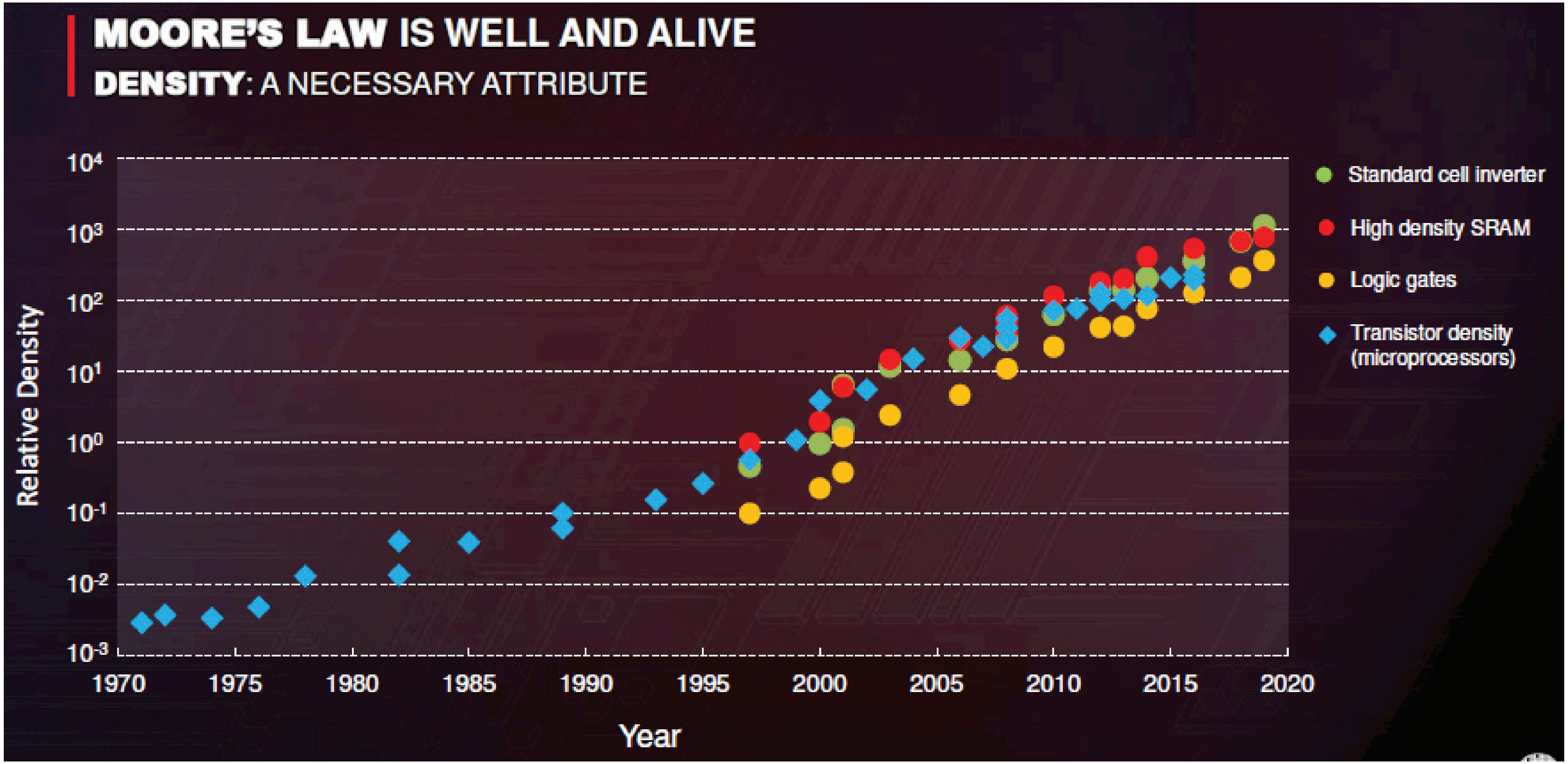
Week 5-1

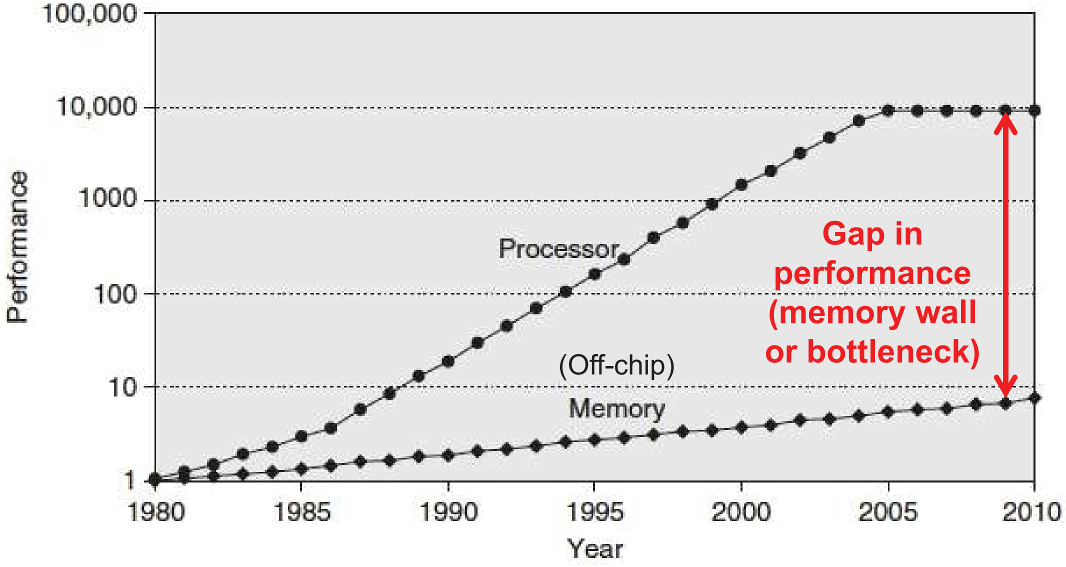
Introduction to Electronic Memories

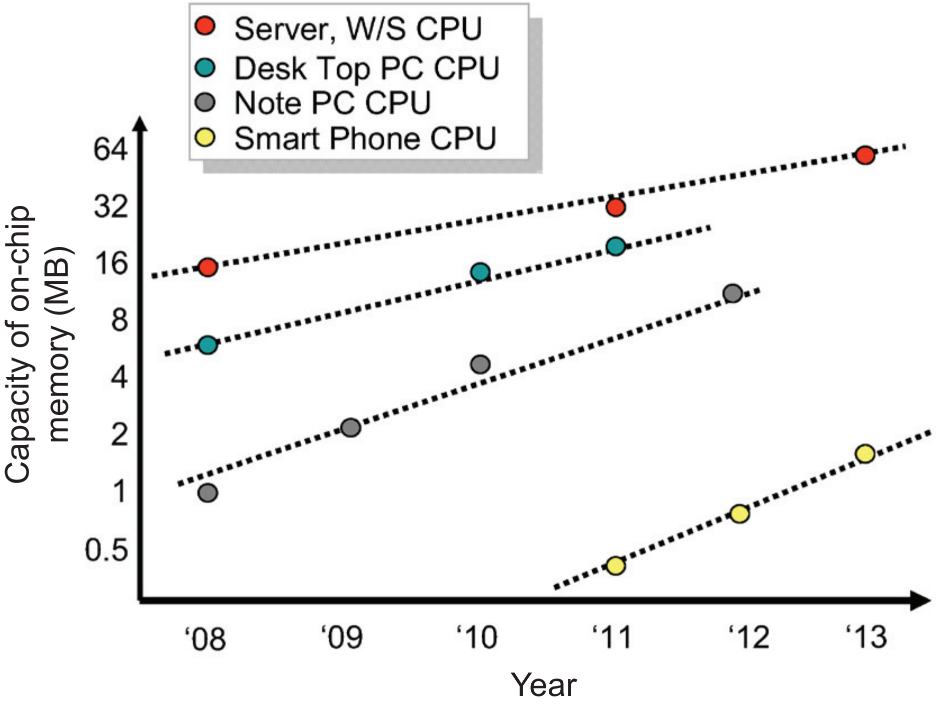
Moore’s Law



Source: H. S. P. Wong, HotChips 2019
see https://www.youtube.com/watch?v=O5UQ5OGOsnM

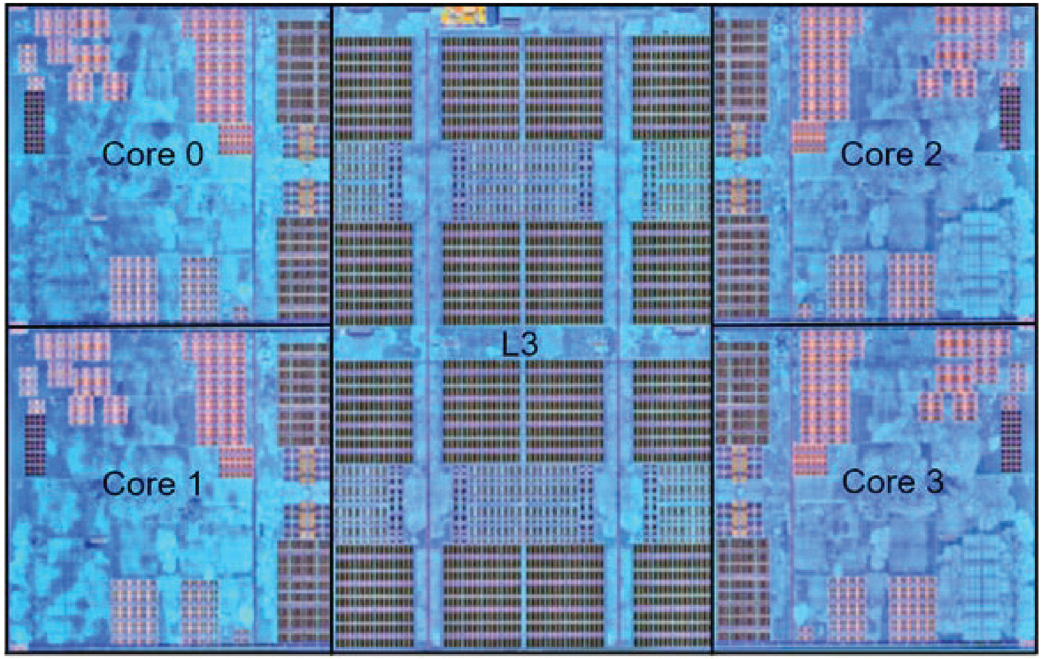
The Memory Wall

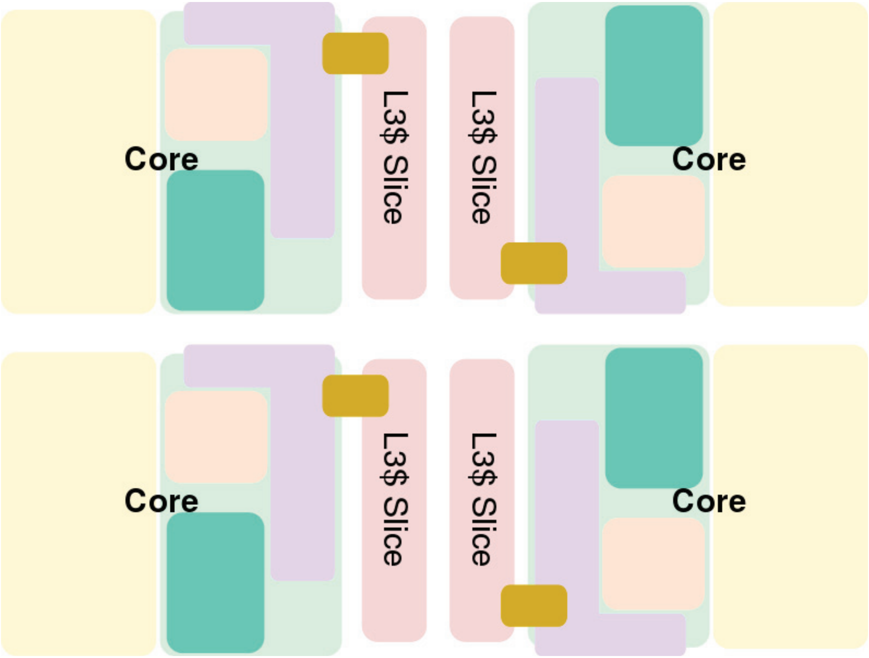




On-chip memory (cache) that is smaller in capacity than off-chip memory but a lot closer in performance and speed to the processor helps to reduce number of slow off-chip memory accesses

