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

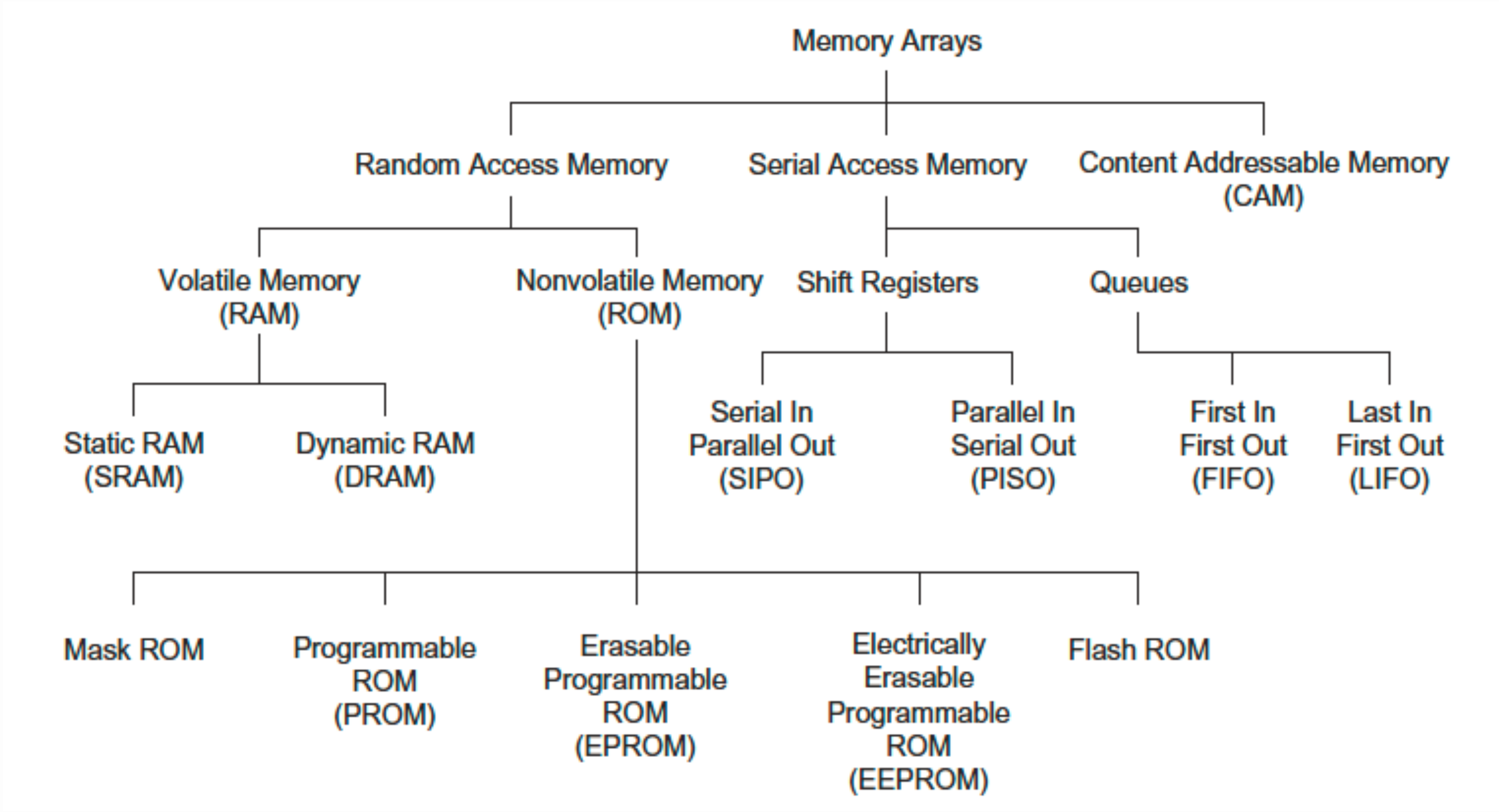
Acknowledgement

- 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

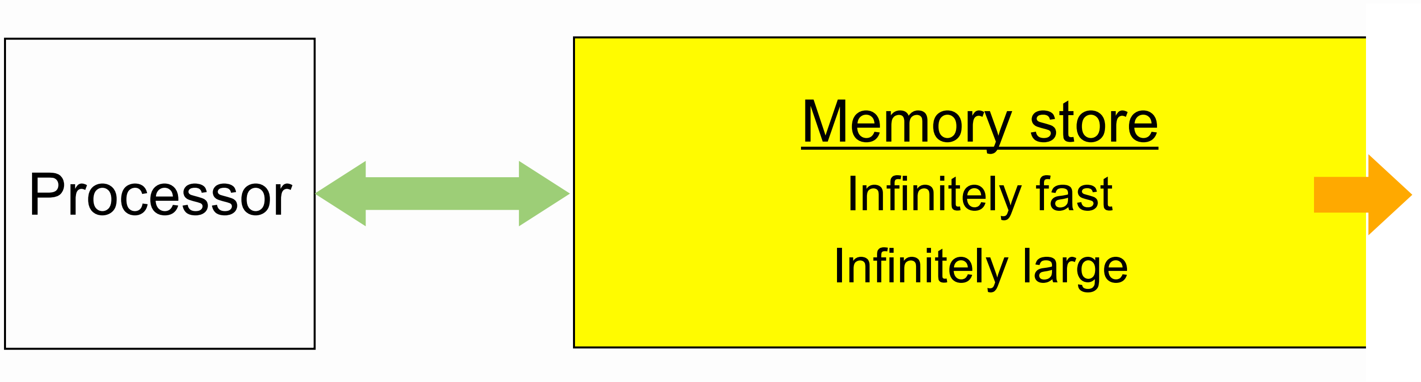
Topics

- □ 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



Motivation for a memory hierarchy – infinite memory



Processor

Memory store

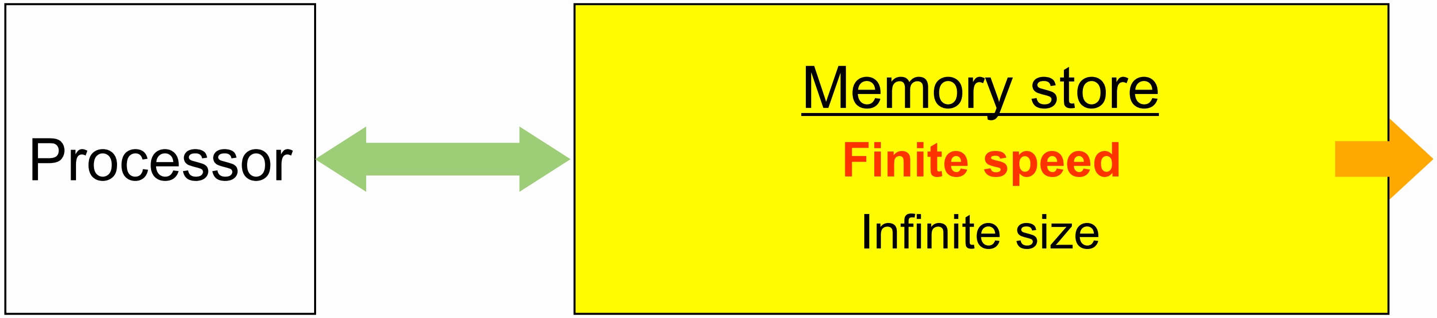
Infinitely fast

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

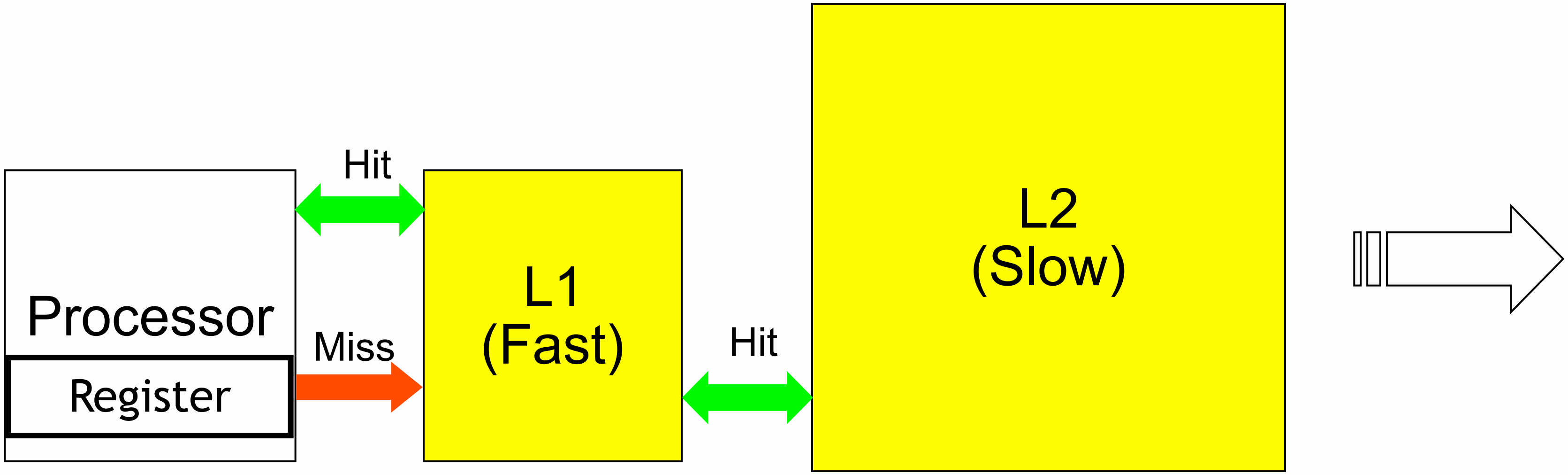
e.g: Loops

Spatial

If you accessed something you'll also need its neighbor

e.g: Arrays

Exploit this to divide memory into hierarchy

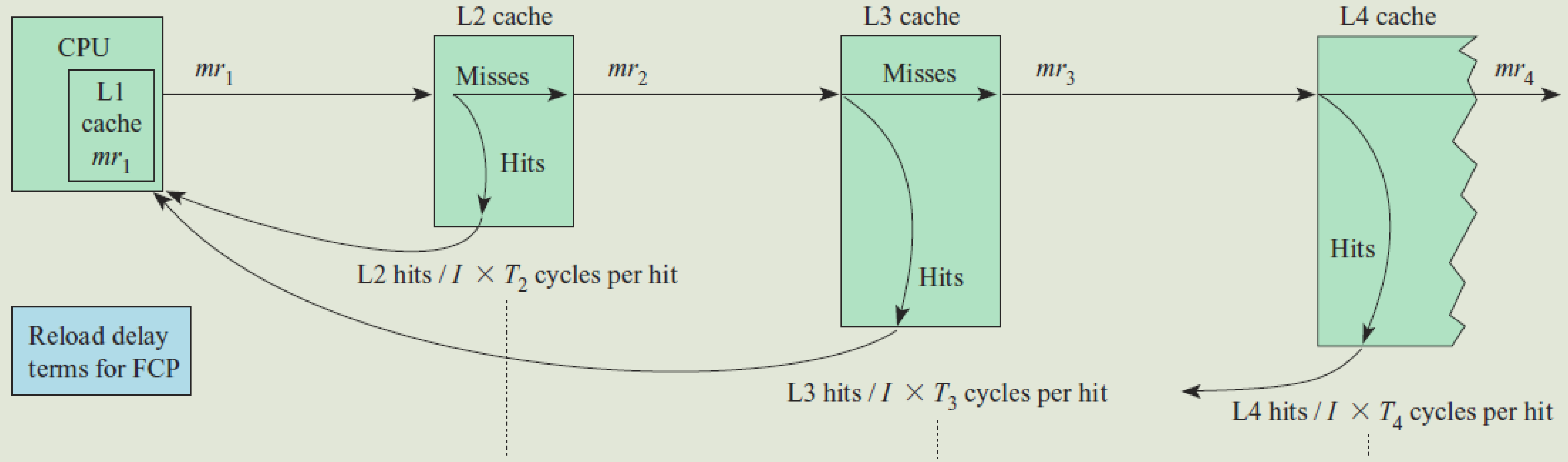


Cache size impacts cycles-per-instruction

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 $mr_3 - mr_4$

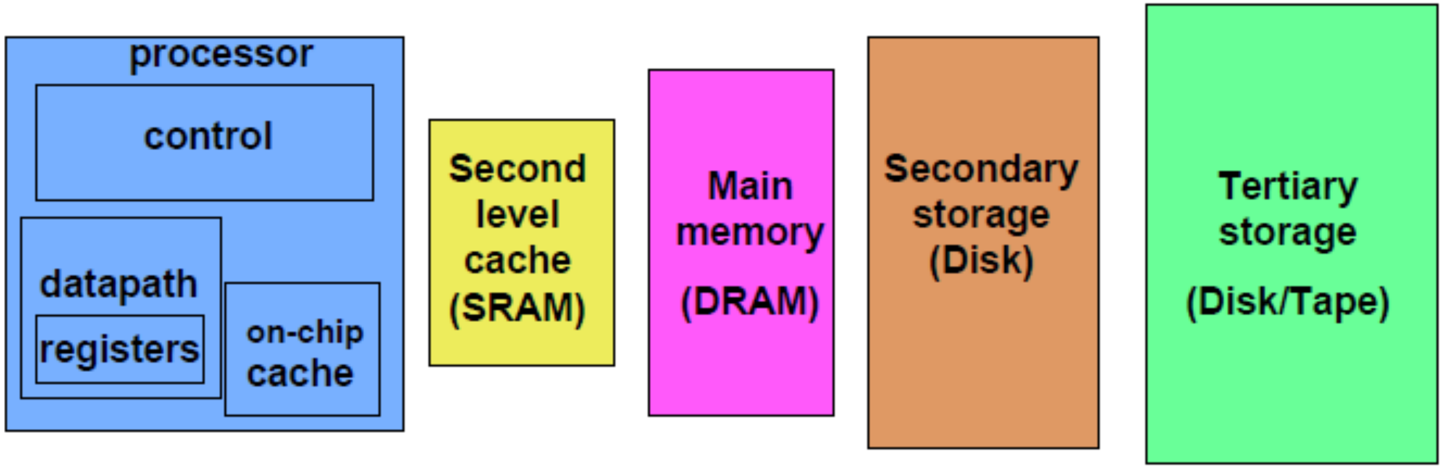
Finite cache penalty = Delay / I for L2 hits + Delay / I for L3 hits + Delay / I for L4 hits ...

