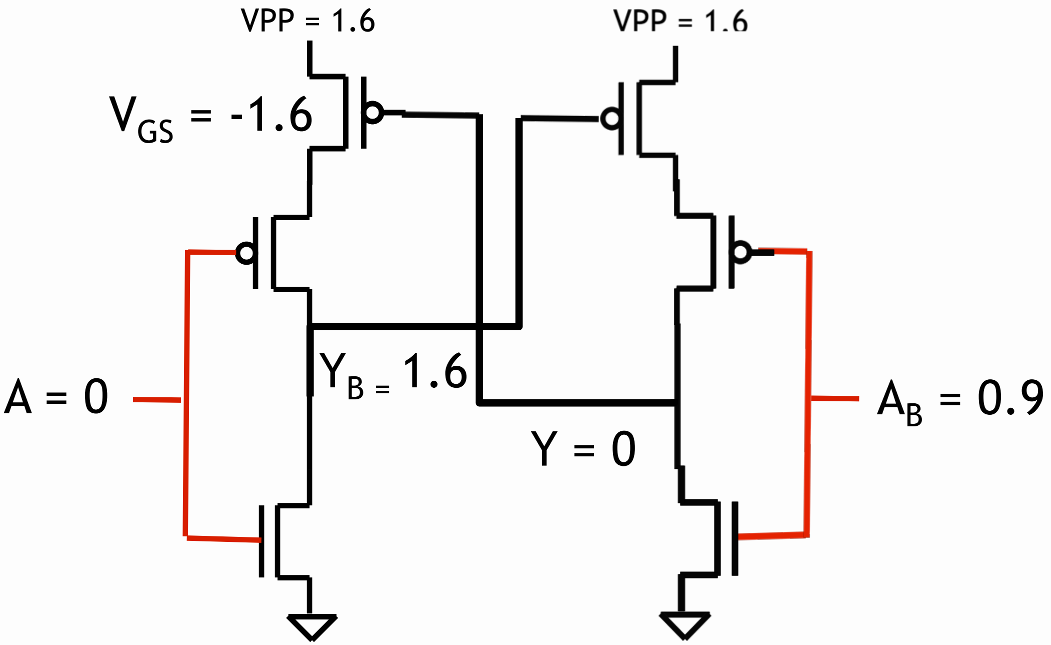
- Need to drive a large load capacitance
- Need to accommodate within y direction of a cell
- Need to level translate to V_{WL} – V_{PP} domain.

Level Translator

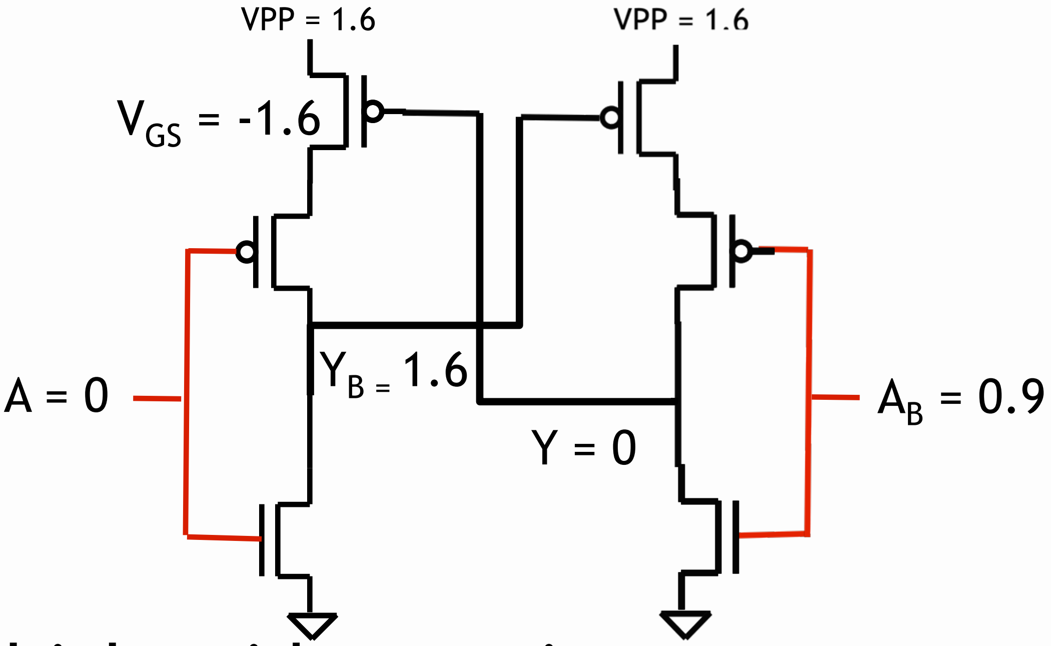




Level Translator

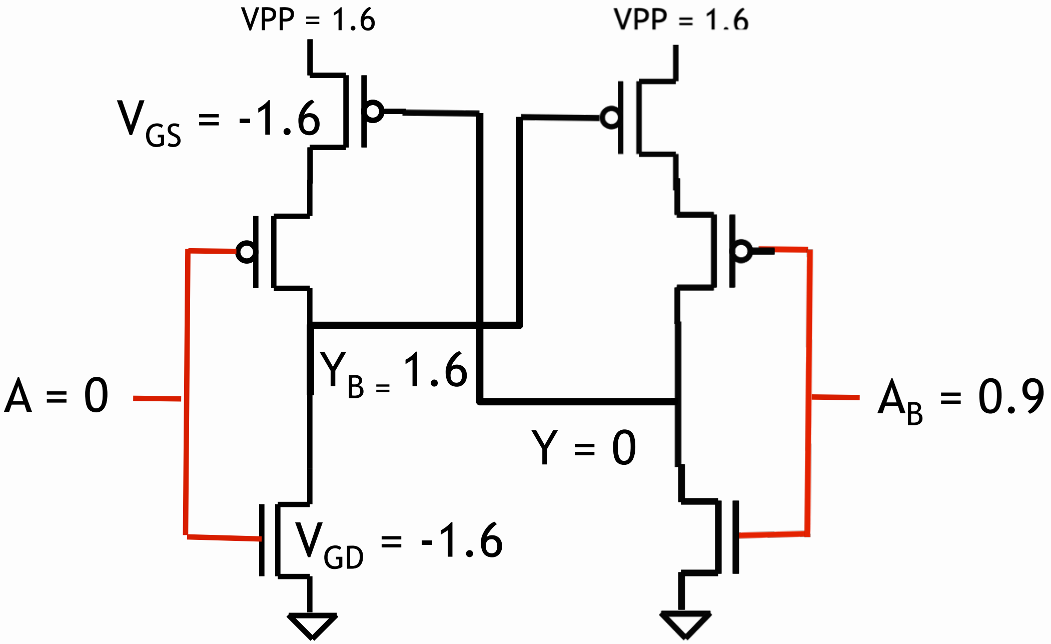


Level Translator - Stress

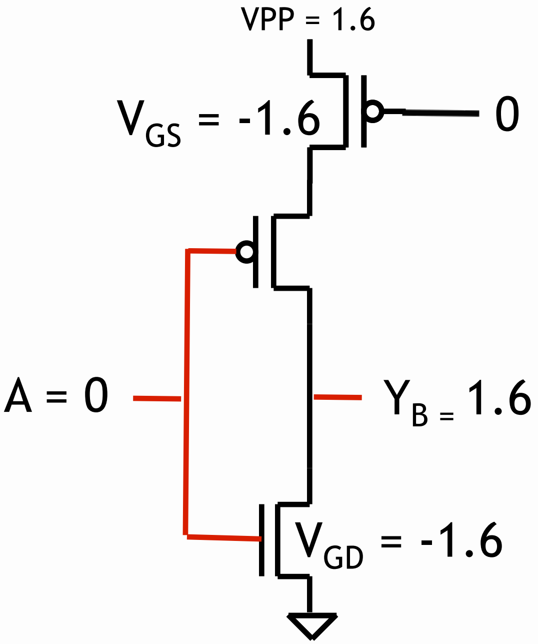


- Need thick oxide transistors
  - Large Area
  - Does not track the logic transistor process
  - Hard to integrate with the array
- Need thin oxide WL drivers

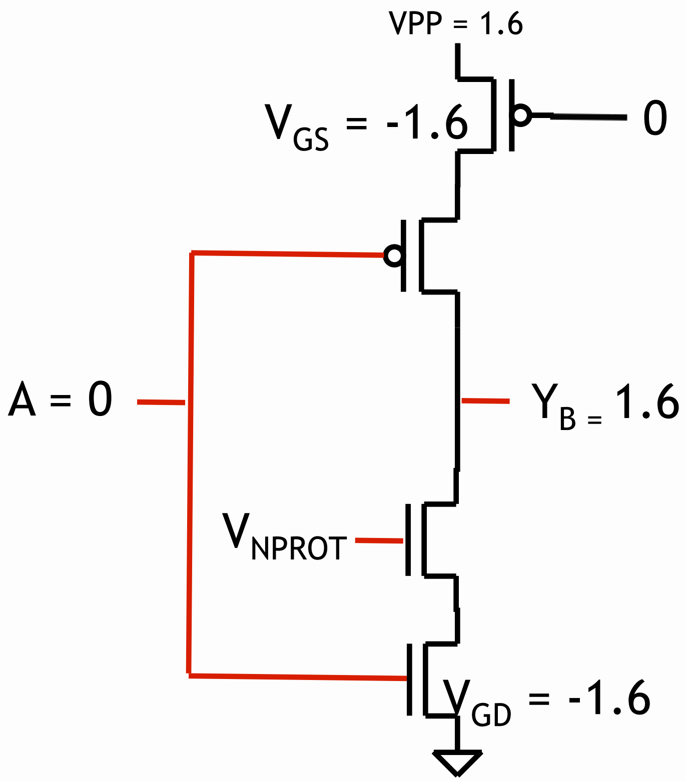
Level Translator - Stress



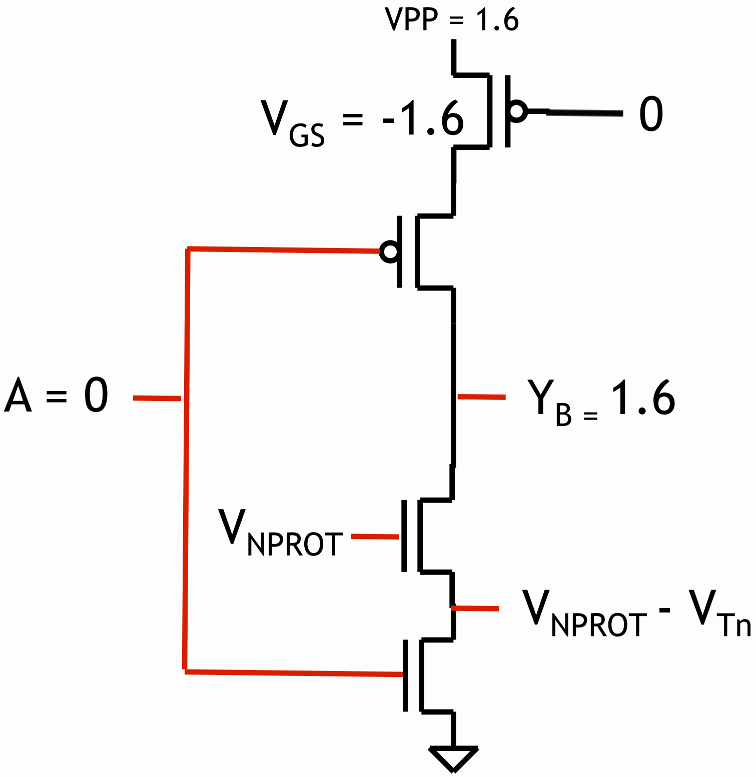
Level Translator - Stress



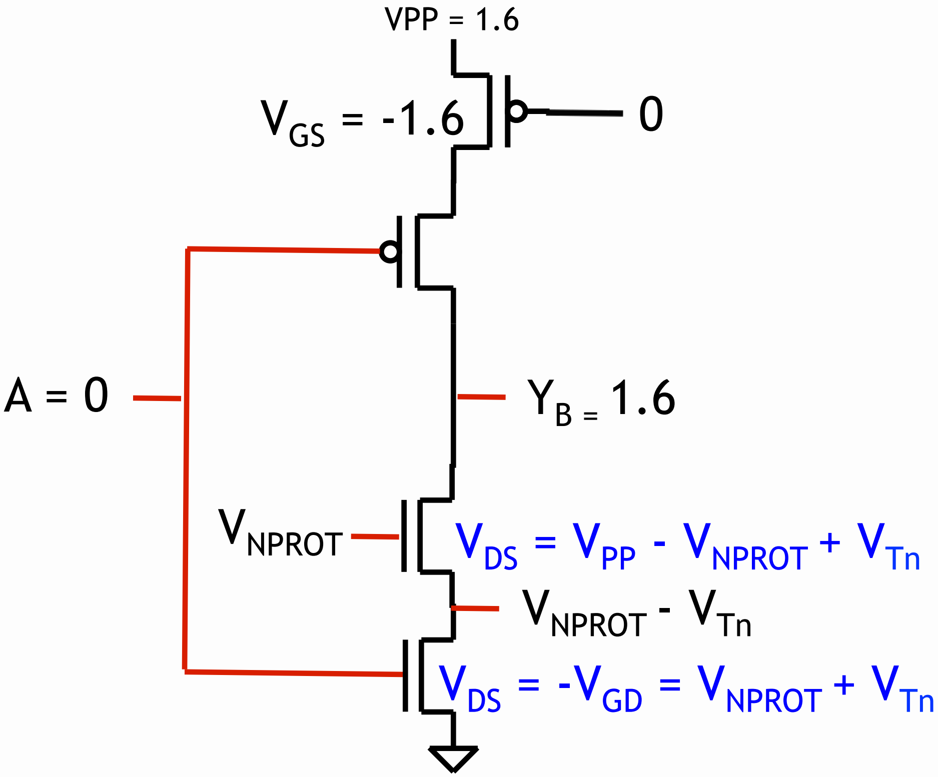
Protect Transistor



Protect Transistor

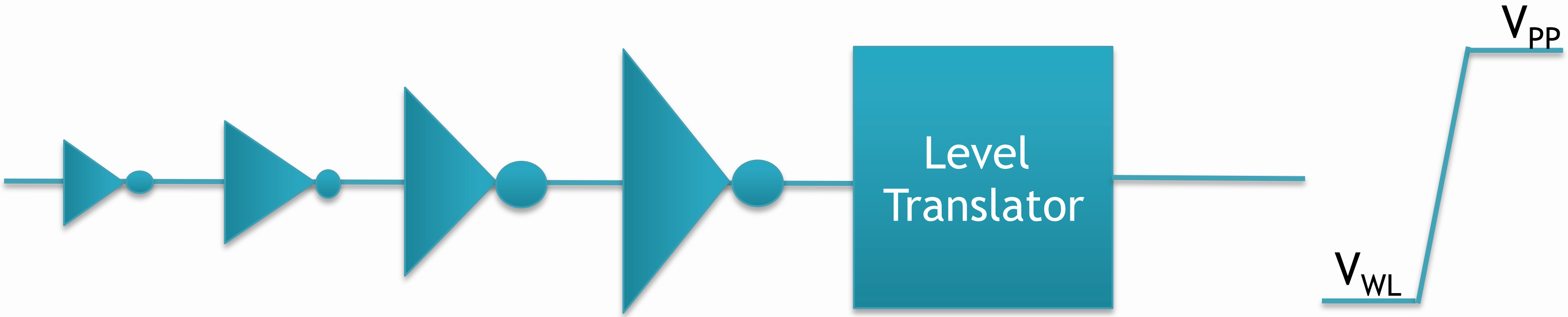


Protect Transistor


VPP = 1.6
V_{GS} = -1.6
0
A = 0
Y_B = 1.6
V_{NPROT}
$V_{DS} = V_{PP} - V_{NPROT} + V_{Tn}$
$V_{NPROT} - V_{Tn}$
$V_{DS} = -V_{GD} = V_{NPROT} + V_{Tn}$