AMD Zen x86-64 Microprocessor

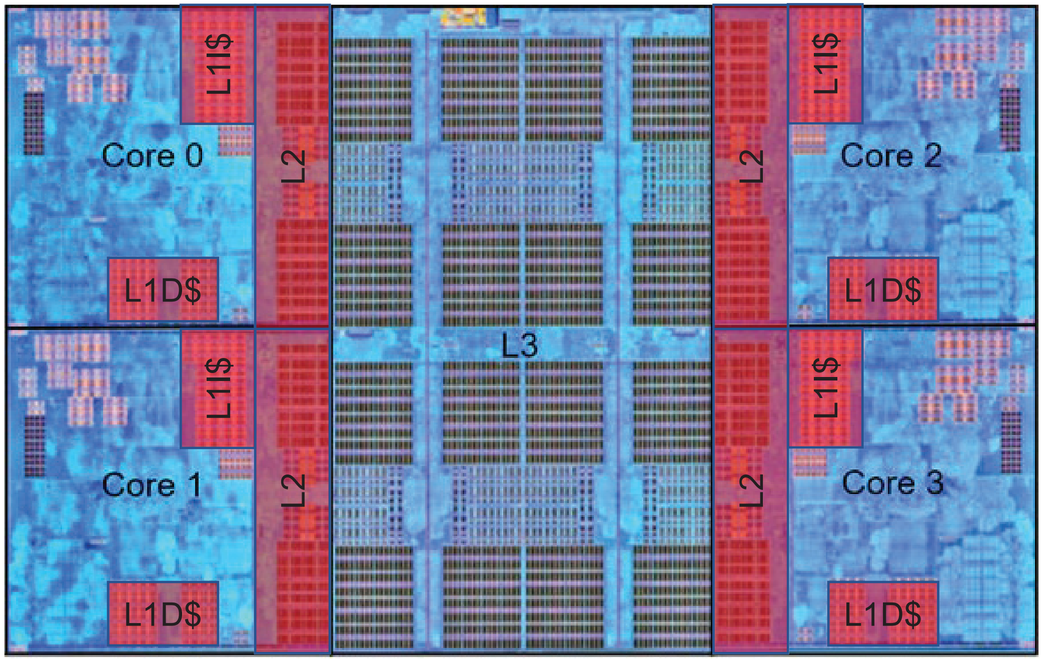


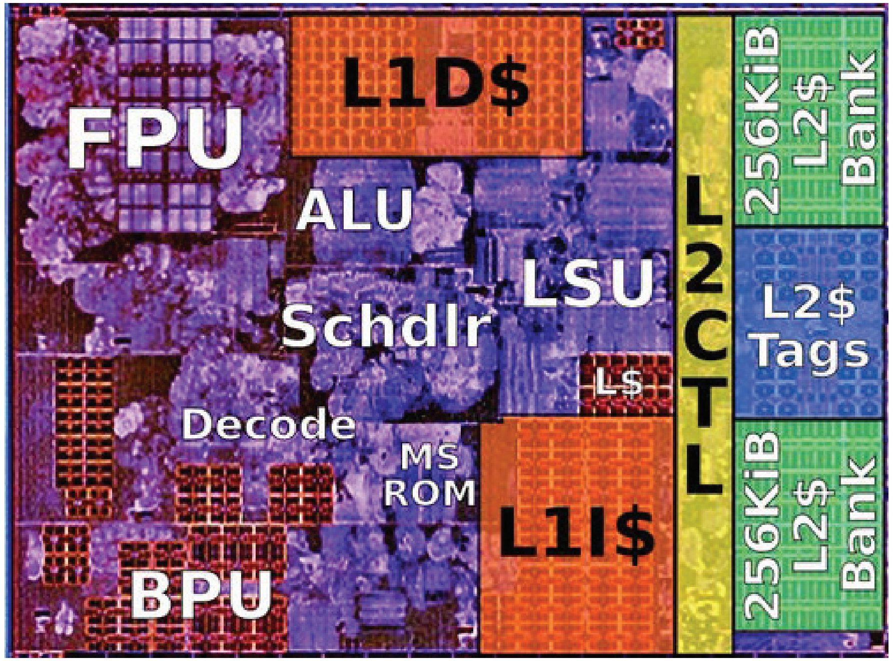


Source: wikichip.org

AMD Zen x86-64 Microprocessor

FPU: Floating-point unit BPU: Branch prediction unit LSU: Load-store unit Schdlr: Instruction scheduler
Decode: Instruction decoder MS ROM: Microcode store ROM L1\$, L2\$: Level 1 and 2 cache