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

Cycles per instruction = Hits per instruction × cycles per hit

Access rate reduces → Slower memory is sufficient

Cache size impacts cycles-per-instruction



processor

control

datapath

registers

on-chip cache

Second level cache (SRAM)

Main memory (DRAM)

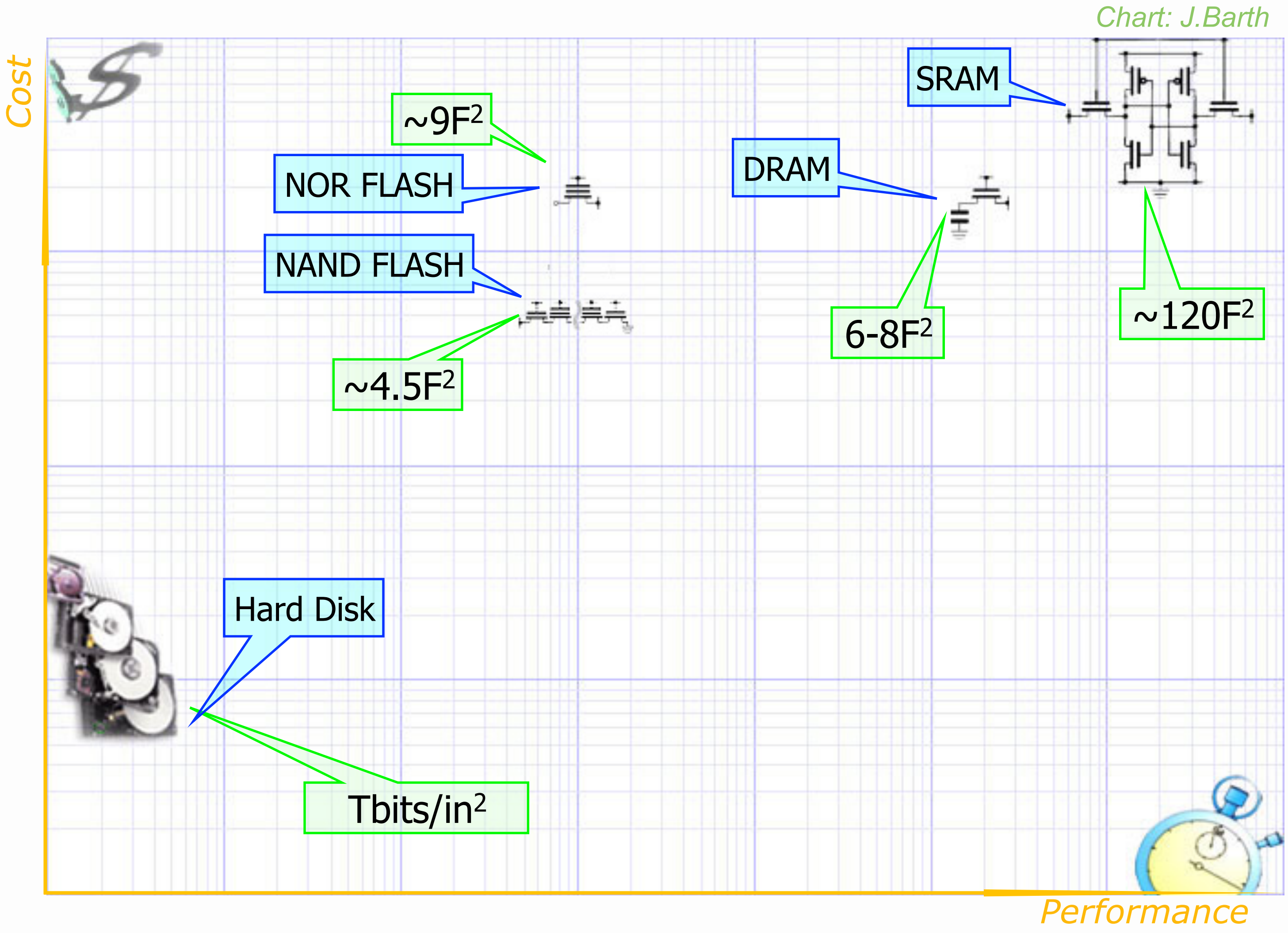
Secondary storage (Disk)

Tertiary storage (Disk/Tape)

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

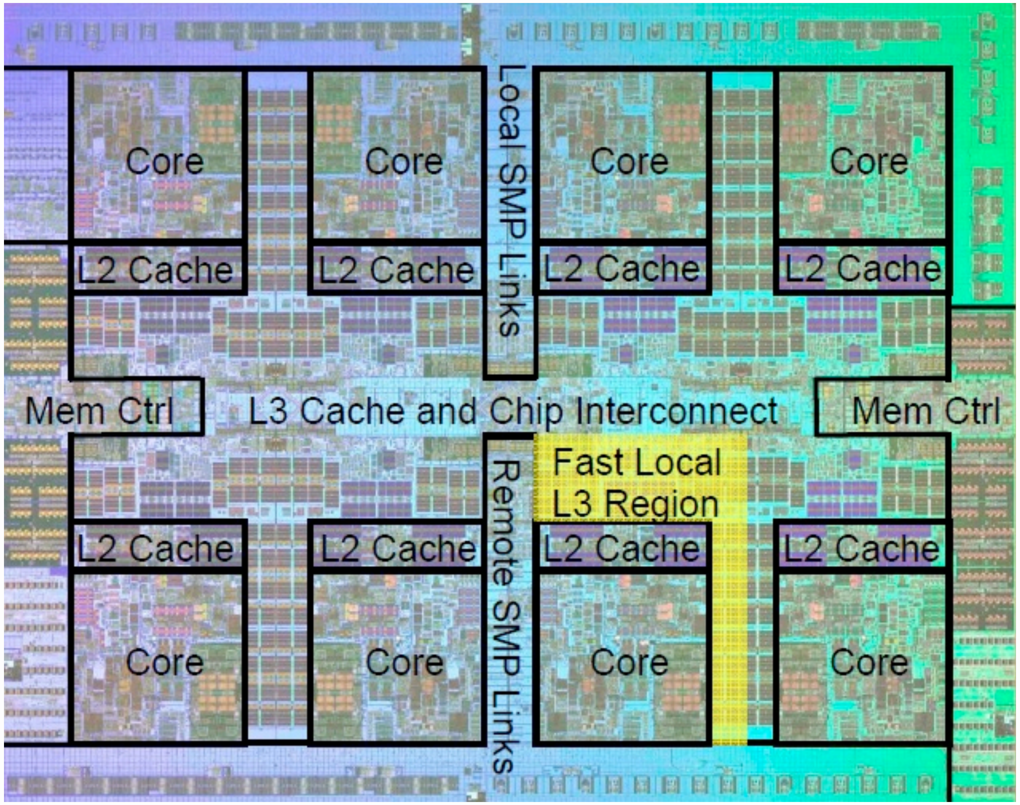
For a 5GHz processor, scale the numbers by 5x

Technology choices for memory hierarchy



eDRAM L3 cache

Power7 processor



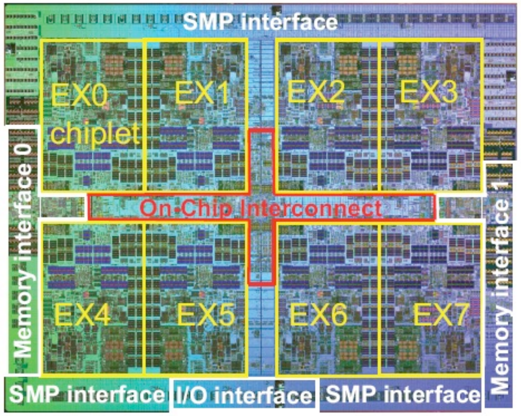
Move L2,L3 Cache inside of the data hungry processor

Higher hit rate → Reduced FCP

JSSCC11

Embedded DRAM Advantages

IBM Power7tm
32MB eDRAM L3

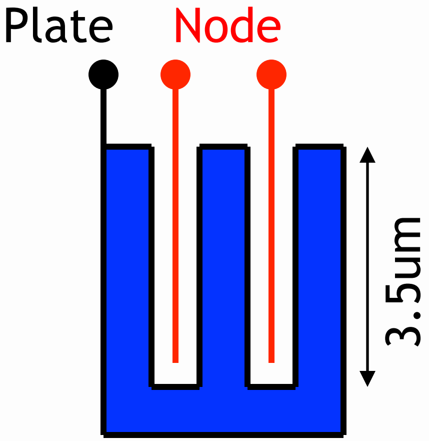


Memory Advantage

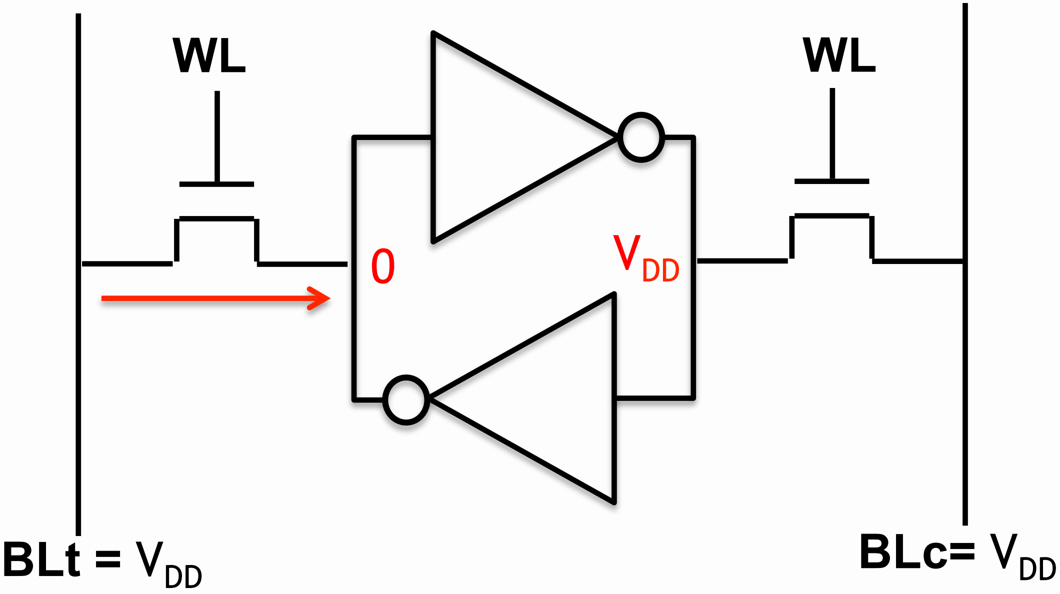
- 2x Cache can provide > 10% Performance
- ~3x Density Advantage over eSRAM
- 1/5x Standby Power Compared to SRAM
- Soft Error Rate 1000x lower than SRAM
- Performance ? DRAM can have lower latency !
- IO Power reduction

Deep Trench Capacitor

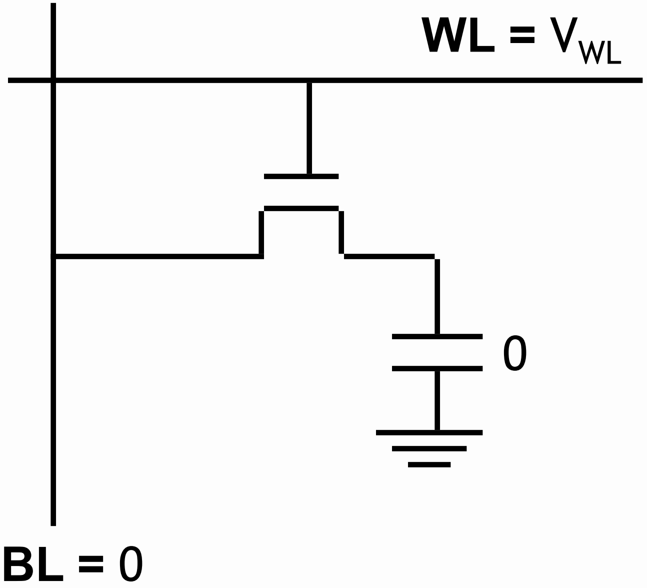
- Low Leakage Decoupling
- 25x more Cap / µm^2 compared to planar
- Noise Reduction = Performance Improvement
- Isolated Plate enables High Density Charge Pump