V_{NPROT} = 0.65V

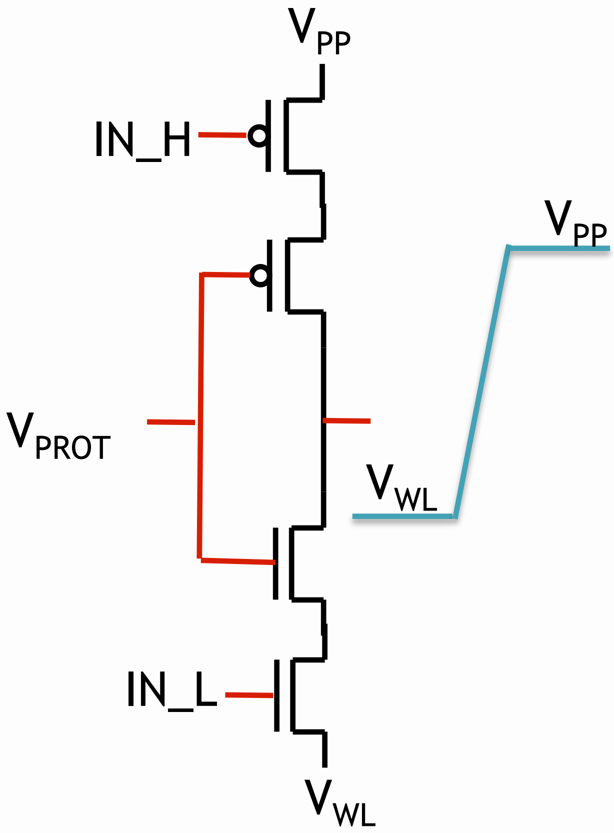
eDRAM - Word Line Driver



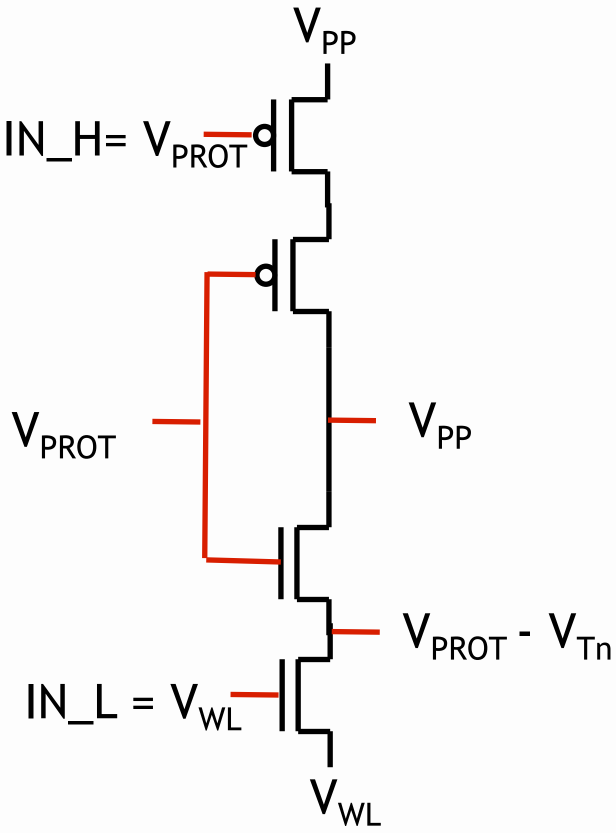
- Start with the Final Stage

WLD – Final Stage

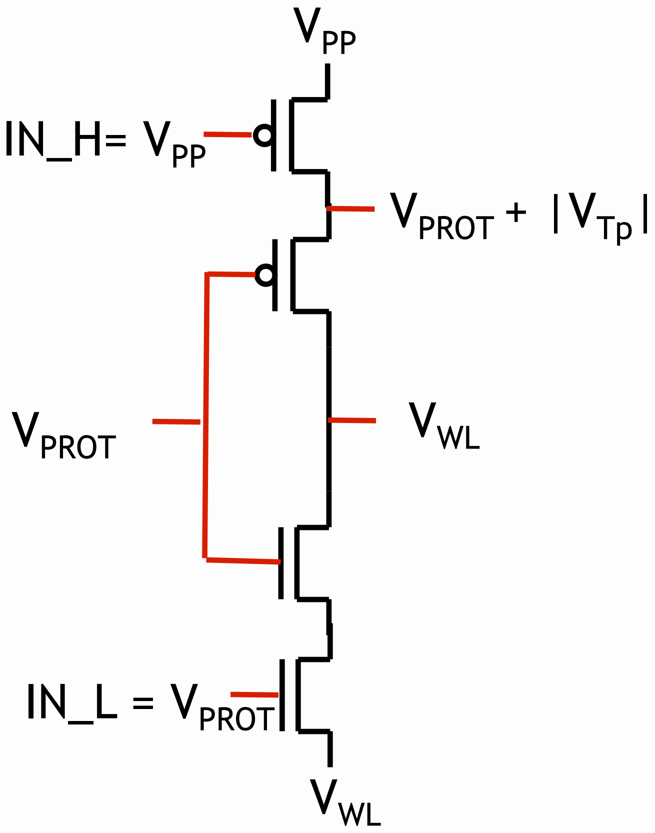
$V_{PROT} = (V_{PP} + V_{WL})/2$



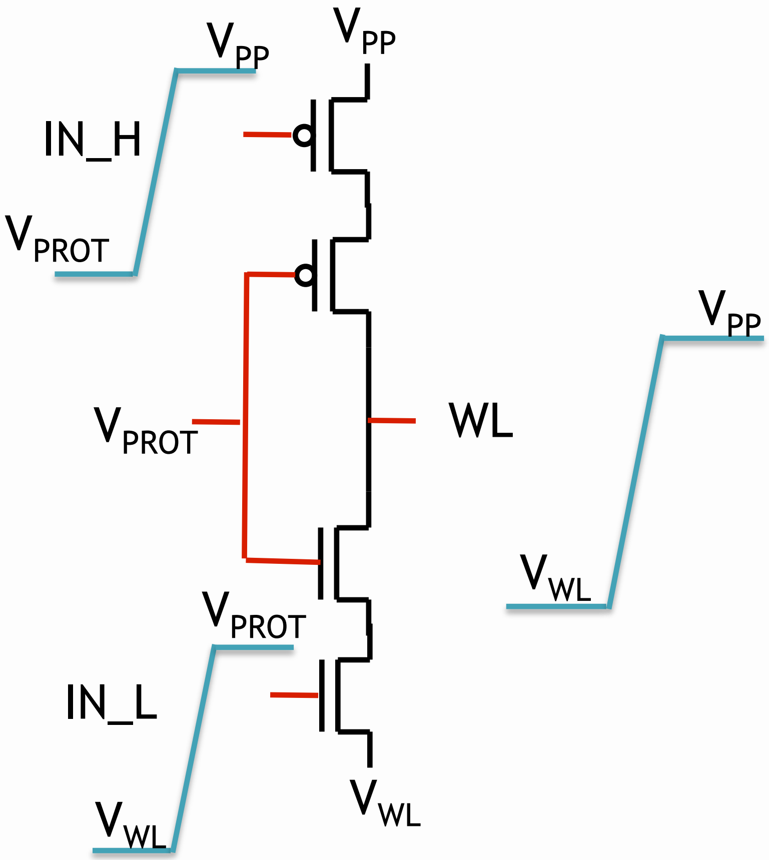
WLD – Rise



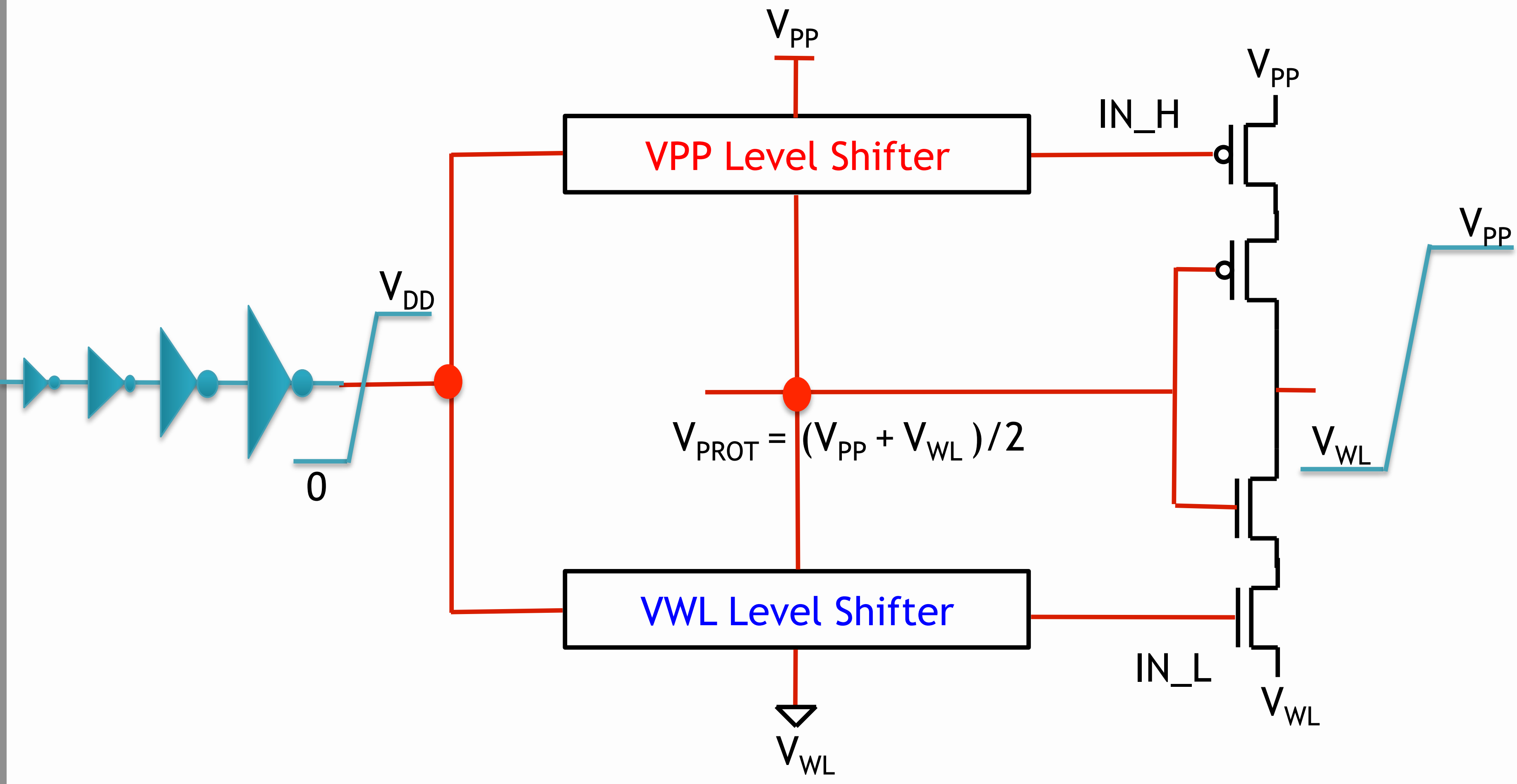
WLD – Fall



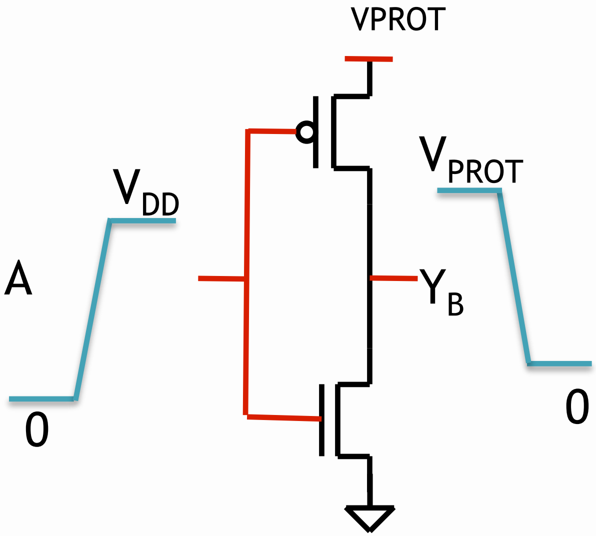
Word-line Driver



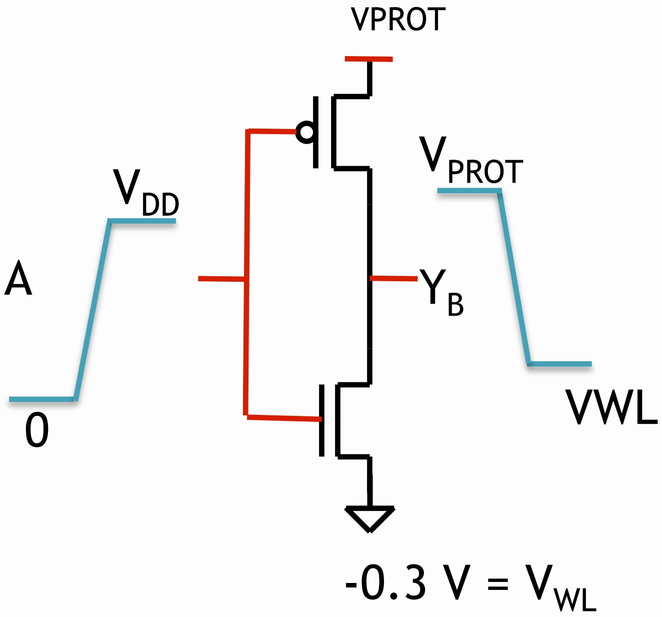
Word-line System



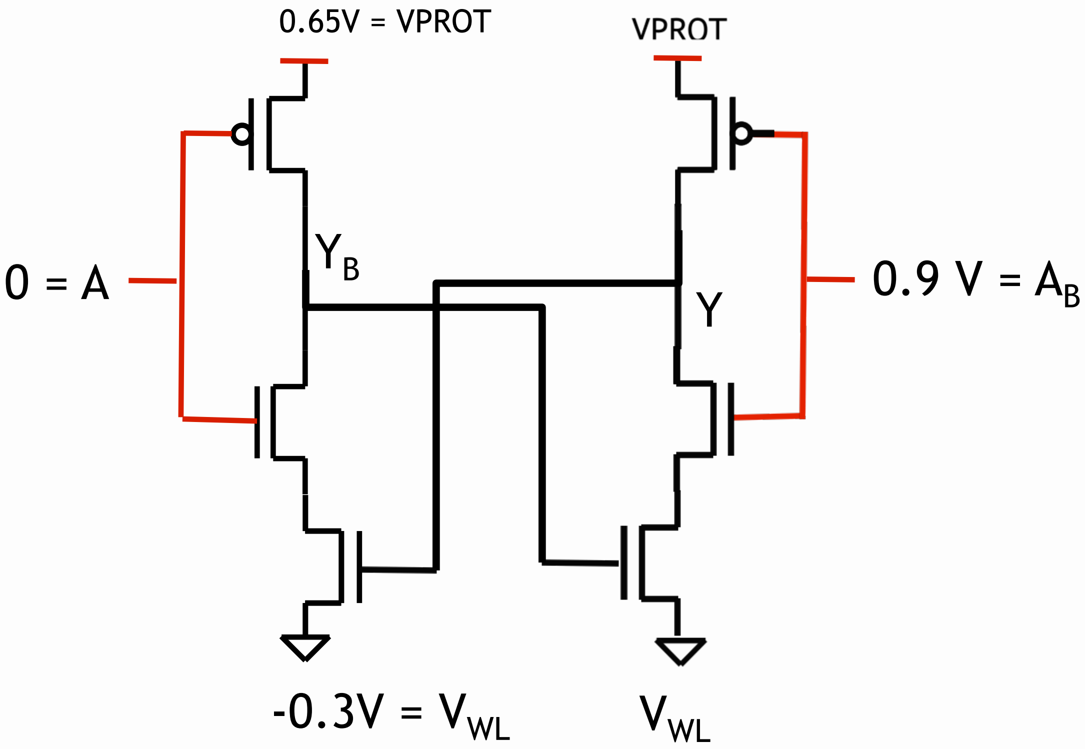
VWL - Level Translator



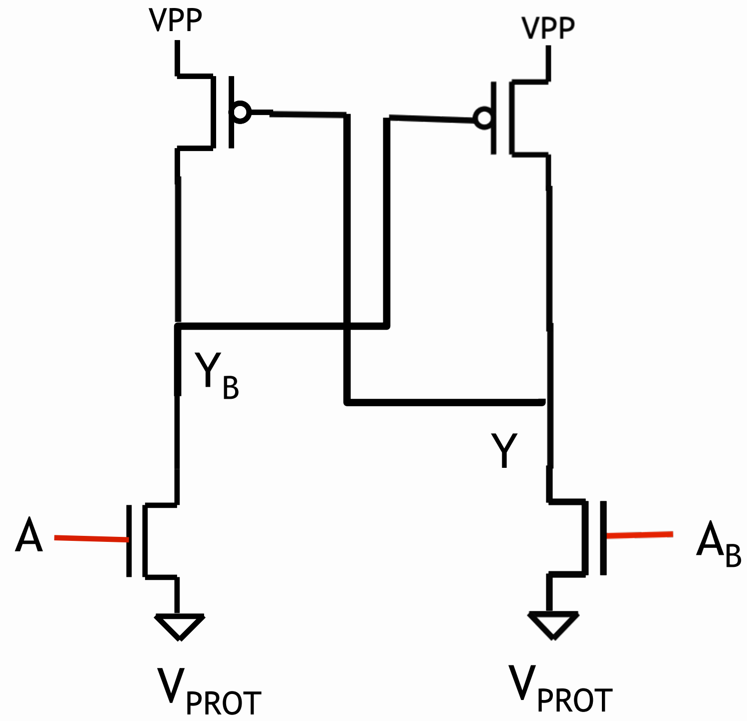
VWL - Level Translator

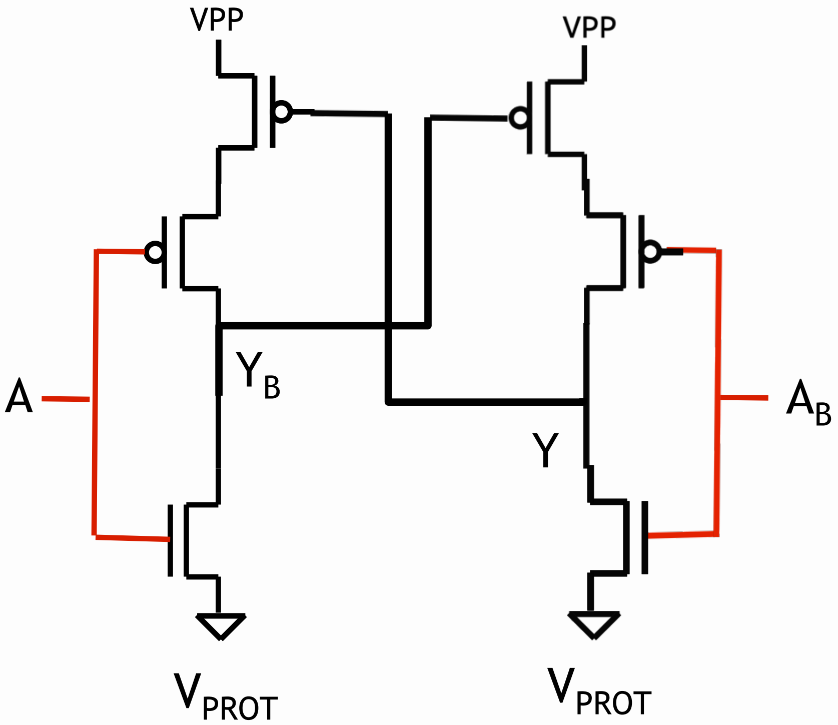


VWL - Level Translator

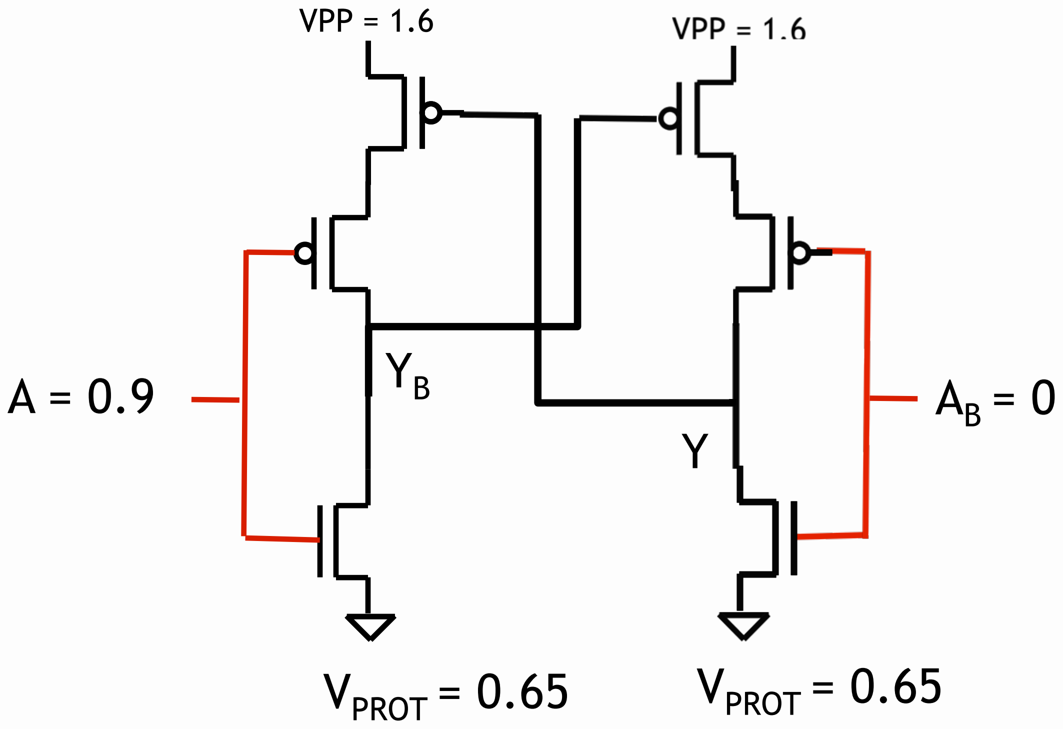


VPP - Level Translator

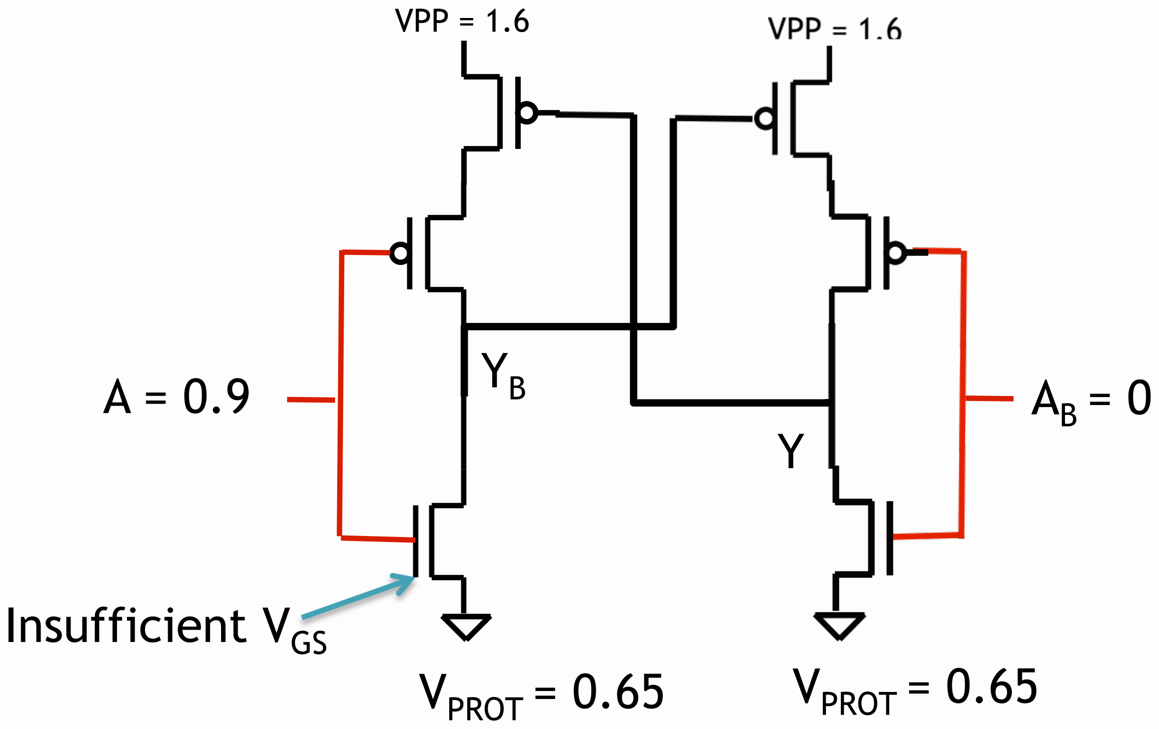




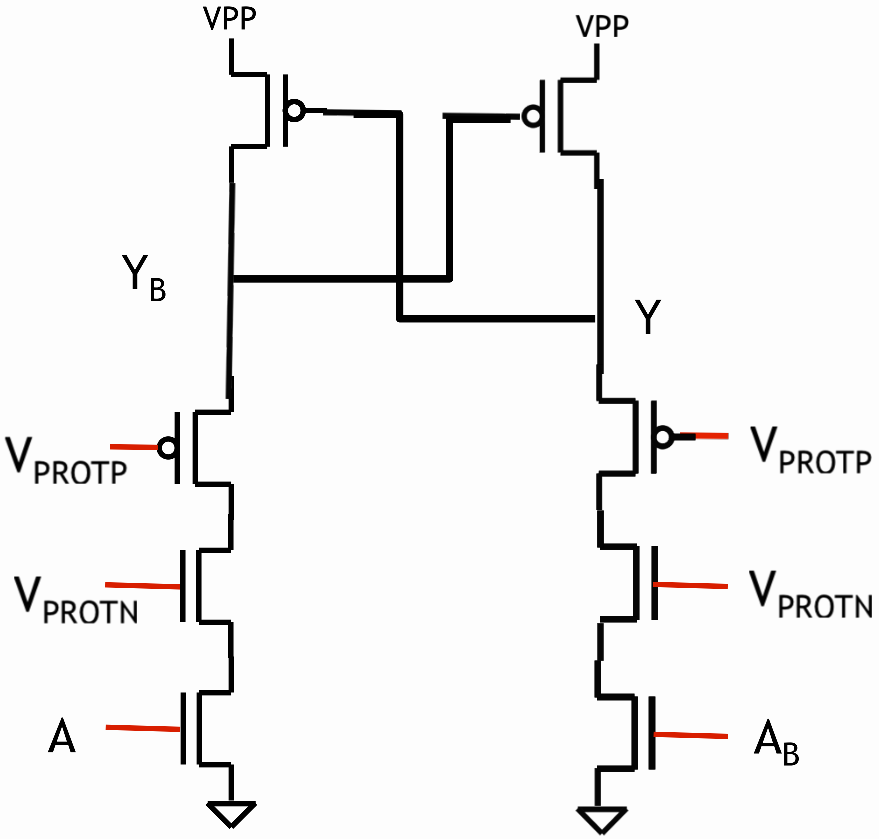
VPP - Level Translator



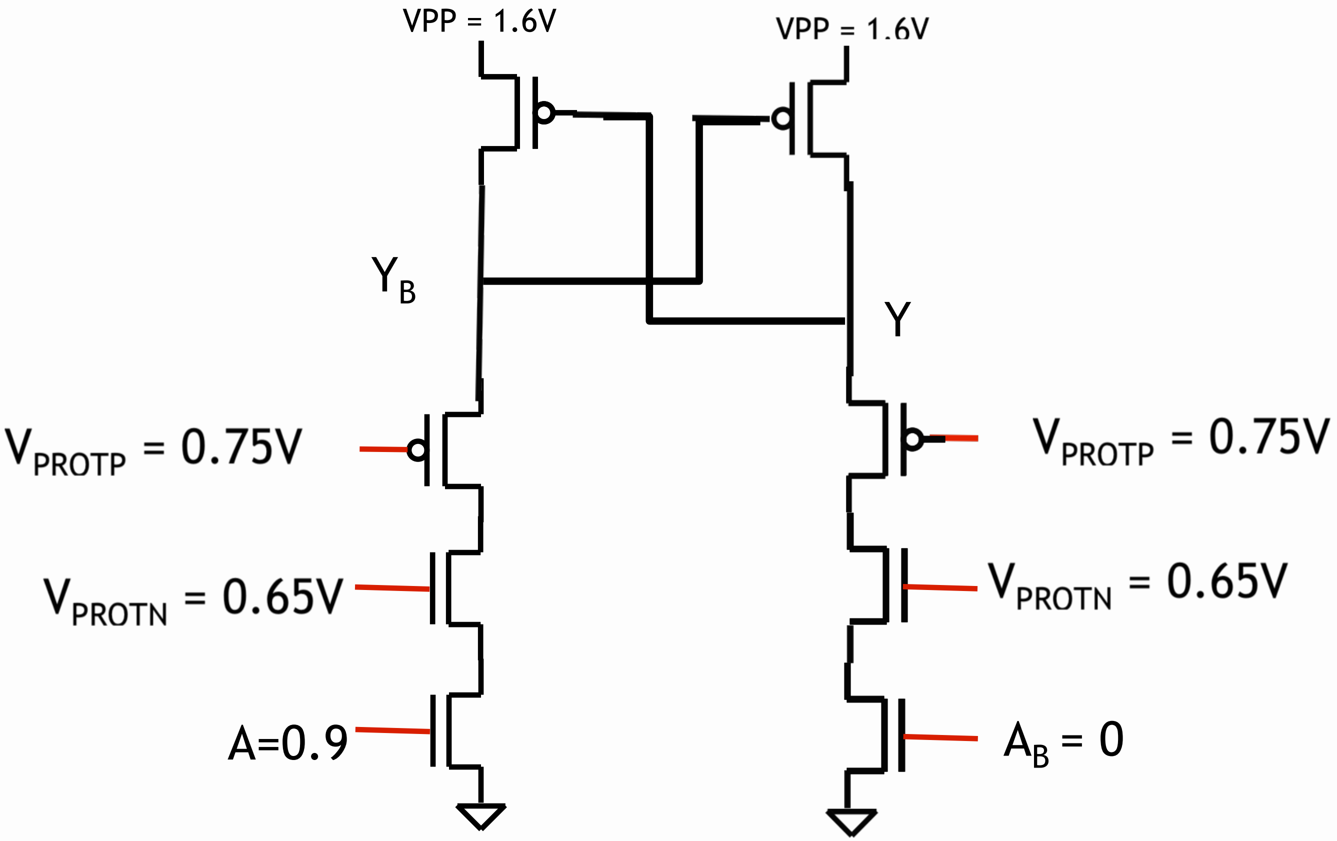
VPP - Level Translator



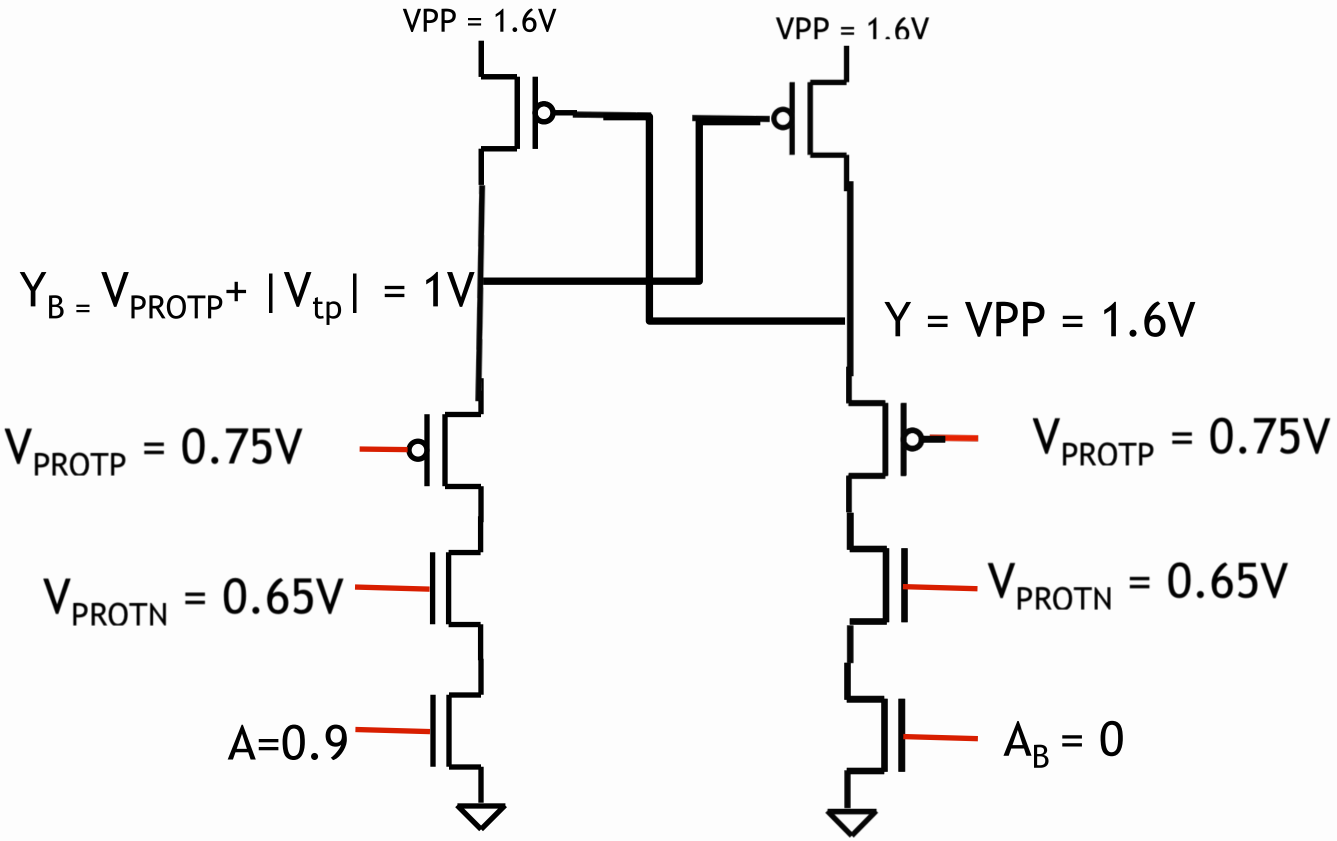
VPP - Level Translator



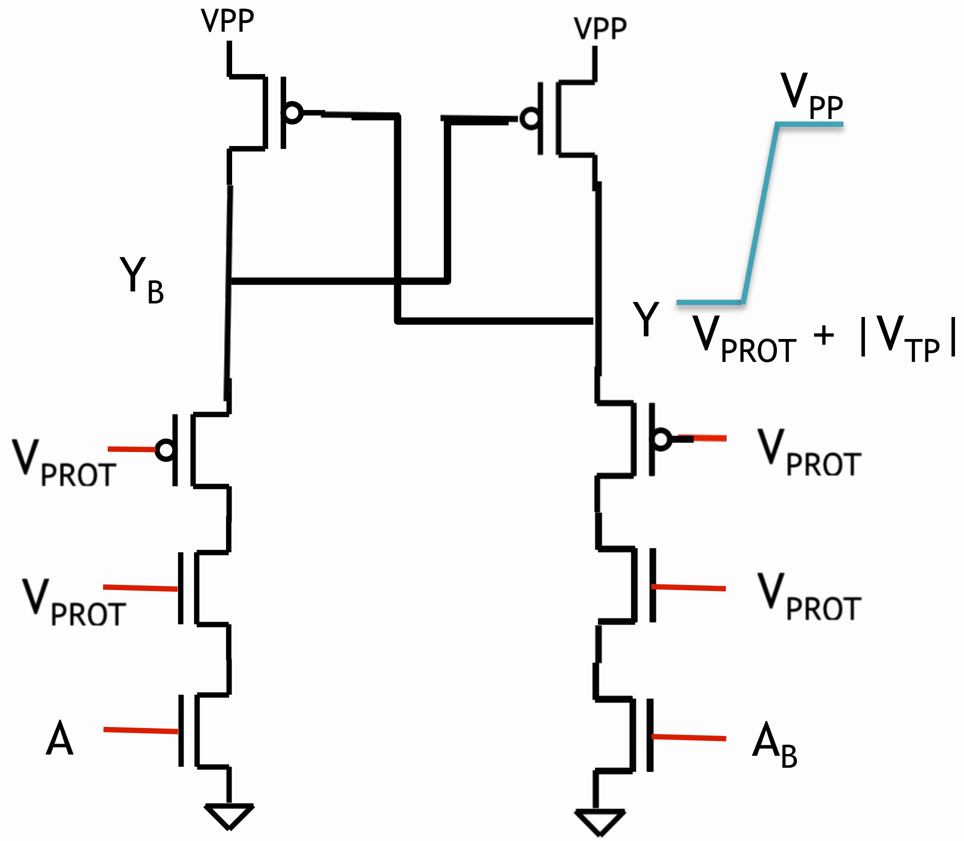
VPP - Level Translator



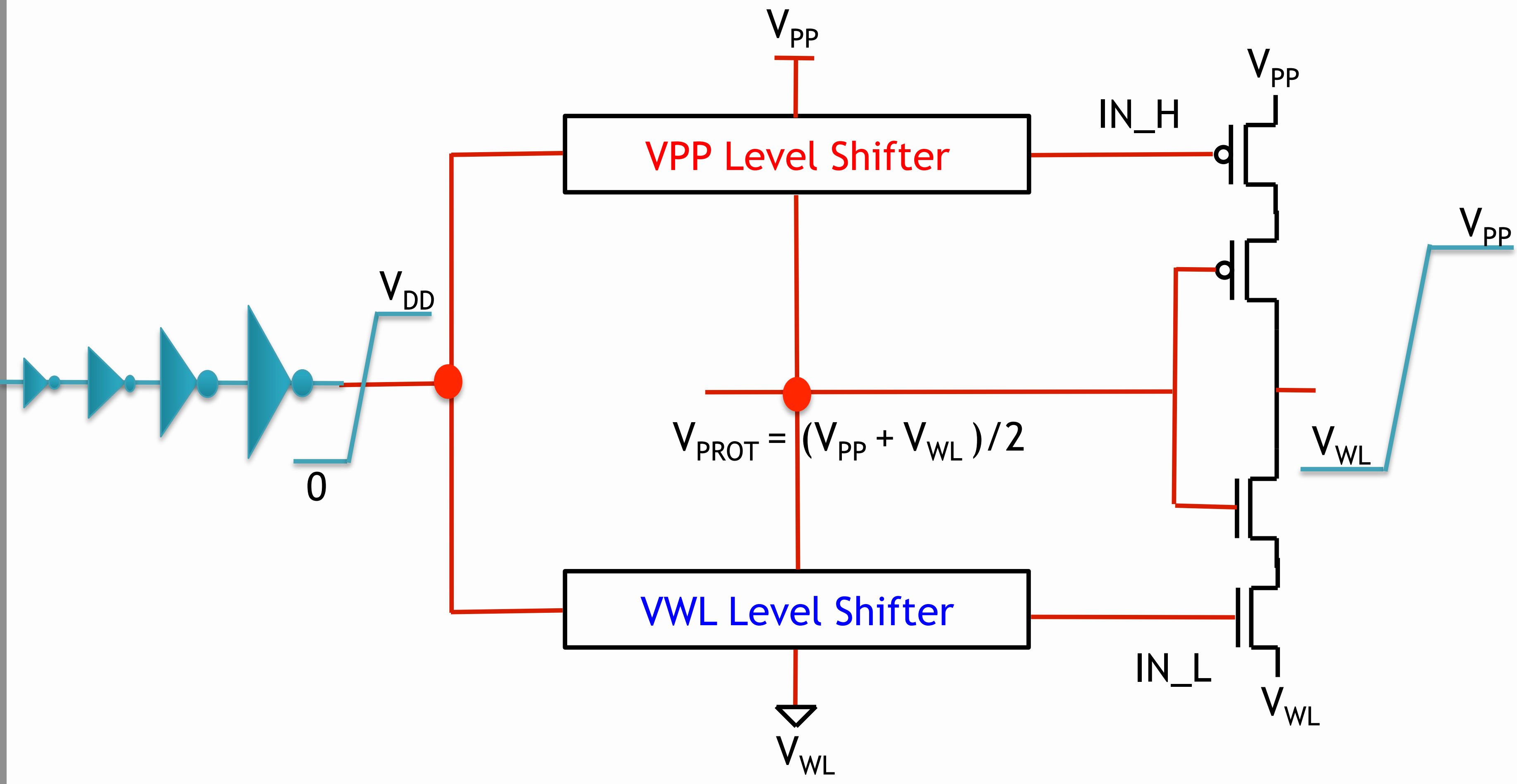
VPP - Level Translator



VPP - Level Translator



Word-line System



WLDRV

Driver with Low voltage transistors → Logic transistors

No thick gate oxide transistors required!!

Voltage across any two terminals should not exceed reliability limits

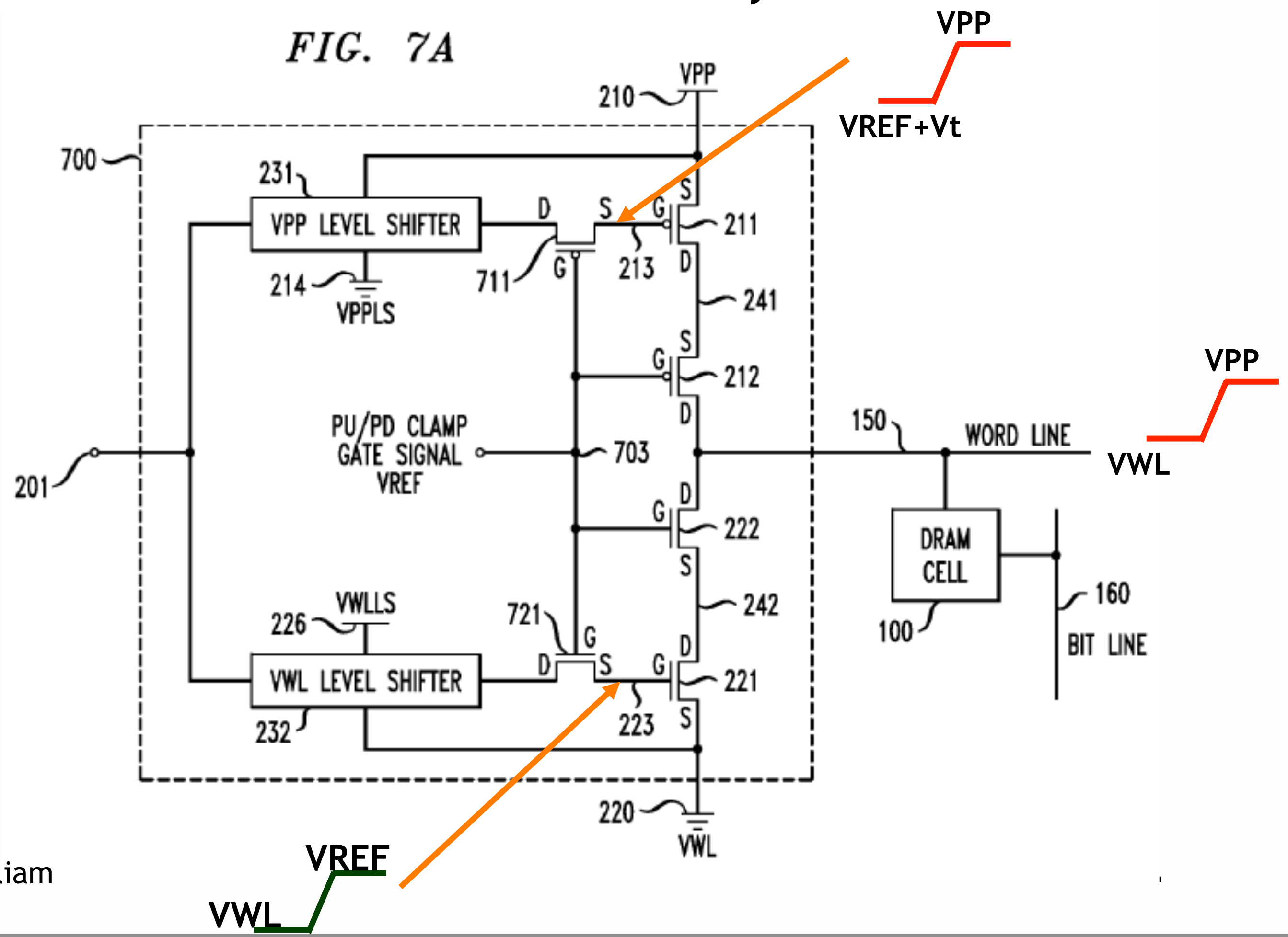
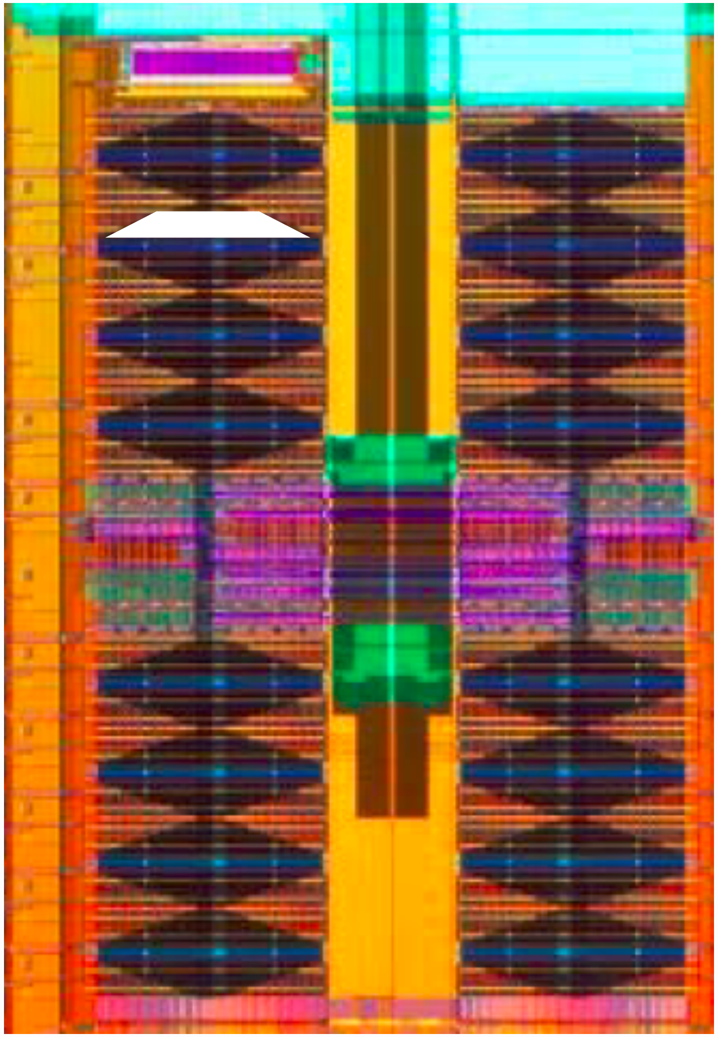
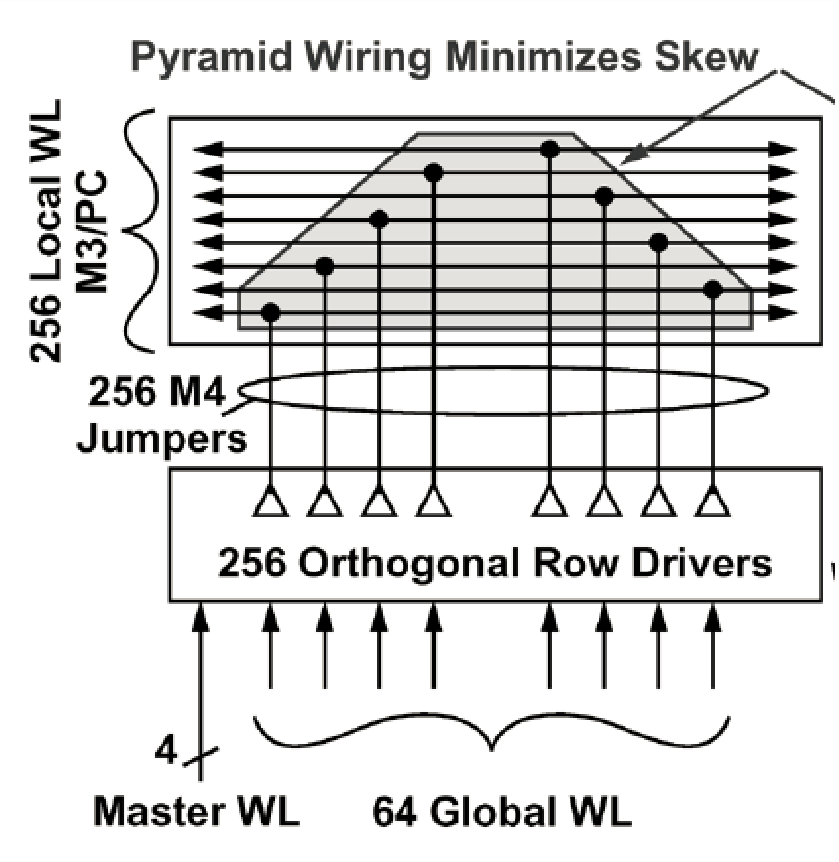


FIG. 7A

1. US patent No: 8,120,968 → William Robert Reohr, John E Barth

Orthogonal WLD and pyramid wiring (M3/ M4)





JSSC08

Topics

- ❑ Introduction to memory
- ❑ DRAM basics and bitcell array
- ❑ eDRAM operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

Retention

Transfer Device and Storage Cap are NOT ideal devices: they LEAK

Leakage Mechanisms include: Ioff, Junction Leakage, GIDL,...