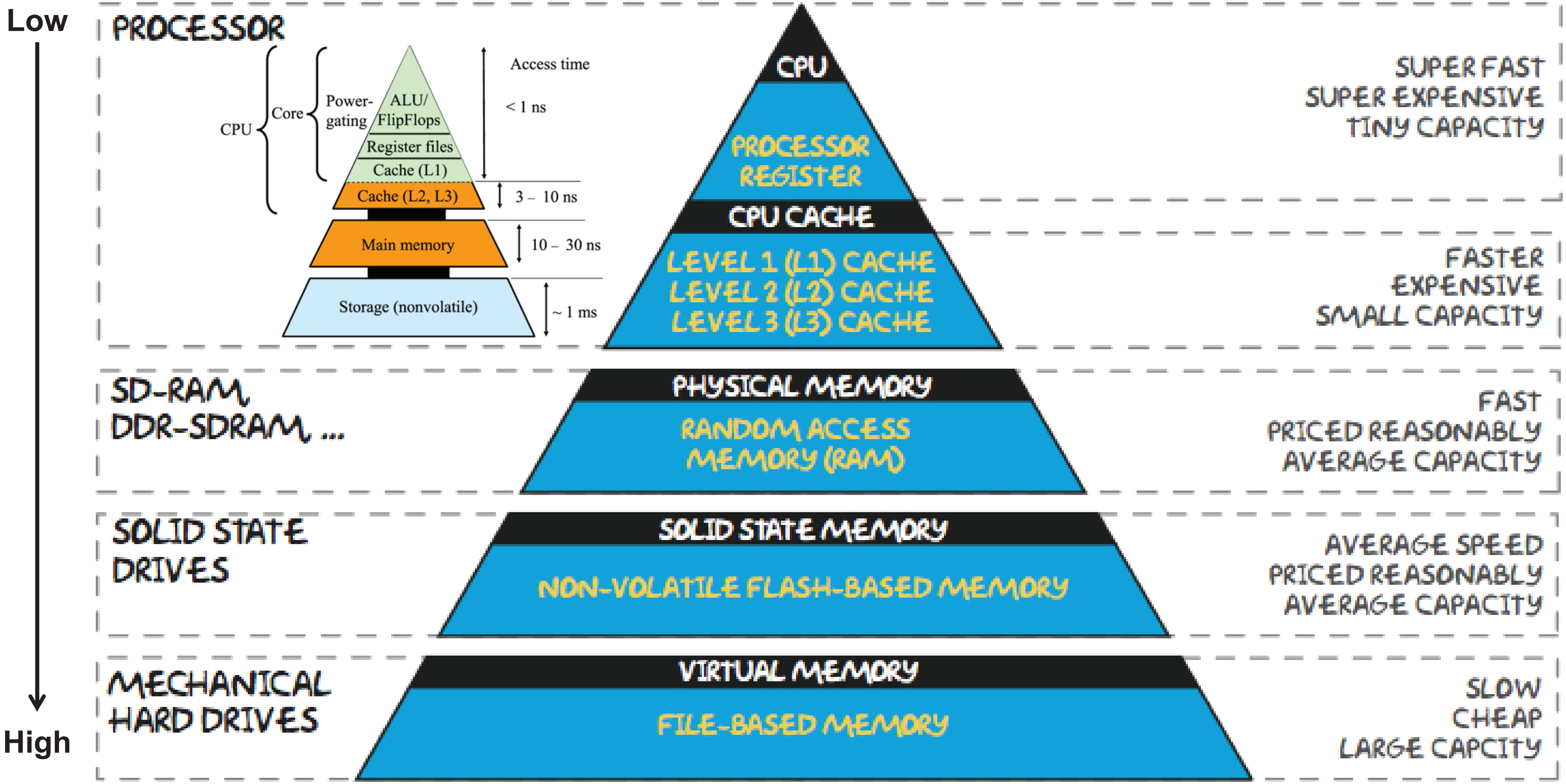




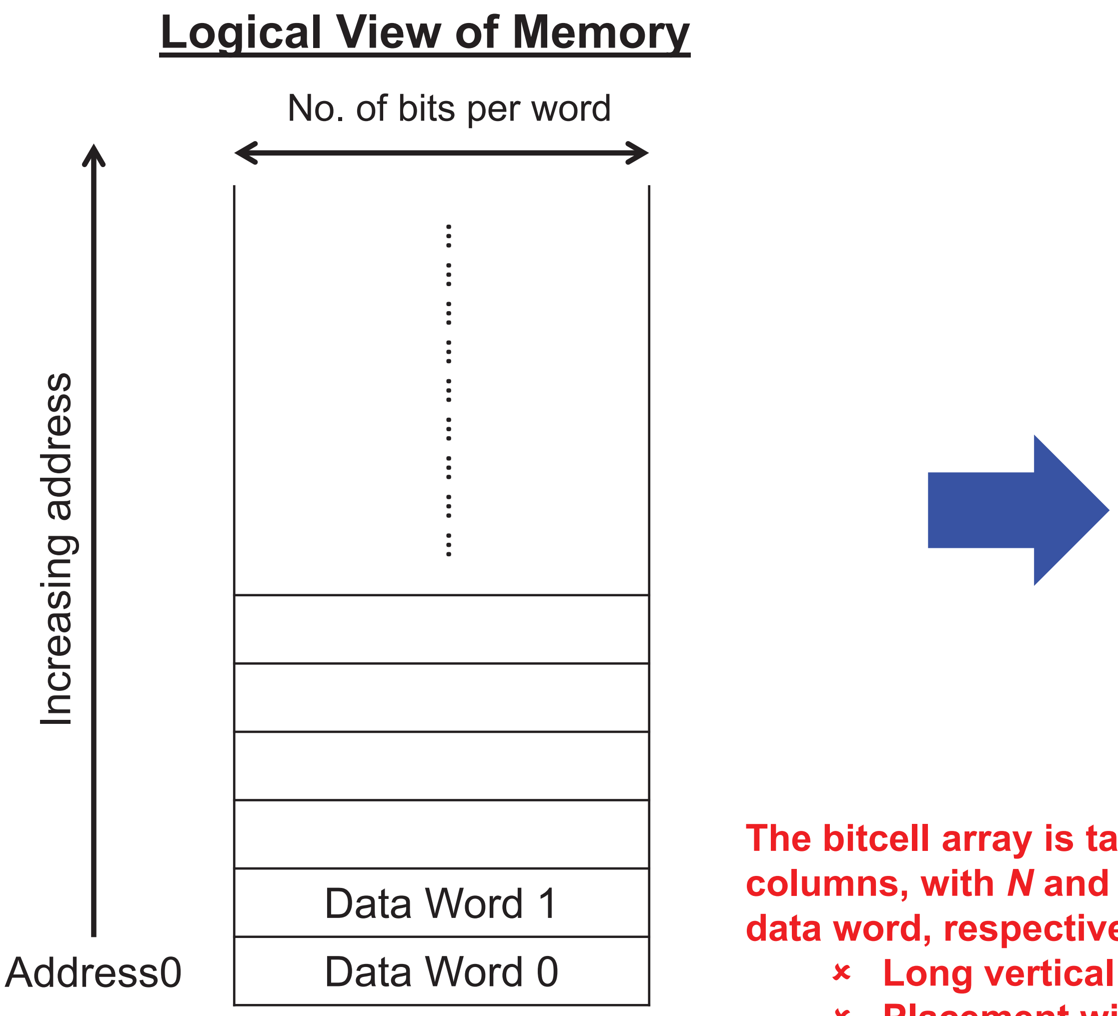
Memory can occupy as much as > 80% of the die area in modern digital processors!
Tackling design issues in memory (especially standby power consumption)
is crucial in advancing processor technology

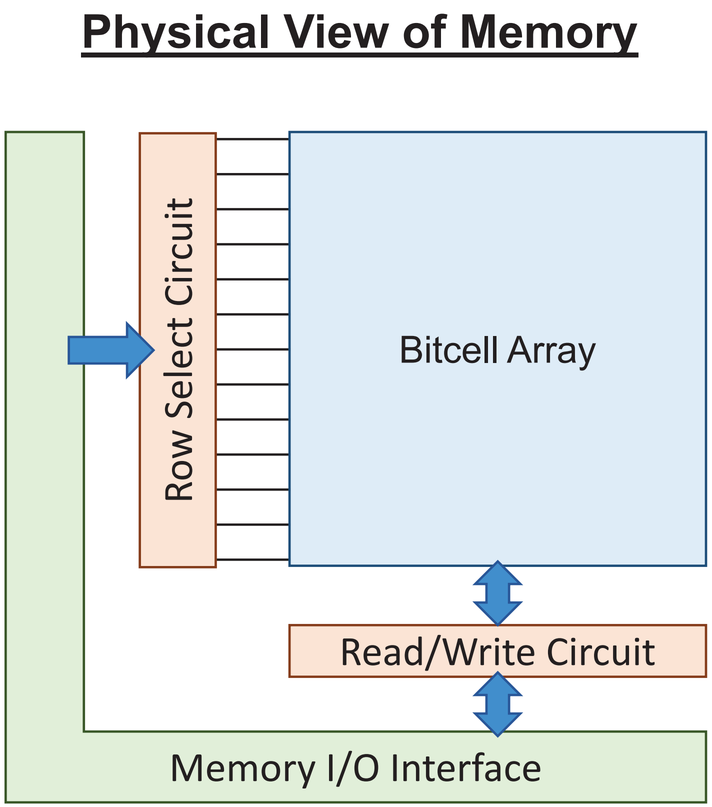
Source: wikichip.org

The Memory Hierarchy



Logical vs Physical Memory – I

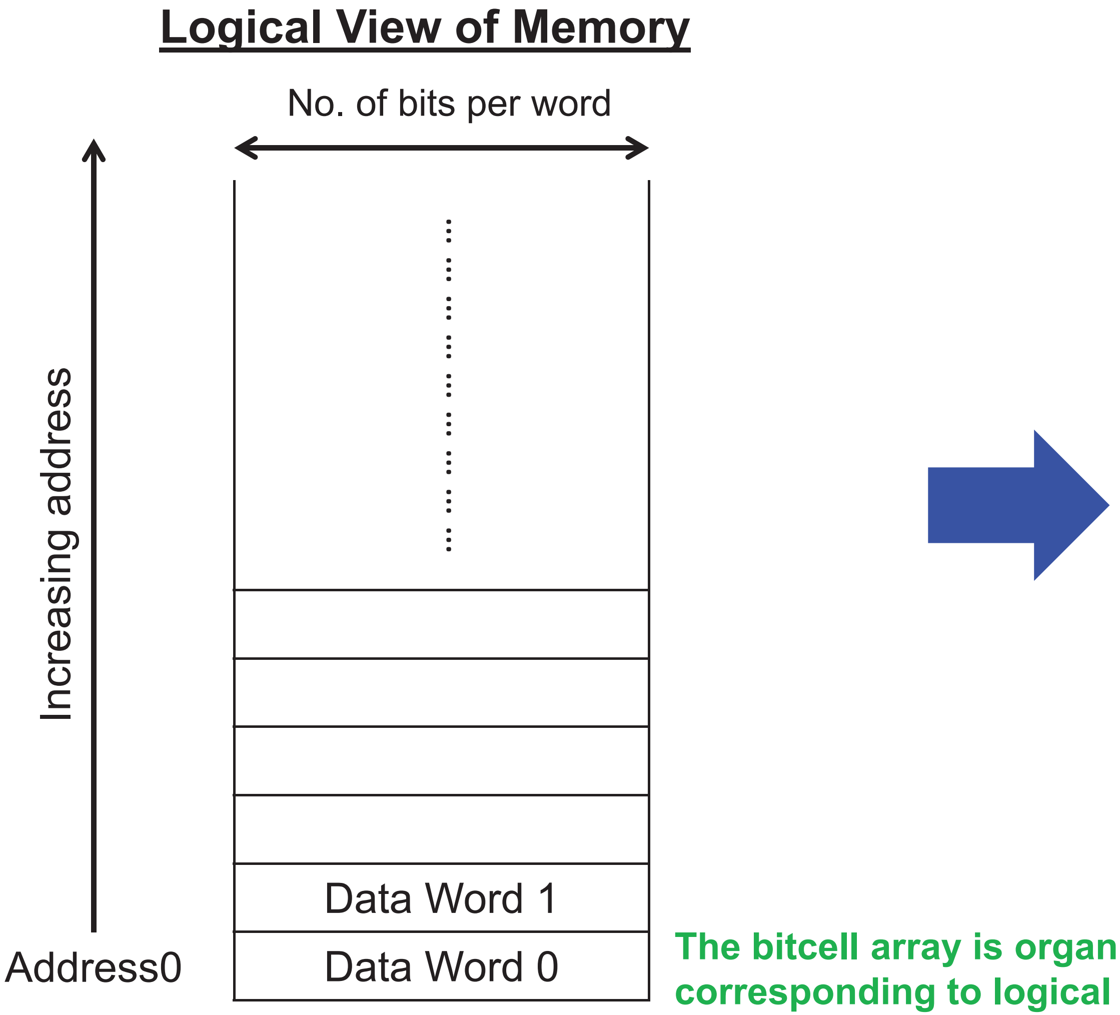


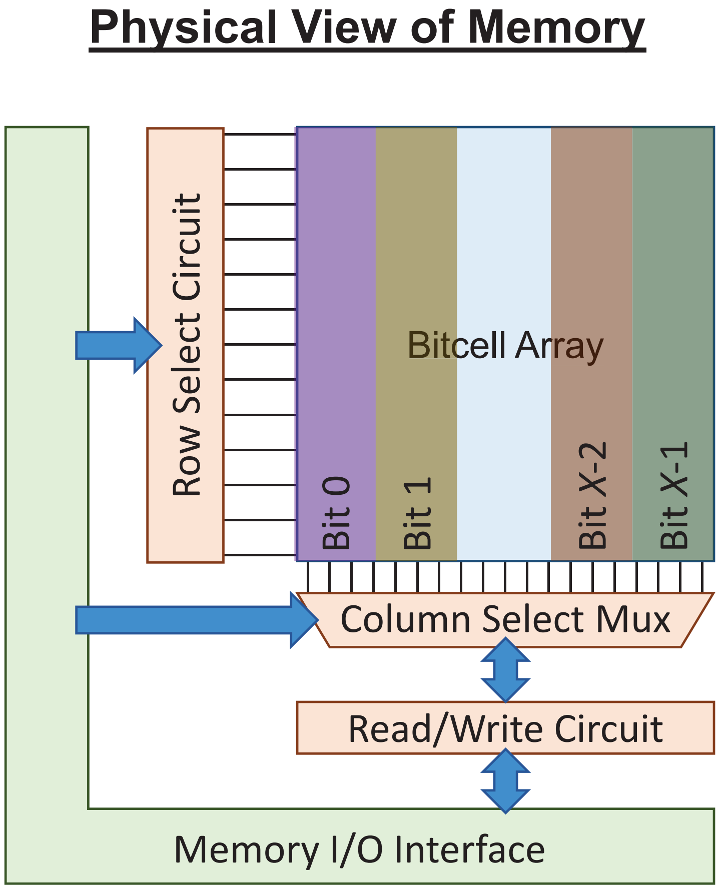


The bitcell array is tall and narrow if it is organized into *N* rows by *M* columns, with *N* and *M* corresponding to logical address and width of data word, respectively)

- ✗ Long vertical wires have large parasitics (bad delay)
- ✗ Placement with other circuitry
- ✗ Yield issues in bitcell array and R/W circuitry