eDRAM Advantages – Stand By Leakage

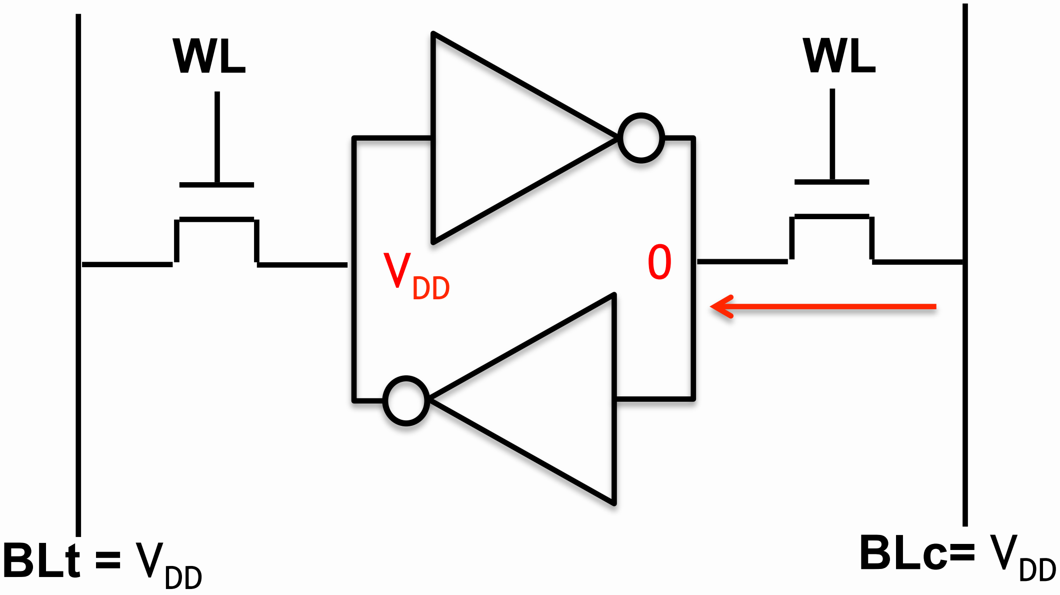


Both inverters leak
True Pass transistor leaks

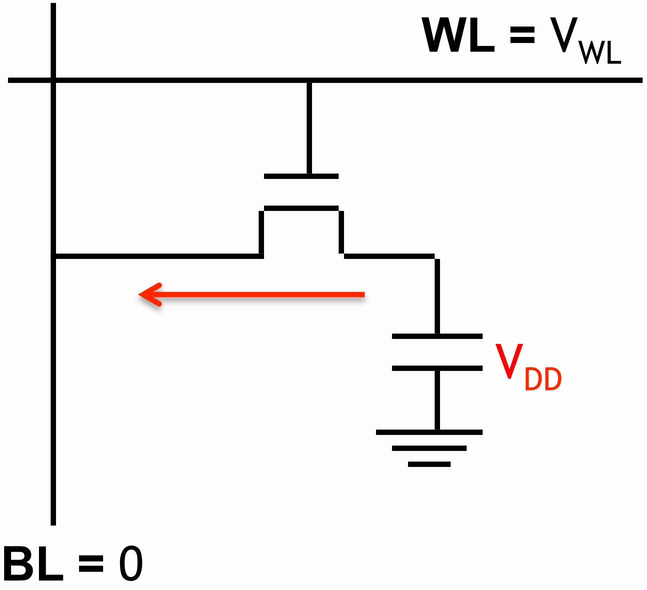


NO Leakage

eDRAM Advantages – Stand By Leakage

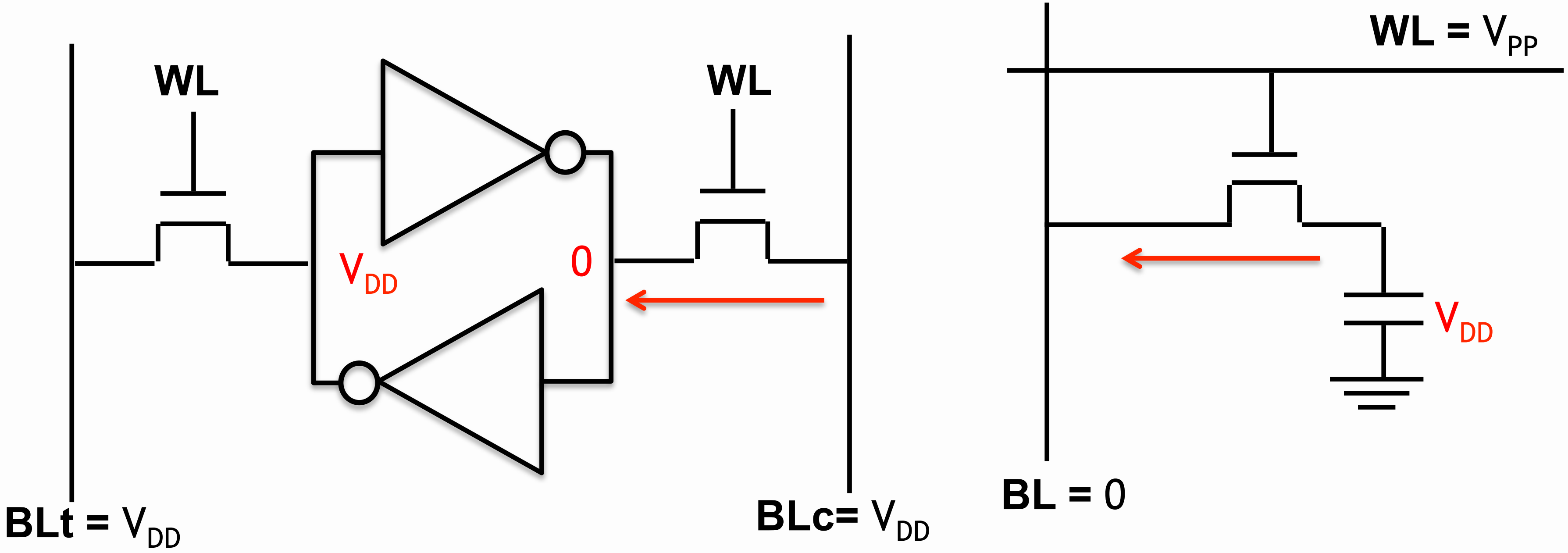


Both inverters leak
Complement Pass transistor leaks



Leakage exists

eDRAM Advantages – Stand By Leakage

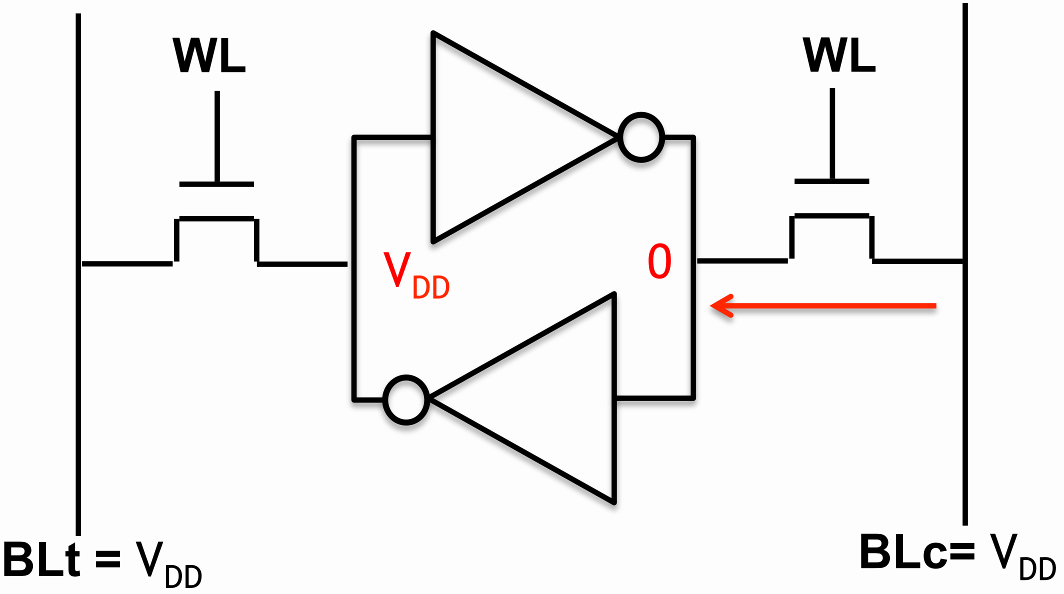


Both inverters leak
Complement Pass transistor leaks

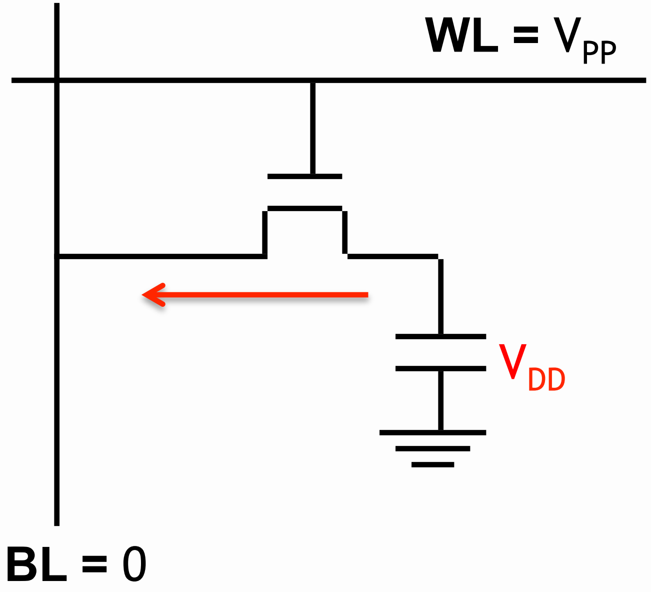
Leakage exists

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

eDRAM Advantages – Performance

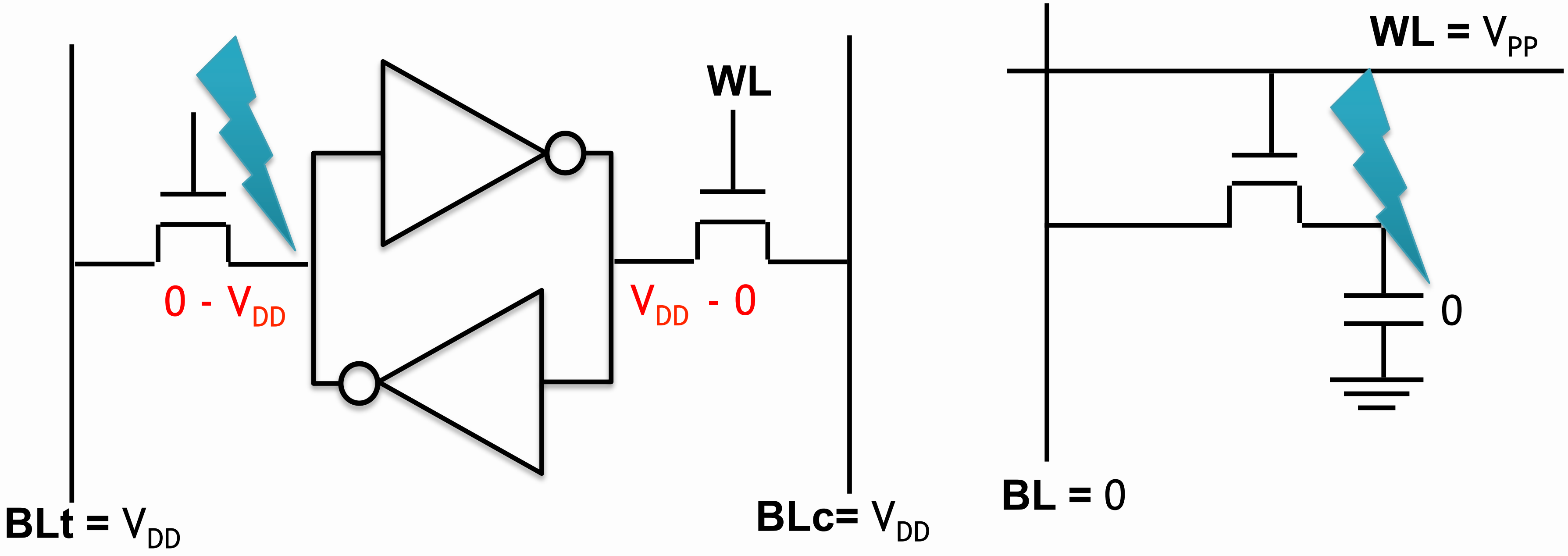


Both inverters leak
Complement Pass transistor leaks



Leakage exists

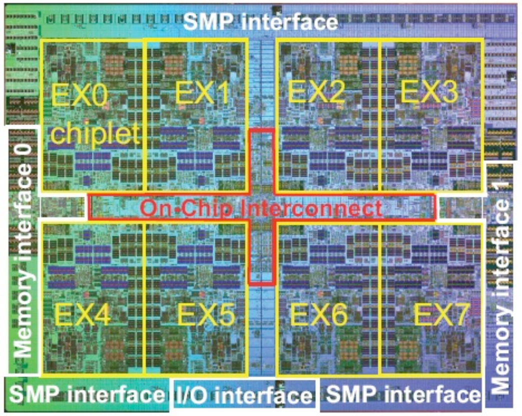
eDRAM Advantages – Soft Error Rate



- 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

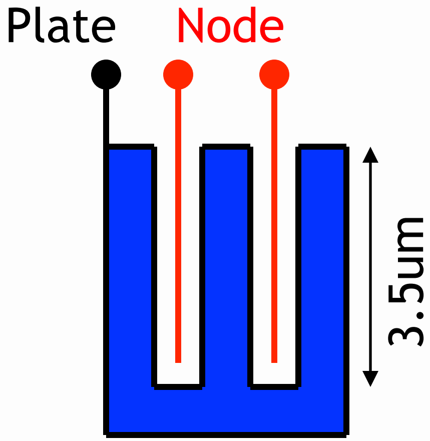
Embedded DRAM Advantages

IBM Power7tm
32MB eDRAM L3



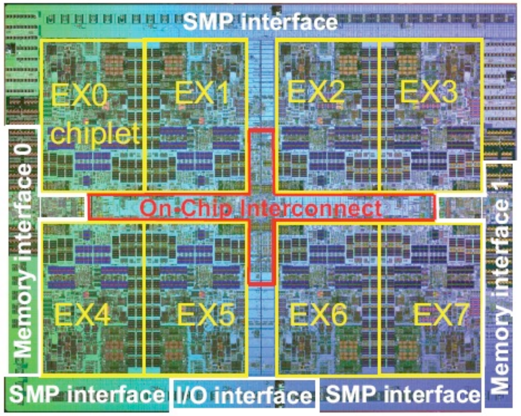
Deep Trench Capacitor

- 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 Power7^{tm}
32MB eDRAM L3