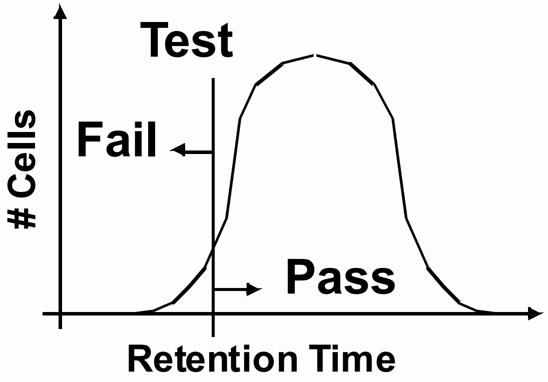
Junction Leakage Temperature Dependence = 2x/10C

Cell Charge needs to be replenished (Refreshed), Median Retention Time:

$$T = \frac{CDV}{I_{leak}} = \frac{35fF \times 400mV}{2fA} = 7 \text{ seconds}$$

Where DV is acceptable loss
C is Cell Capacitance
I_{leak} is Total Leakage

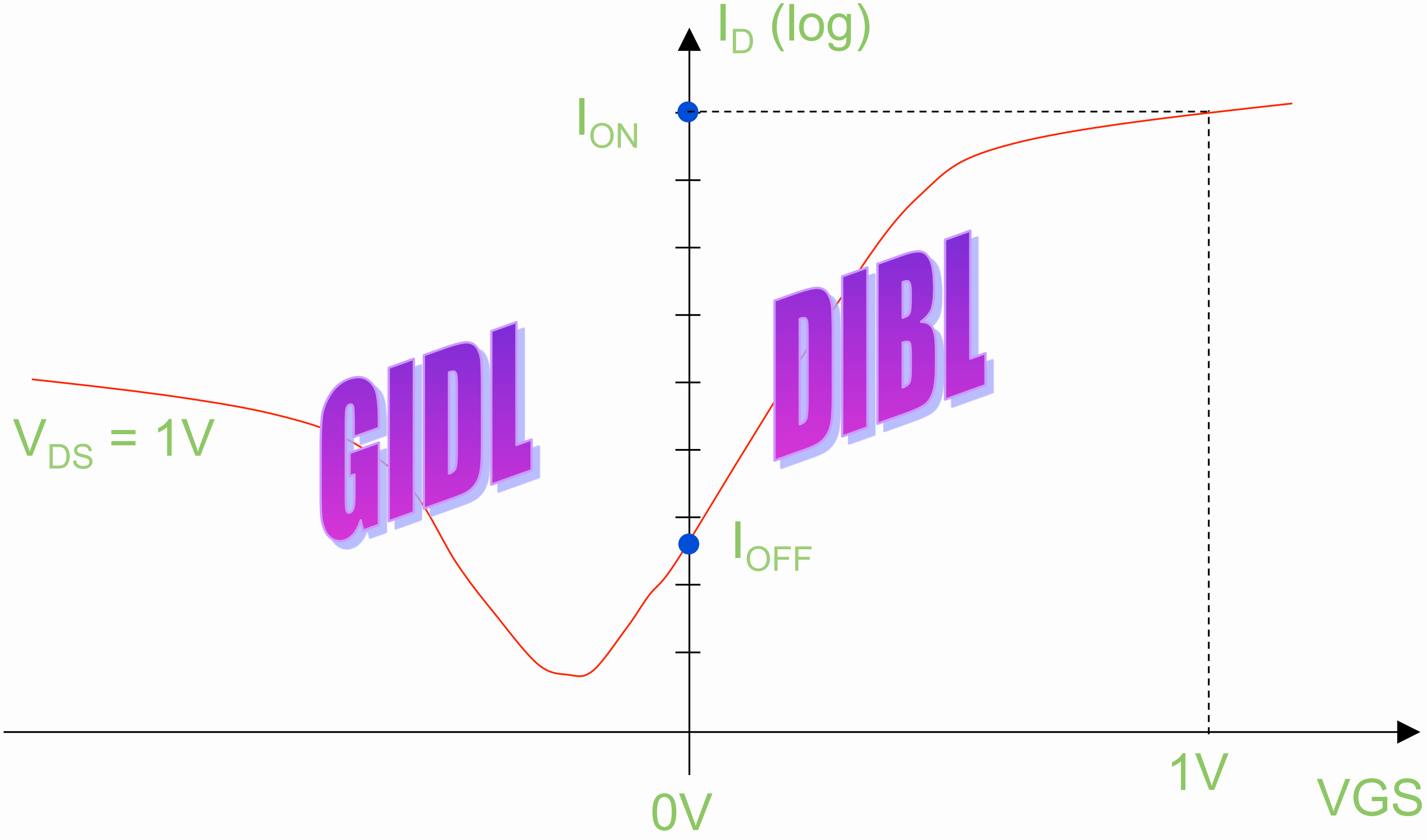
Retention Distribution has Tails created by Defects and Leaky Cells



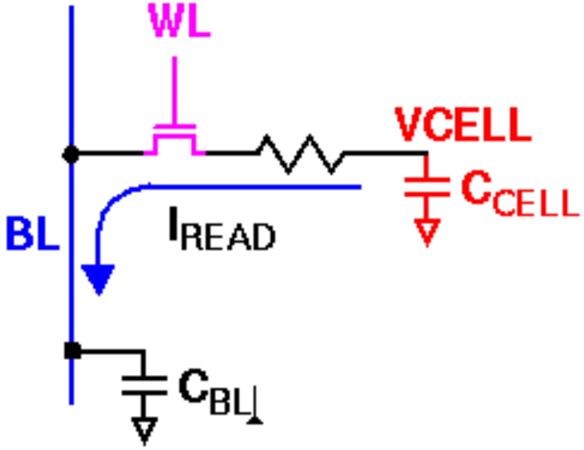
Weak Cells Tested out (5x Guardband) and replaced with Redundancy

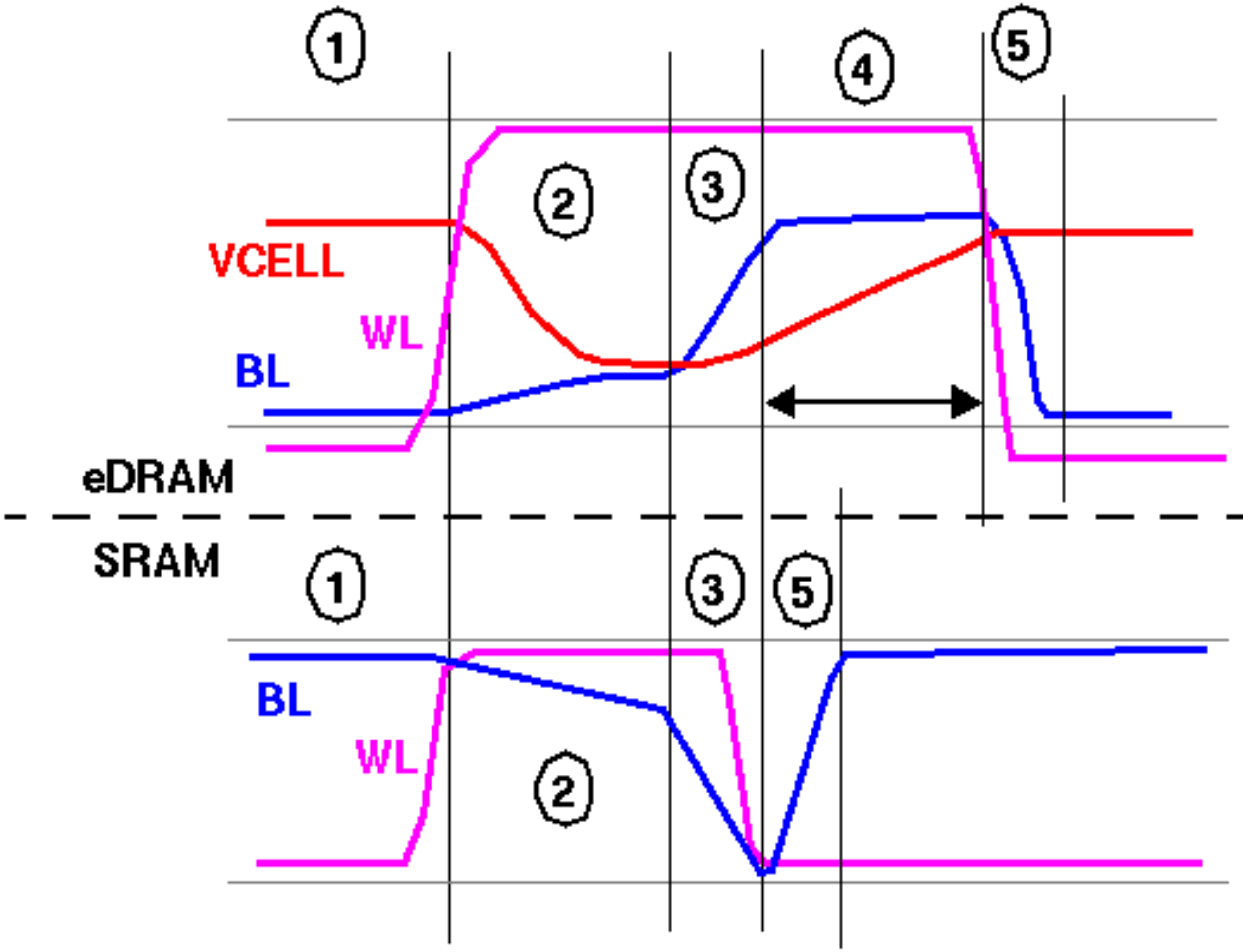
Customer issues periodic Refresh Cycle

Pass transistor leakage



eDRAM vs. SRAM Cycle-Time Comparison





1. WL Activation
2. Charge Transfer to Bit-Line

   (I_{READ} Similar to SRAM)
3. Amplification
4. Write-Back
5. Precharge

NET: SRAM Random Cycle will continue to lead!

Topics

❑ Introduction to memory

❑ DRAM basics and bitcell array

❑ eDRAM operational details (case study)

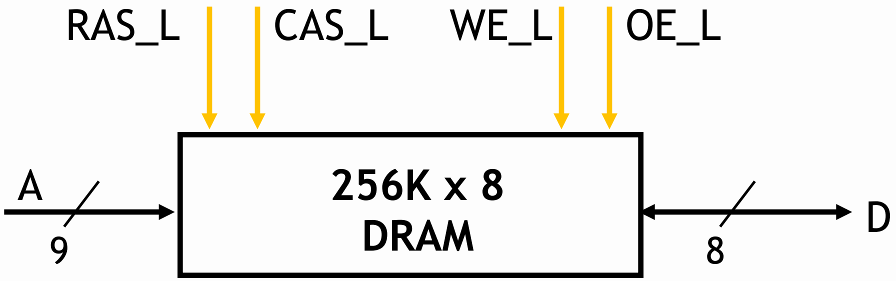
❑ Noise concerns

❑ Wordline driver (WLDRV) and level translators (LT)

❑ Challenges in eDRAM

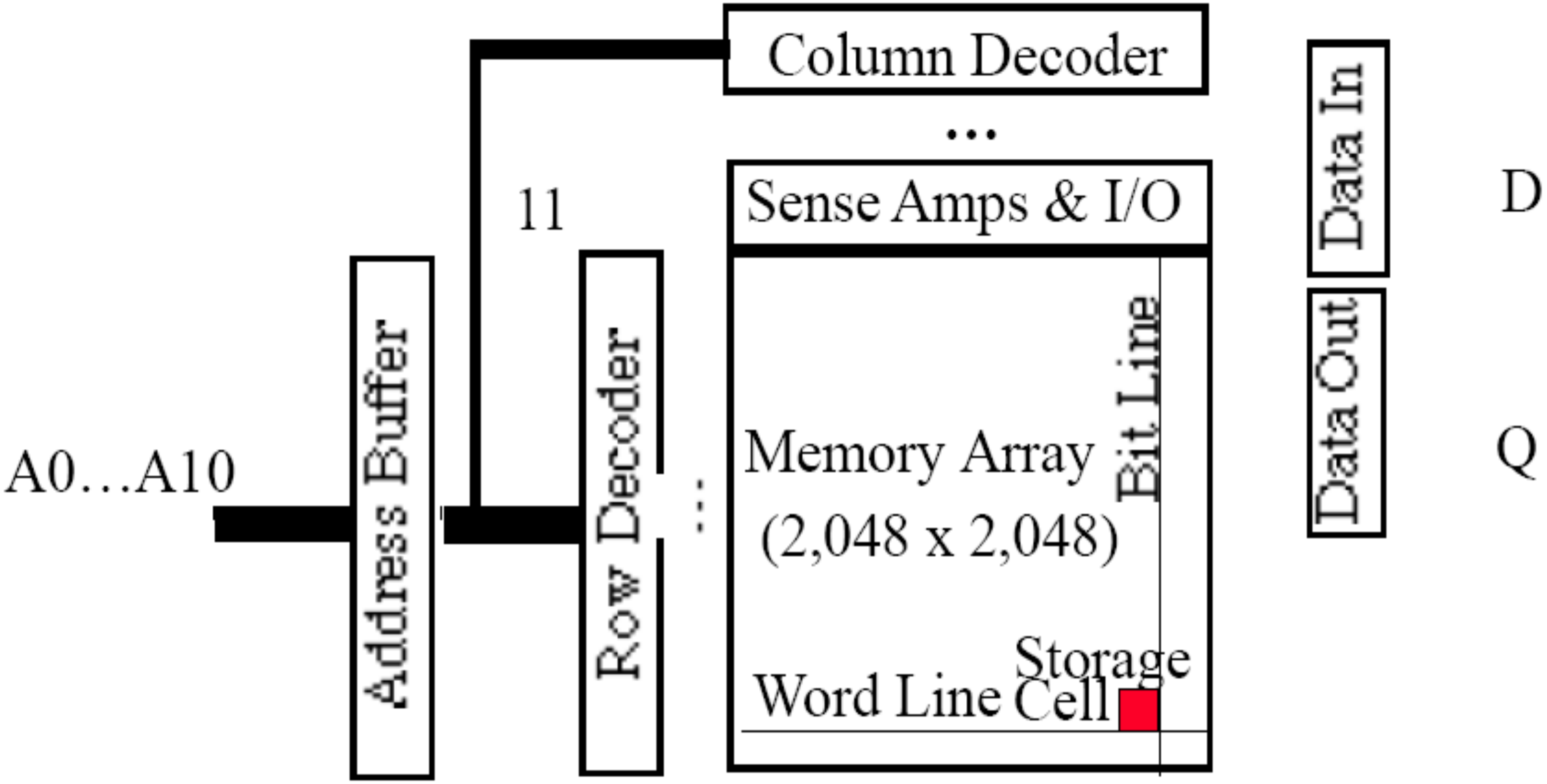
❑ Understanding Timing diagram – An example

Logic Diagram of a Typical DRAM



- Control Signals (RAS_L, CAS_L, WE_L, OE_L) are all active low
- Din and Dout are combined (D):
  - WE_L is asserted (Low), OE_L is disasserted (High)
    - D serves as the data input pin
  - WE_L is disasserted (High), OE_L is asserted (Low)
    - D is the data output pin
- Row and column addresses share the same pins (A)
  - RAS_L goes low: Pins A are latched in as row address
  - CAS_L goes low: Pins A are latched in as column address
  - RAS/CAS edge-sensitive

DRAM logical organization (4 Mbit)



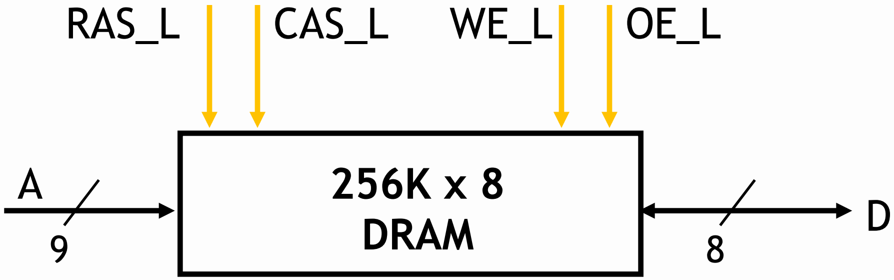
- Square root of bits per RAS/CAS

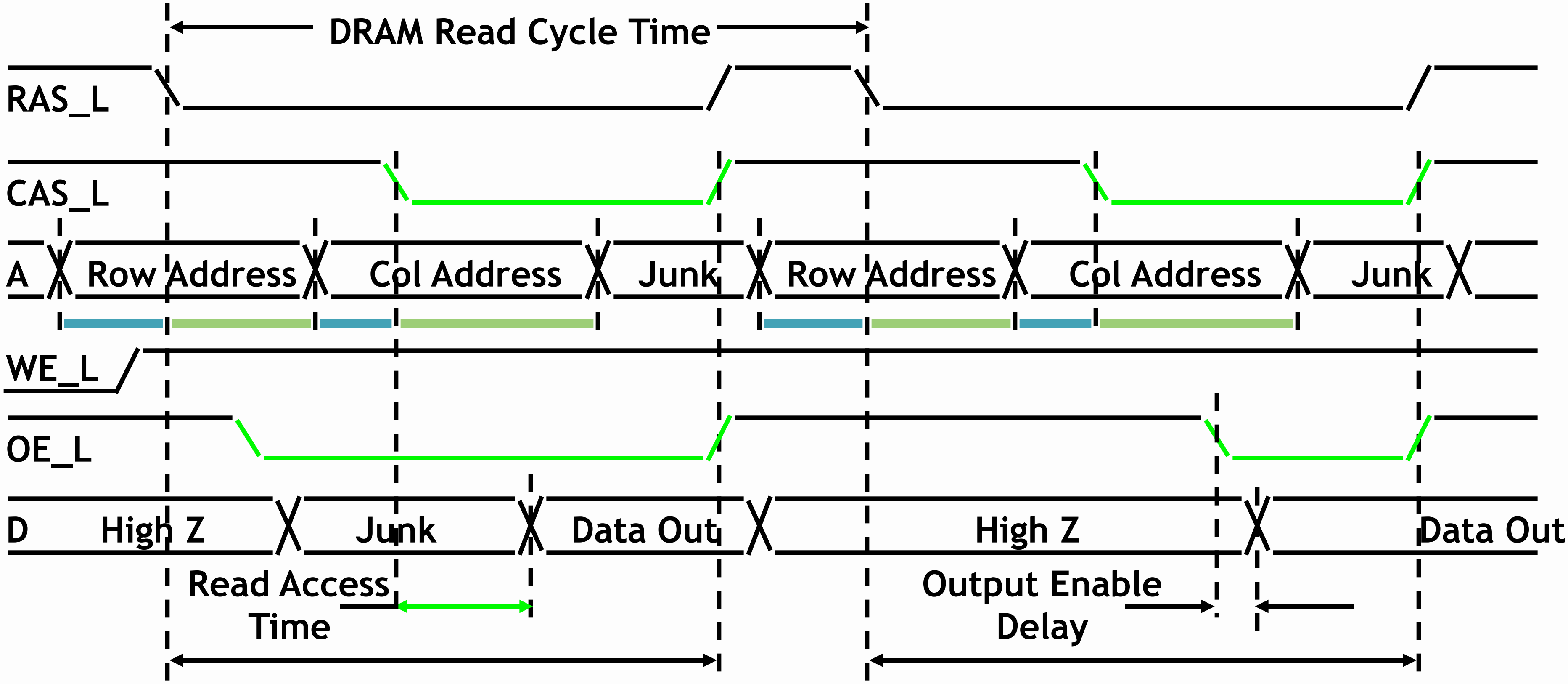
Din Dout can be clubbed together with a BiDi buffer

DRAM Read Timing

- Every DRAM access begins at:
  - The assertion of the RAS_L
  - 2 ways to read:

    early or late v. CAS

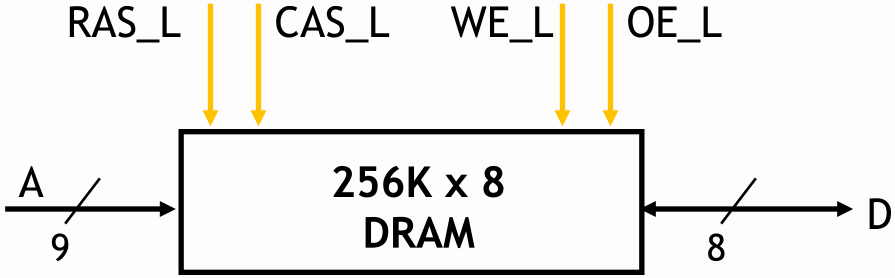


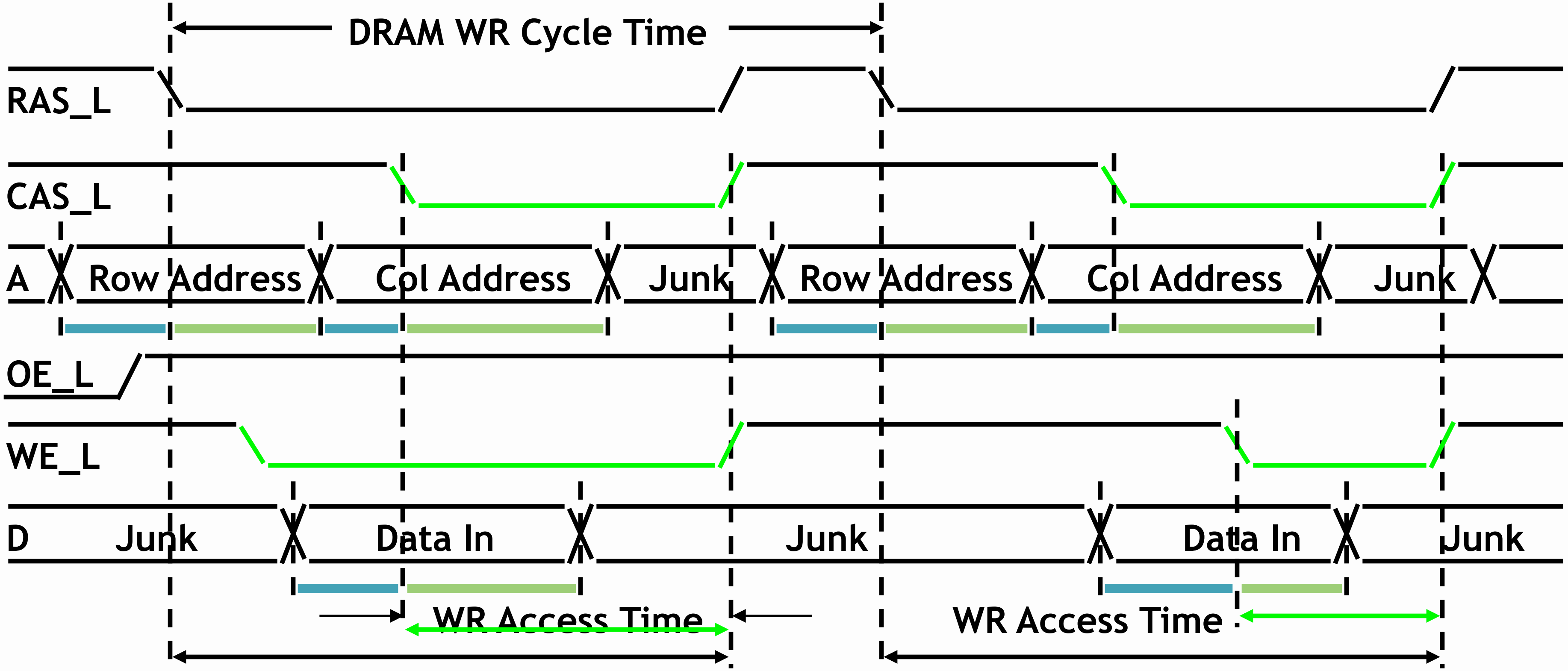


Early Read Cycle: OE_L asserted before CAS_L Late Read Cycle: OE_L asserted after CAS_L

DRAM Write Timing

- Every DRAM access begins at:
  - The assertion of the RAS_L
  - 2 ways to write: early or late v. CAS





Early Wr Cycle: WE_L asserted before CAS_L

Late Wr Cycle: WE_L asserted after CAS_L

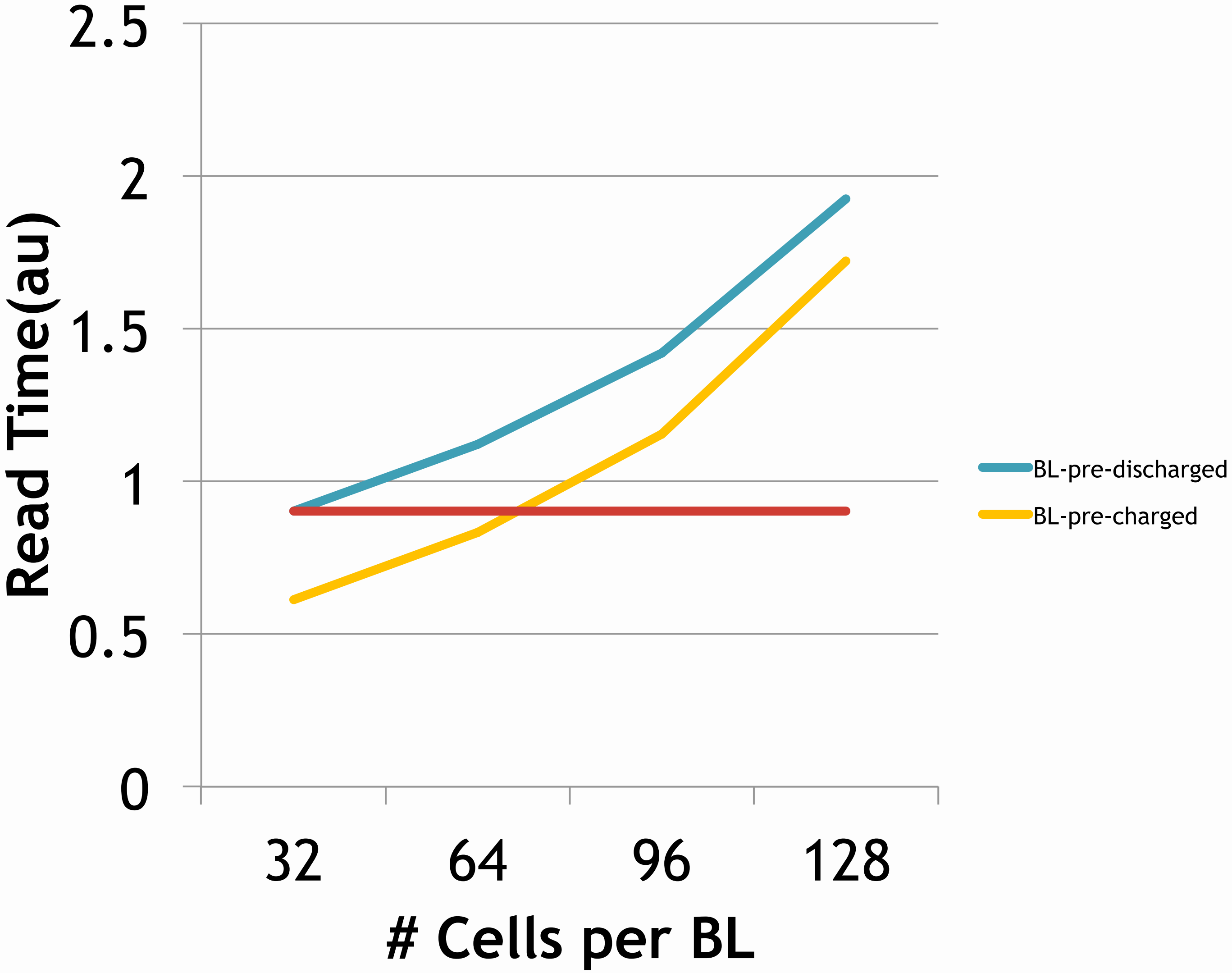
A Fast Area Efficient Sense-Amp (Case Study)