Logical vs Physical Memory – II



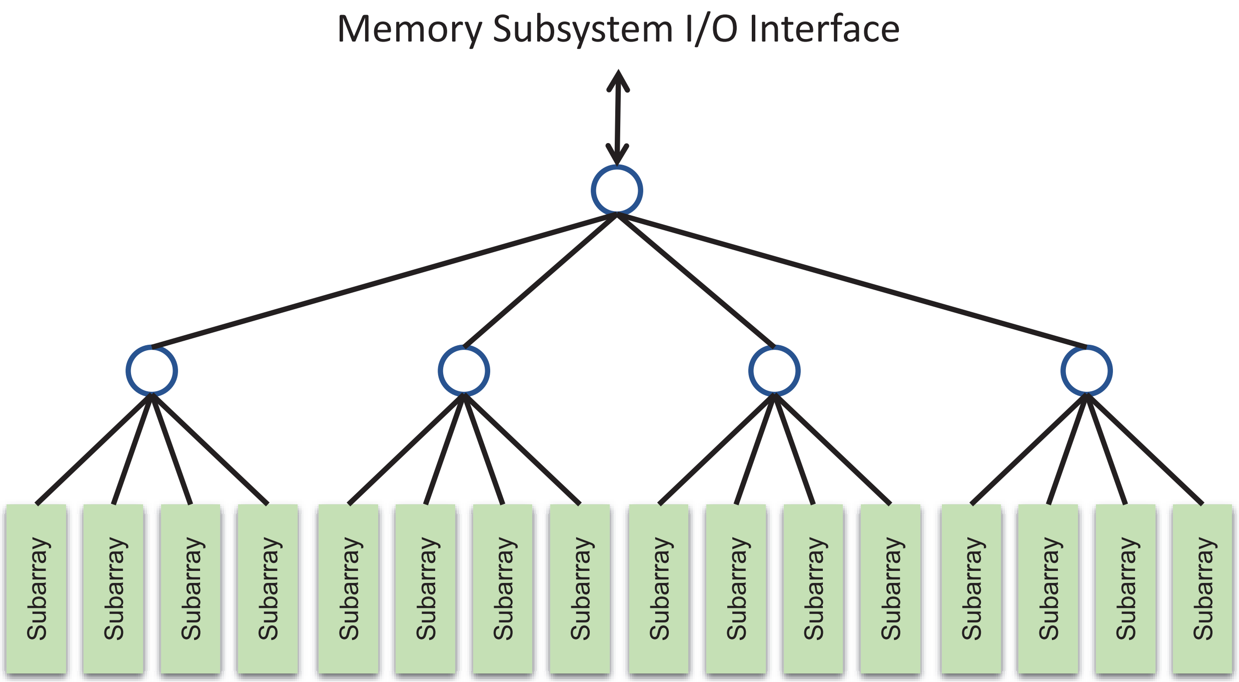


The bitcell array is organized into *N* rows by *M* columns, with *N* and *M* corresponding to logical address and multiple (usually 8) times the width of data word, respectively

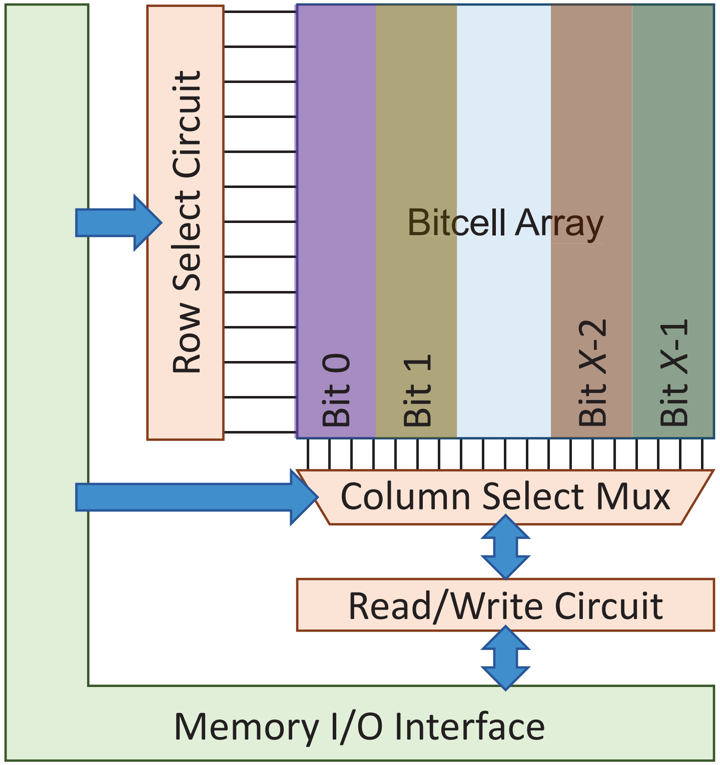
- **Bit-interleaved to guard against errors using error-correcting codes (ECC)**

Physical Organization of Memory

Memory Subsystem

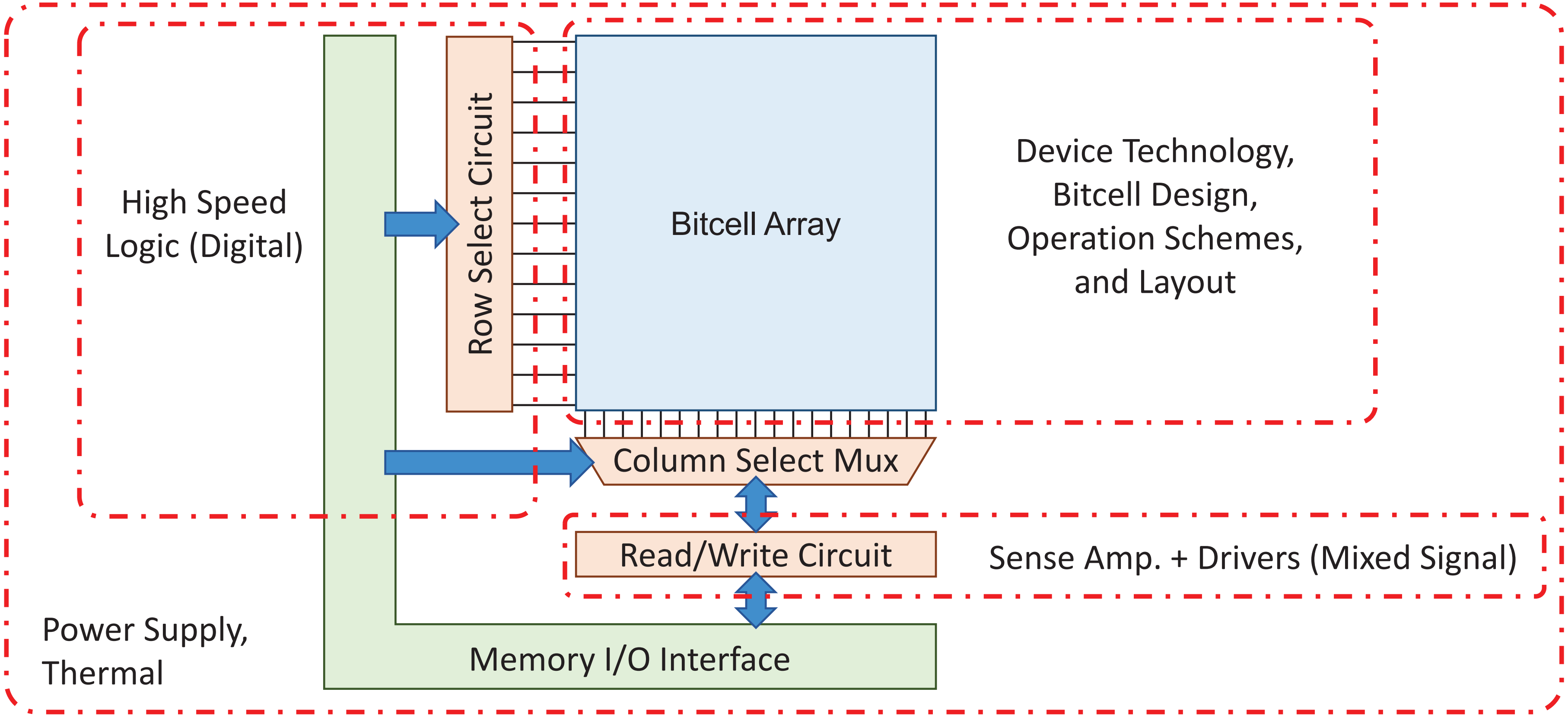


Single Subarray



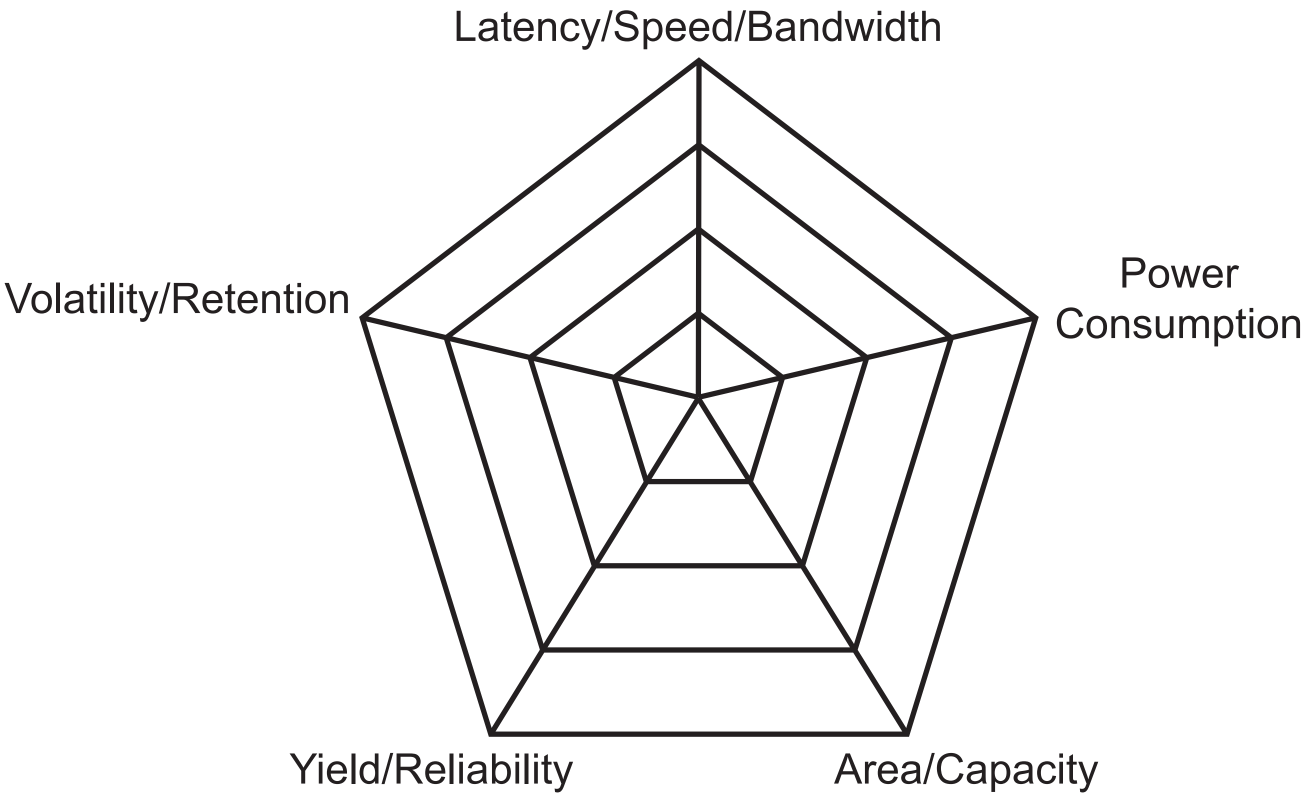
Memory subsystem is stitched together from many subarrays in a tree like fashion to tradeoff between memory capacity, bandwidth, complexity of control circuitry, power consumption and speed