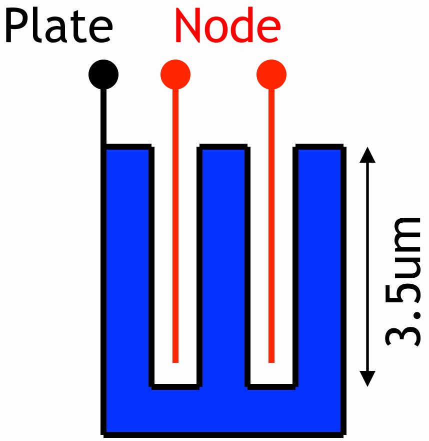


Memory Advantage

- 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

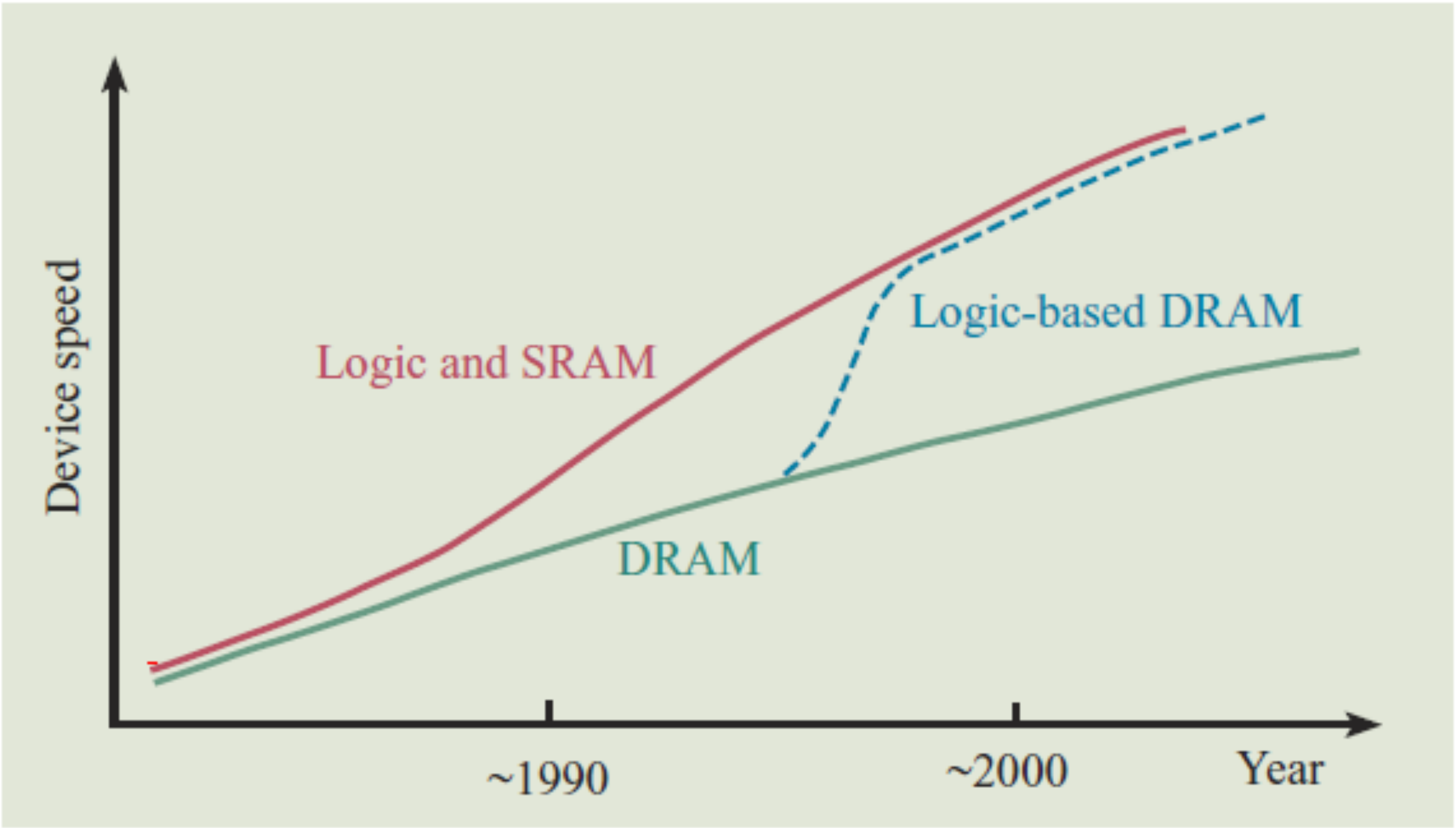
Deep Trench Capacitor

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

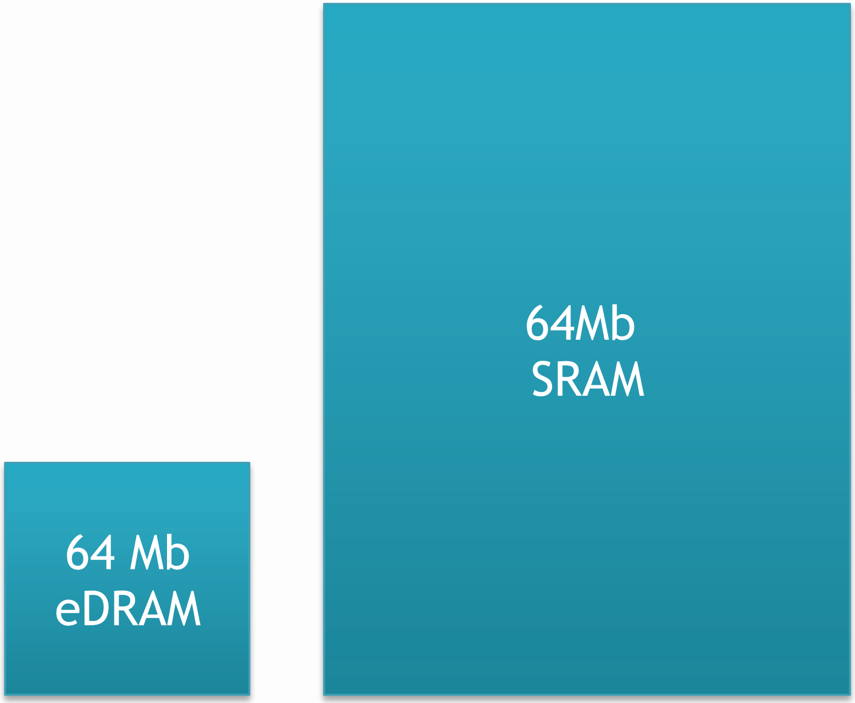


Cache performance – SRAM vs. DRAM

Chart: Matick & Schuster, op. cit.

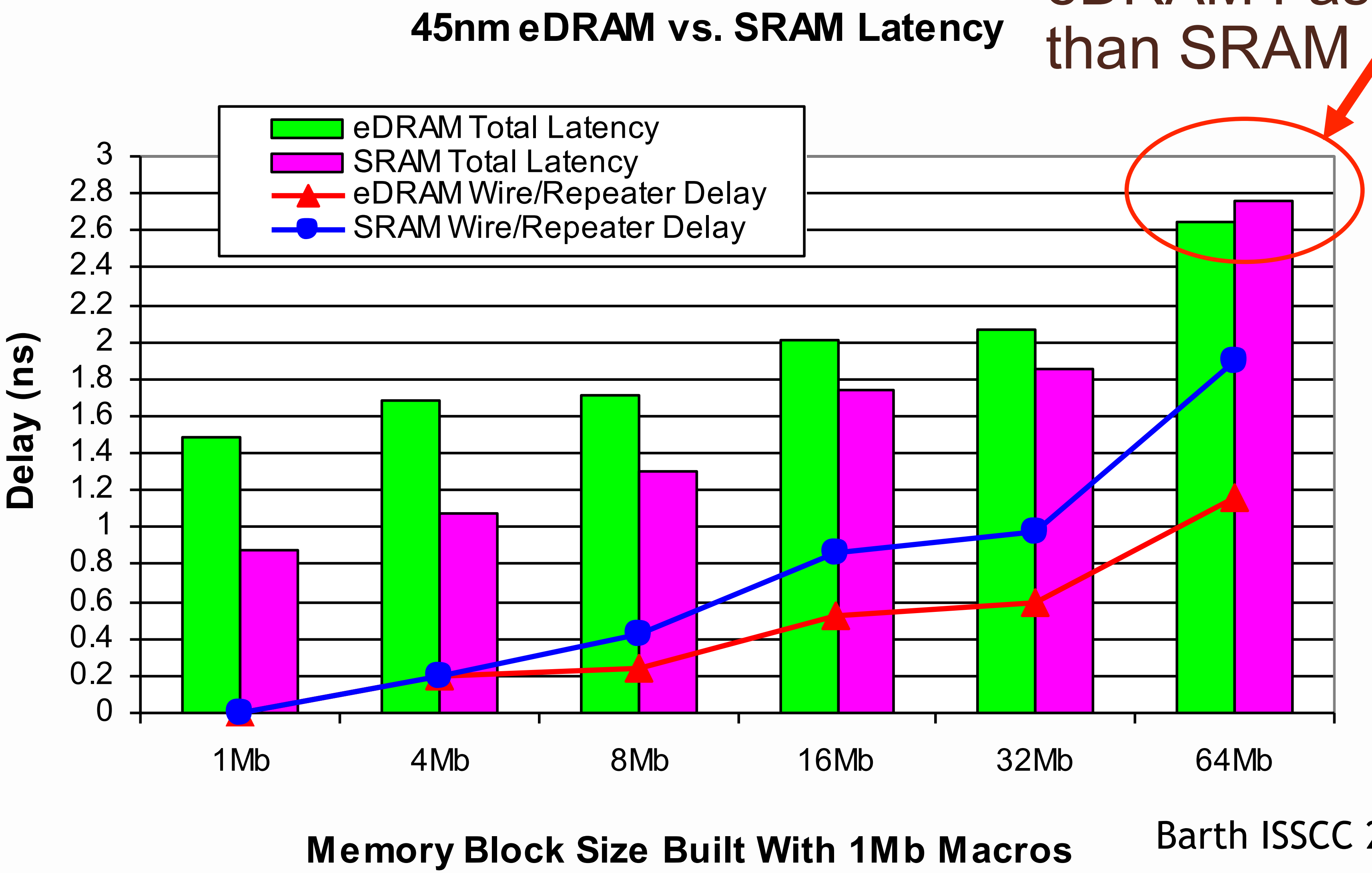


Cache performance – SRAM vs. DRAM



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 operational details (case study)
- ❑ Noise concerns
- ❑ Wordline driver (WLDRV) and level translators (LT)
- ❑ Challenges in eDRAM
- ❑ Understanding Timing diagram – An example

Fundamental DRAM Operation

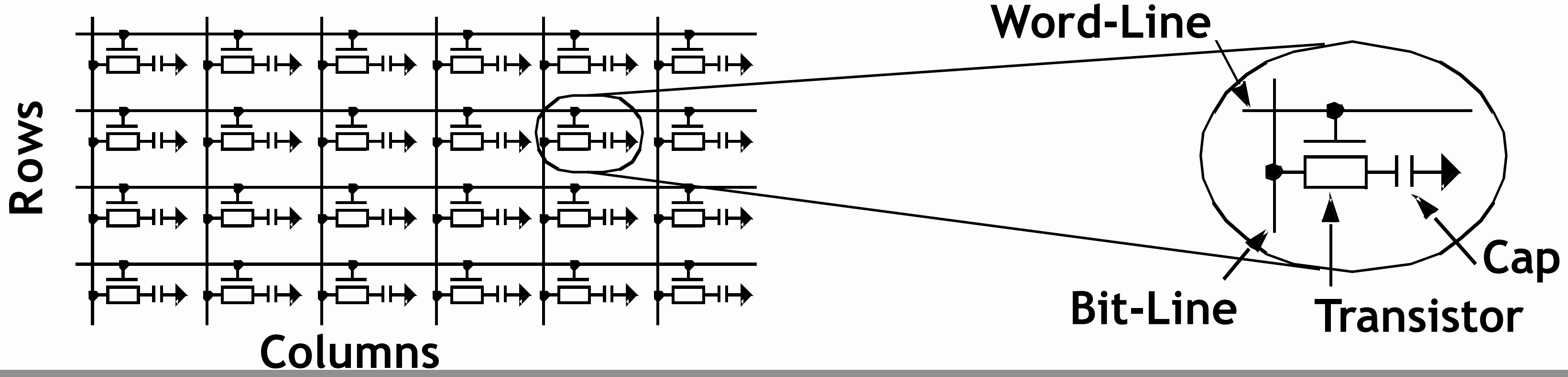
Memory Arrays are composed of Row and Columns

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

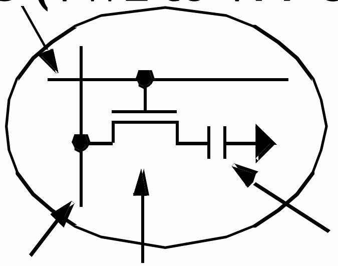
Single Cell Accessed by Decoding One Row / One Column (Matrix)

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

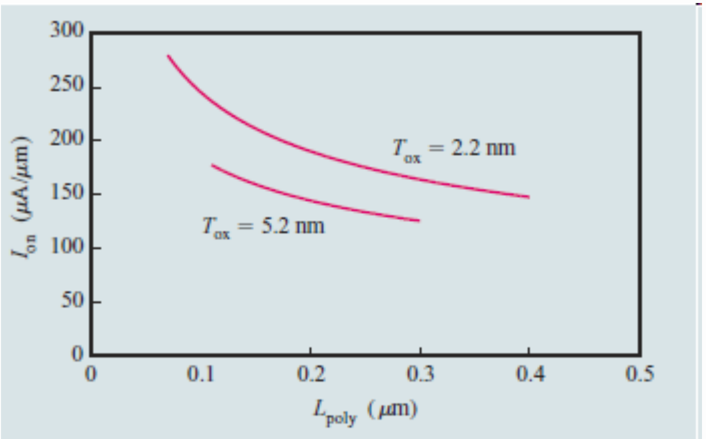


Word-Line (VWL to VPP Swing)

Bit-Line (0 to VDD)

Back Bias (VBB - Bulk Only)

Cap(0 to VDD)



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

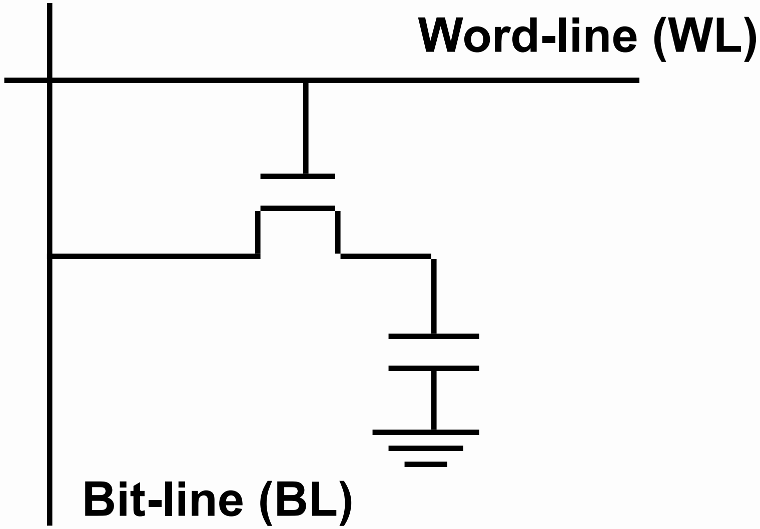
VPP: Word-Line High Supply, 1.8V up to 3.5V depending on Technology
Required to be at least a Vt above VDD to write full VDD