G. Fredeman et al., “A 14 nm 1.1 Mb Embedded DRAM Macro With 1 ns Access,”
in IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 230-239,
Jan. 2016. doi: 10.1109/JSSC.2015.2456873

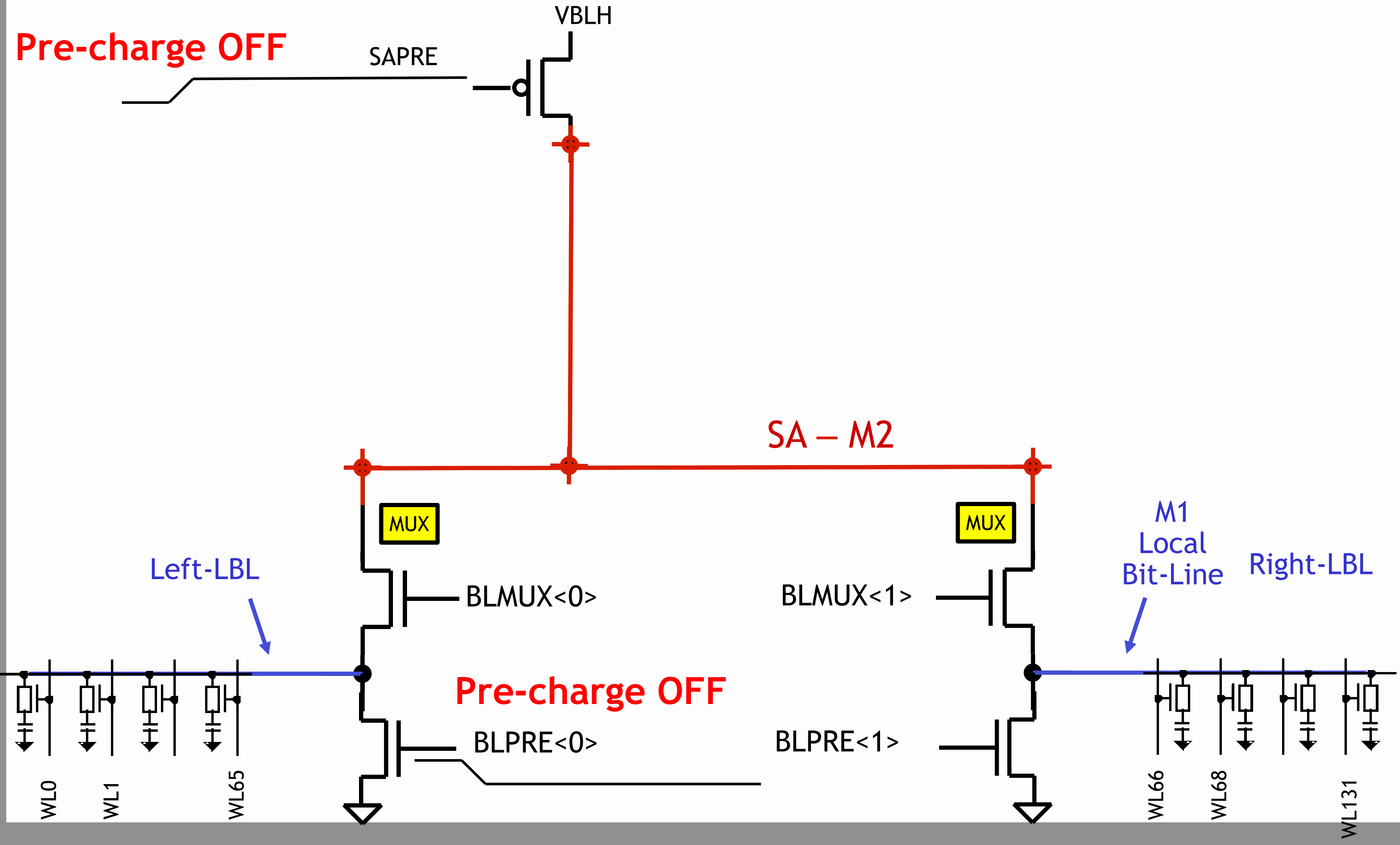
Problems with Micro Sense Amp

- ❑ By default the Sense Amp reads a 0
- ❑ Access transistor has to pull the LBL HIGH to read 1
  - ❑ Asymptotic charge up to High since Vgs keeps reducing
  - ❑ Very slow by nature
  - ❑ Need to minimize the WLs per BL(33) for performance reasons
- ❑ Cannot pre-charge LBL to High
  - ❑Floating Body Effect affects retention
- ❑ NMOS (Access Device) is very fast when pulling down to zero
  - ❑ Can we make a Sense Amp that reads a one by default?
  - ❑ This will allow more WLs per BL