Parts of the Subarray (Macro)


High Speed Logic (Digital)
Row Select Circuit
Bitcell Array
Device Technology, Bitcell Design, Operation Schemes, and Layout
Column Select Mux
Read/Write Circuit
Sense Amp. + Drivers (Mixed Signal)
Power Supply, Thermal
Memory I/O Interface

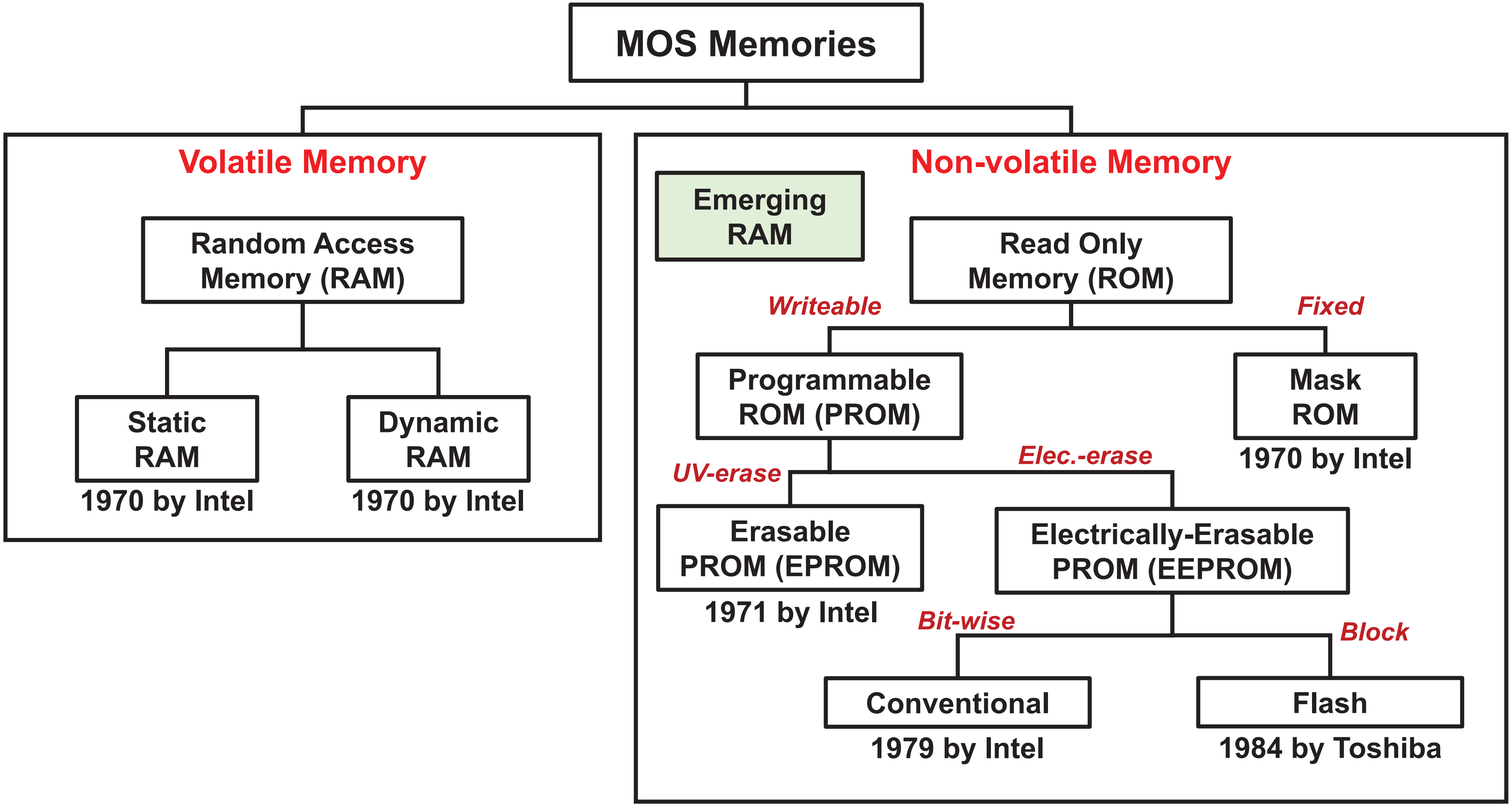
Memory design covers topics ranging from device physics, many areas of circuit design and computer architecture, to information theory, data analytics and statistics (improving yield)

Tradeoffs in Memory Design



Addressing all challenges in memory design requires understanding of interactions across all levels of design abstraction from devices, circuits, architectures, software to applications