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

IBM J RES & DEV 2005

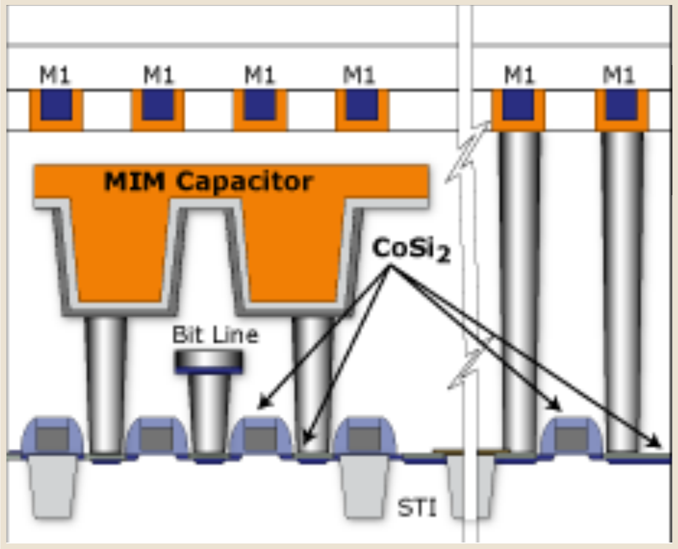
Choice of Access Transistor

- **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

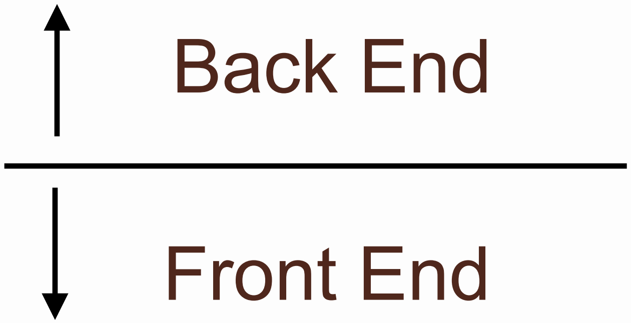


IBM J RES & DEV 2005

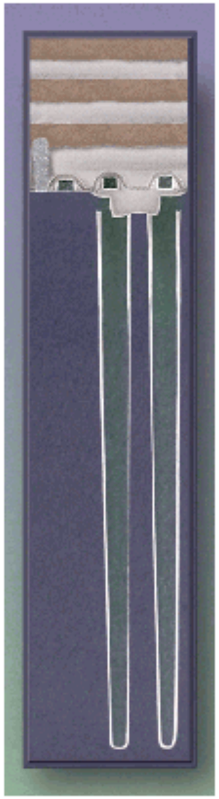
MIM Cap v/s Trench



MIM eDRAM Process



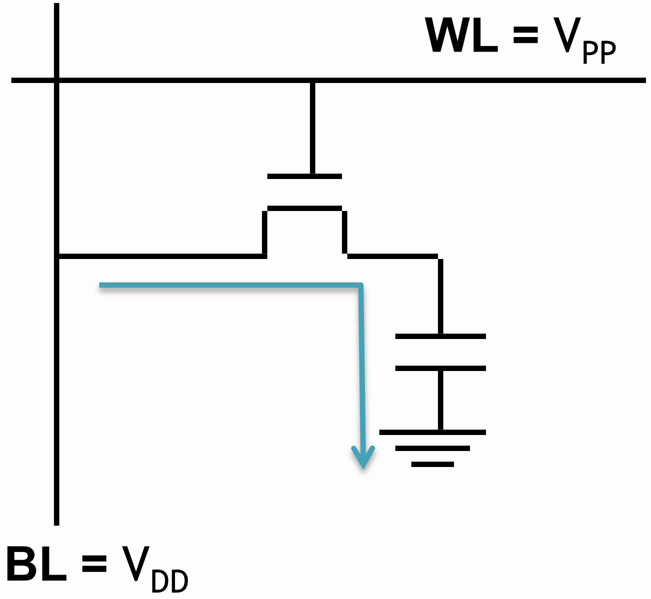
Trench eDRAM Process



- 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

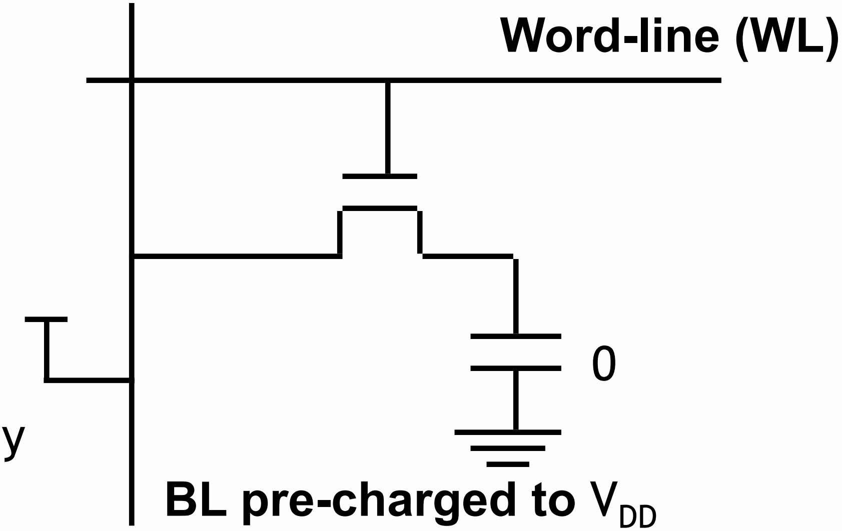
- WL High Voltage – ON State
  - Technology maximum high voltage?
    - Access device is thick oxide
    - Can handle a swing of more than V_{DD}.
  - How high?
    - We wish to write a logic-1 completely
    - Logic-1 = V_{DD}
    - Access device is an NMOS transistor
      - Cannot pass V_{DD} fully 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

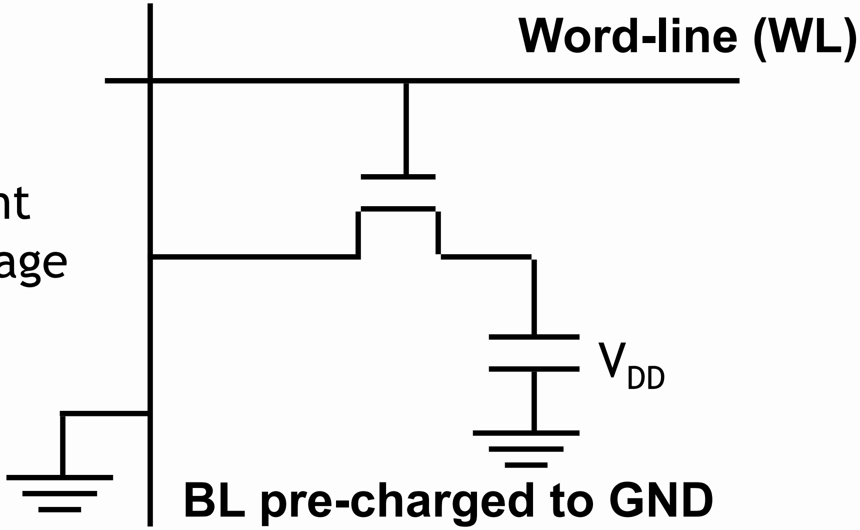


IBM J RES & DEV 2005

Word-line Swing - Low

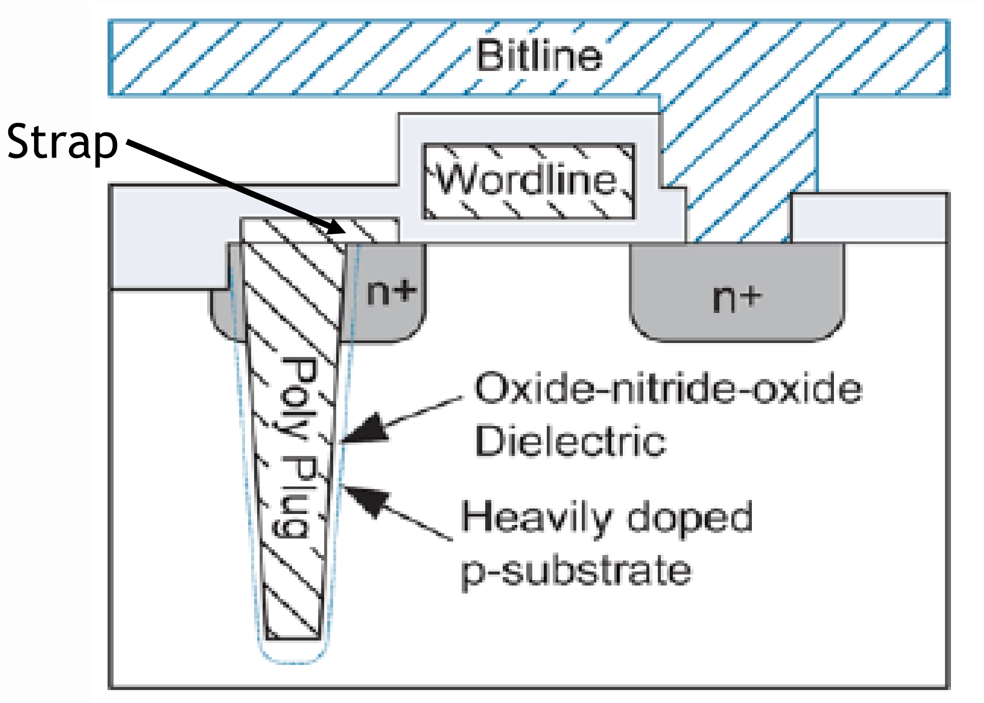
- WL Low Voltage (V_{WL})– OFF State
- BL will be pre-charged to either 0 or V_{DD}
  - BL pre-charged to 0
    - Cell storing a 1 causes leakage
  - BL pre-charged to 1
    - Cell storing a 0 causes leakage
- Need to minimize leakage current either way
- V_{GS} of access device needs to be as low as 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