LBL Pre-charge vs Pre-discharge

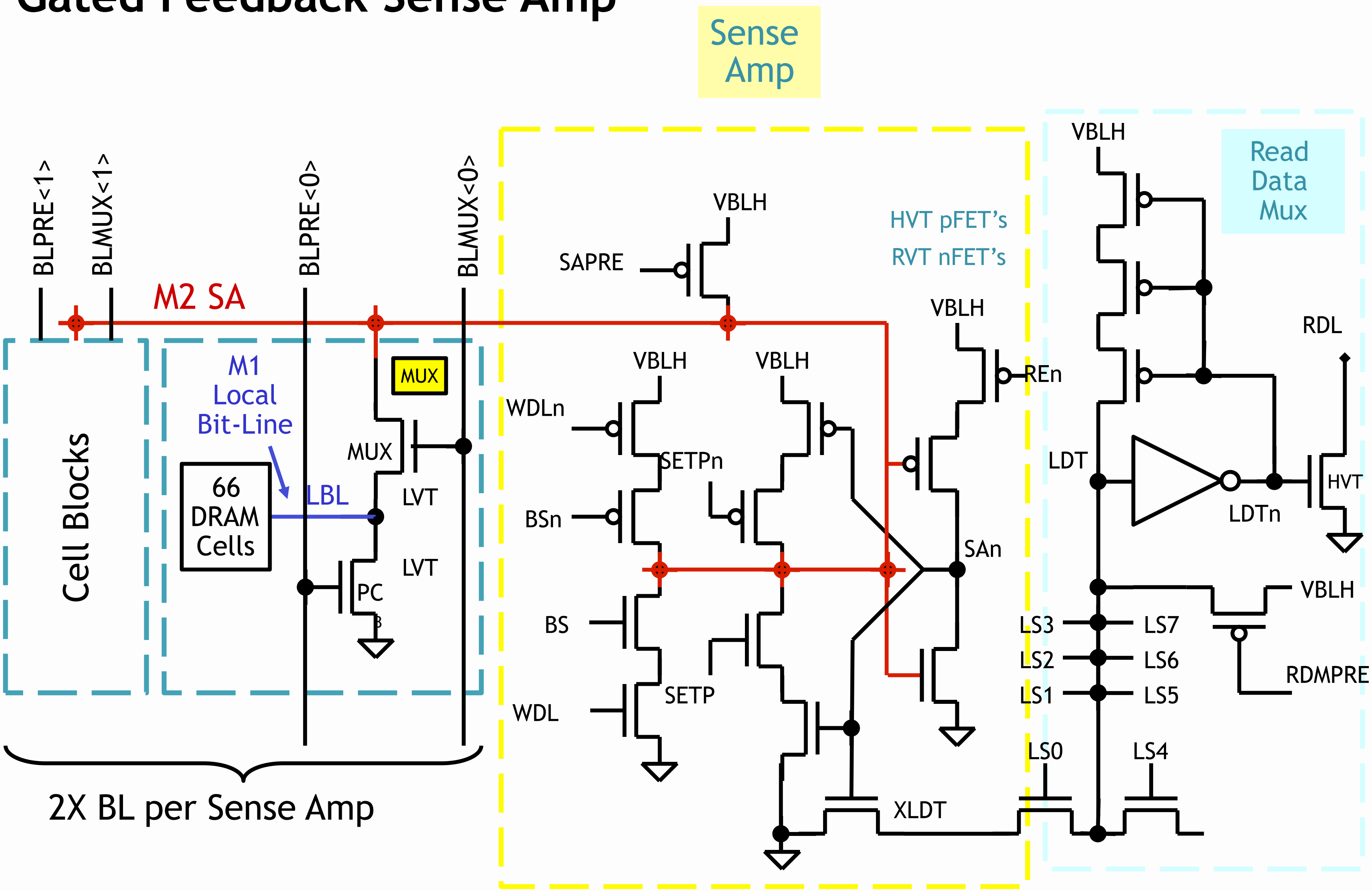


Basic Structure



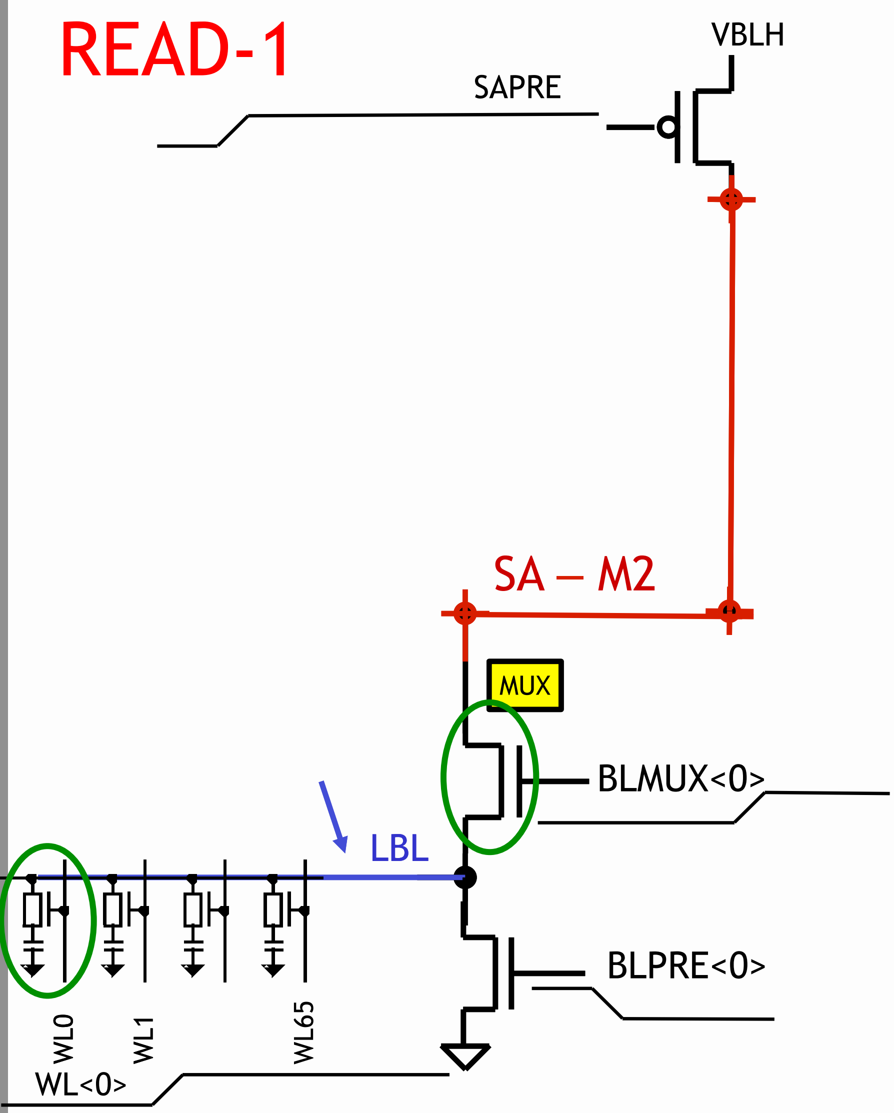
Animation

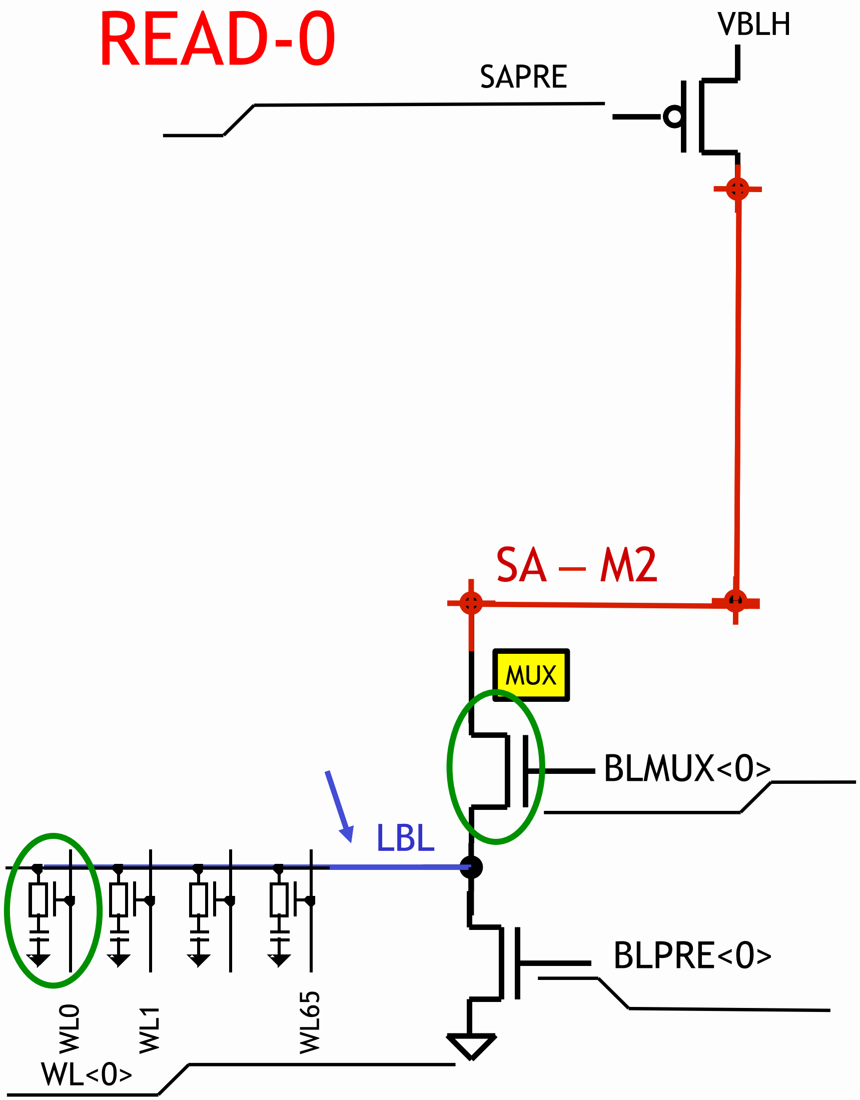
Gated Feedback Sense Amp



2X BL per Sense Amp

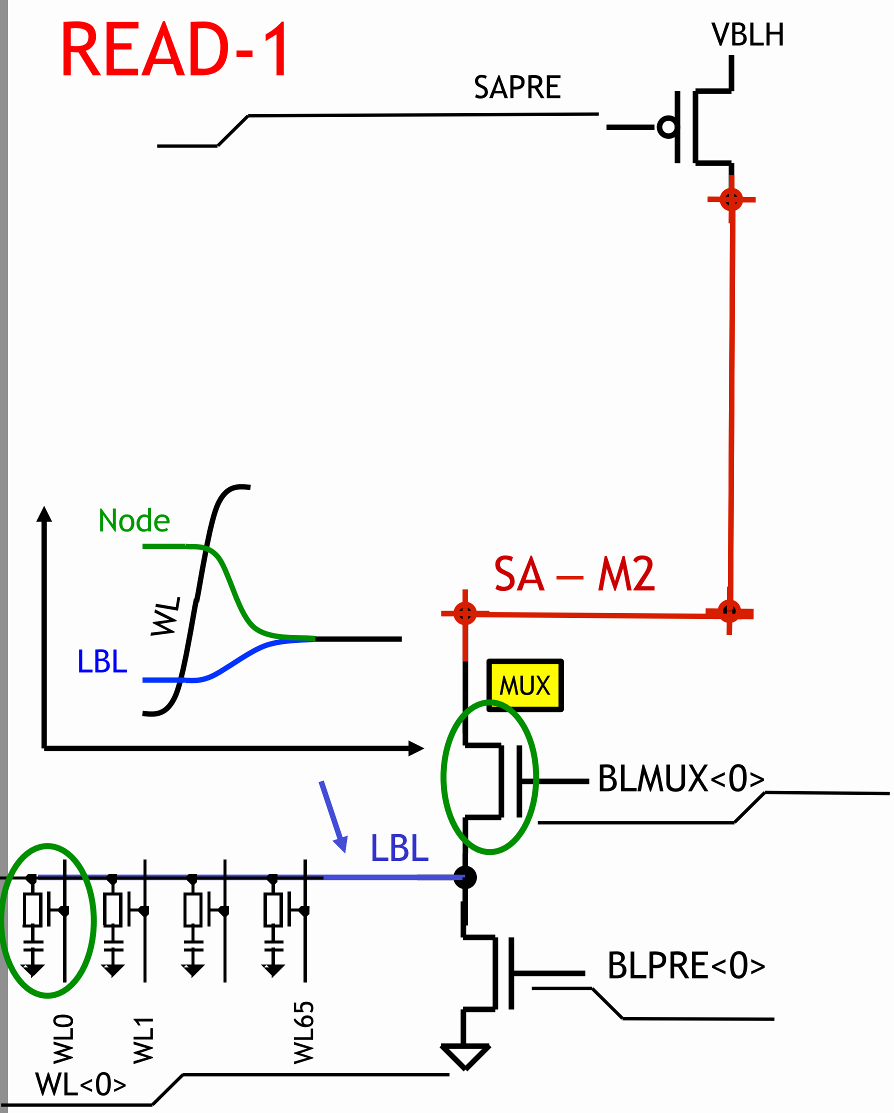
Gated Feedback Sense Amp – Construction

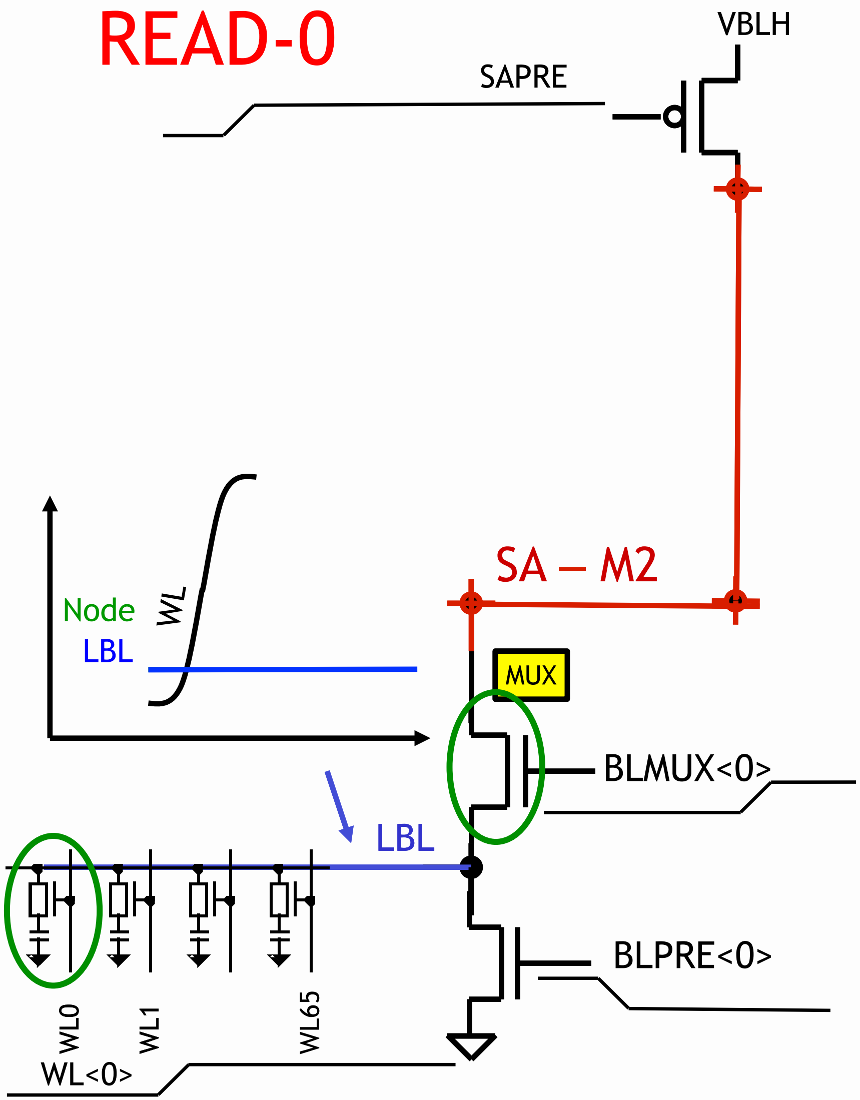




Animation

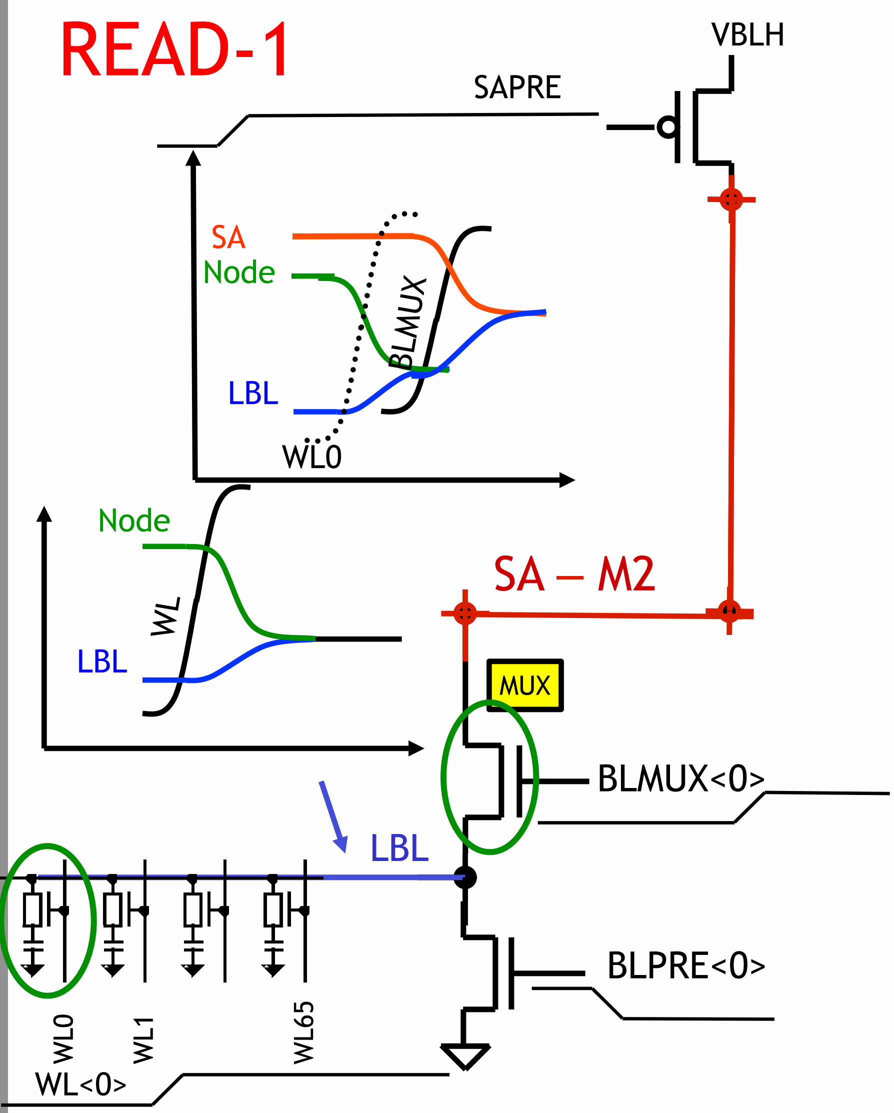
Gated Feedback Sense Amp – Construction

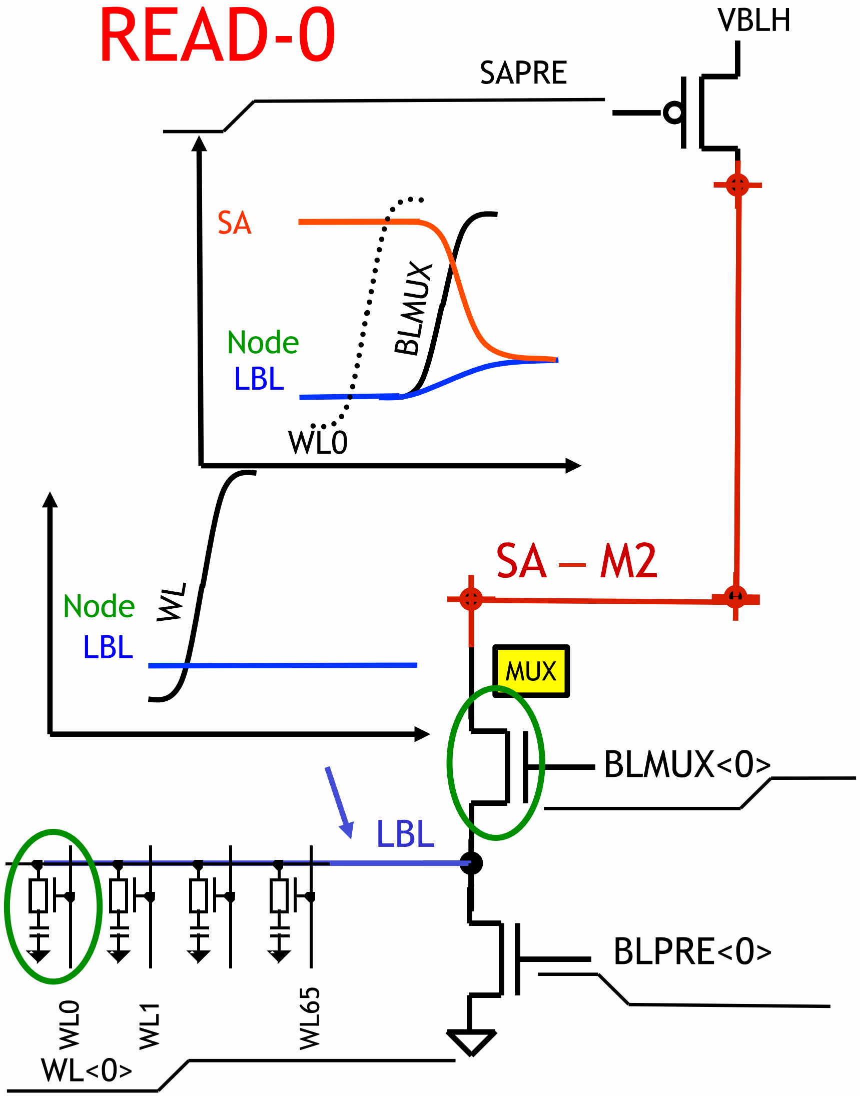




Animation

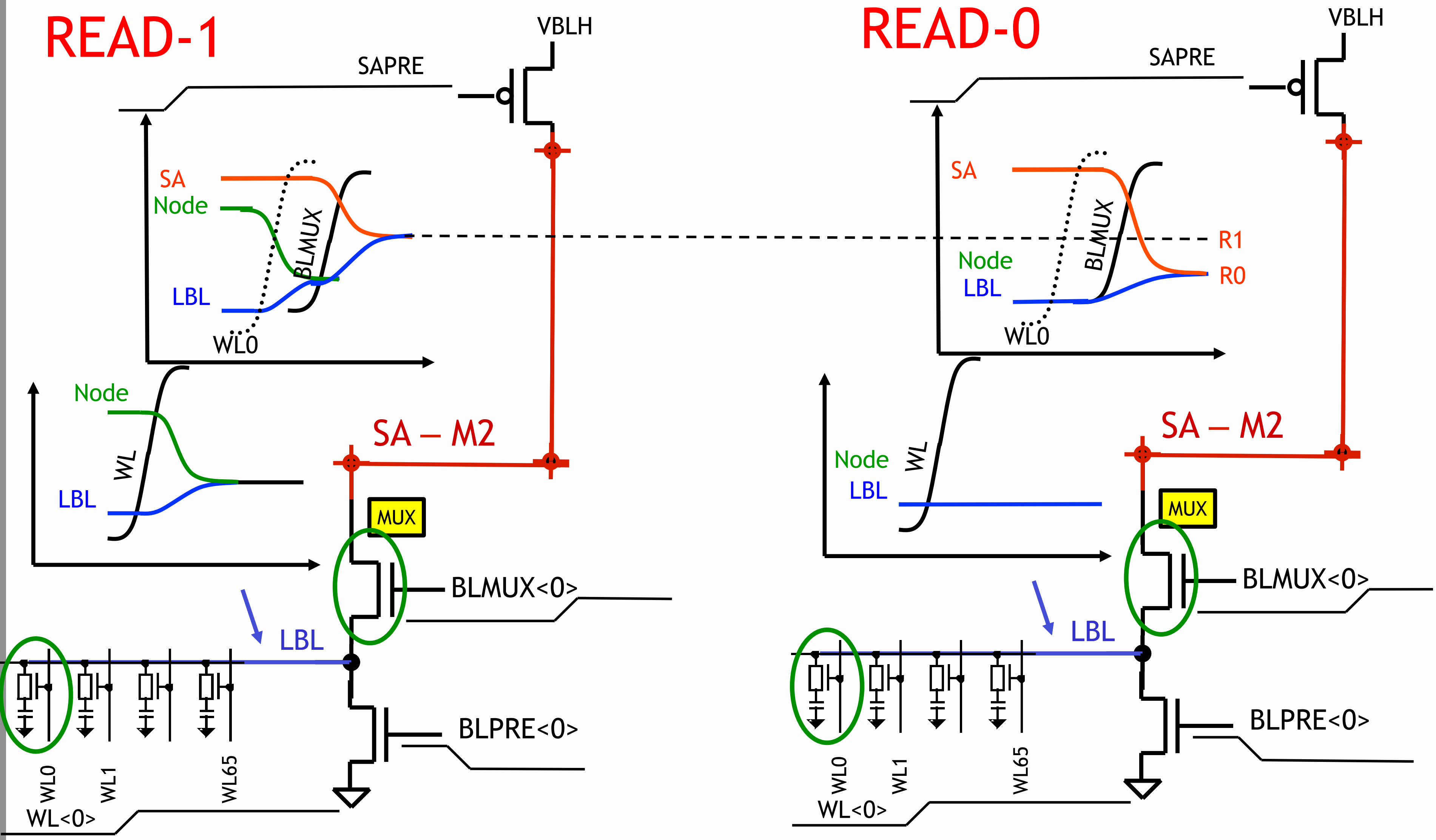
Gated Feedback Sense Amp - Construction





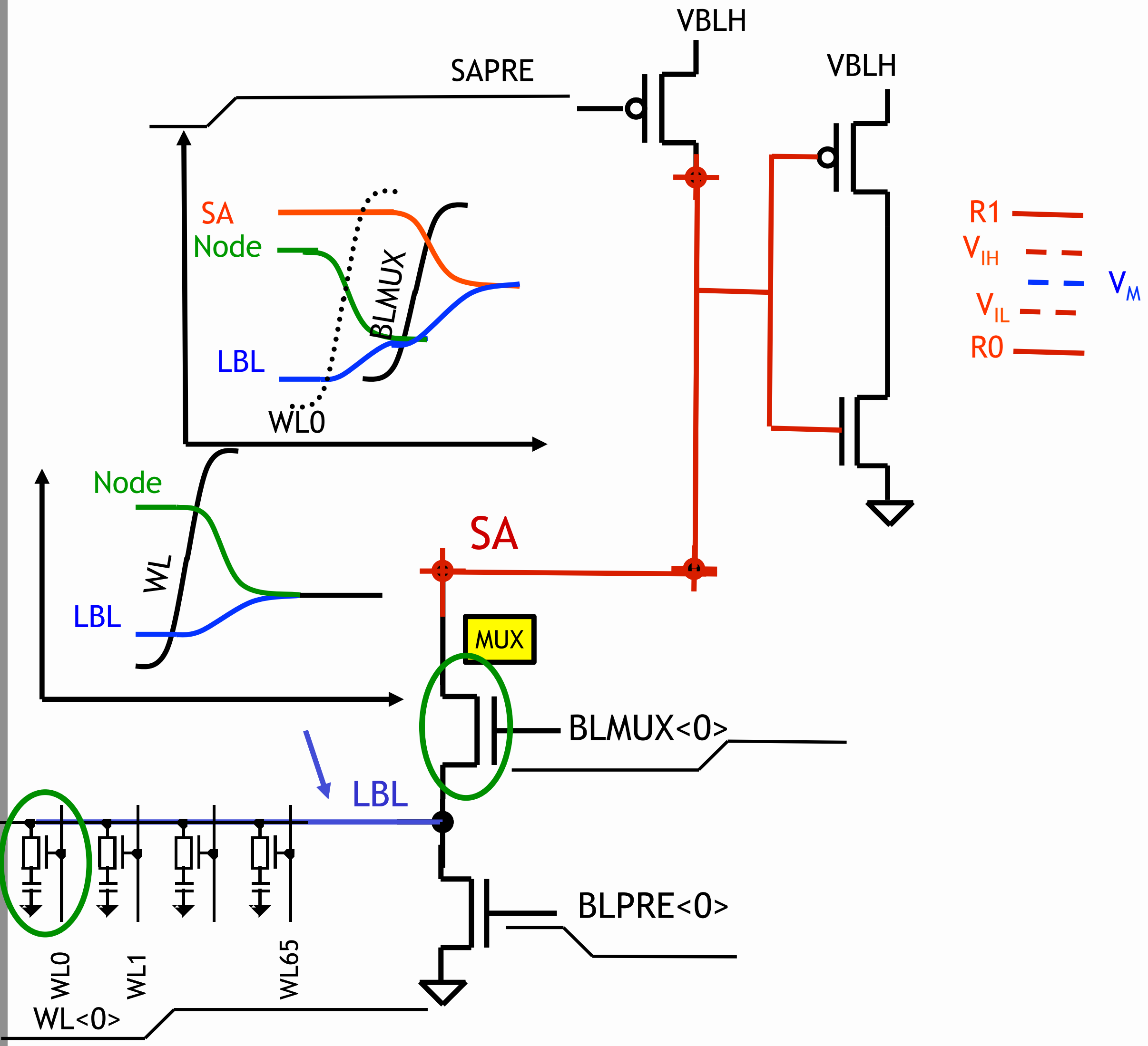
Animation

Gated Feedback Sense Amp - Construction



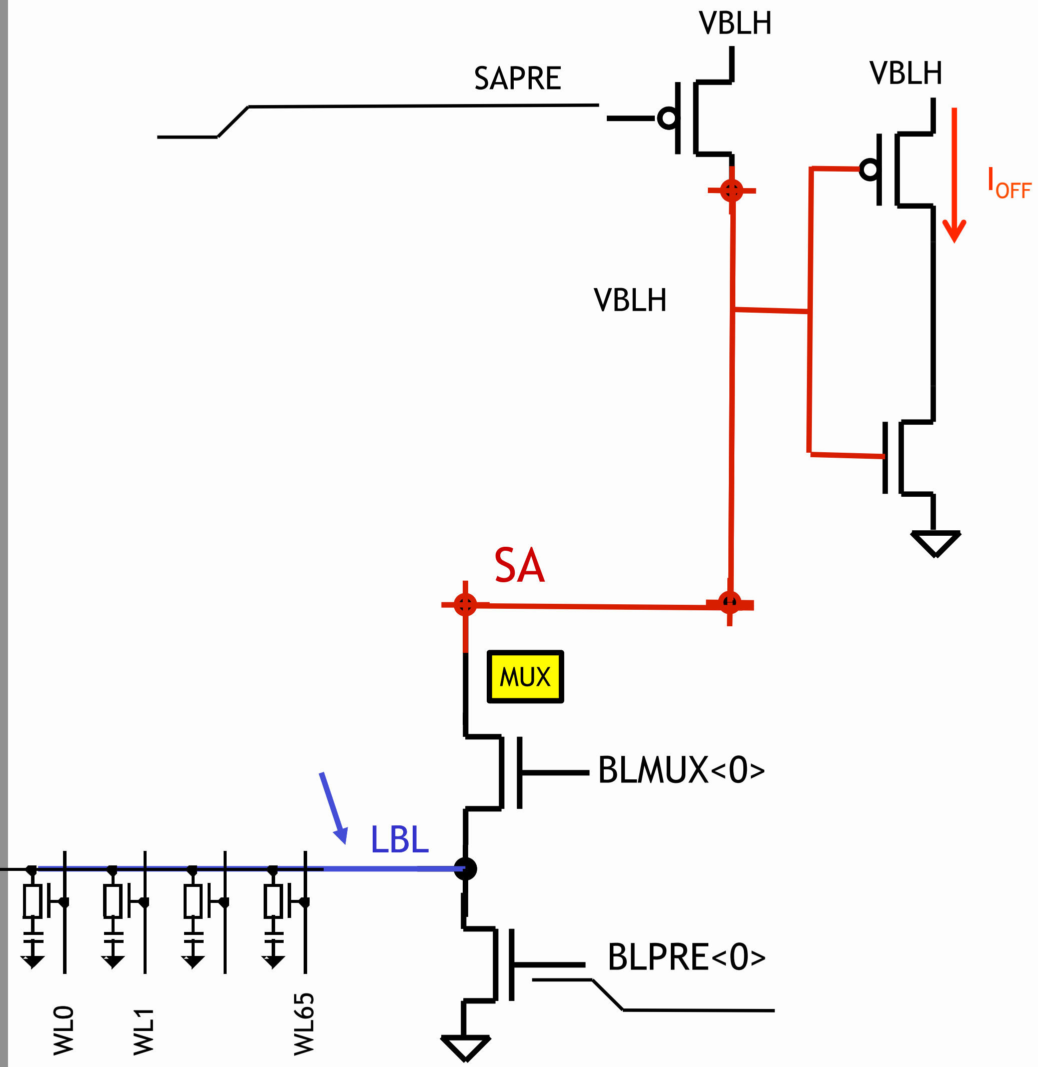
Animation

Gated Feedback Sense Amp - Construction

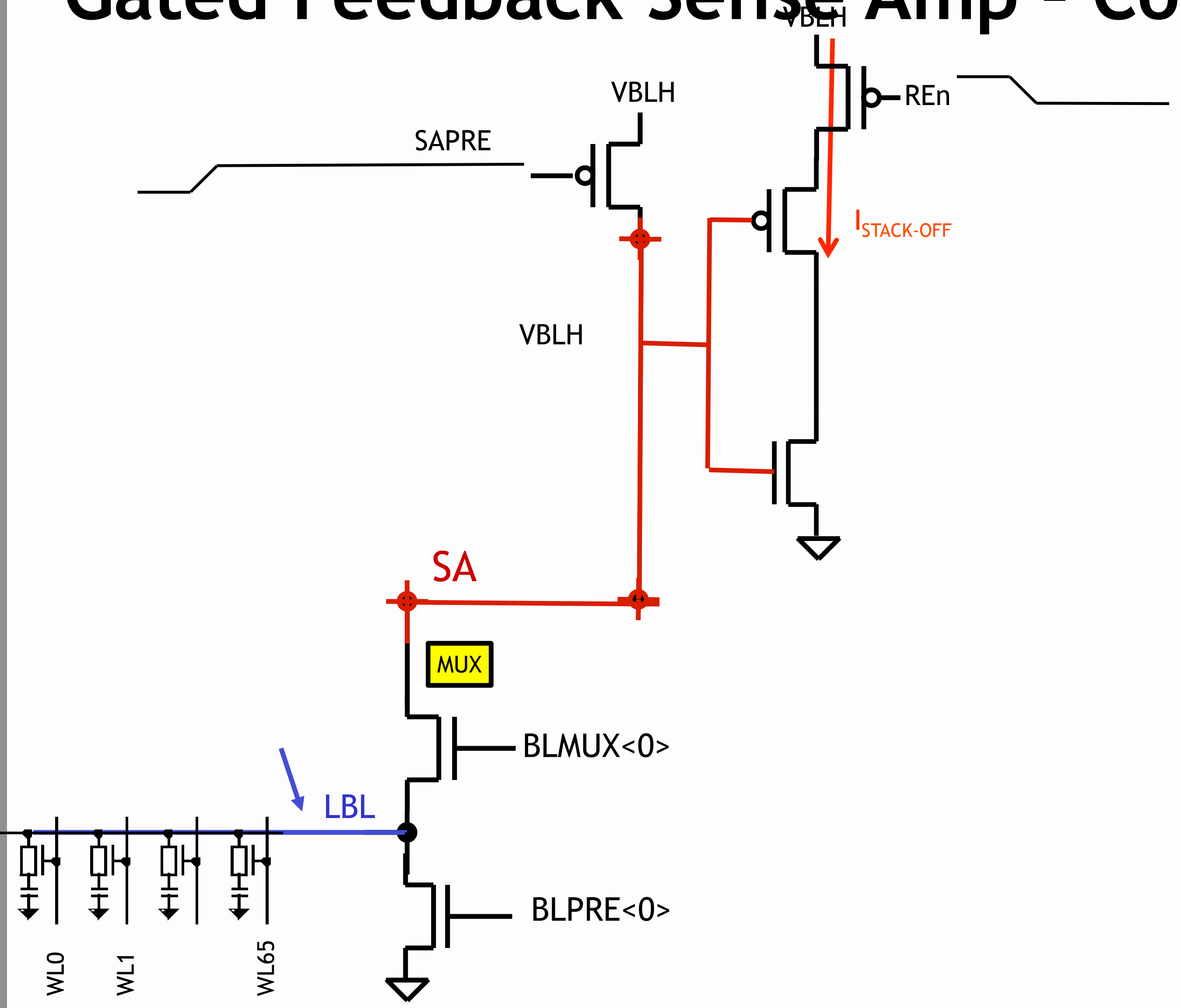


Animation

Gated Feedback Sense Amp - Construction