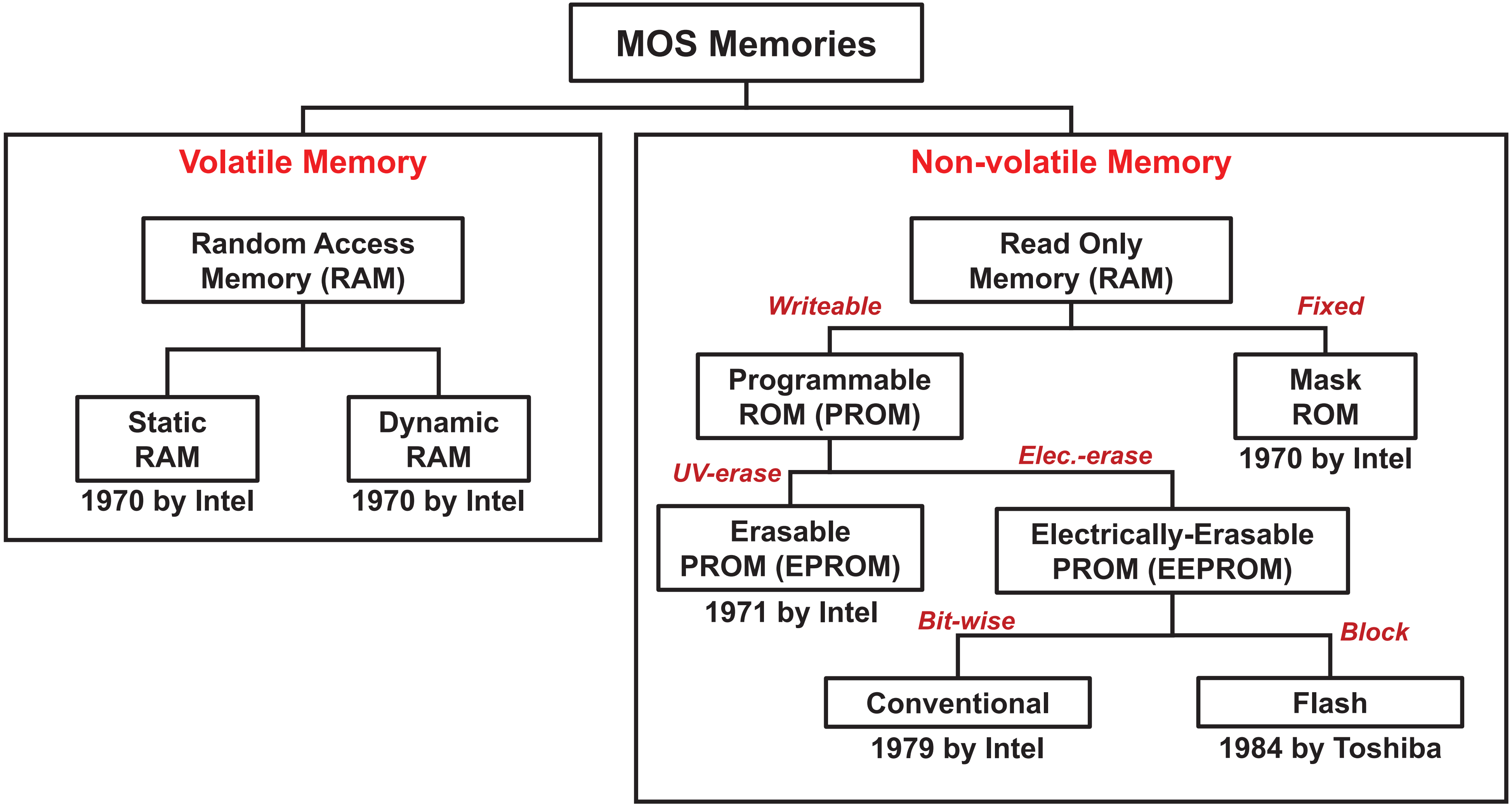
MOS Memory Classification



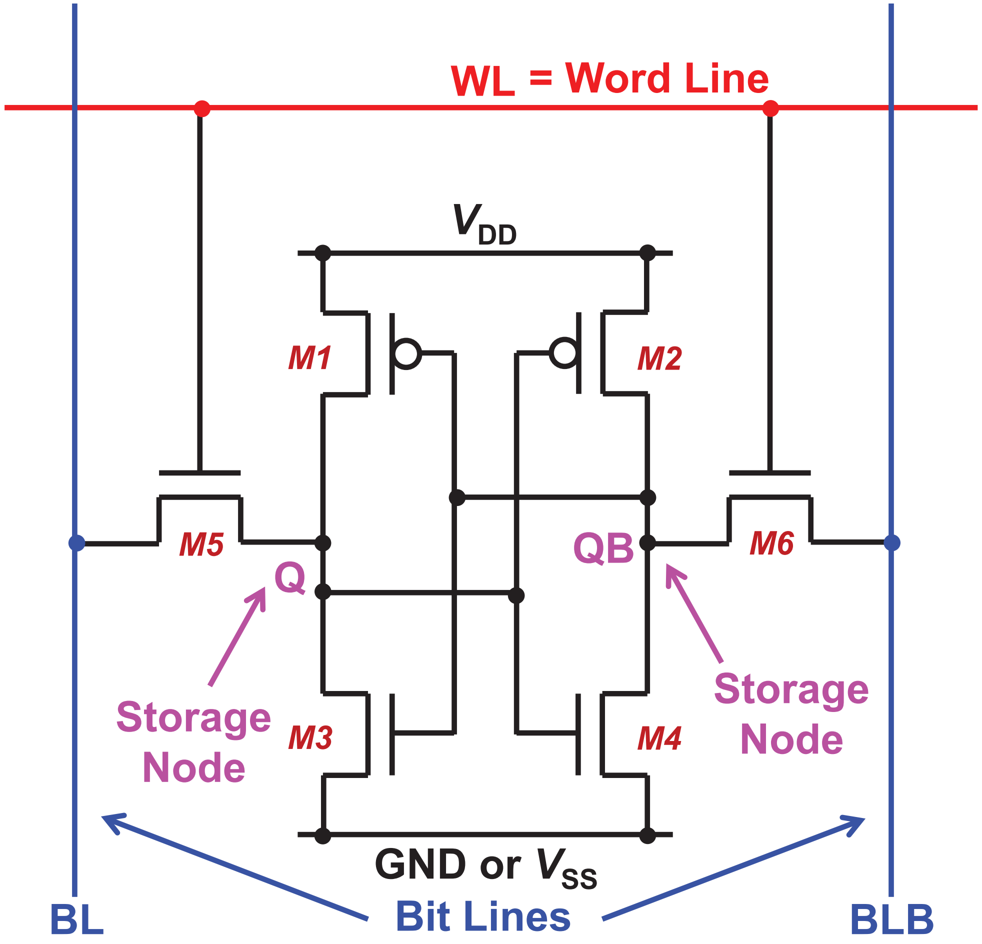
Week 5-2

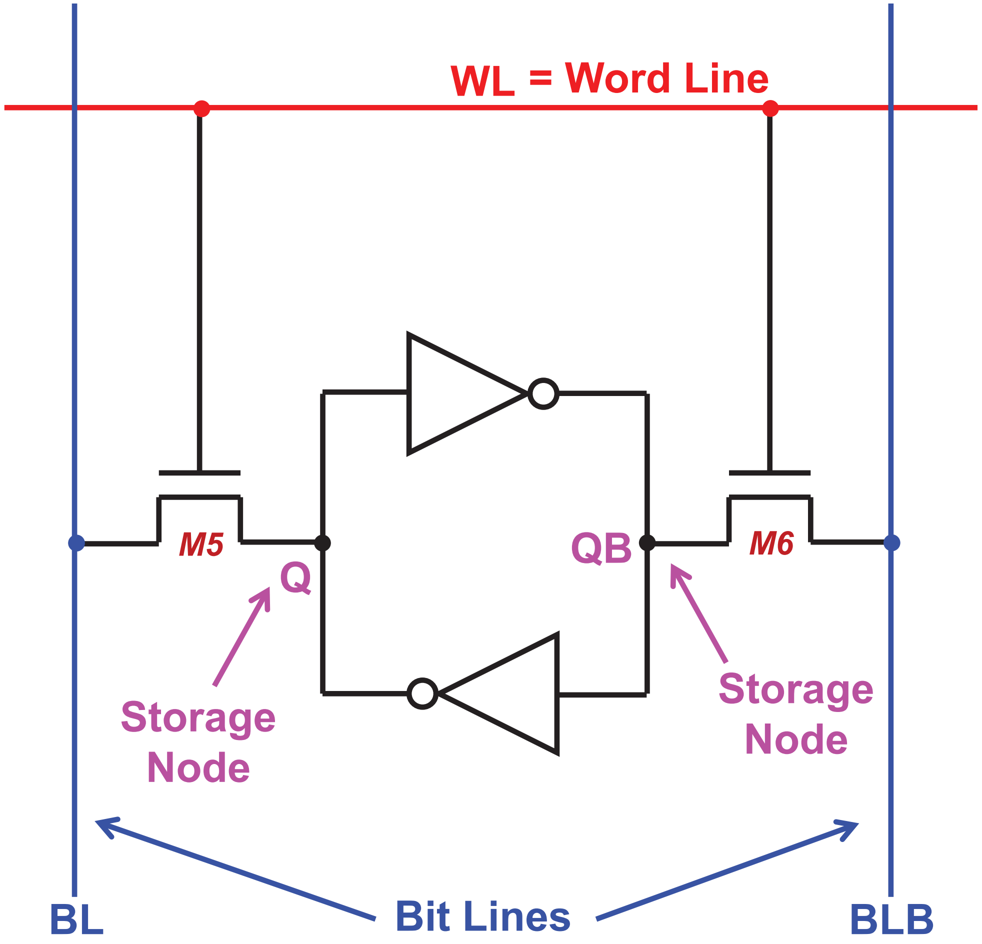
Introduction to the 6-Transistor Static RAM (6T SRAM)

MOS Memory Classification



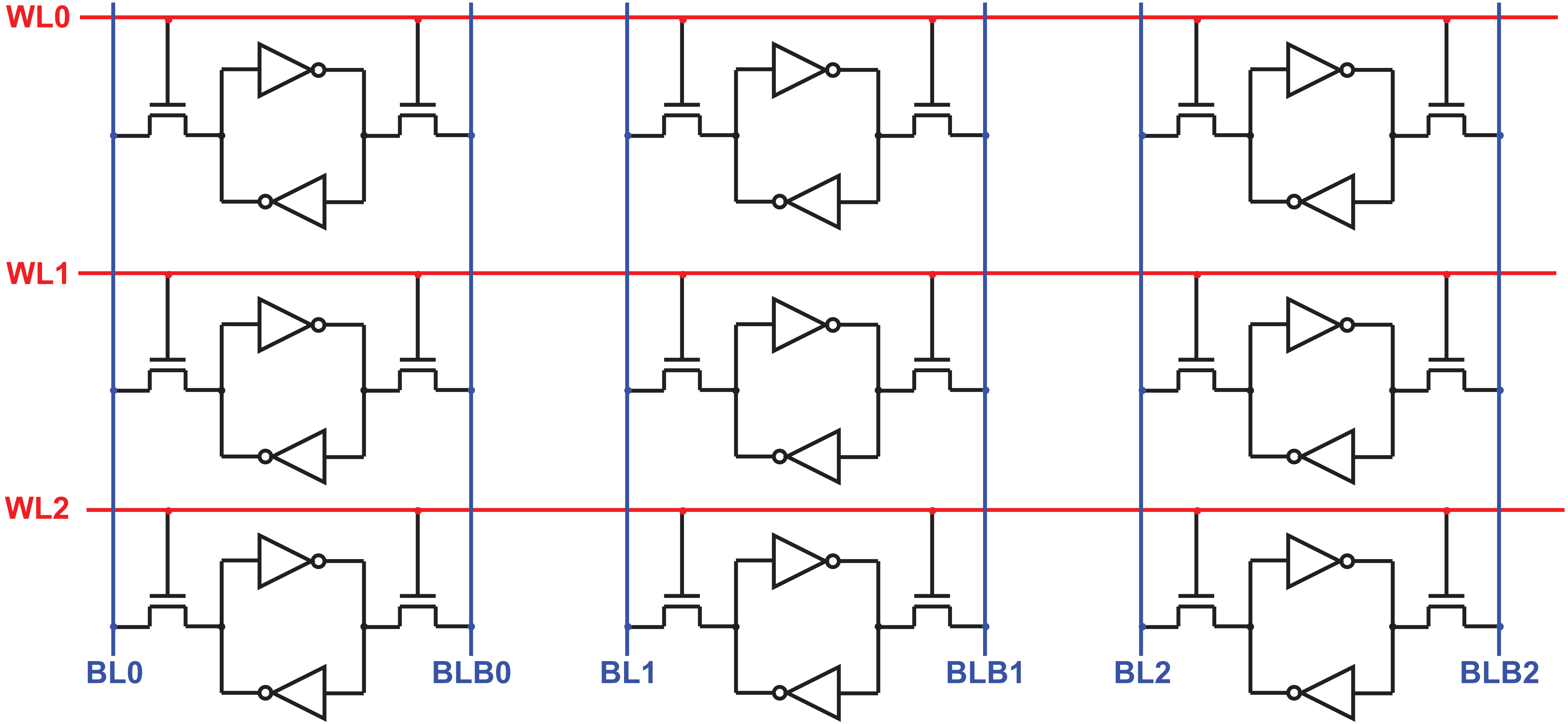
Bitcell Structure of the 6-Transistor Static RAM (6T SRAM)