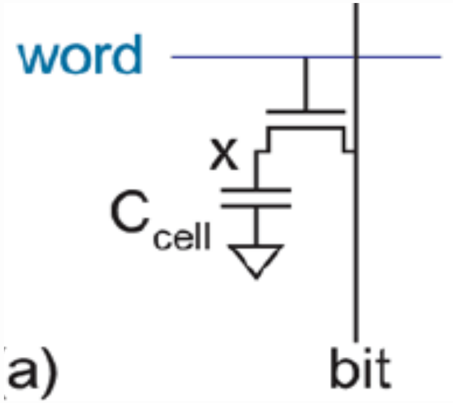


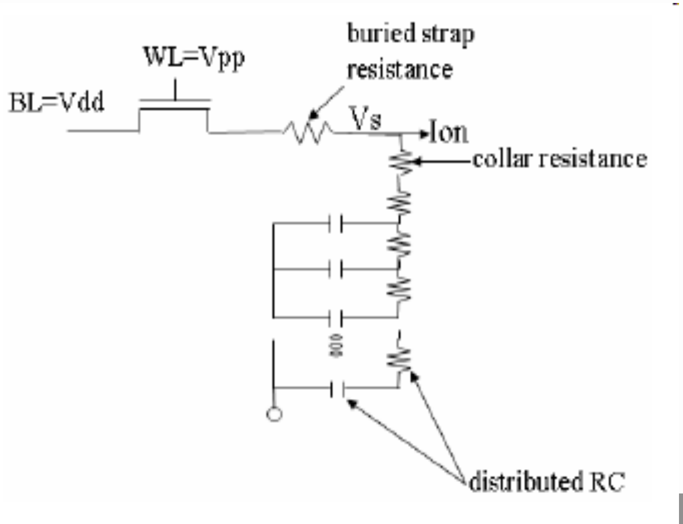
DRAM cell Cross section

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

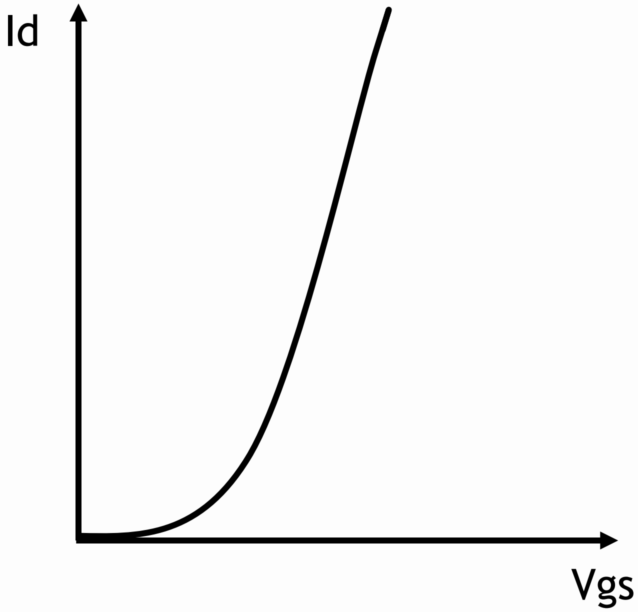


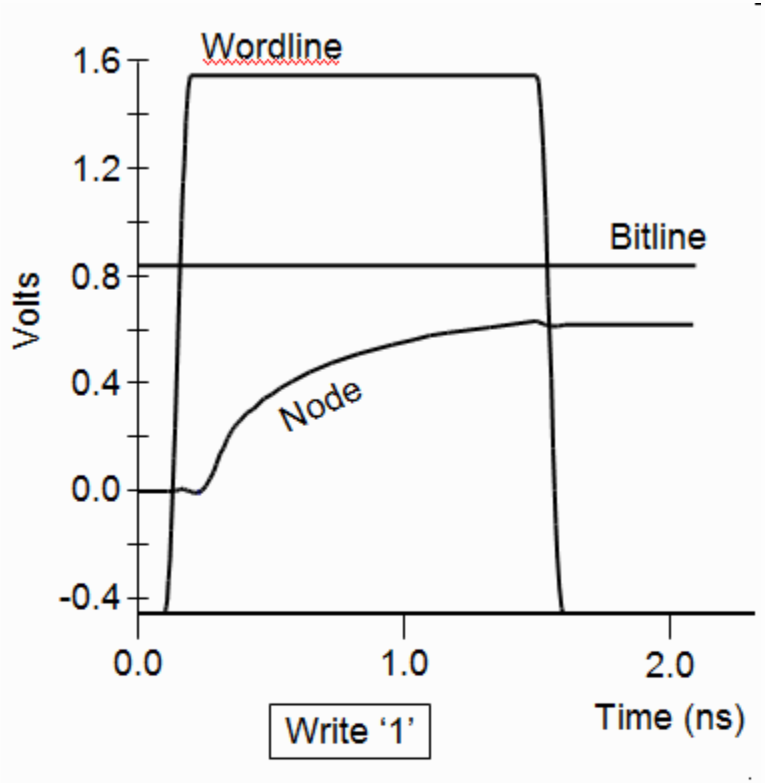
CMOS VLSI design - PEARSON





Storing data ‘1’ in the cell

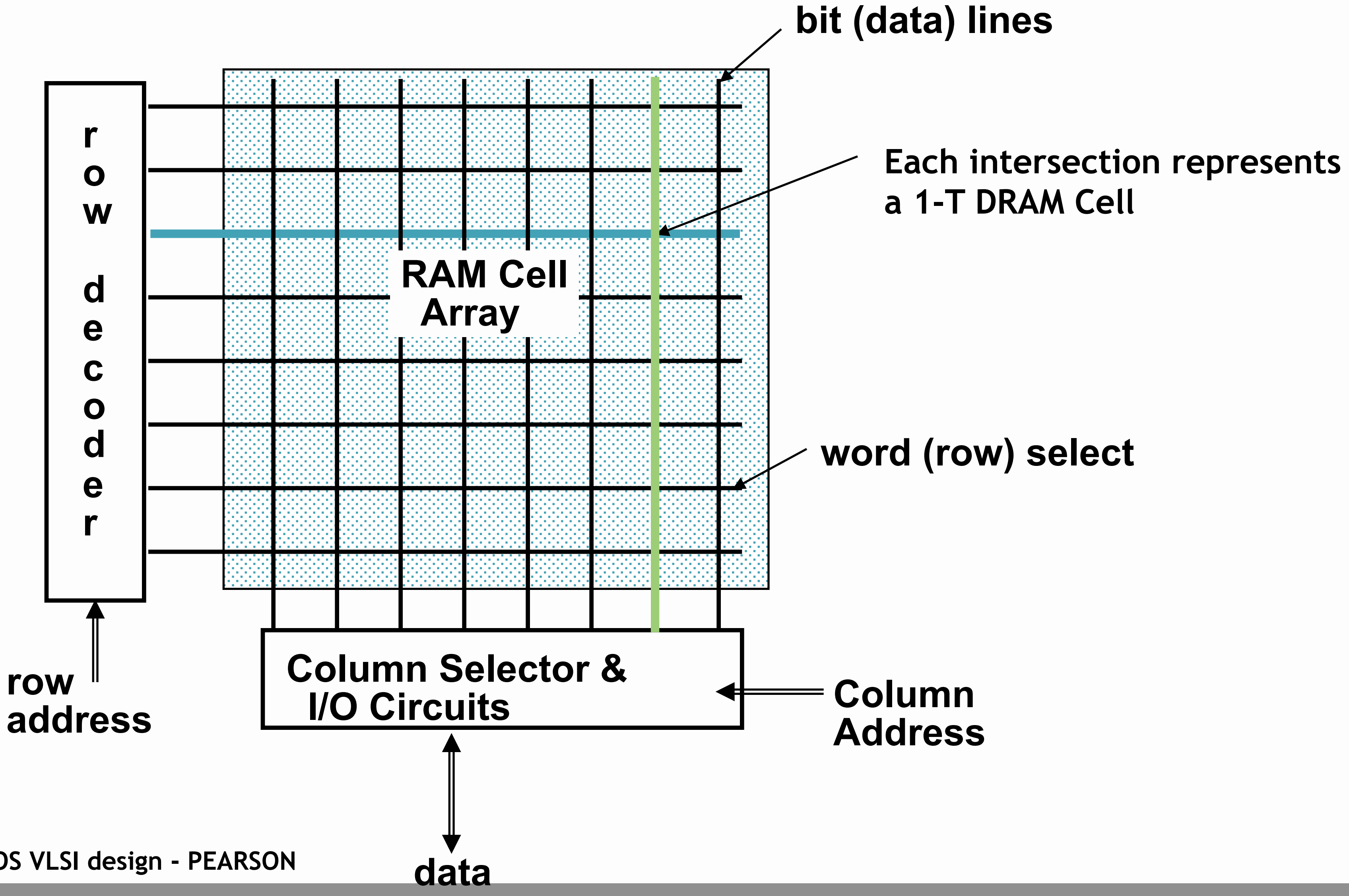




Vgs for pass transistor reduces as bitcell voltage rises, increasing Ron

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

Classical DRAM Organization



CMOS VLSI design - PEARSON

DRAM Subarray

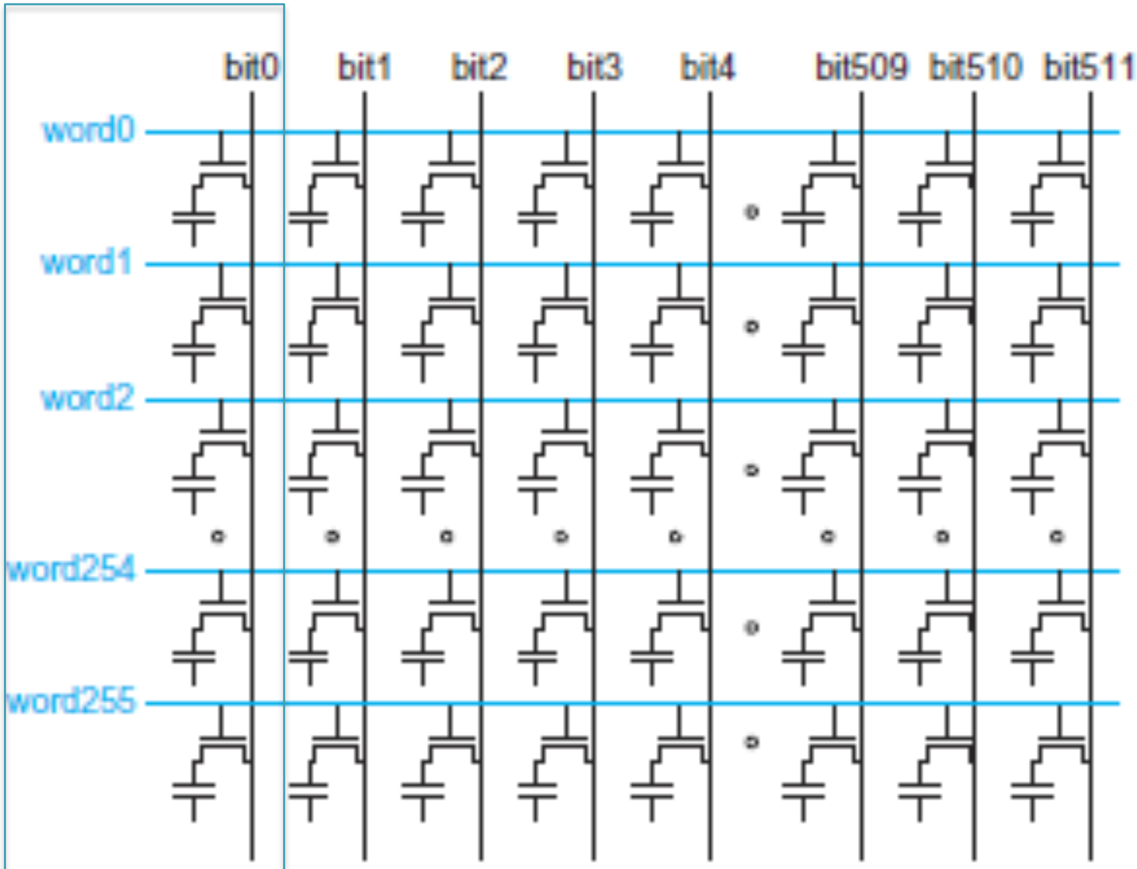
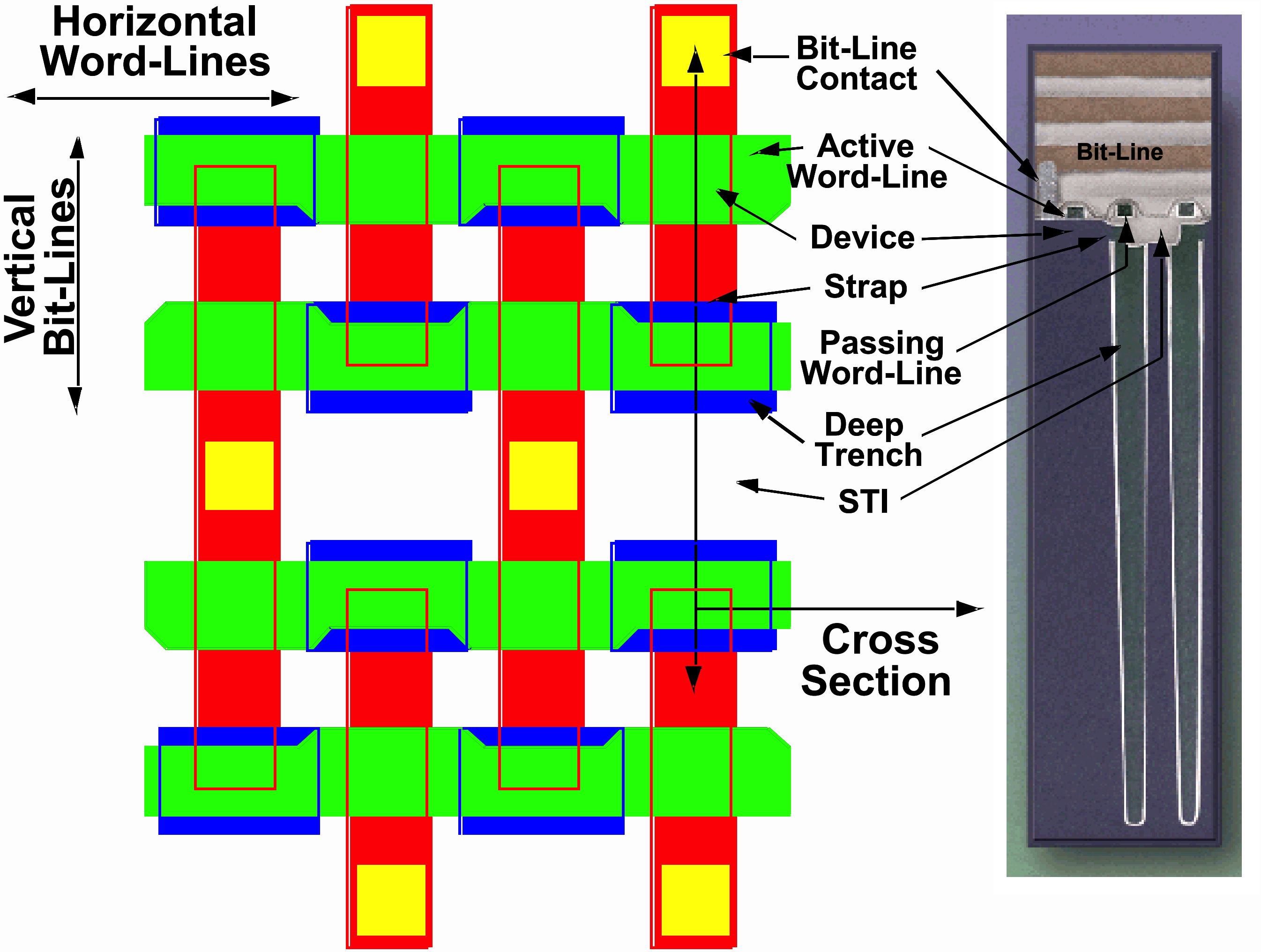
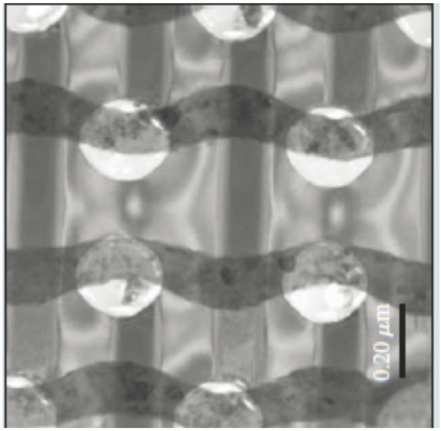


FIGURE 12.43 DRAM subarray

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.

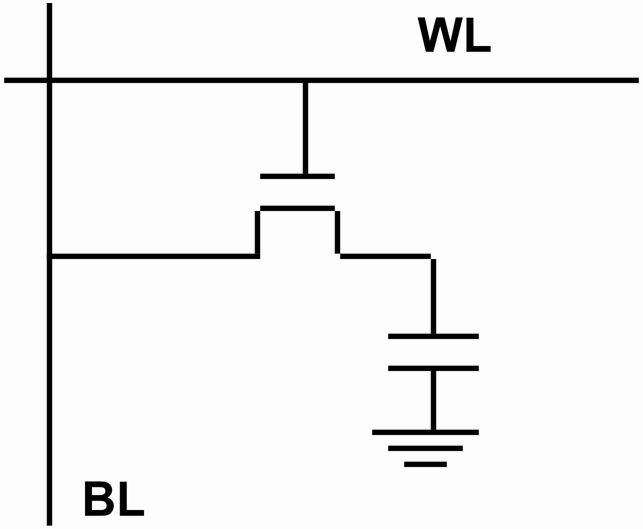
Barth, J. et al., “A 45nm SOI Embedded DRAM Macro for POWER7TM 32MB On-Chip L3 Cache,” ISSCC Dig. Tech. Papers, pp. 342-3, Feb. 2010.

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.