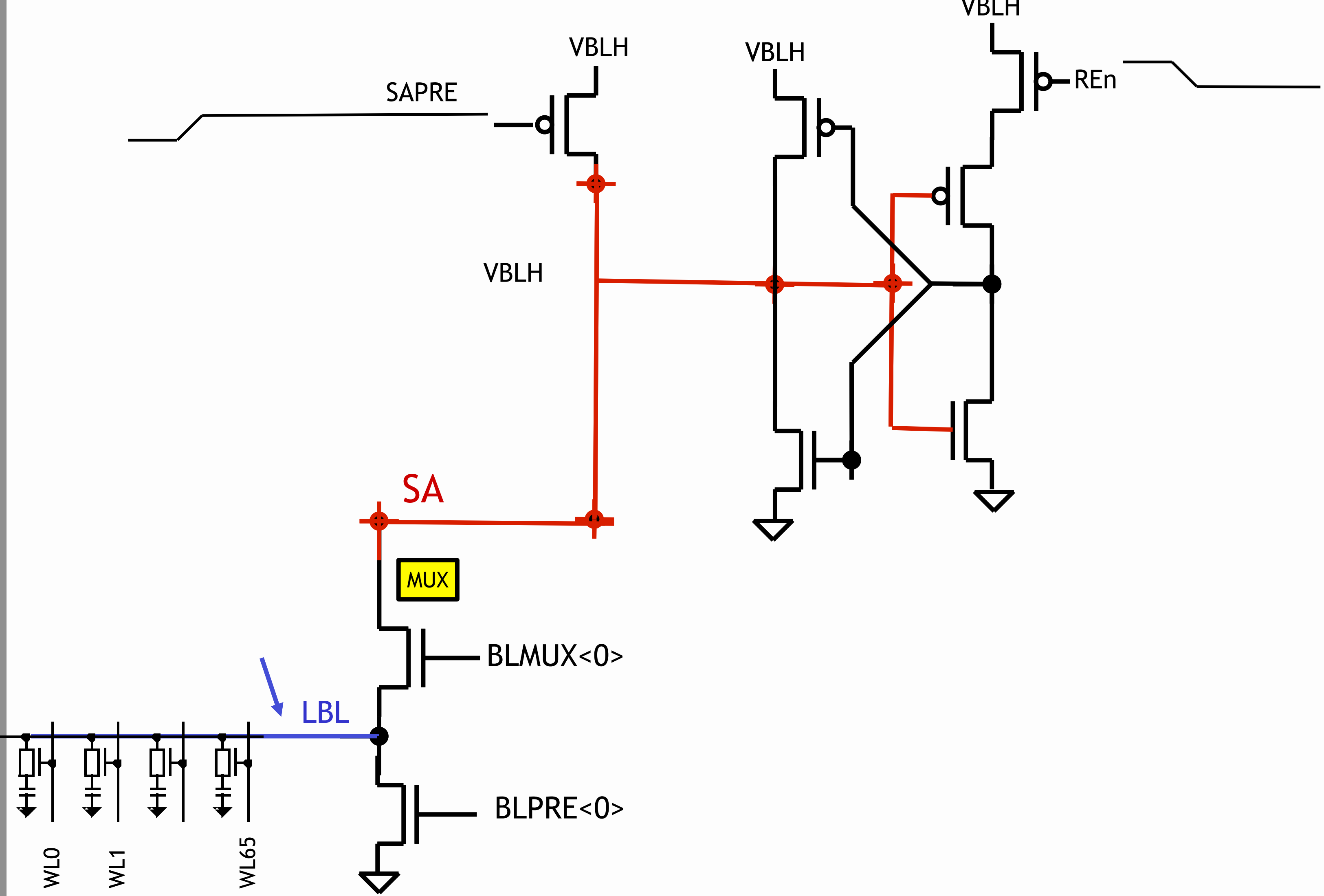


Gated Feedback Sense Amp - Construction



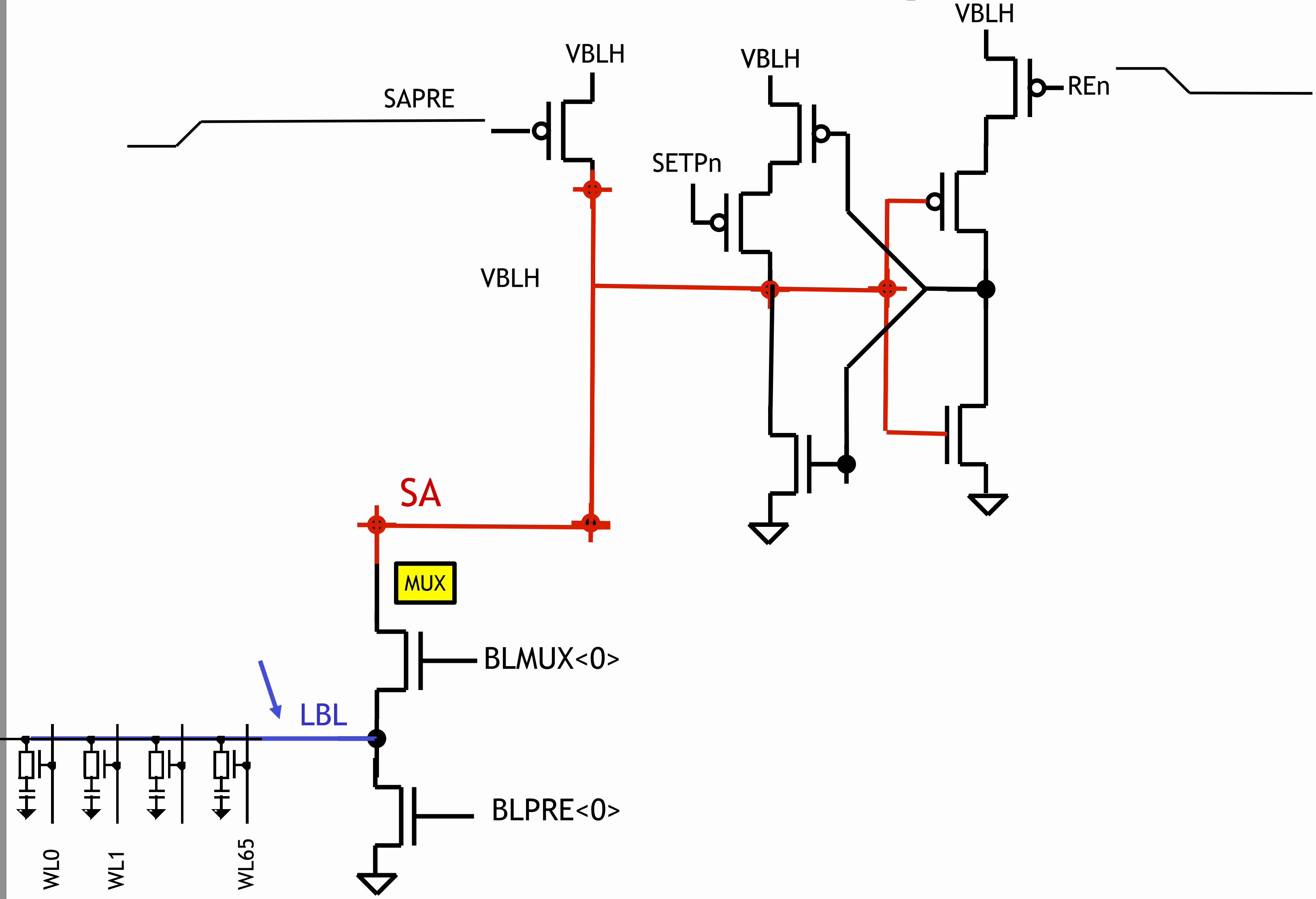
Animation

Gated Feedback Sense Amp - Construction



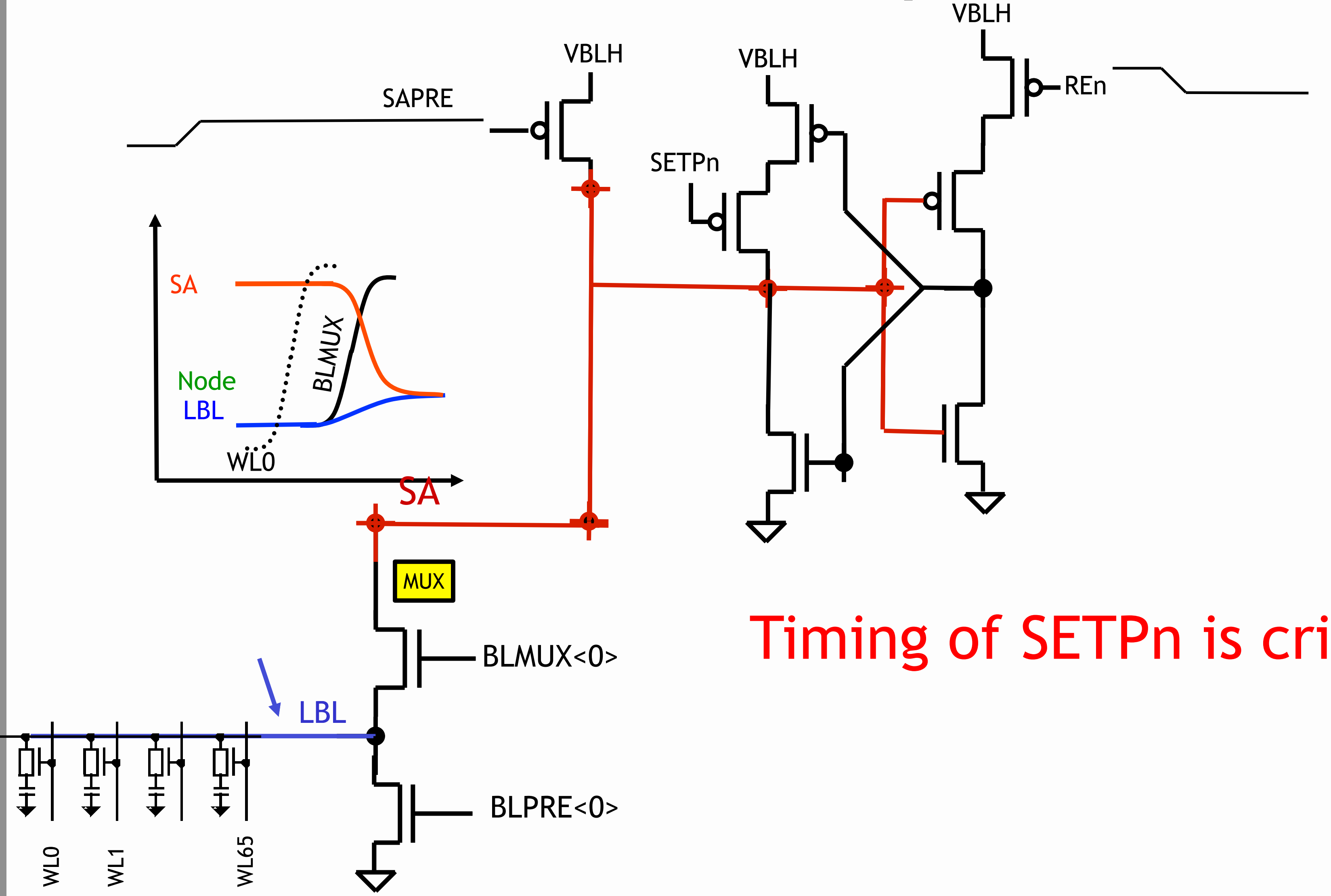
Animation

Gated Feedback Sense Amp - Construction



Animation

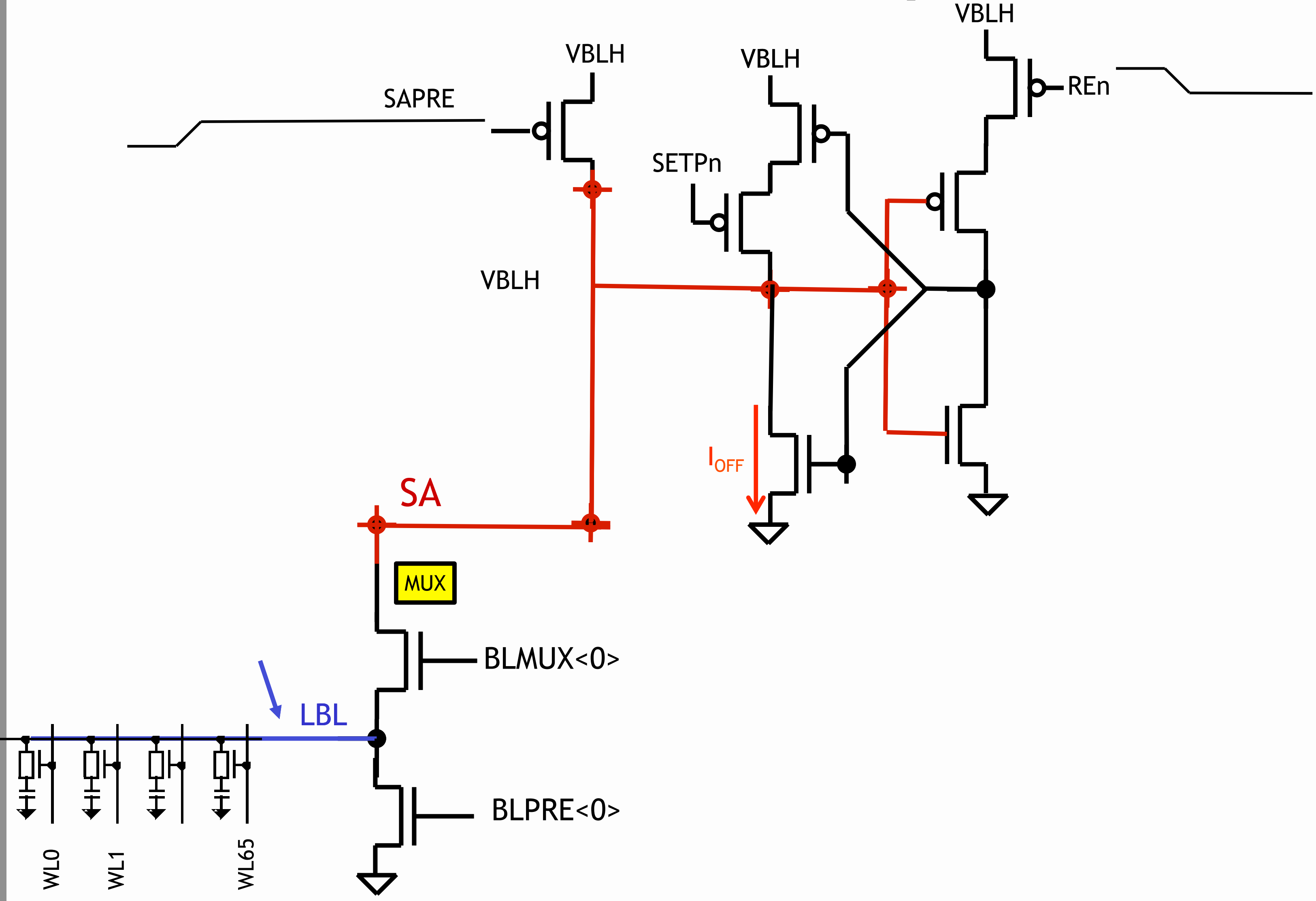
Gated Feedback Sense Amp - Construction



Timing of SETPn is critical

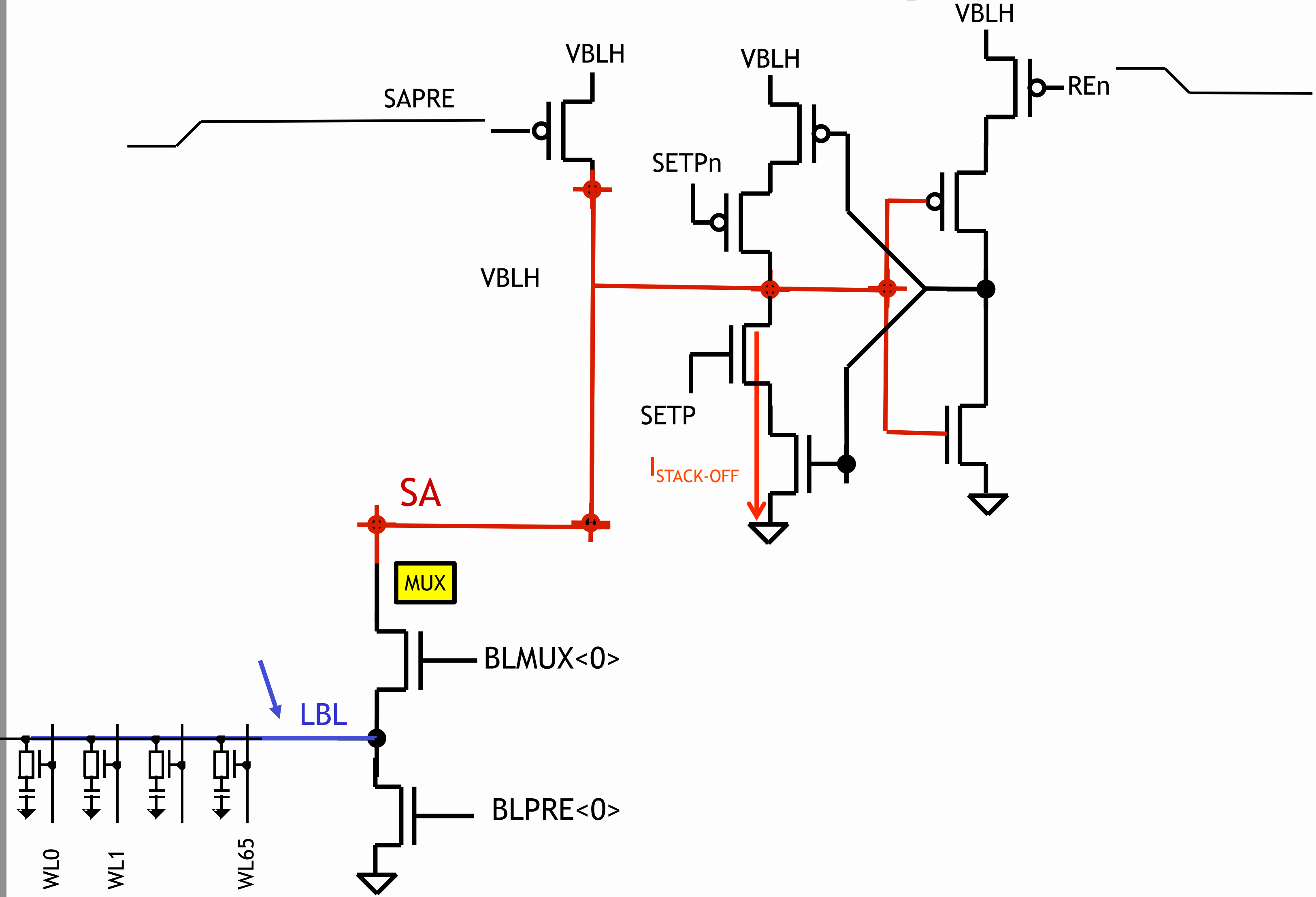
Animation

Gated Feedback Sense Amp - Construction



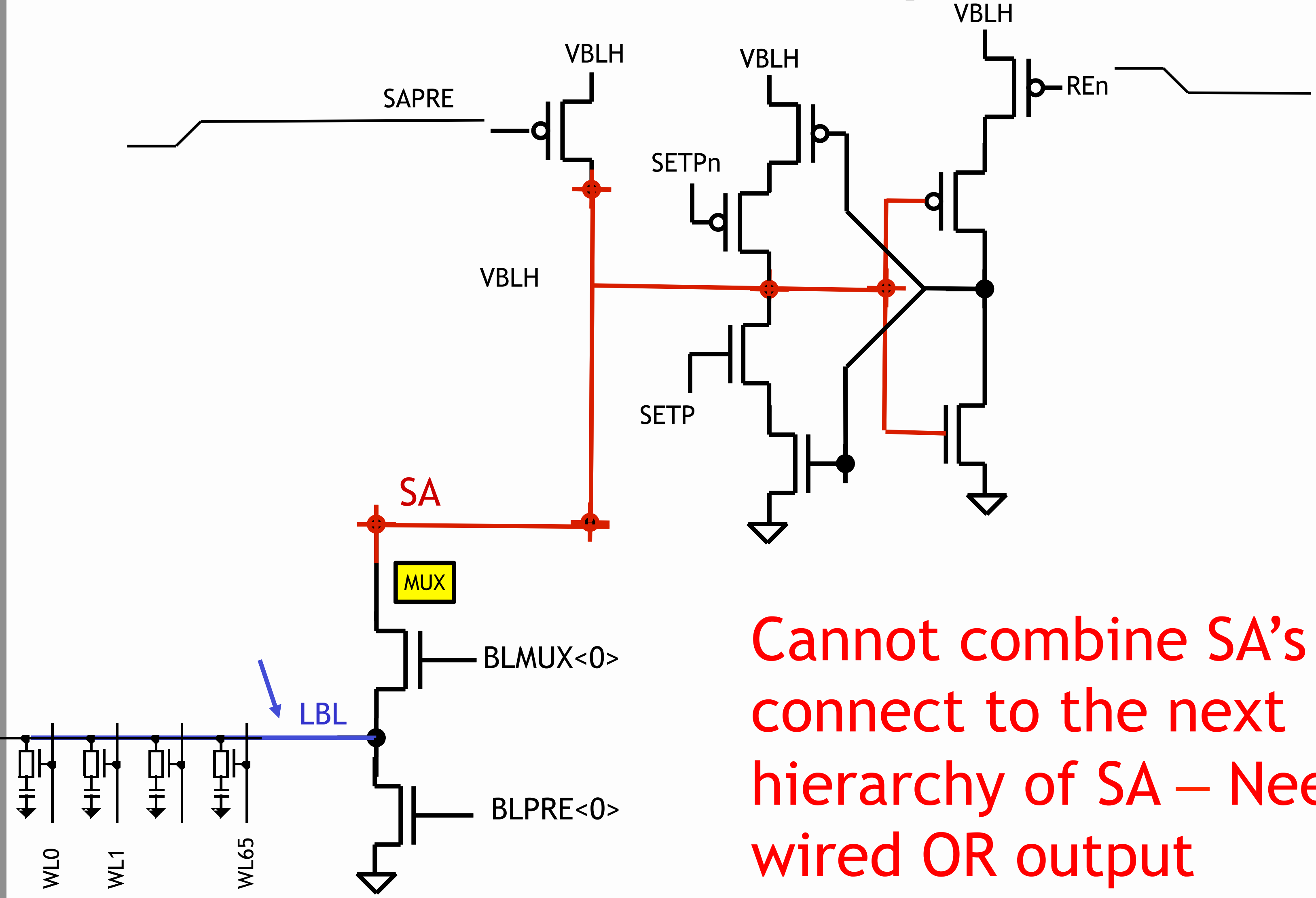
Animation

Gated Feedback Sense Amp - Construction



Animation

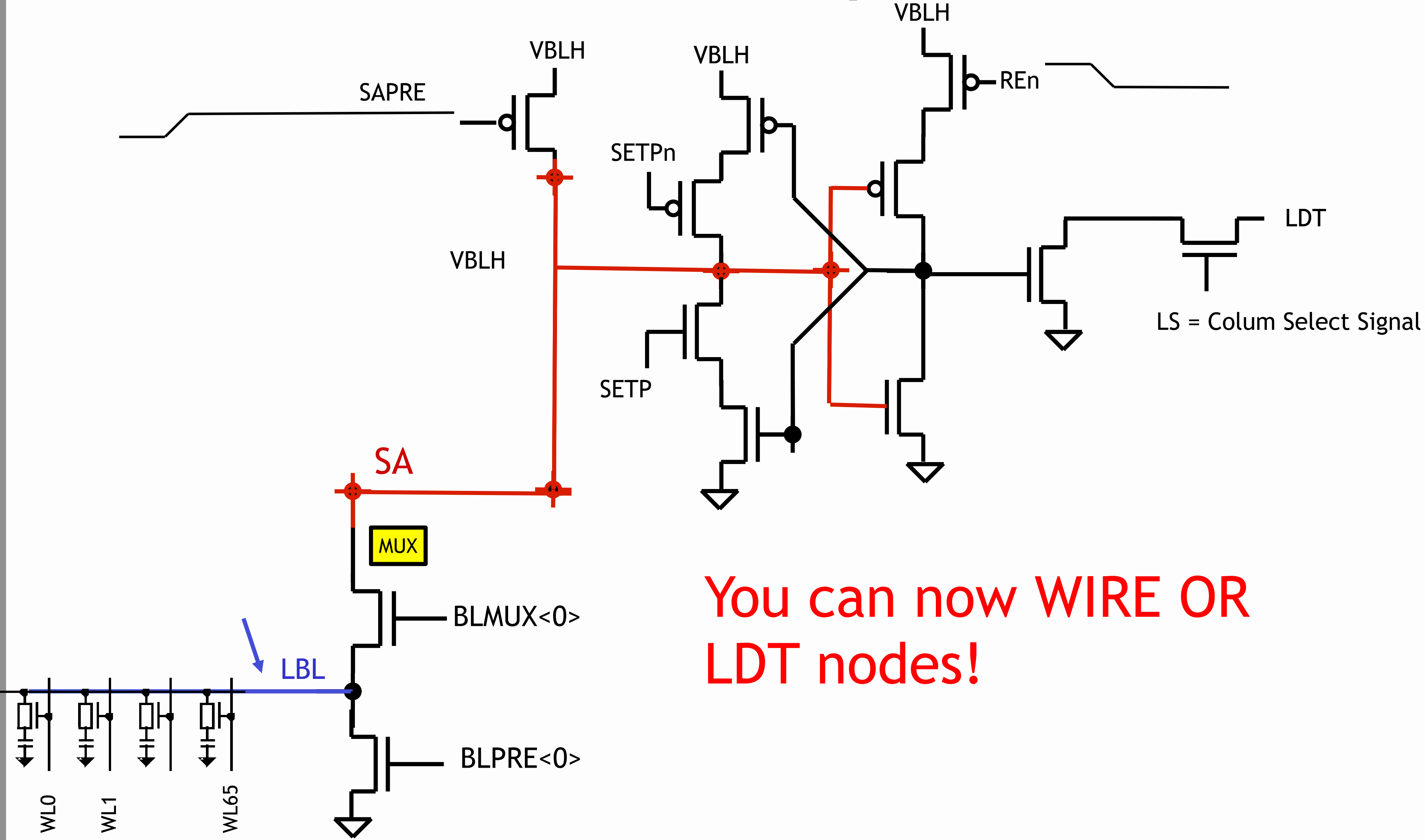
Gated Feedback Sense Amp - Construction



Cannot combine SA’s to connect to the next hierarchy of SA – Need a wired OR output

Animation

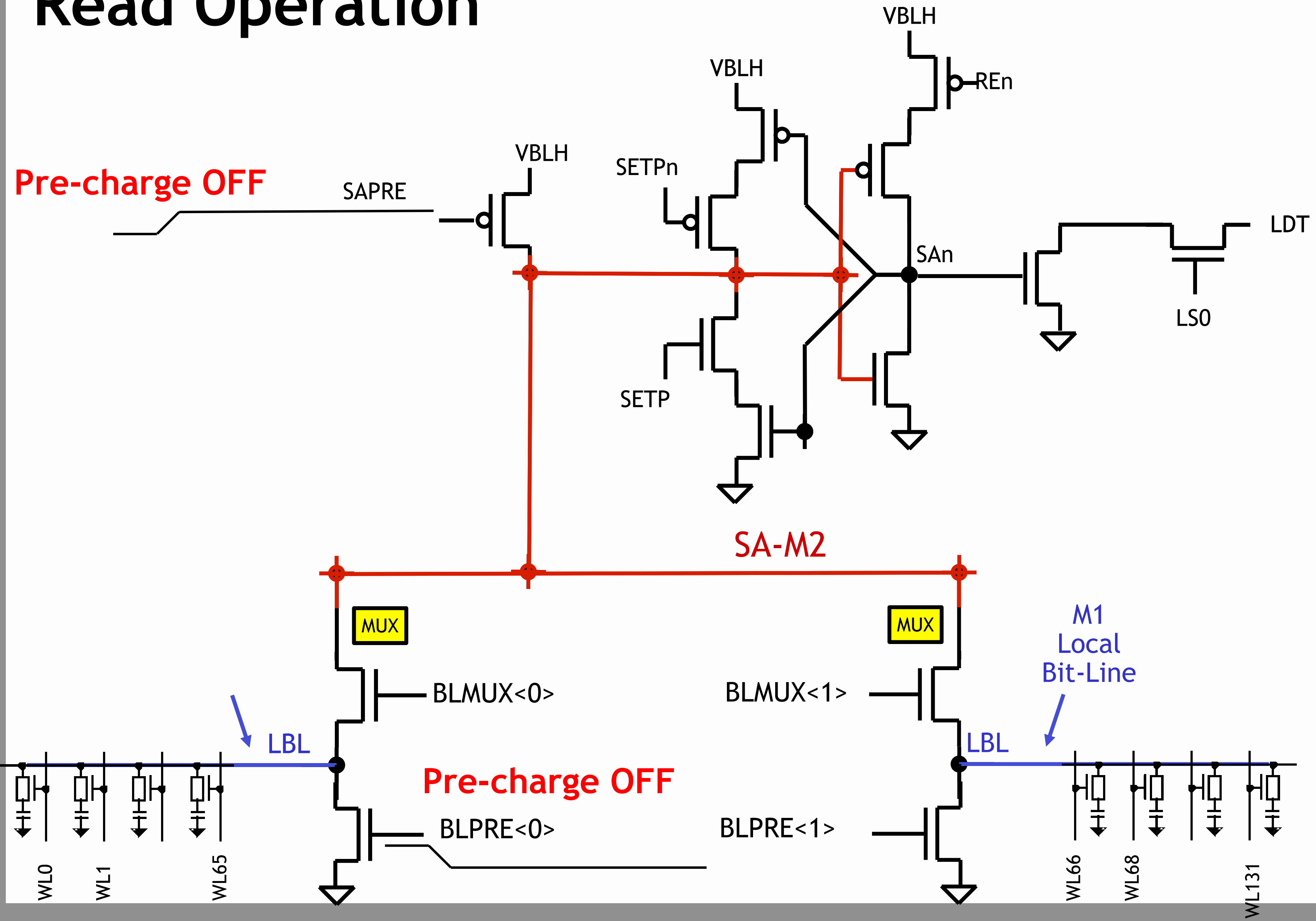
Gated Feedback Sense Amp - Construction



You can now WIRE OR LDT nodes!