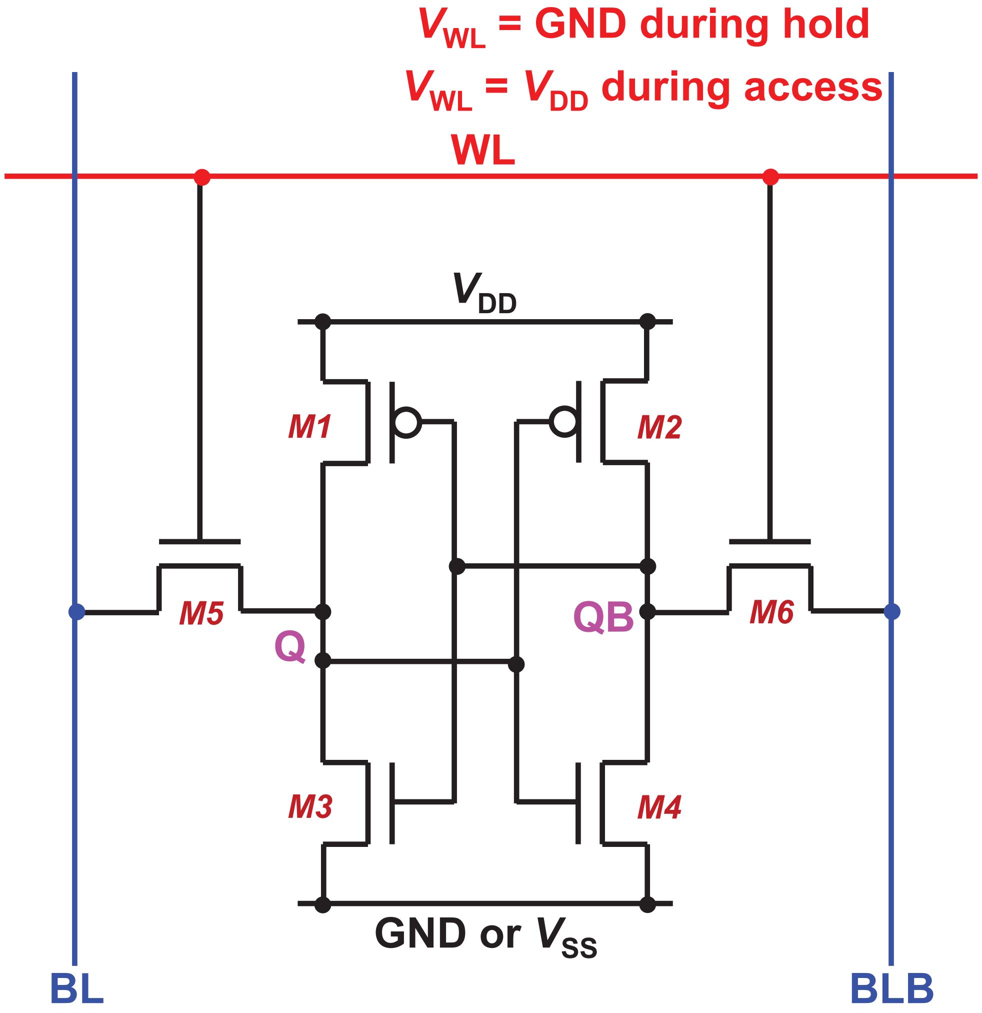


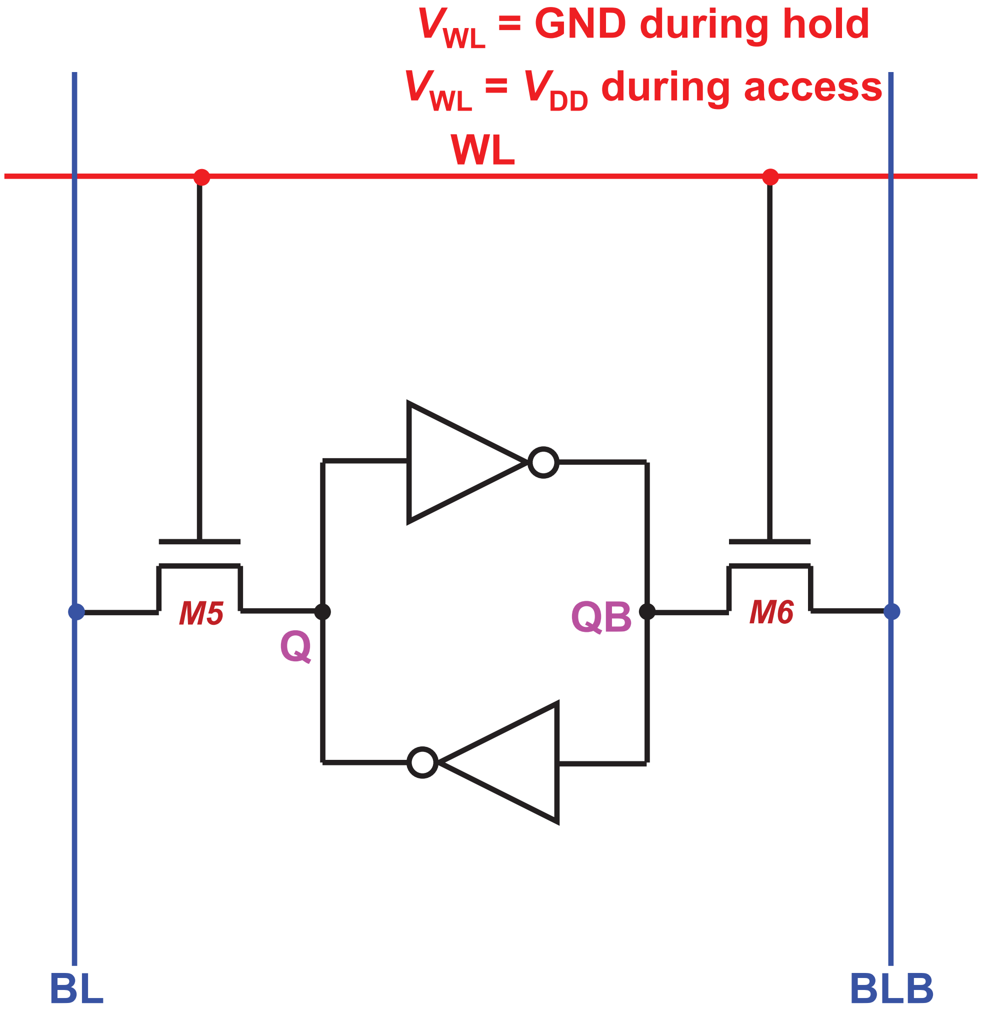
Due to symmetry, $W_{M1}/L_{M1} = W_{M2}/L_{M2}$; $W_{M3}/L_{M3} = W_{M4}/L_{M4}$; $W_{M5}/L_{M5} = W_{M6}/L_{M6}$

Array of 6T SRAM Bitcells (Bitcell Array)



Bitcell Structure of the 6-Transistor Static RAM (6T SRAM)



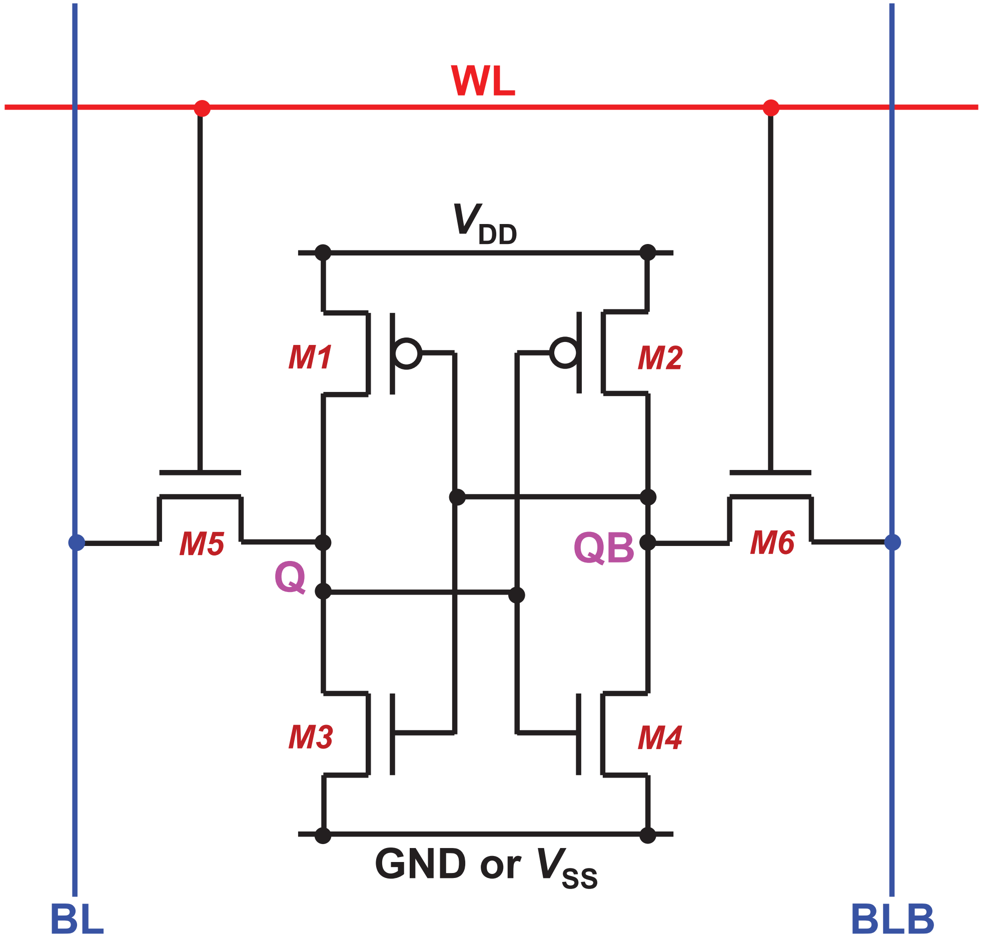


- The bitcell is accessed to read from or write into the bitcell
- Otherwise, bitcell is in hold or hold mode and retains its stored data