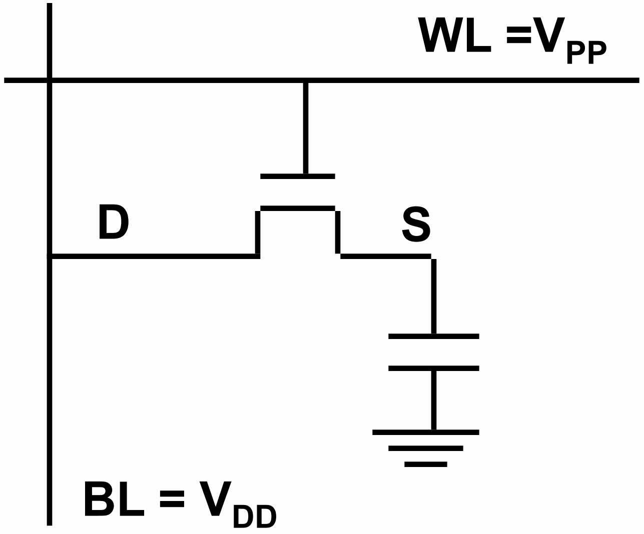
DRAM Operations

- Write
- Read
- Refresh



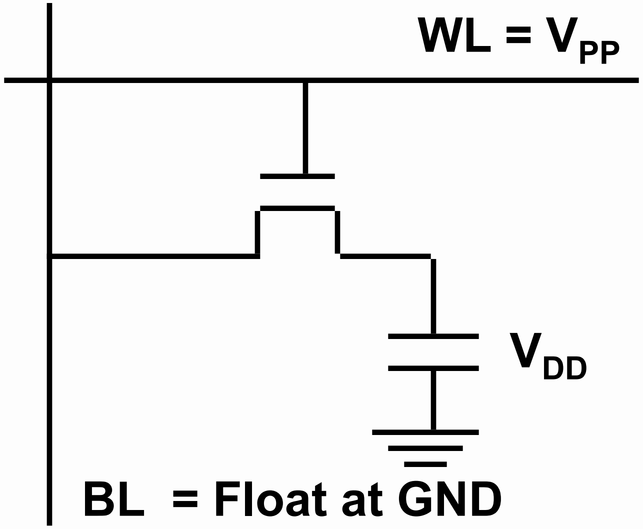
DRAM Read, Write and Refresh

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

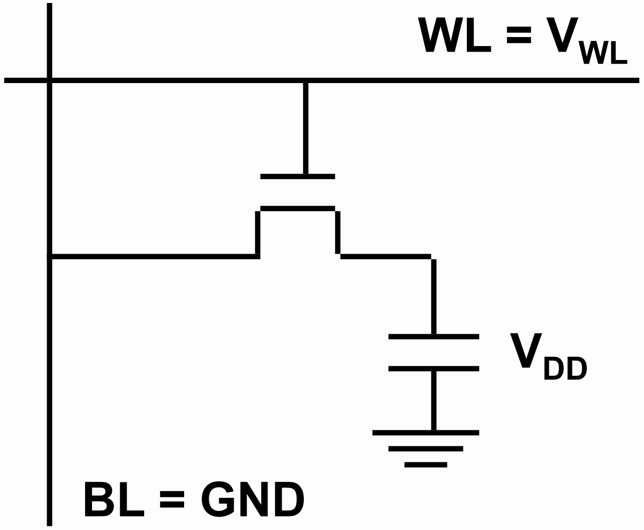


DRAM Read, Write and Refresh

- 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**

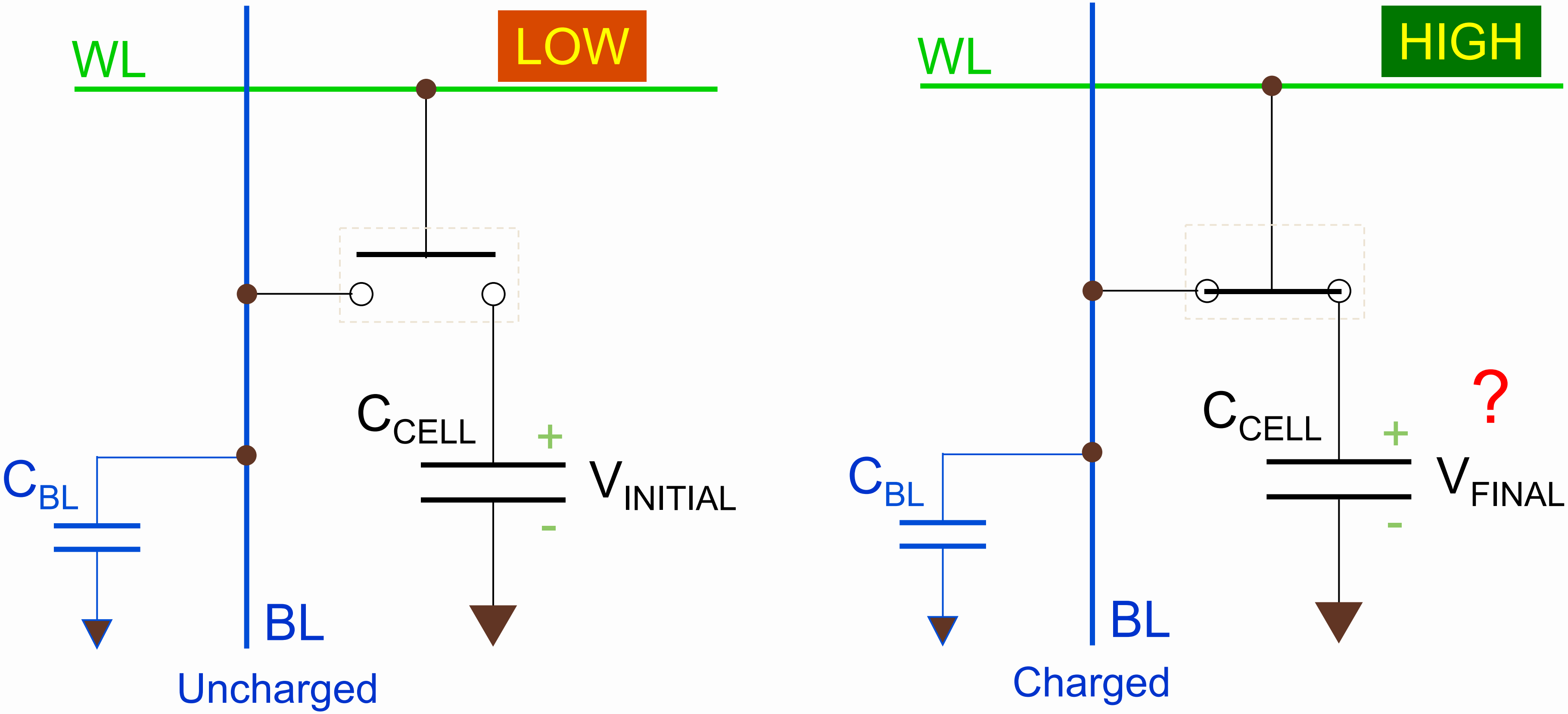


DRAM Read, Write and Refresh



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

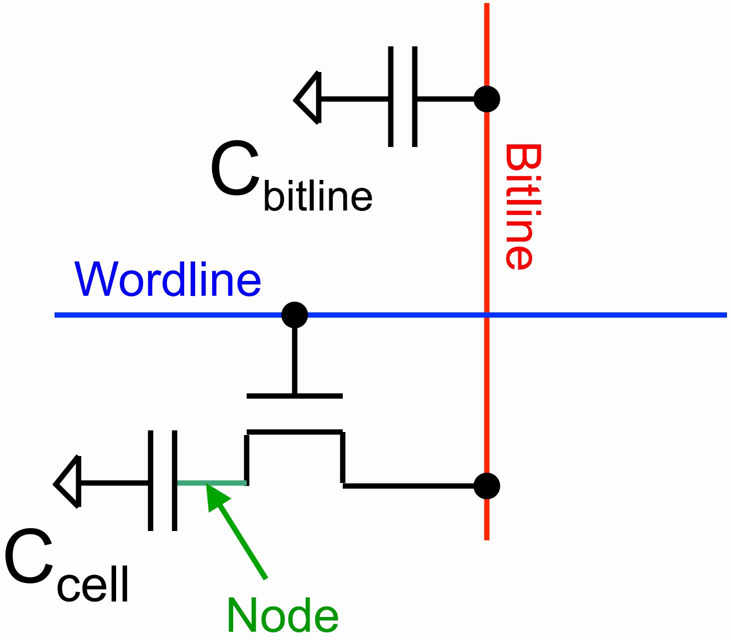
Read - Cell transfer ratio

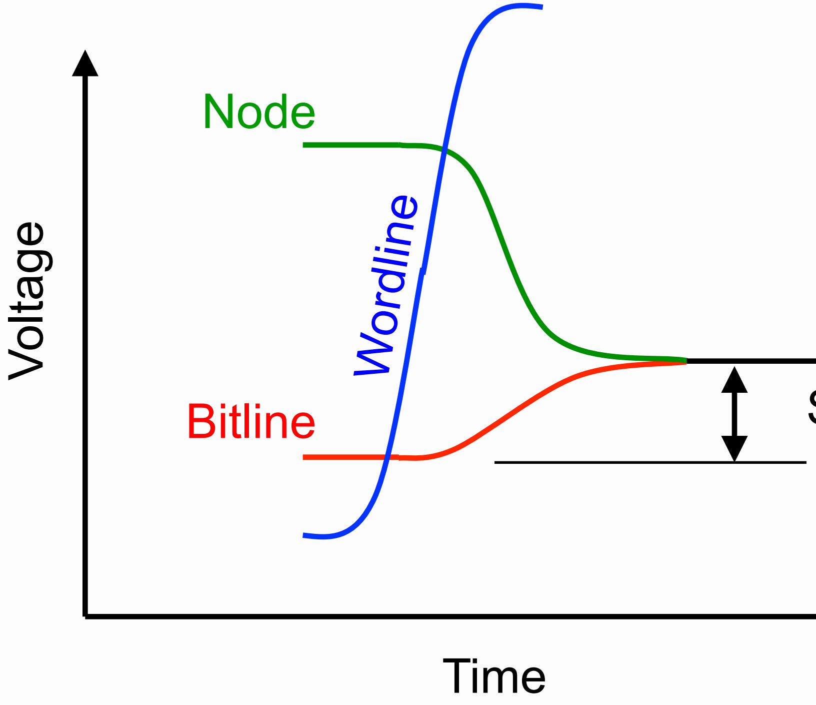


$$C_{CELL} \times V_{INITIAL} = (C_{CELL} + C_{BL}) \times V_{FINAL}$$

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

Cell Charge Transfer



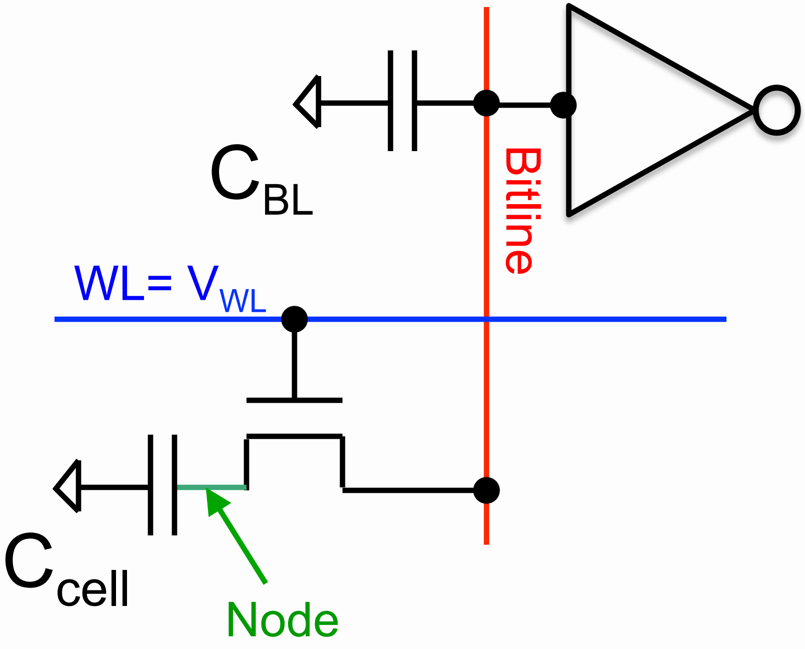


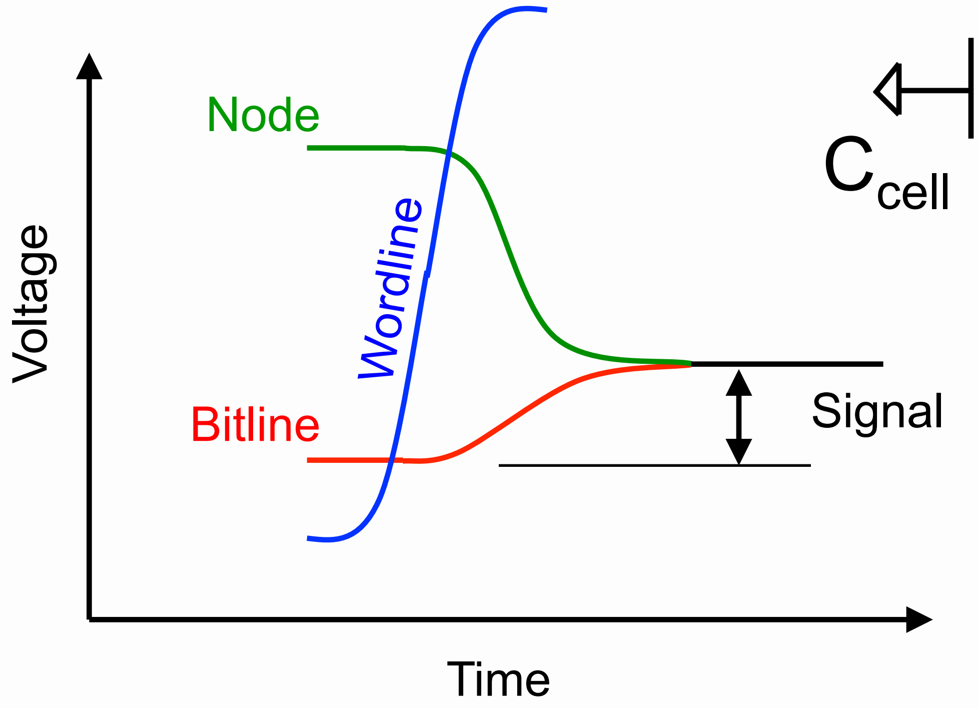
$$\Delta V = (V_{bl} - V_{cell}) \left[\frac{C_{cell}}{C_{bl}+C_{cell}} \right]$$

Transfer ratio

Sensing

Signal ΔV > V_M (Trip Point)



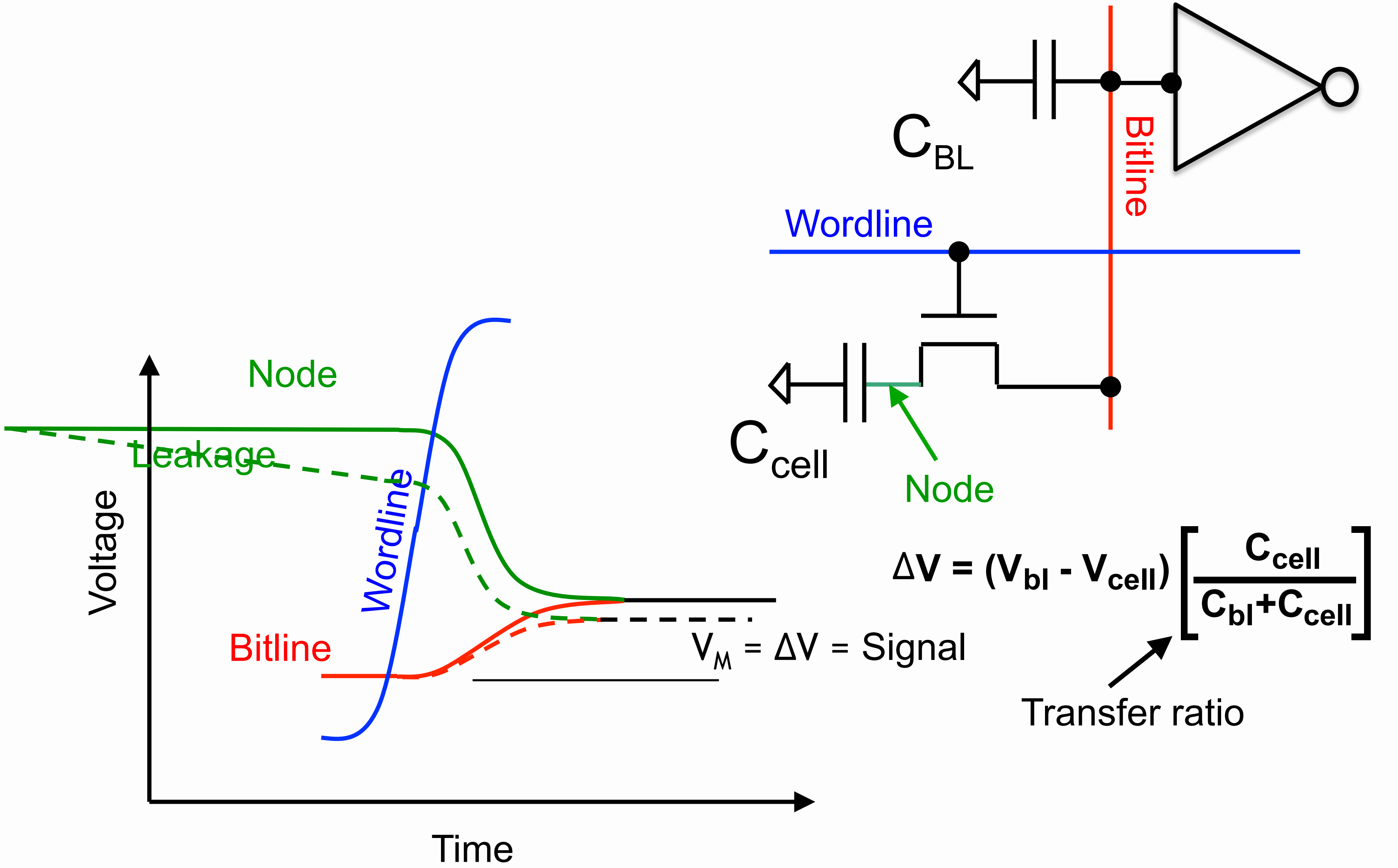


$$\Delta V = (V_{bl} - V_{cell}) \left[\frac{C_{cell}}{C_{bl}+C_{cell}}\right]$$

Transfer ratio

Retention

Signal ΔV > V_M (Trip Point)