Animation

Read Operation

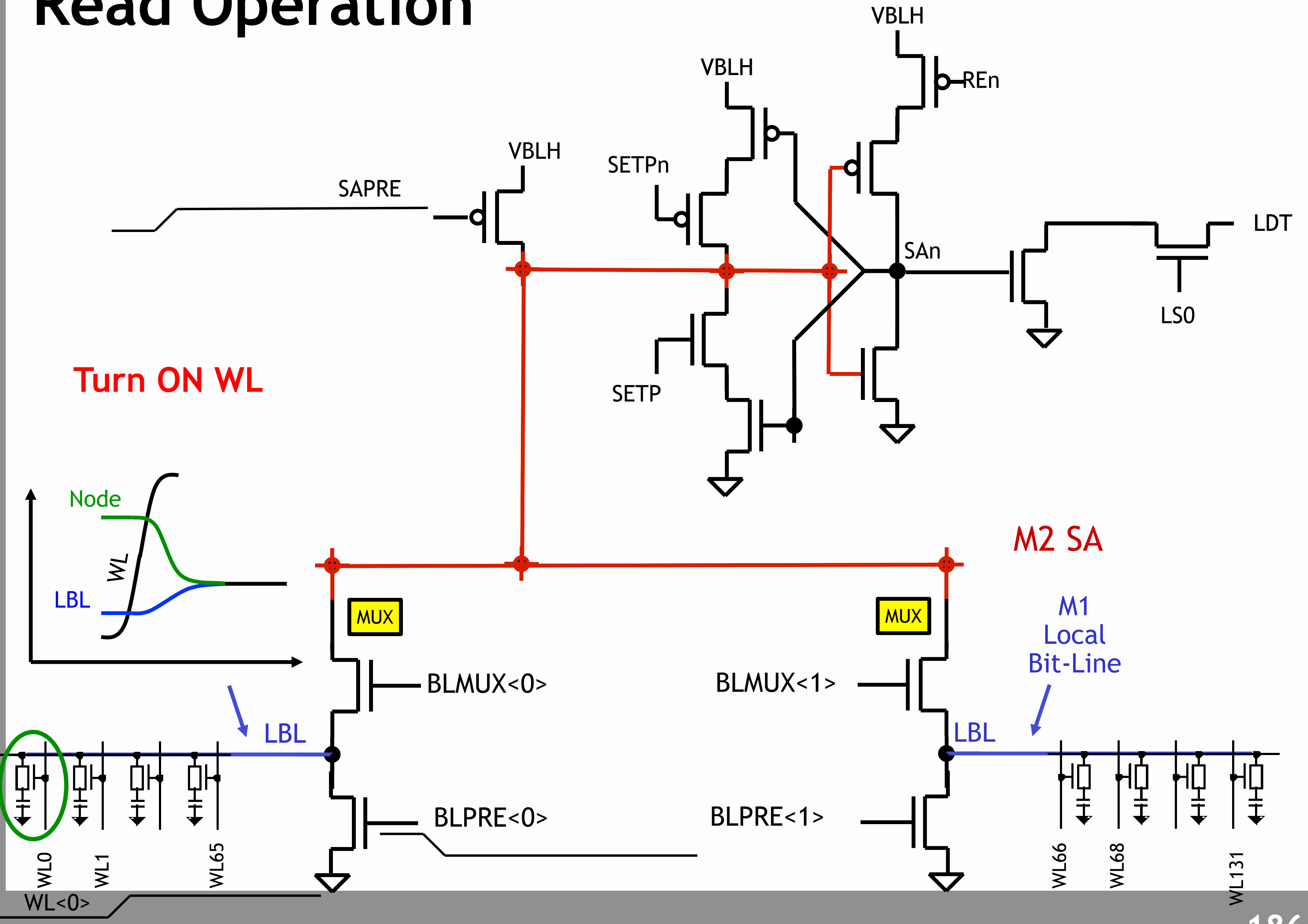


Pre-charge OFF

Pre-charge OFF

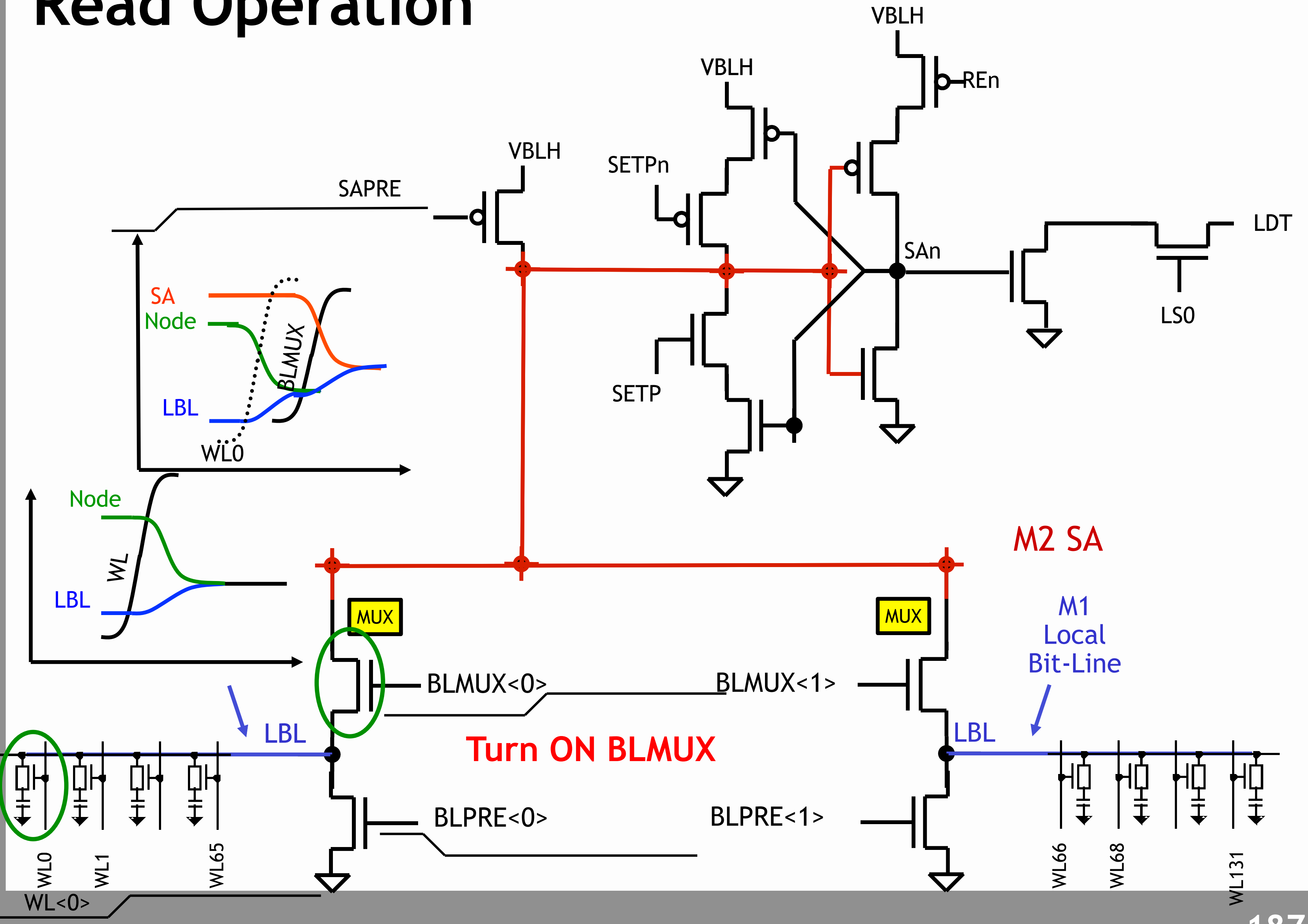
Animation

Read Operation



Animation

Read Operation

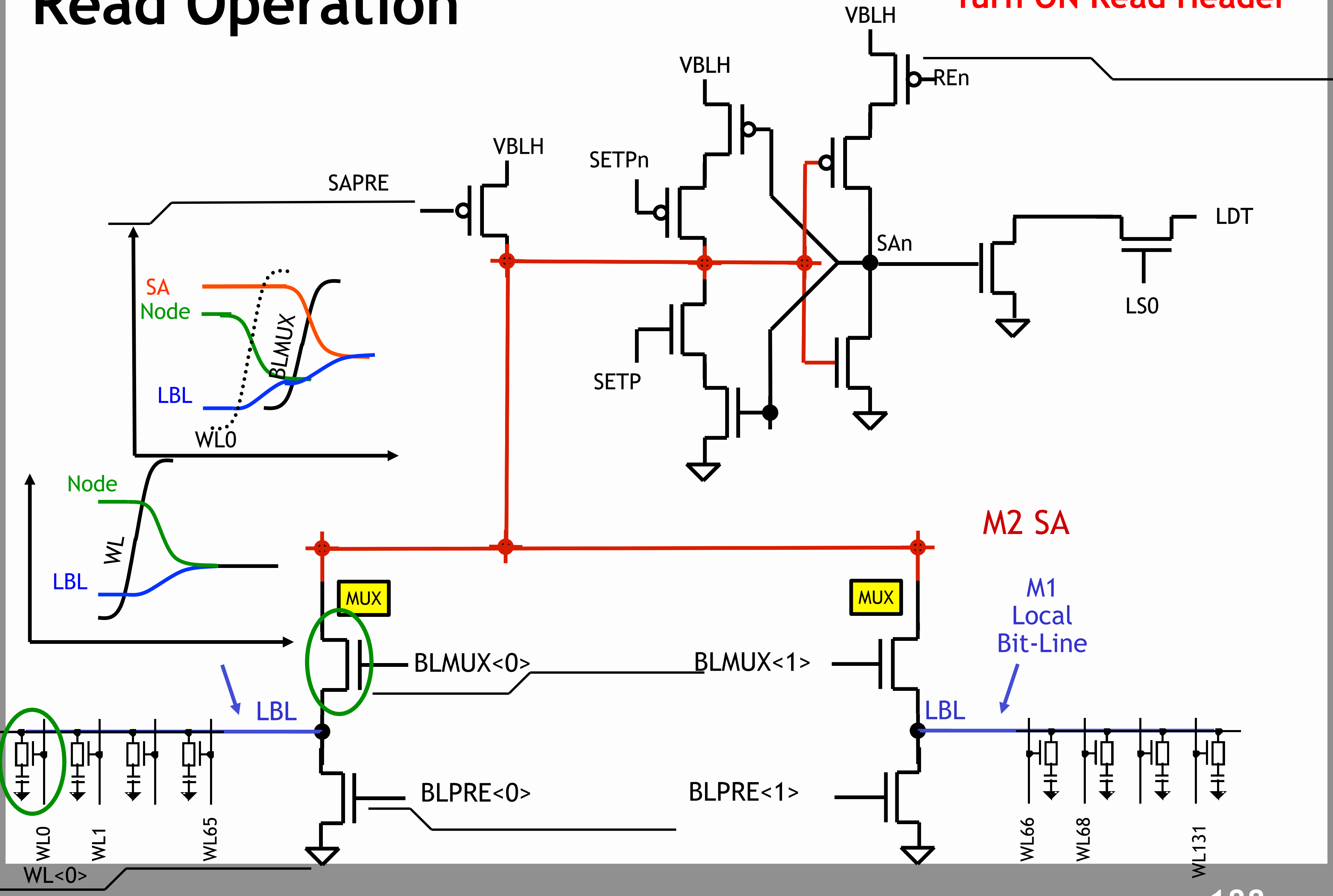


Turn ON BLMUX

Animation

Read Operation

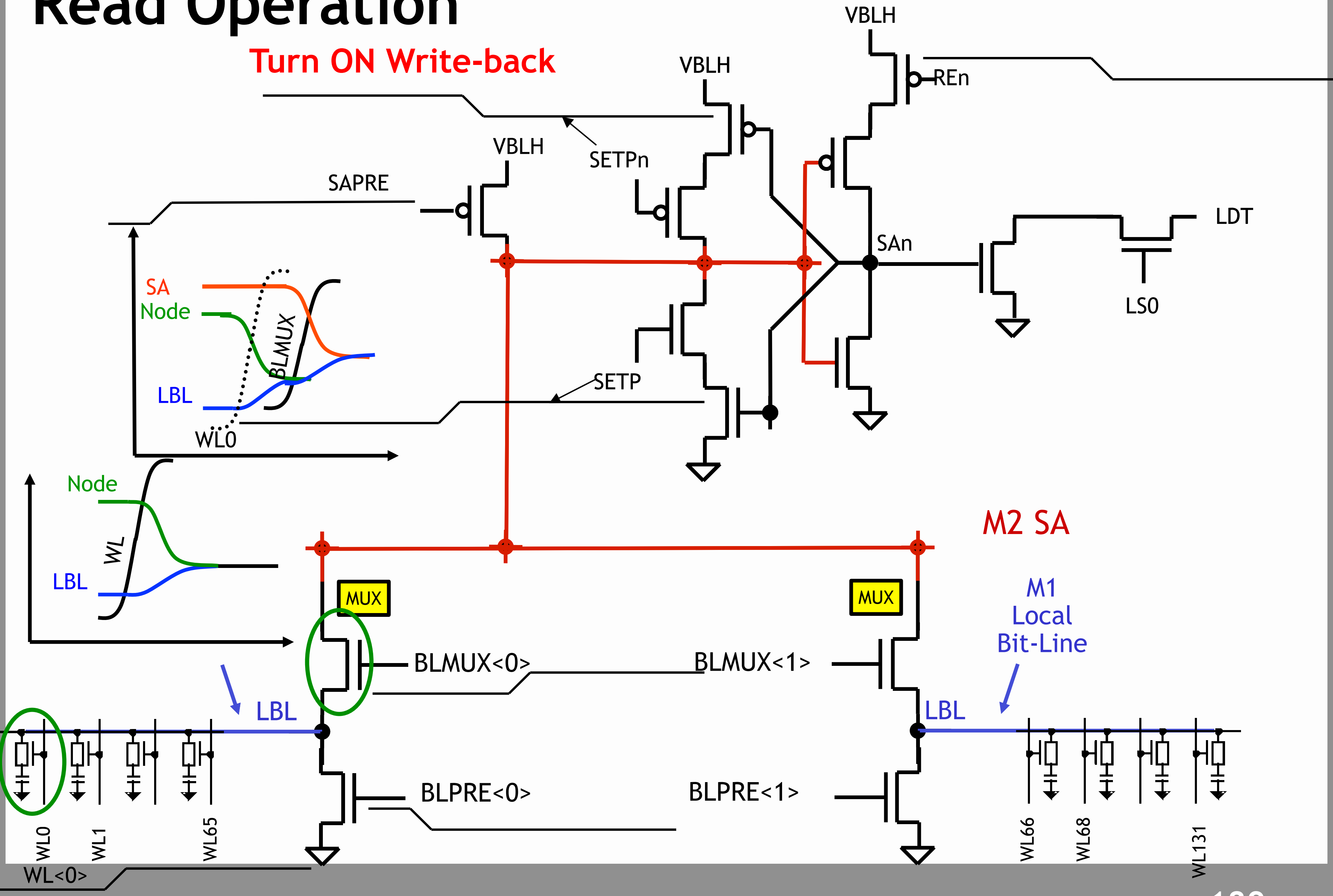
Turn ON Read Header



Animation

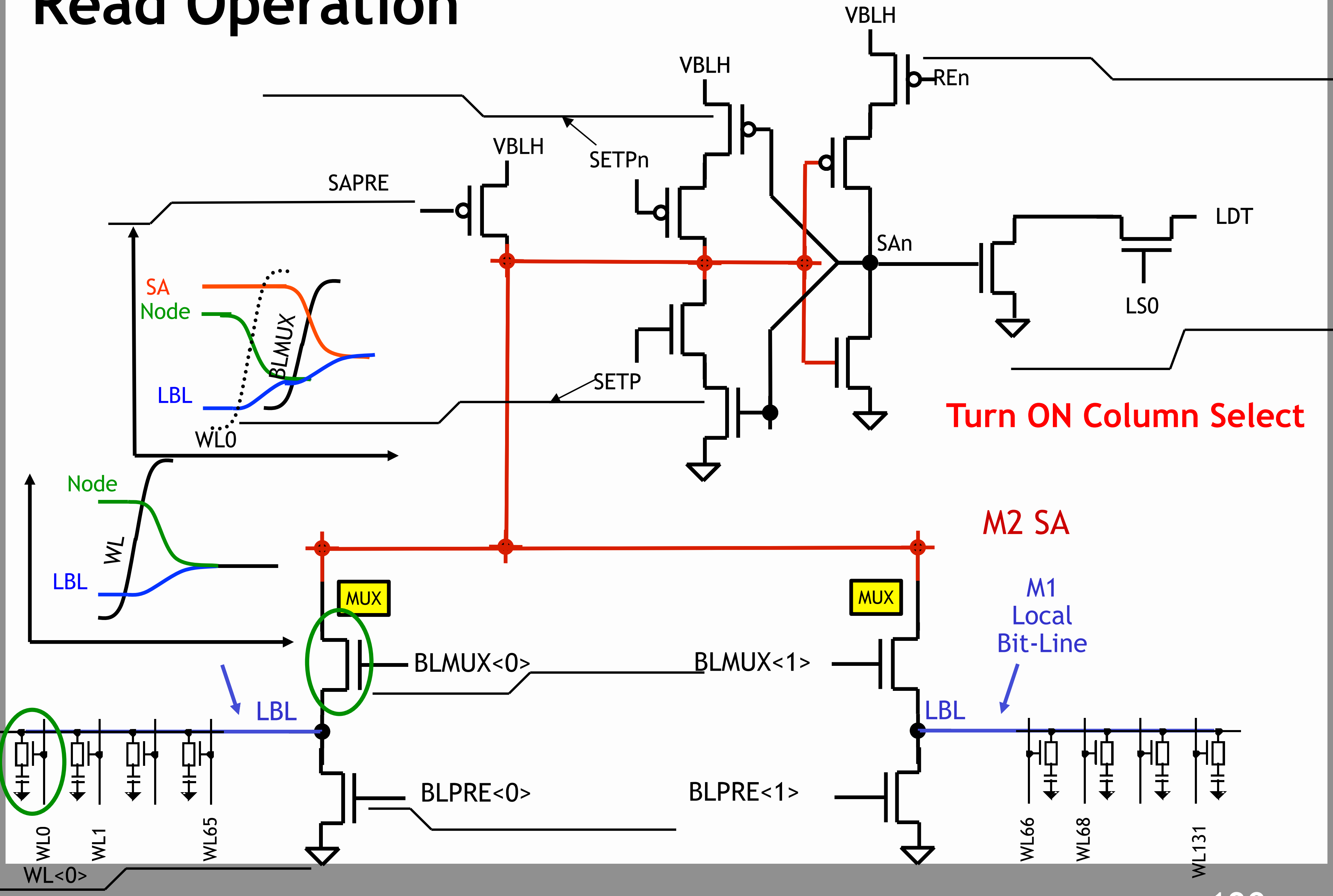
Read Operation

Turn ON Write-back



Animation

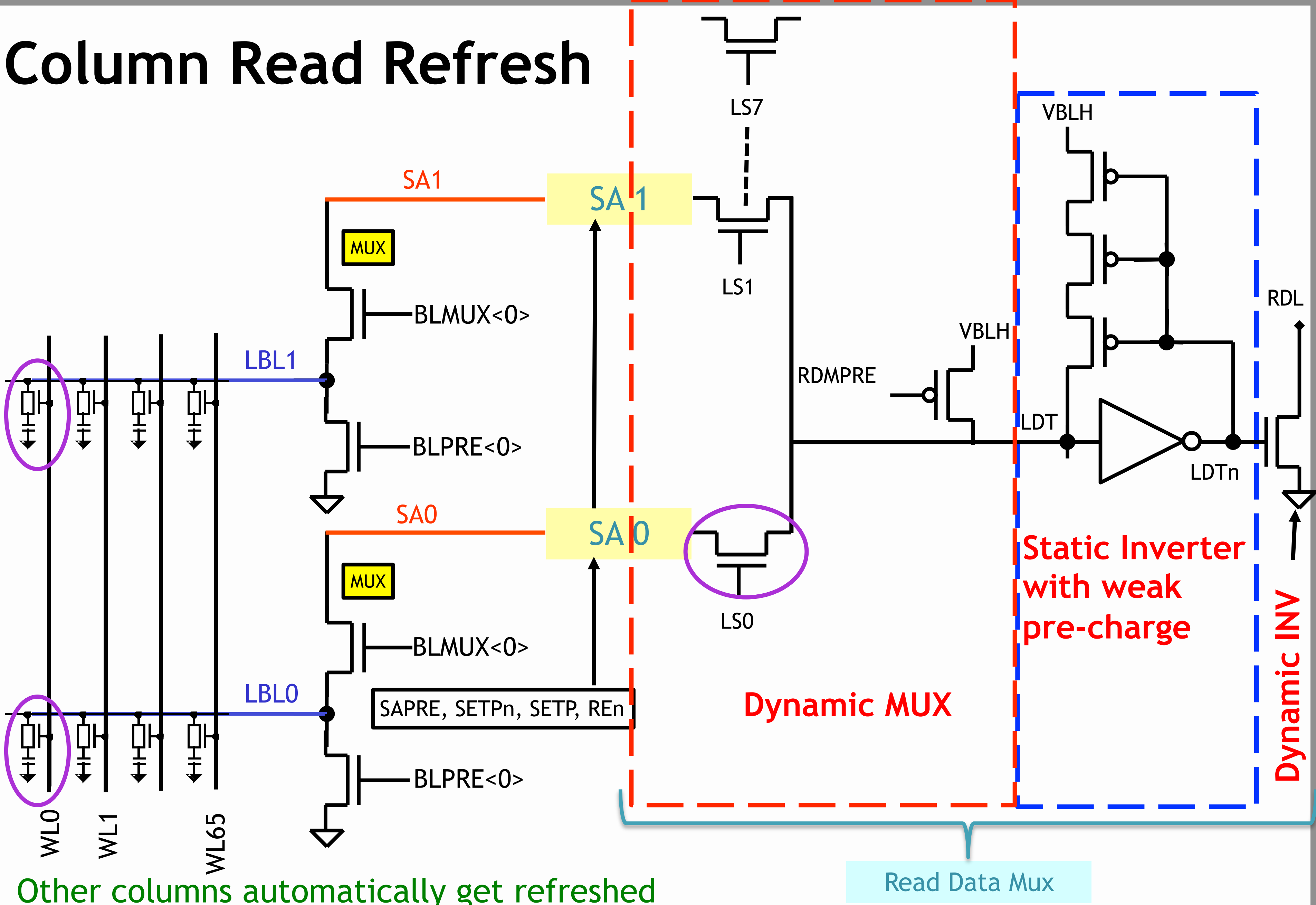
Read Operation



Turn ON Column Select

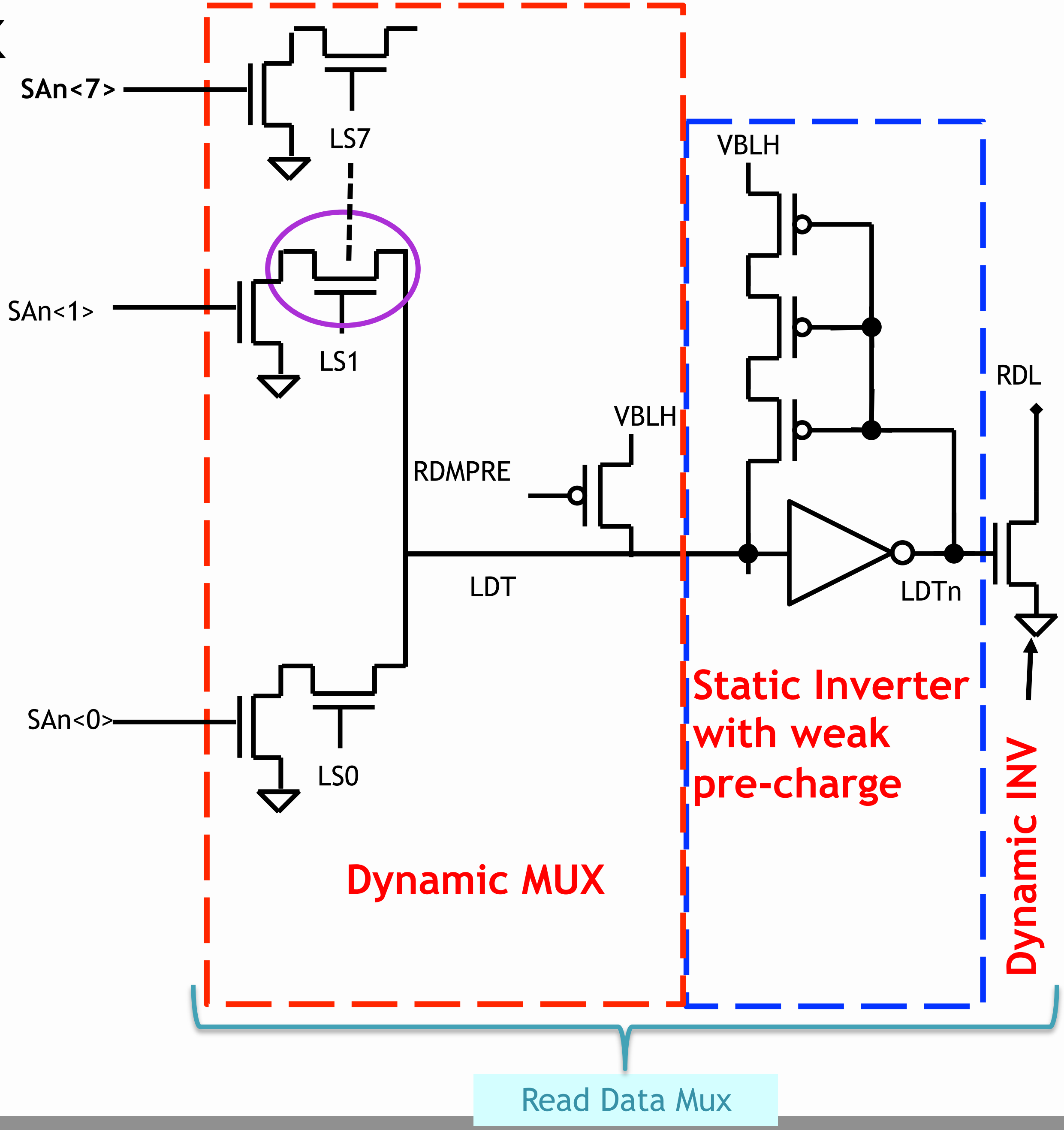
Animation

Column Read Refresh

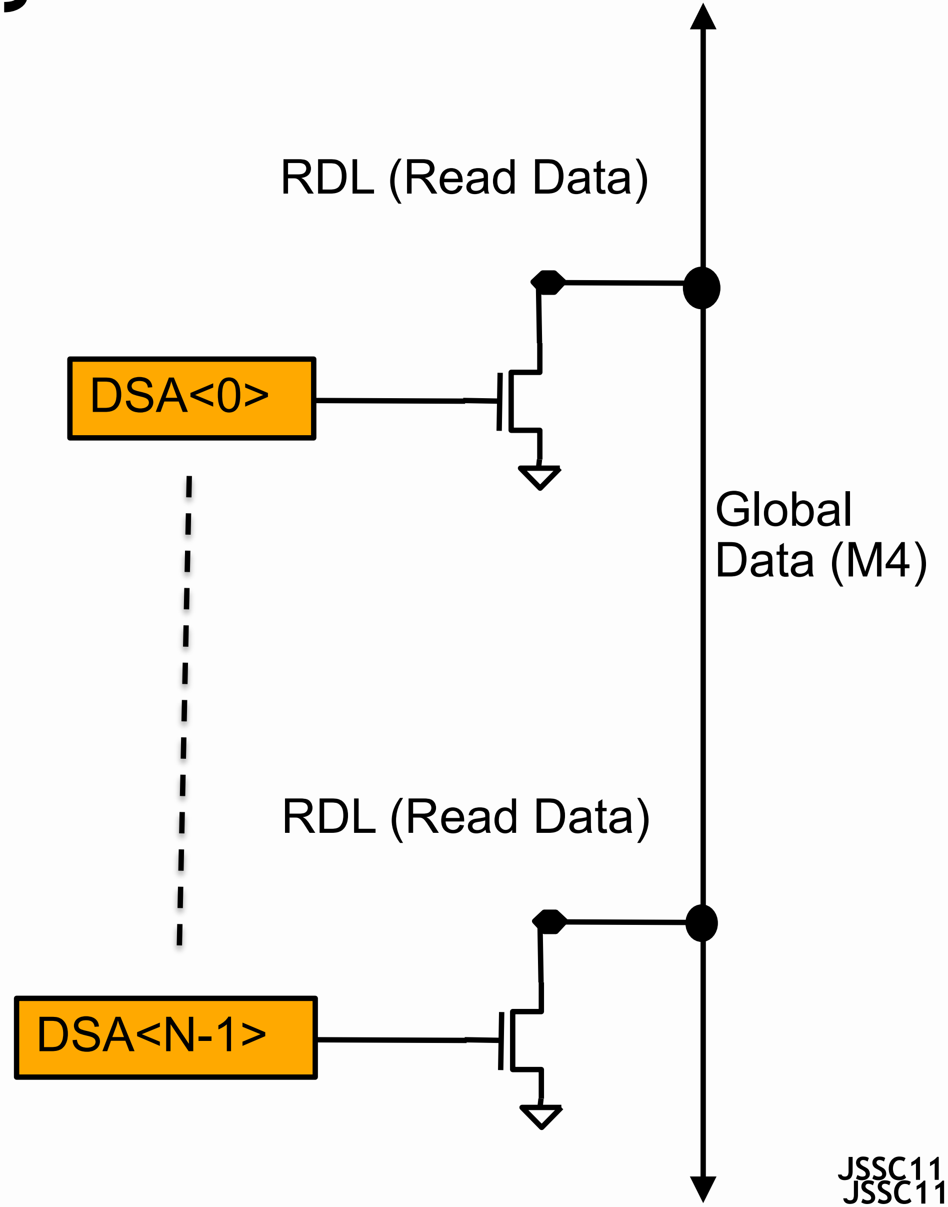


Other columns automatically get refreshed

Read Data Mux



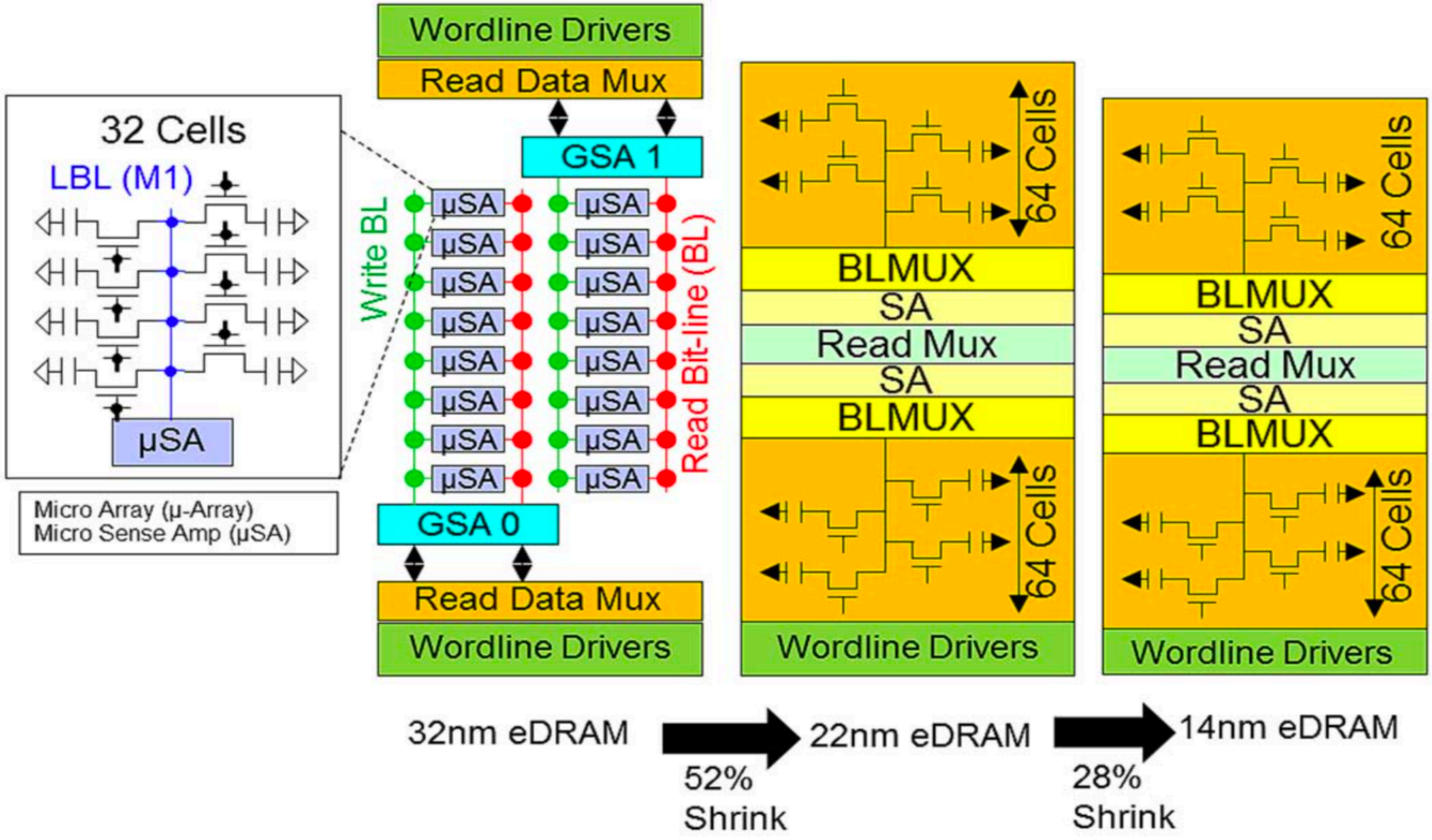
Combining RDM's- Dynamic NOR Gate



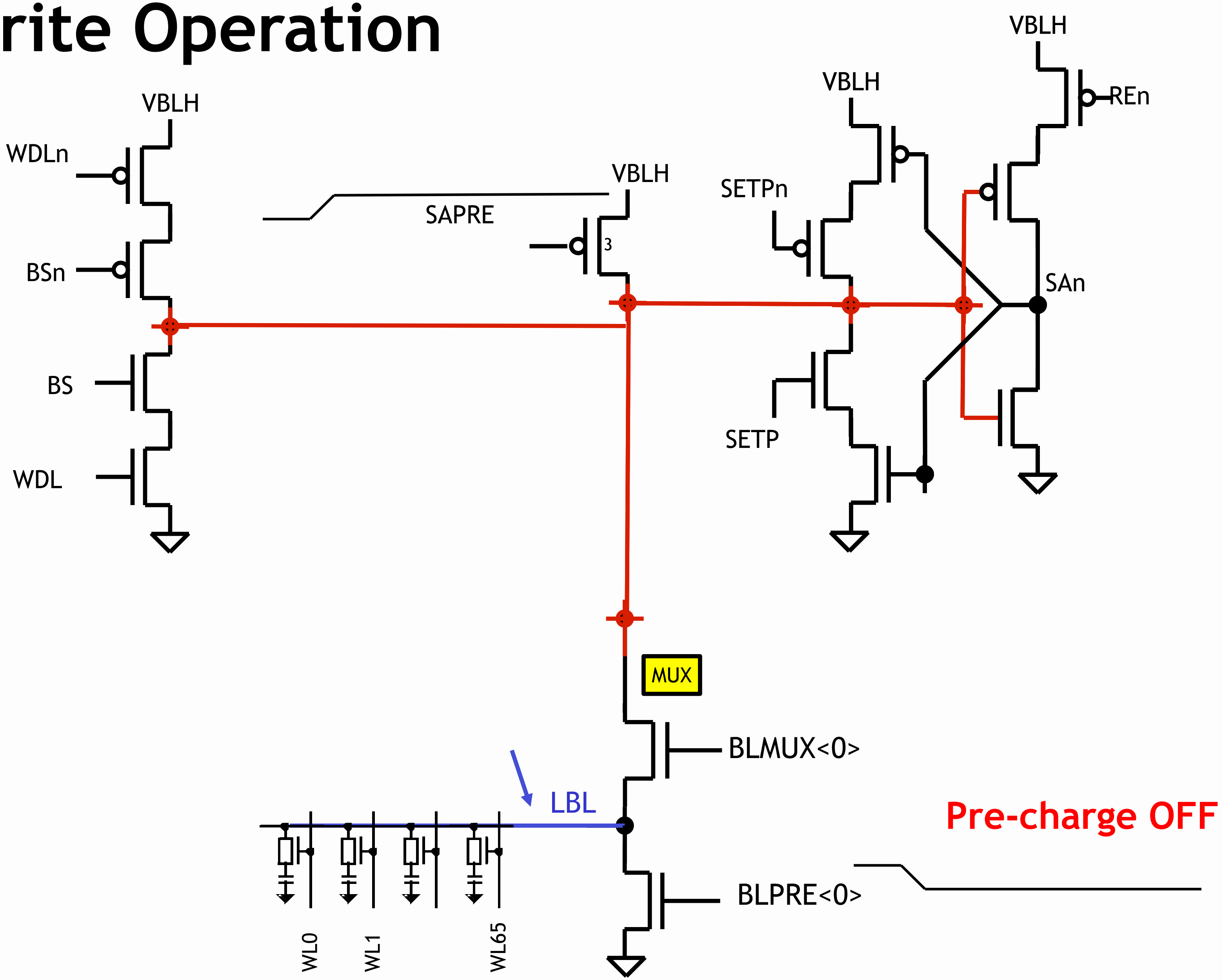
JSSC11

JSSC11

Area Savings and Comparison with 3T uSA



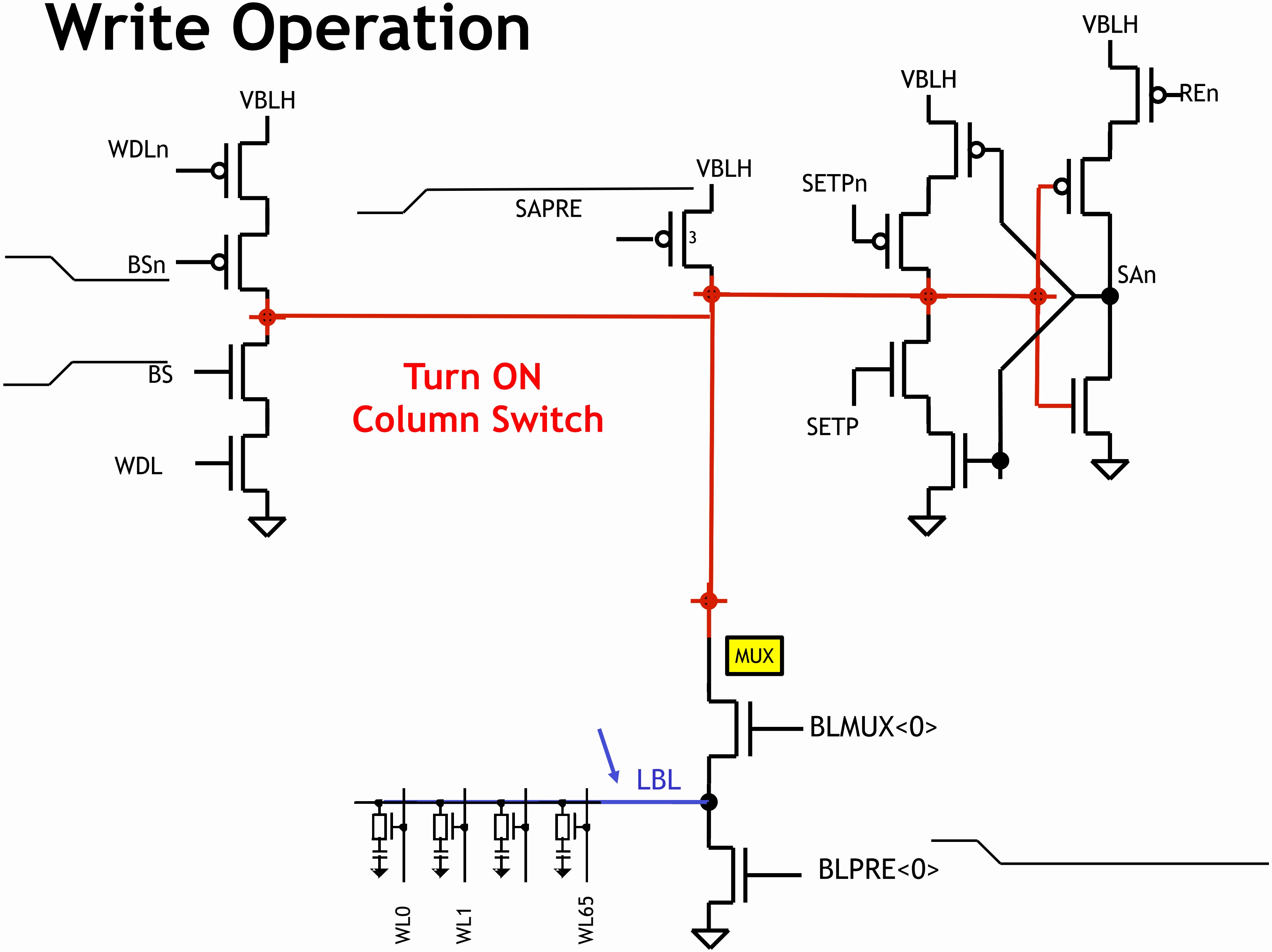
Write Operation



Pre-charge OFF

Animation

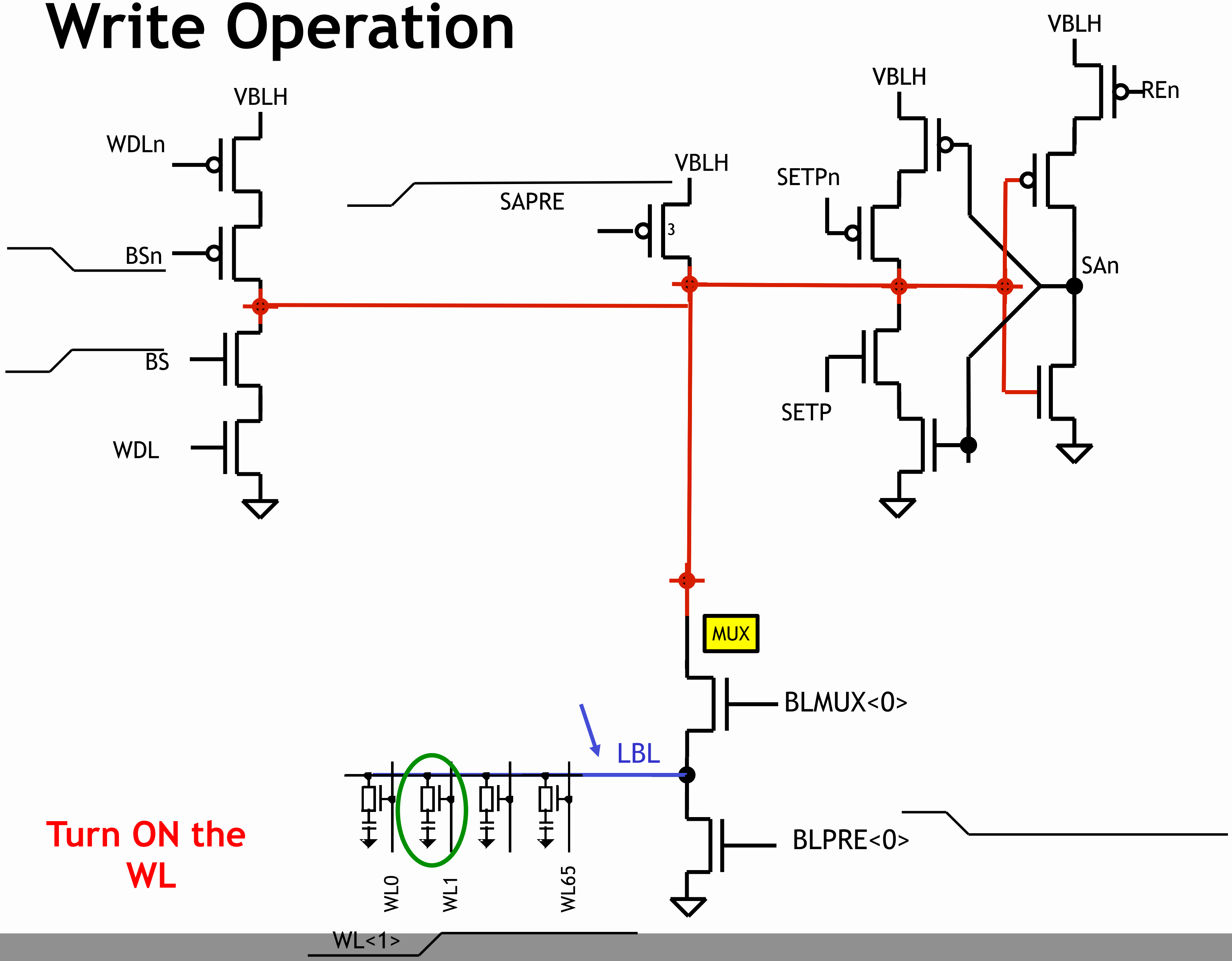
Write Operation



Turn ON
Column Switch

Animation

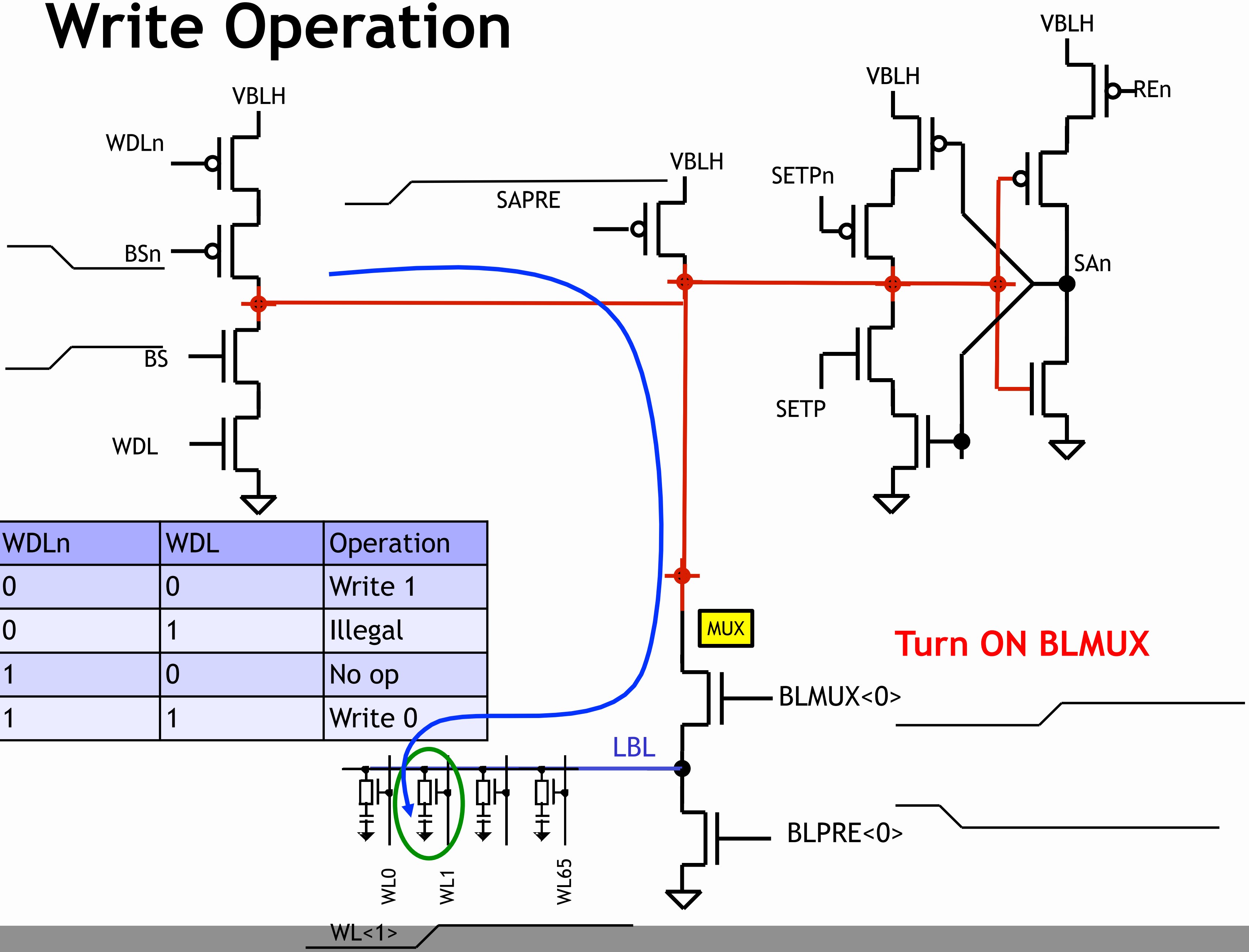
Write Operation



Turn ON the WL

Animation

Write Operation

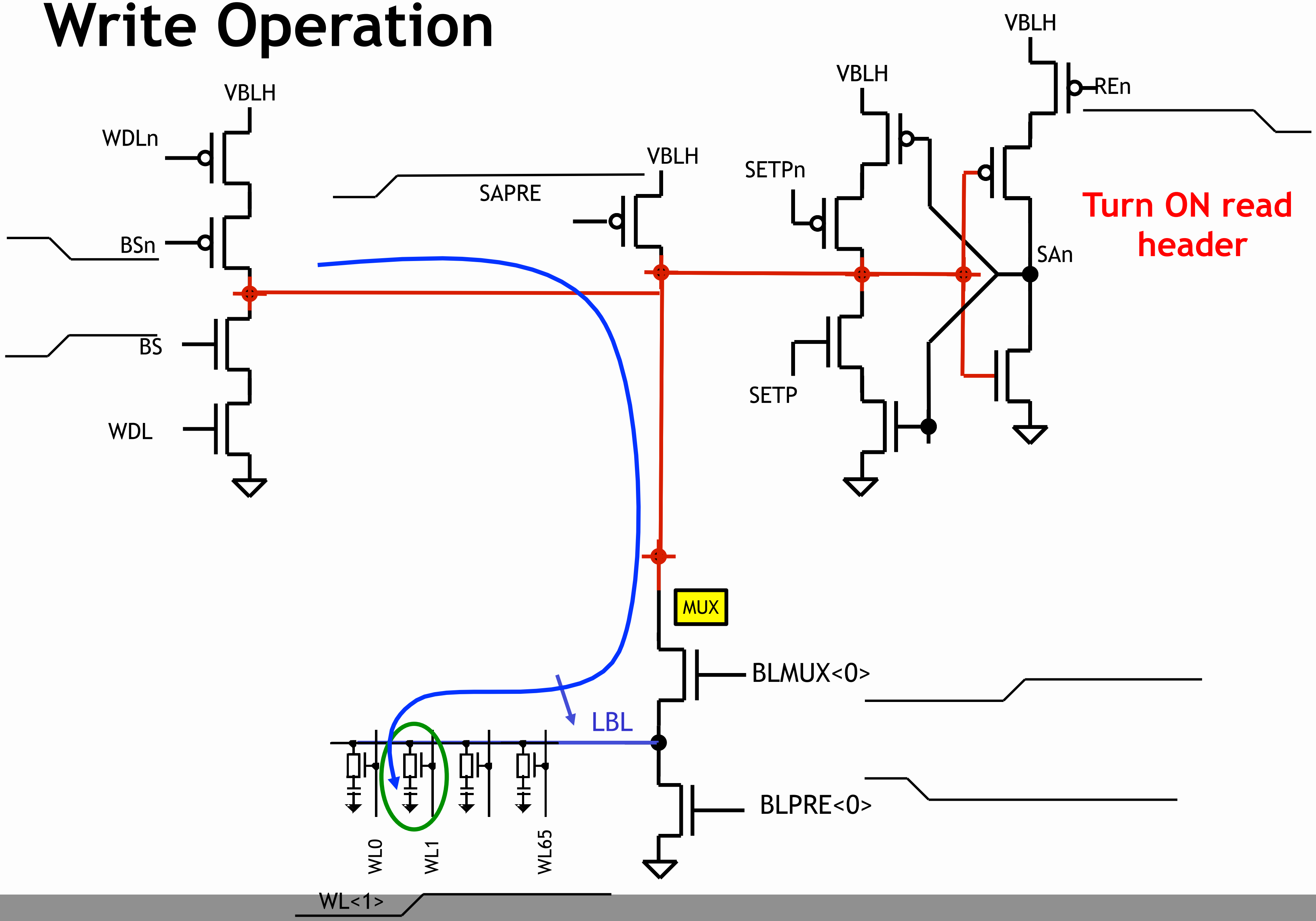


WDLn	WDL	Operation
0	0	Write 1
0	1	Illegal
1	0	No op
1	1	Write 0

Turn ON BLMUX

Animation

Write Operation

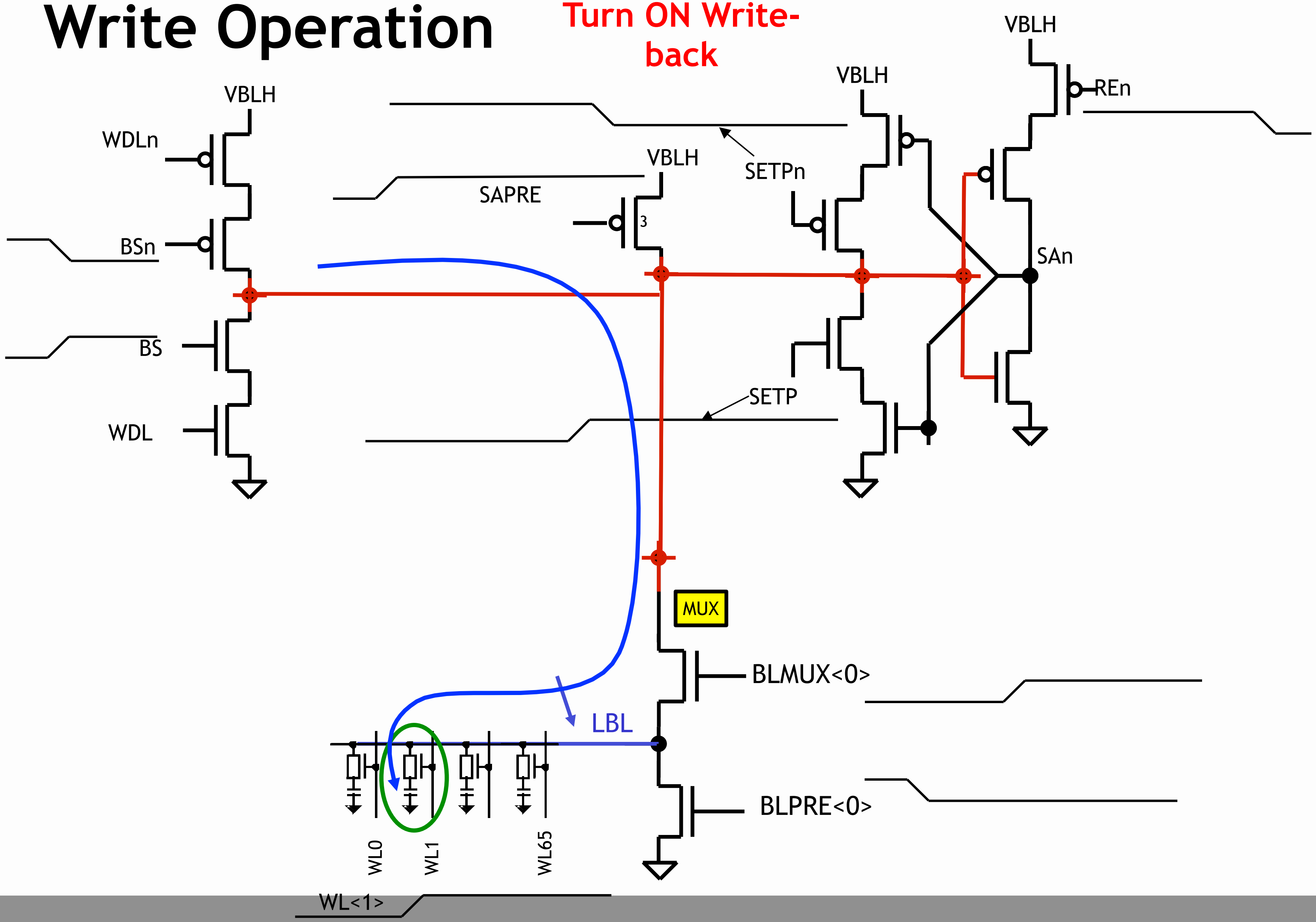


Turn ON read header

Animation

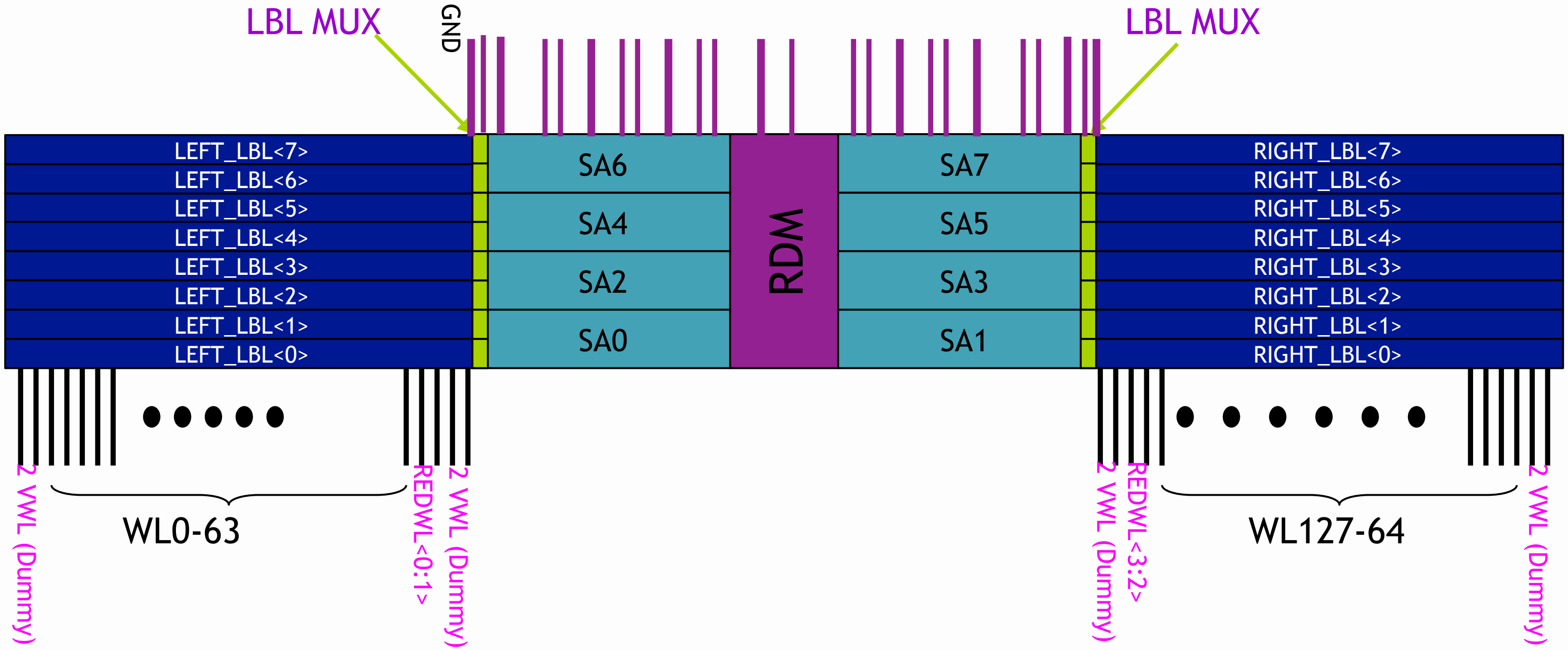
Write Operation

Turn ON Write-back



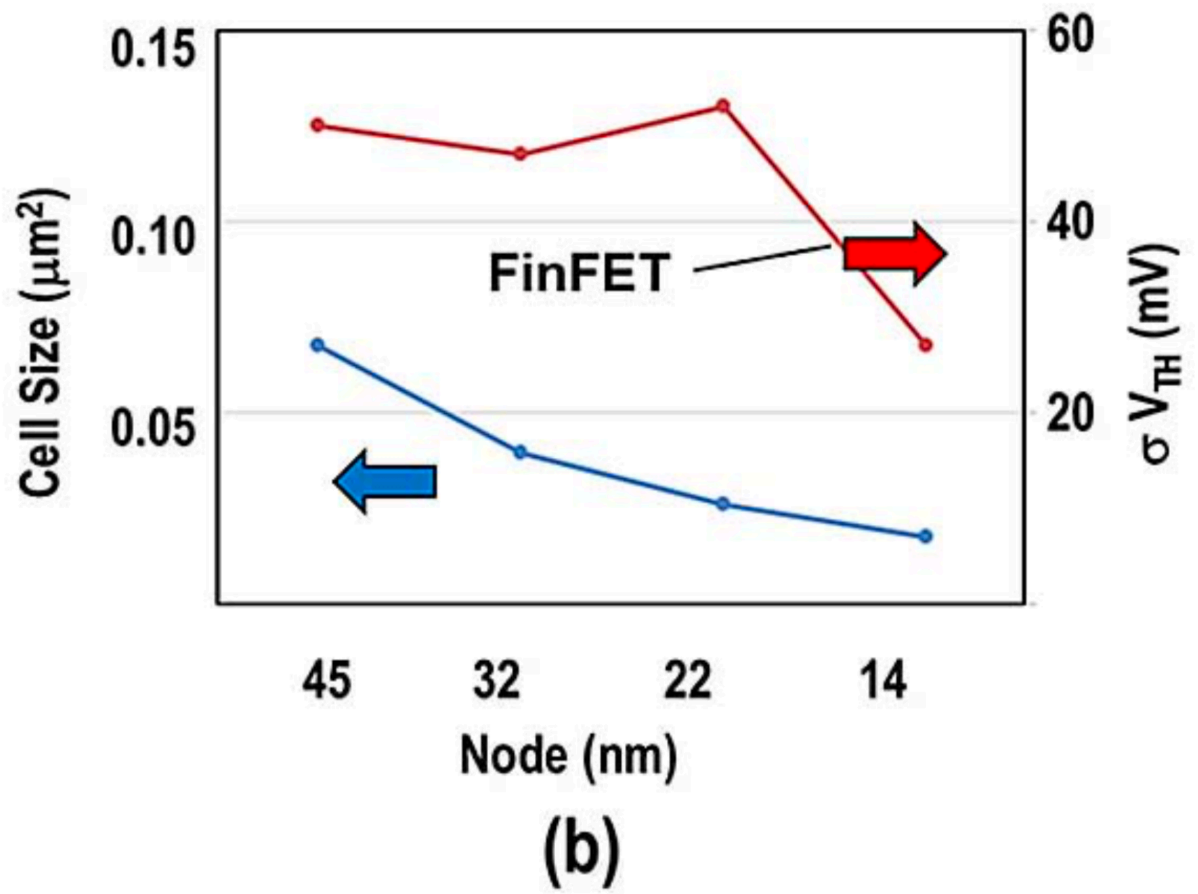
Animation

One Data Line Organization



- Single bit can be read out/ written into by selecting one of 128 rows and one of 8 columns
- The components are sized and arranged to make the layout nice and rectangular
- Repeat this structure as many as there are Data-lines

14nm FinFET Advantage

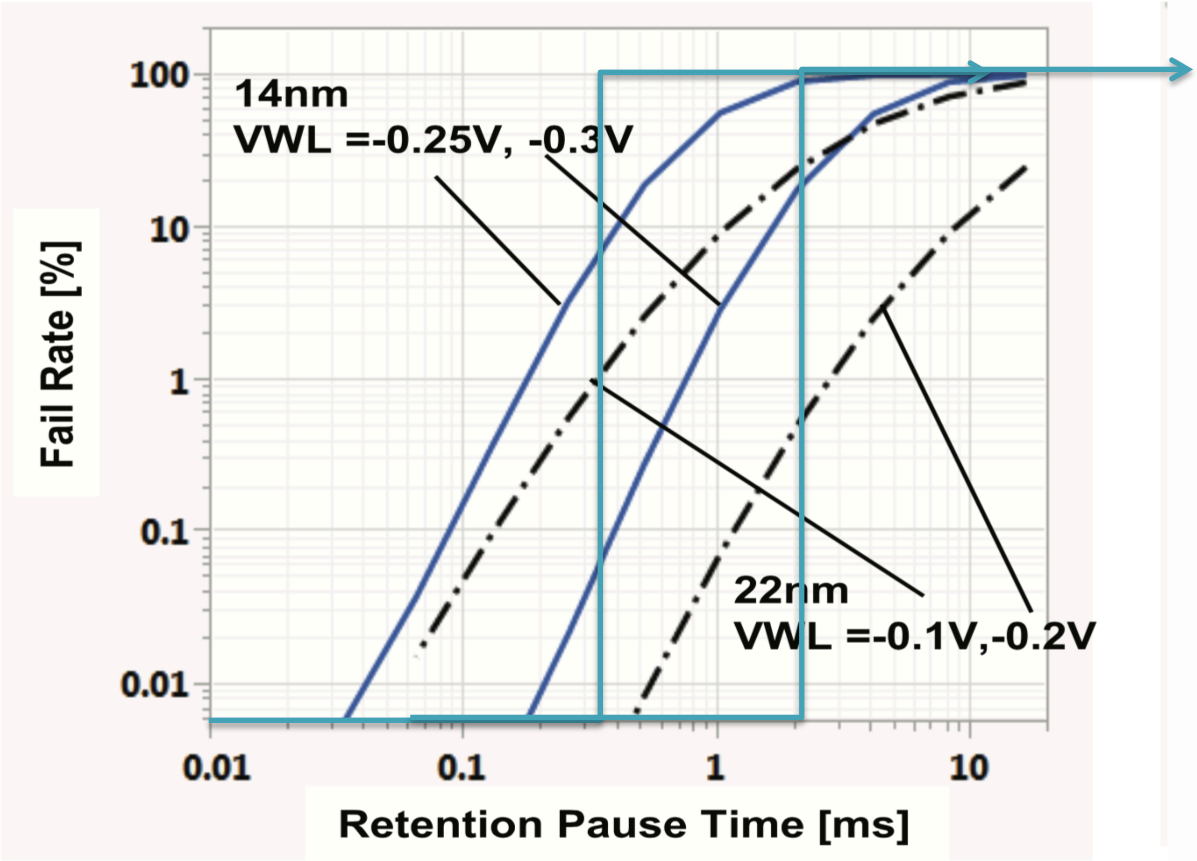


14nm Access Device is 2.5X stronger than the 22nm planar device due to

- 50% more effective width
- 42% shorter channel length
- Lower target Vth

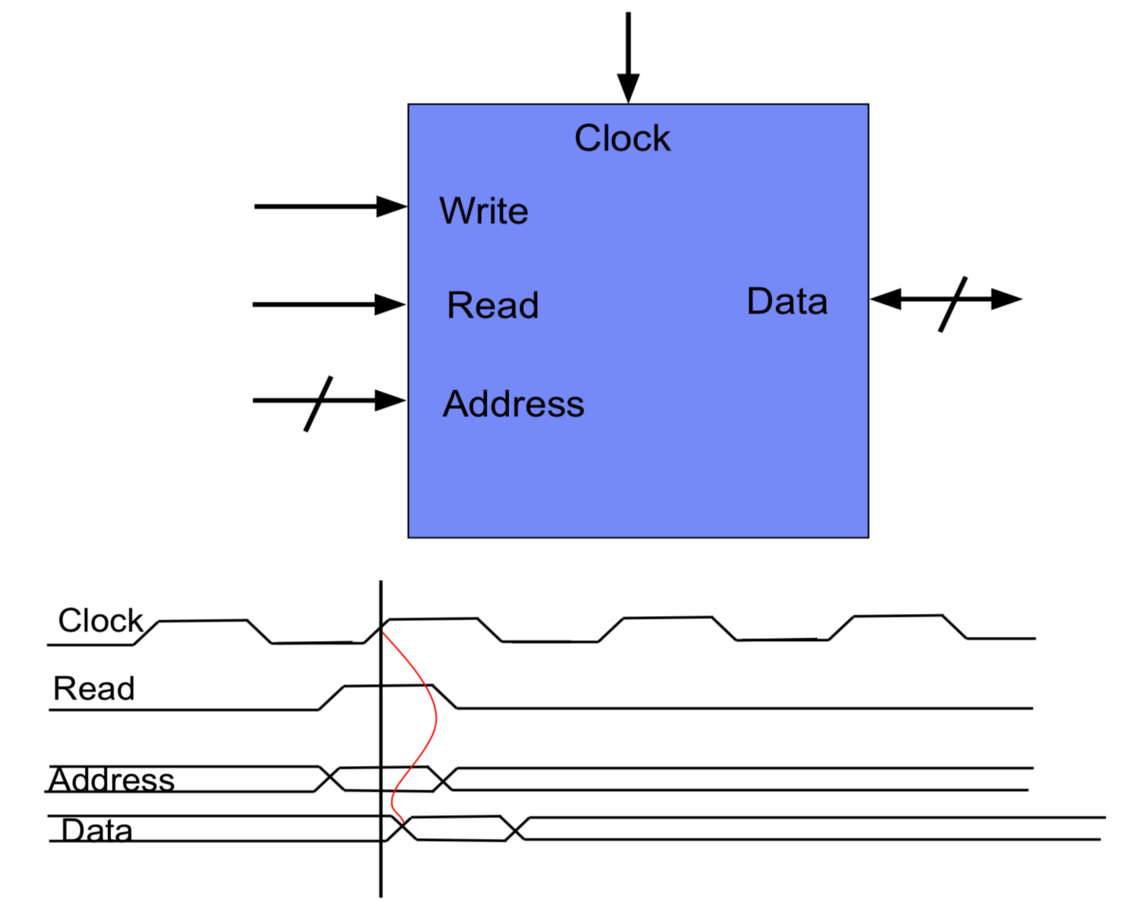
Lower VT variation due to undoped channel

Lower Vth Variation Effect on Retention



- Write a 1 into all the cells
- Read the cells after a pause time
- Ideally (with no local variations) there should be an step jump in the #fail
  - With variations, steeper the slope lesser the variations

Memory Testing



- No access to internal signals
- Need to detect and fix faults from outside

Memory Faults

- Data faults – Pattern dependent
- Address faults
- Technology Specific faults
- Need to come up with test patterns to detect various faults.

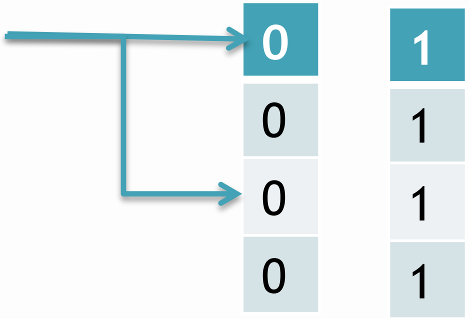
Address Faults

- Write zero to all addresses in increasing order: ↑{W0}
- Read zero from all addresses in increasing order: ↑{R0}
- Write one to all addresses in decreasing order: ↓{W1}
- Read one from all addresses in decreasing order: ↓{R1}

- Decoder error
  - Multiple WLs might fire at the same time.
  - No WL might fire

Address Faults

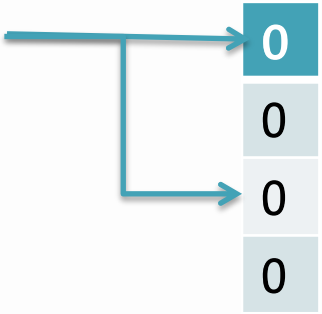
- Test pattern:
  - ↑{W0}
  - ↑{R0}
  - ↑{W1}
  - ↑{R1}



- Decoder error
  - Not sufficient to just read OR write in a single operation

Decoder Address Fault

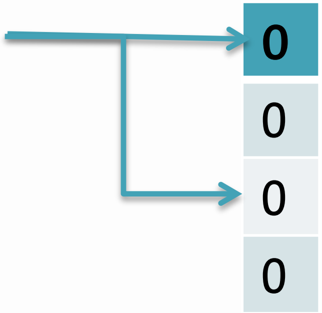
- Test pattern:
  - ↑{W0}
  - ↑{R0 W1}
  - ↑{R1}



- Decoder error
  - – Not sufficient to just read OR write in a single operation

Decoder Address Fault

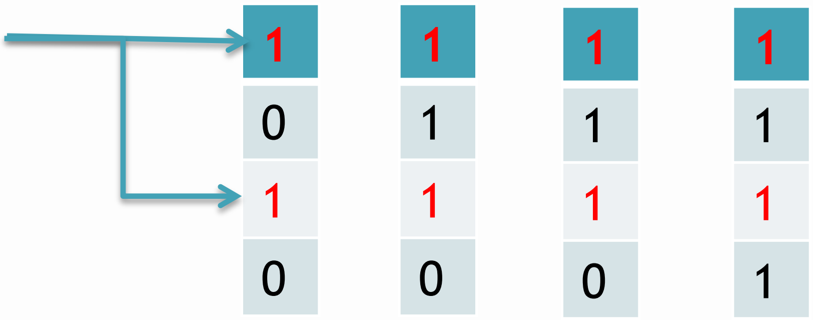
- Test pattern:
  - ⇑{W0}
  - ⇑{R0 W1}
  - ⇑{R1}



- Decoder error
  - – Not sufficient to just read OR write in a single operation

Decoder Address Fault

- **MARCH Pattern**
  - ↑{W0}
  - ↑{R0 W1}
  - ↑{R1}

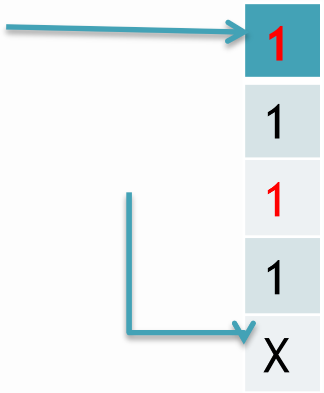


1	1	1	1
0	1	1	1
1	1	1	1
0	0	0	1

- Decoder error
  - – Fault: Row-0 and row-2 fire together
  - – Not sufficient to just read OR write in a single operation

Decoder Address Fault

- MARCH Pattern
  - ↑{W0}
  - ↑{R0 W1}
  - ↑{R1}



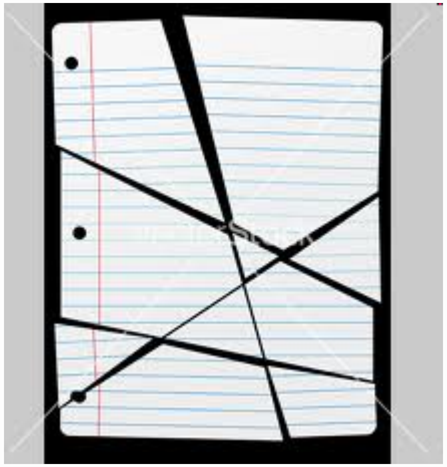
1
1
1
1
X

- Decoder error
  – MARCH Patterns

Redundancy

Notebook

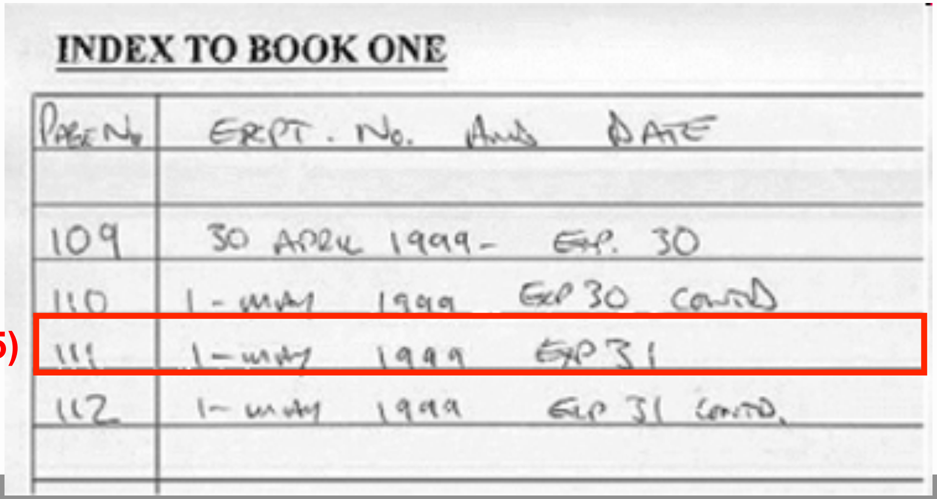
Page 111



Extra Page
R05

eFuse based repair table

(see page R05)



INDEX TO BOOK ONE

PAGE No	EXPT. No. AND DATE
109	30 APRIL 1999 - EXP. 30
110	1 - MAY 1999 EXP 30 CONTD
111	1 - MAY 1999 EXP 31
112	1 - MAY 1999 EXP 31 CONTD.