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

Transfer ratio

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

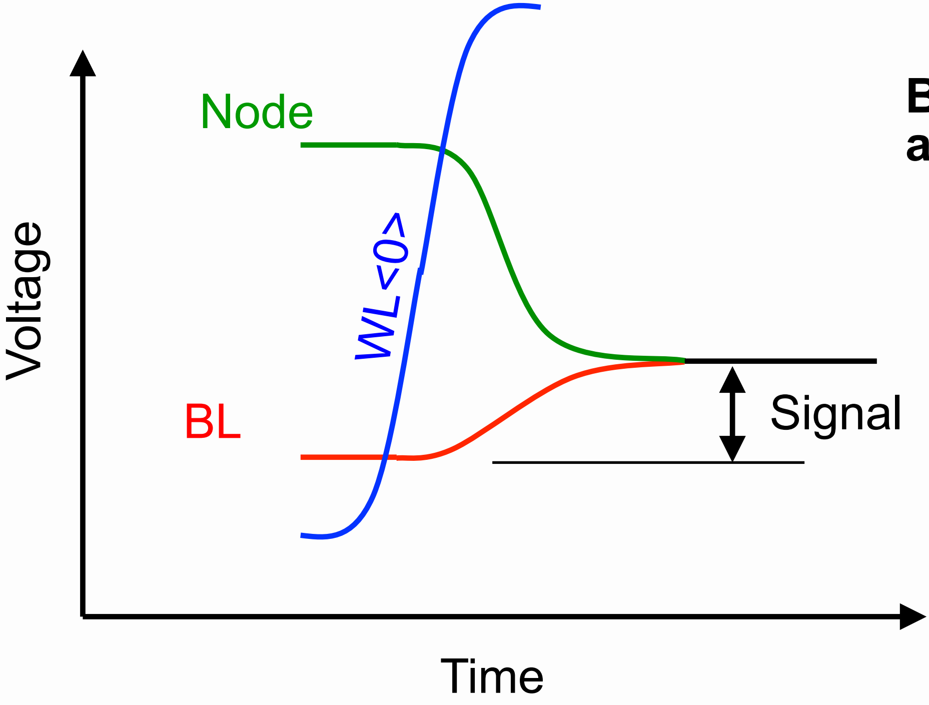
Bit-Line Voltage - Reference = Signal

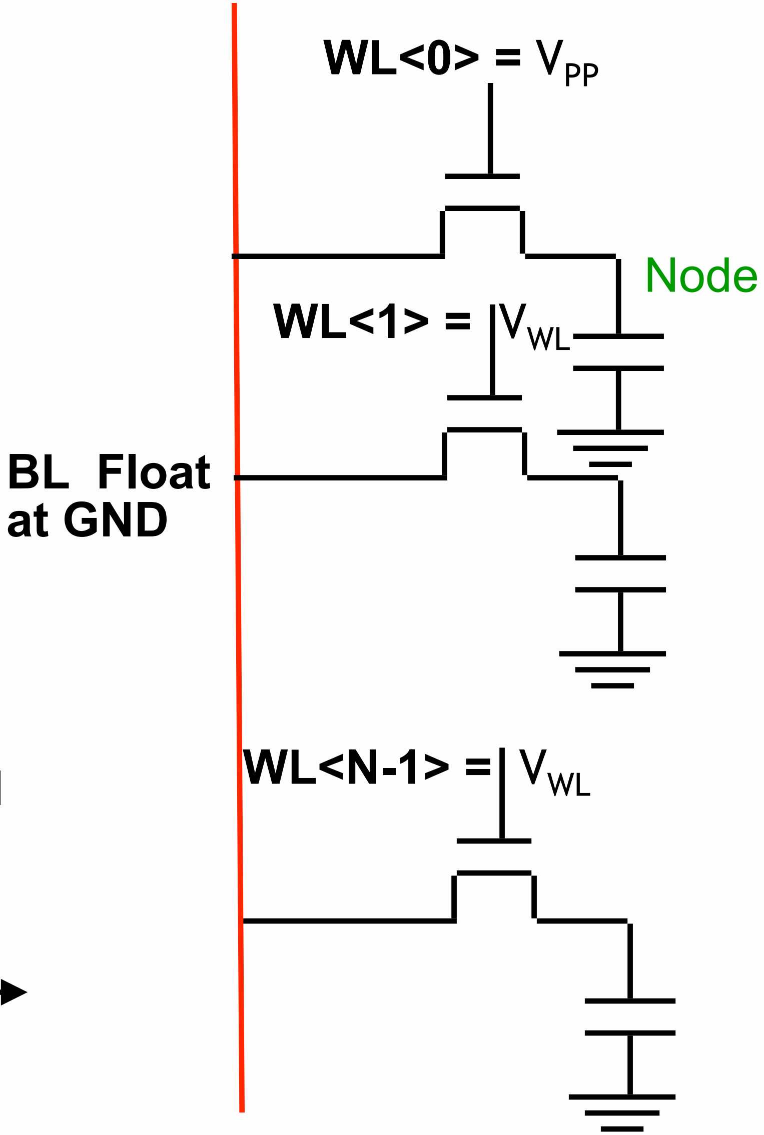
Pos Signal Amplifies to Logical ‘1’, Neg Signal Amplifies to Logical ‘0’

Signal: # WLs on a BL

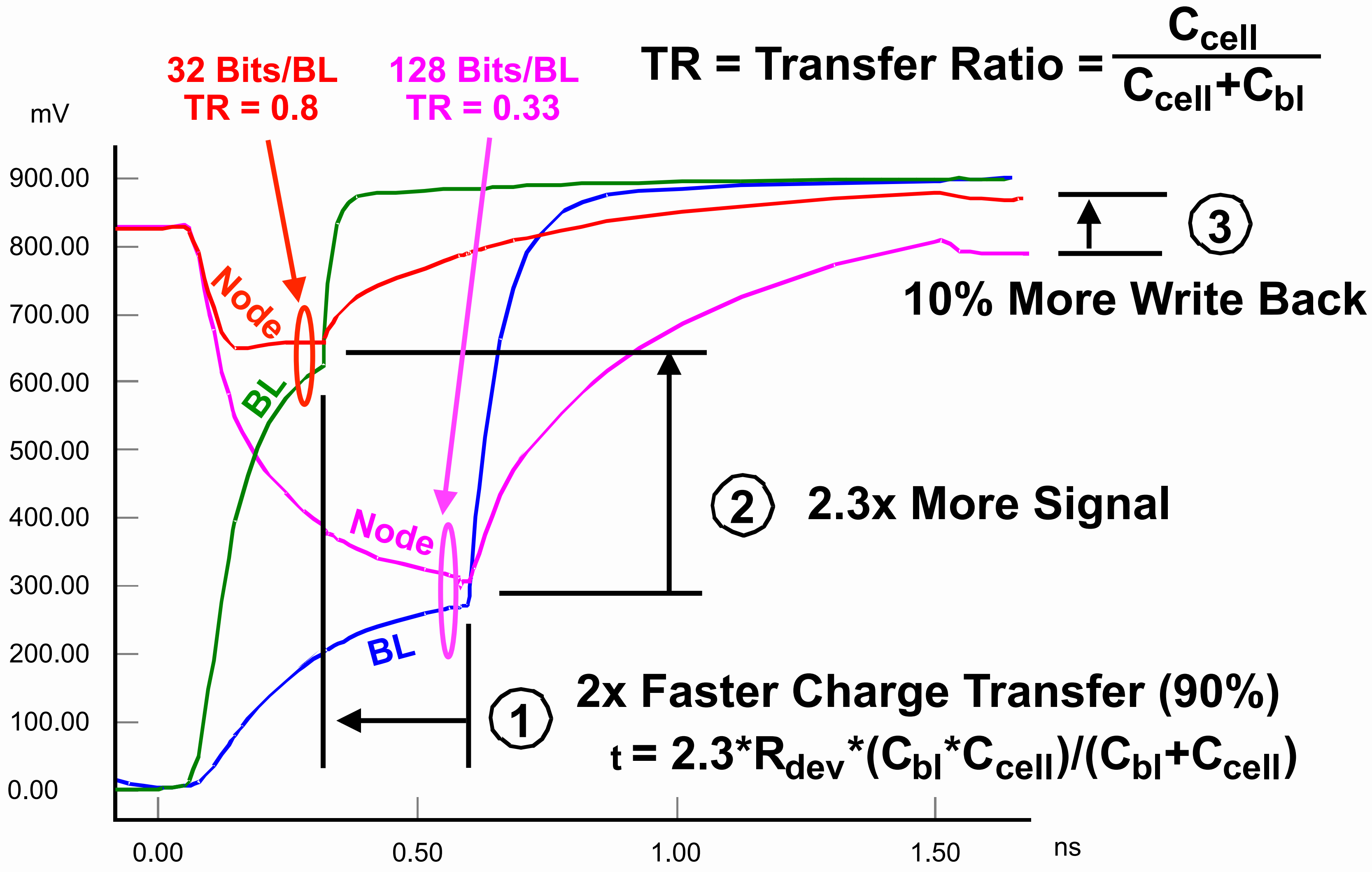
$C_{BL} = N\ C_{DIFF} + NC_{M1}$

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





Bits per Bit-Line v/s Transfer Ratio



32 Bits/BL
TR = 0.8

128 Bits/BL
TR = 0.33

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

10% More Write Back

2.3x More Signal

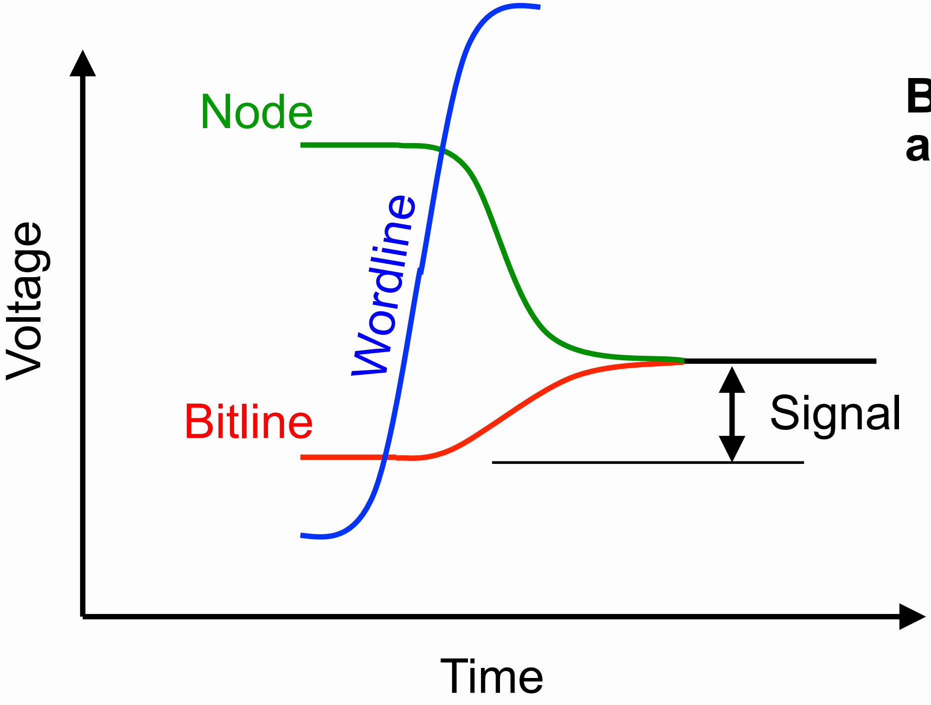
2x Faster Charge Transfer (90%)

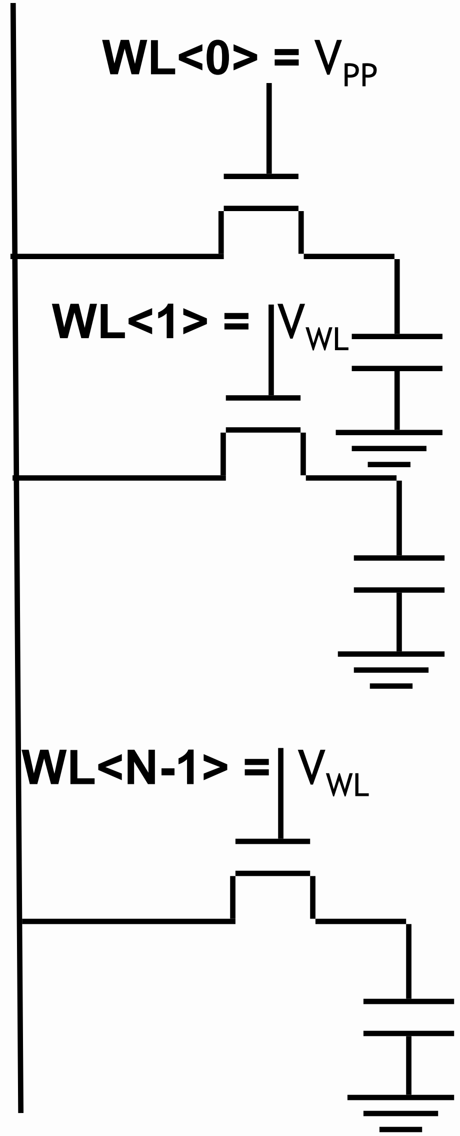
$$t = 2.3*R_{dev}*(C_{bl}*C_{cell})/(C_{bl}+C_{cell})$$

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

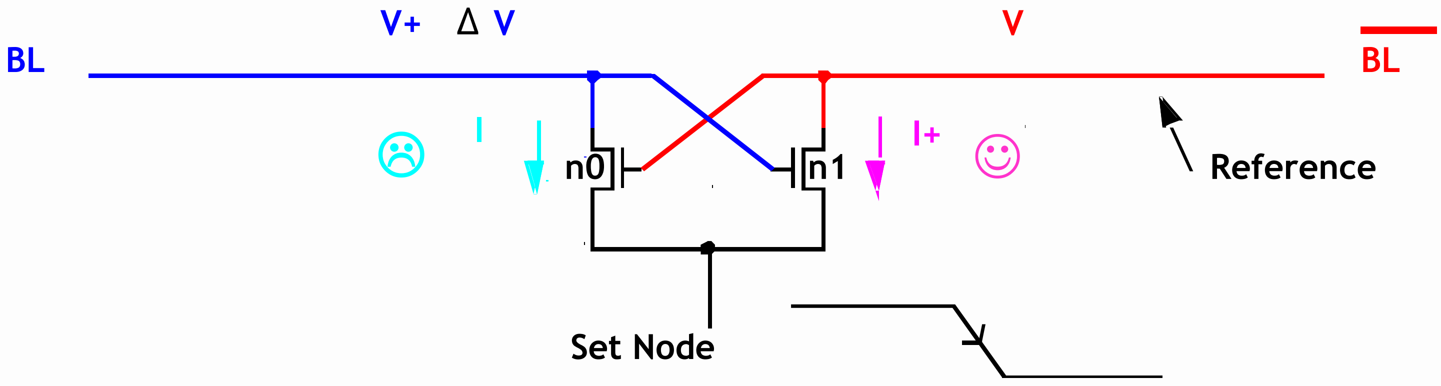
Differential sense amp

- 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