Fault Detection and Correction

- Run various patterns and detect failed rows and columns
- Mark the addresses as failed ones
- Map them to redundant rows/ columns
- Compare incoming address with failed rows and columns
  - If match found – Change address to redundant row/ column
  - Else let the address through
- Errors on the field are fixed using error correcting codes

Conclusion

- Pulling more DRAM cache (L2,L3) inside the processor improves overall performance
- eDRAM design using logic process is a challenge
- Case study is done, covering many of the eDRAM design aspects
- Sense amp has to read a 1 by default to provide performance improvement
  - Achieved in the Gated Feedback Sense Amp

Learning Objectives for EDRAM

❑ Explain the working of a (e)DRAM. What does Embedded mean?

❑ Explain the working of a feedback sense amplifier and modify existing designs to improve performance

❑ Calculate the voltage levels of operation of various components for an eDRAM

❑ Introduce stacked protect devices to reduce voltage stress of the WL driver

References

Matick, R. et al., “Logic-based eDRAM: Origins and Rationale for Use,” IBM J. Research Dev., vol. 49, no. 1, pp. 145-165, Jan. 2005.

Barth, J. et al., “A 500MHz Random Cycle 1.5ns-Latency, SOI Embedded DRAM Macro Featuring a 3T Micro Sense Amplifier,” ISSCC Dig. Tech. Papers, pp. 486-487, Feb. 2007.

Barth, J. et al., “A 500 MHz Random Cycle, 1.5 ns Latency, SOI Embedded DRAM Macro Featuring a Three-Transistor Micro Sense Amplifier,” IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 43, NO. 1, JANUARY 2008.

Barth, J. et al., “A 45nm SOI Embedded DRAM Macro for POWER7TM 32MB On-Chip L3 Cache,” ISSCC Dig. Tech. Papers, pp. 342-3, Feb. 2010.

Barth, J. et al., “A 45 nm SOI Embedded DRAM Macro for the POWER™ Processor 32 MByte On-Chip L3 Cache,” IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 46, NO. 1, JANUARY 2011.

S. Iyer et al., “Embedded DRAM: Technology Platform for BlueGene/L Chip,” IBM J. Res. & Dev., Vol. 49, No. 2/3, MARCH/ MAY 2005, pp.333-50.

Barth, J. et al., “A 300MHz Multi-Banked eDRAM Macro Featuring GND Sense, Bit-line Twisting and Direct Reference Cell Write,” ISSCC Dig. Tech. Papers, pp. 156-157, Feb. 2002.

Barth, J. et. al., “A 500-MHz Multi-Banked Compilable DRAM Macro With Direct Write and Programmable Pipelining,” IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 40, NO. 1, JANUARY 2005.

Butt,N., et al., “A 0.039um2 High Performance eDRAM Cell based on 32nm High-K/Metal SOI Technology,” IEDM pp. 27.5.1-2, Dec 2010.

Bright, A. et al., “Creating the BlueGene/L Supercomputer from Low-Power SoC ASICs,” ISSCC Dig. Tech. Papers, pp. 188-189, Feb. 2005.

Blagojevic, M. et al., “SOI Capacitor-Less 1-Transistor DRAM Sensing Scheme with Automatic Reference Generation,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 182-183, Jun. 2004.

References

Karp, J. et al., “A 4096-bit Dynamic MOS RAM” ISSCC Dig. Tech. Papers, pp. 10-11, Feb. 1972.

Kirihata, T. et al., “An 800-MHz Embedded DRAM with a Concurrent Refresh Mode,” IEEE Journal of Solid State Circuits, pp. 1377-1387, Vol. 40, Jun. 2003.

Luk, W. et al., “2T1D Memory Cell with Voltage Gain,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 184-187, Jun. 2004.

Luk, W. et al., “A 3-Transistor DRAM Cell with Gated Diode for Enhanced Speed and Retention Time,” Symposium on VLSI Circuits Dig. Tech. Papers, pp. 228-229, Jun. 2006.

NEC eDRAM Cell Structure (MIM Capacitor): http://www.necel.com/process/en/edramstructure.html

Ohsawa, T. et al., “Memory Design using One-Transistor Gain Cell on SOI,” ISSCC Dig. Tech. Papers, pp. 152-153, Feb. 2002.

Pilo, H. et al., “A 5.6ns Random Cycle 144Mb DRAM with 1.4Gb/s/pin and DDR3-SRAM Interface,” ISSCC Dig. Tech. Papers, pp. 308-309, Feb. 2003.

Taito, Y. et al., “A High Density Memory for SoC with a 143MHz SRAM Interface Using Sense-Synchronized-Read/Write,” ISSCC Dig. Tech. Papers, pp. 306-307, Feb. 2003.

Wang, G. et al., A 0.127 mm2 High Performance 65nm SOI Based embedded DRAM for on-Processor Applications,” International Electron Devices Meeting, Dec. 2006.

G. Fredeman et al., “A 14 nm 1.1 Mb Embedded DRAM Macro With 1 ns Access,” in IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 230-239, Jan. 2016. doi: 10.1109/JSSC.2015.2456873