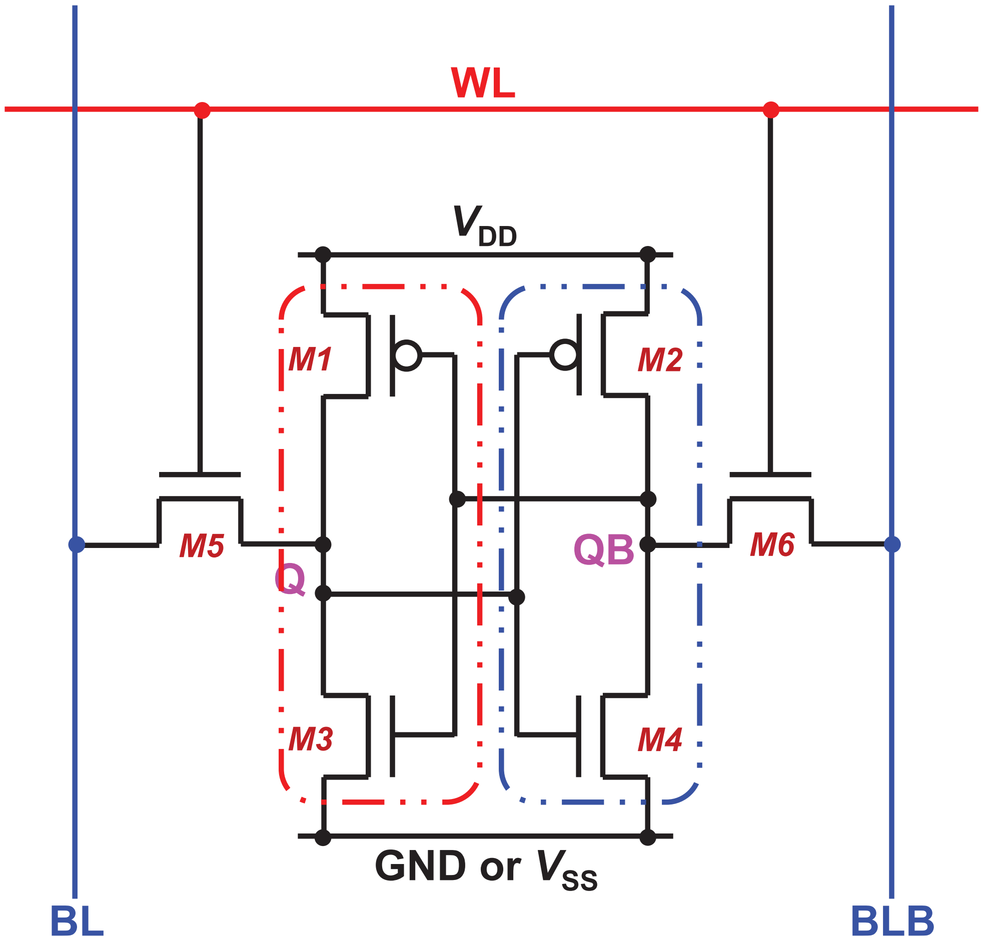
Week 5-3

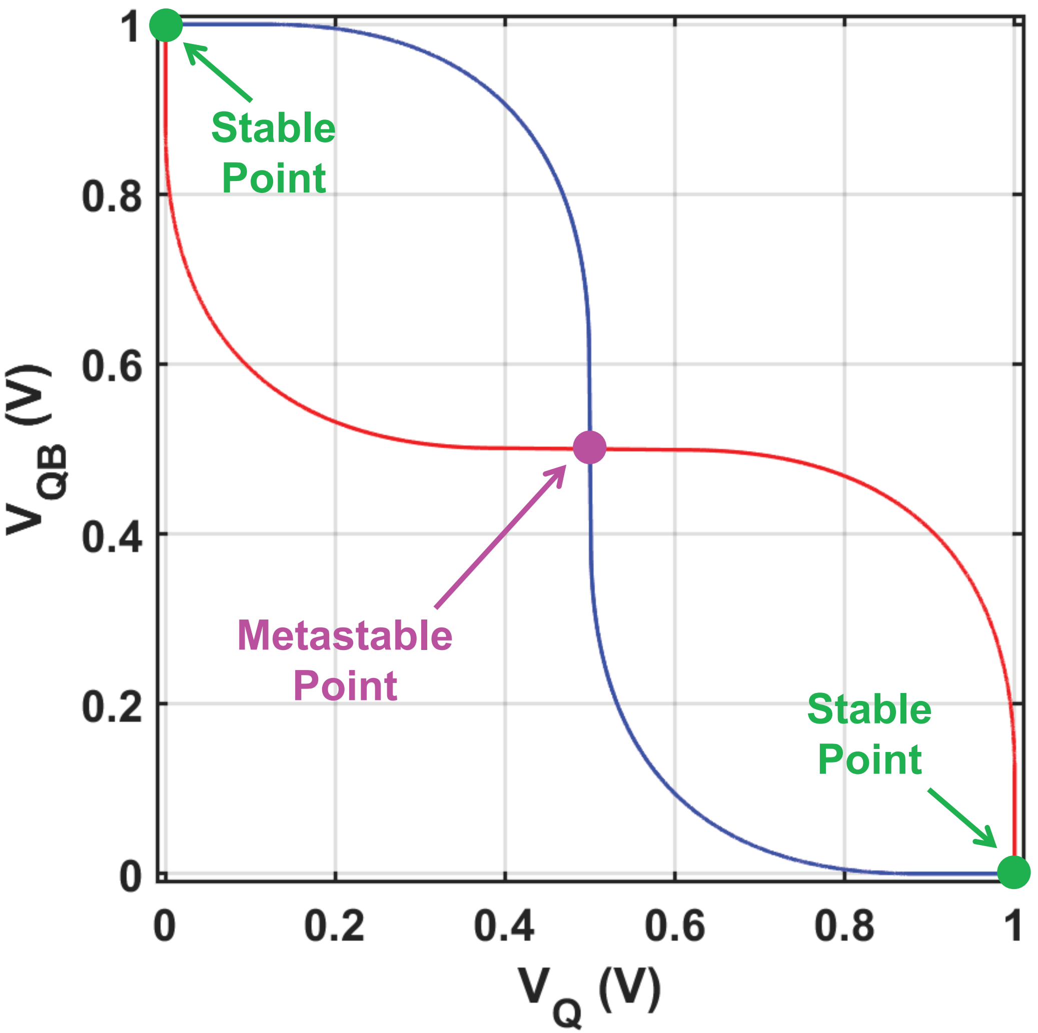
SRAM Read and Write Operations

Bitcell Structure of the 6T SRAM

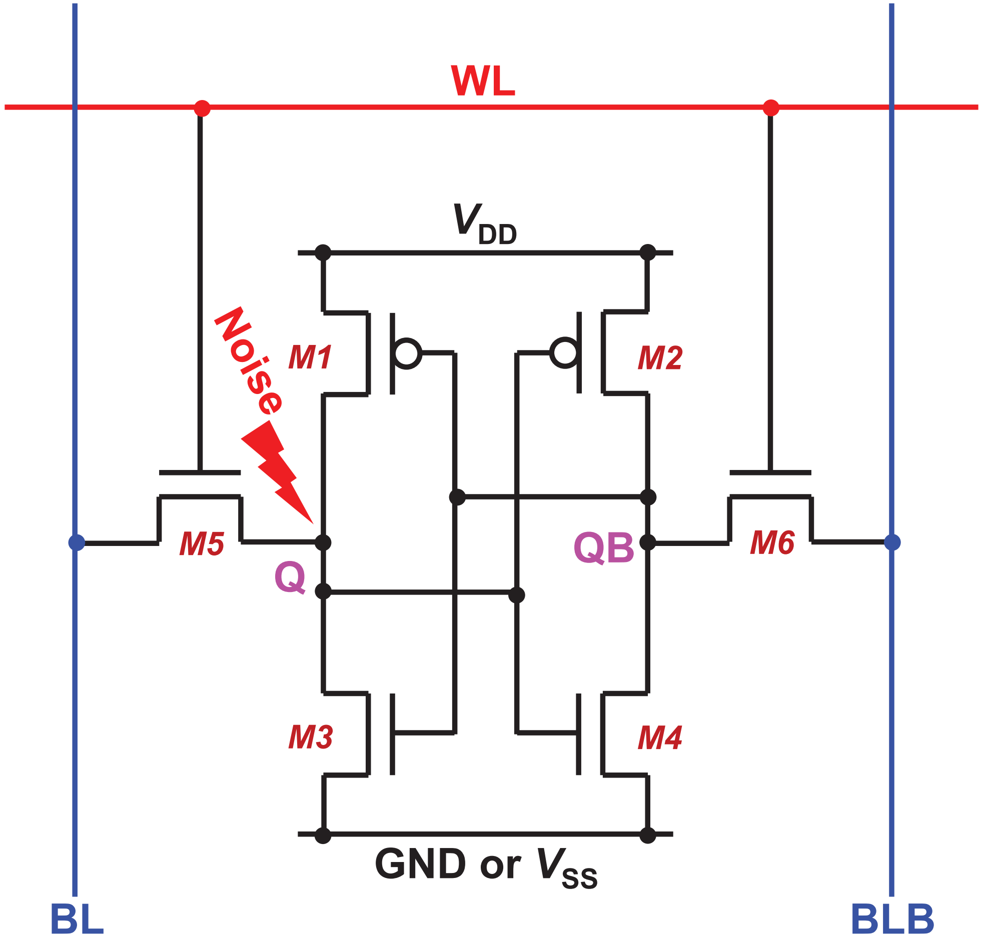


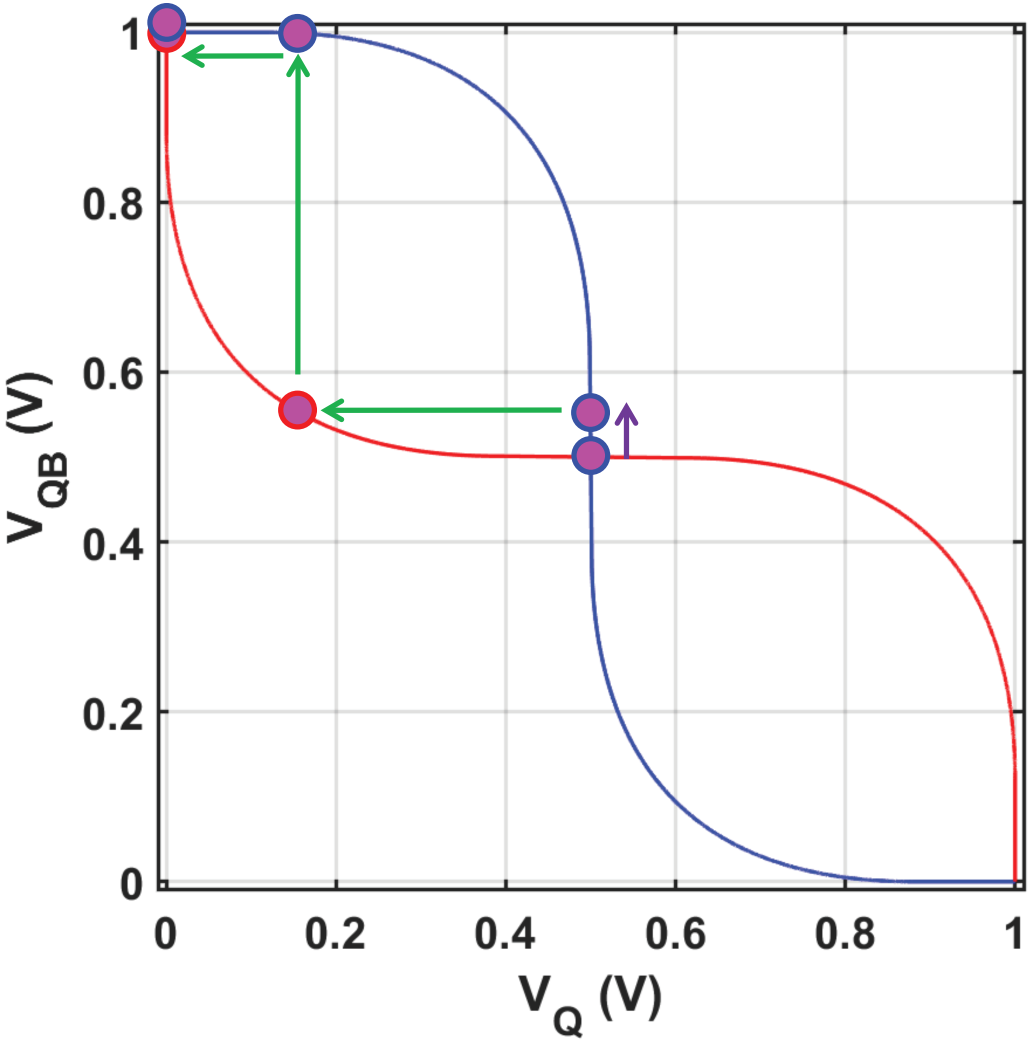
The Bistability Principle



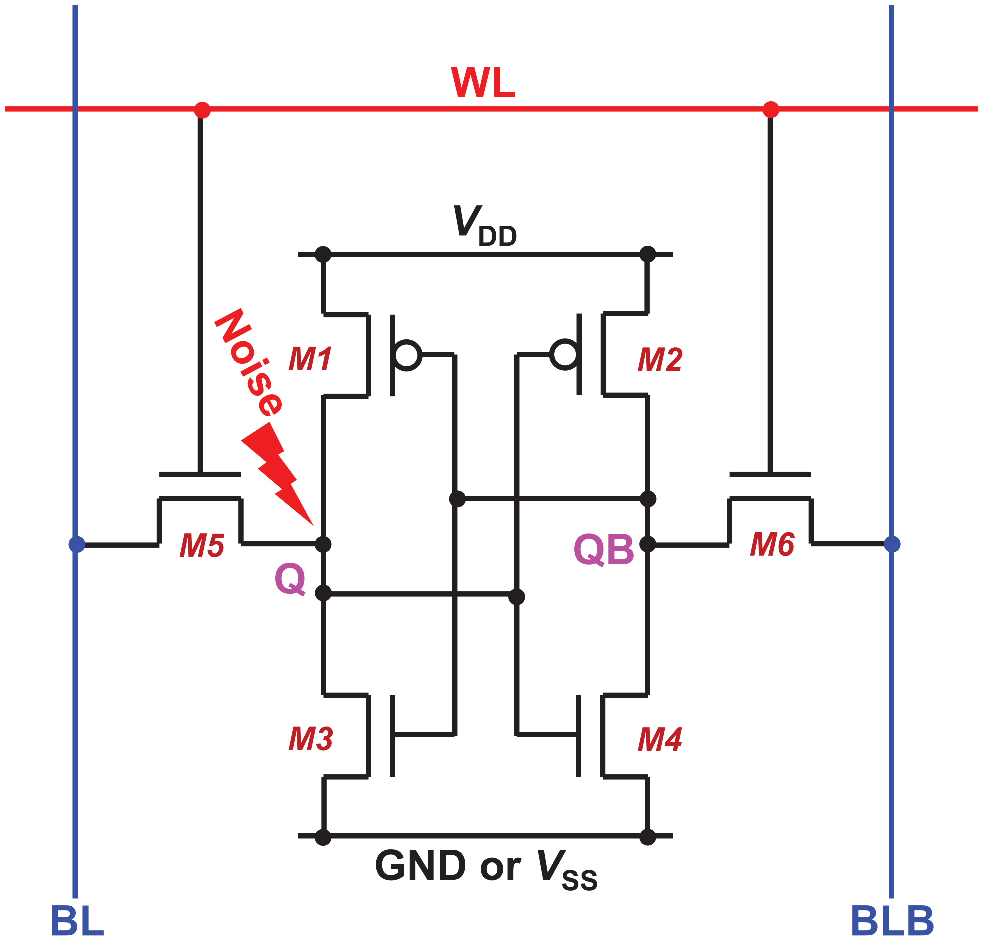


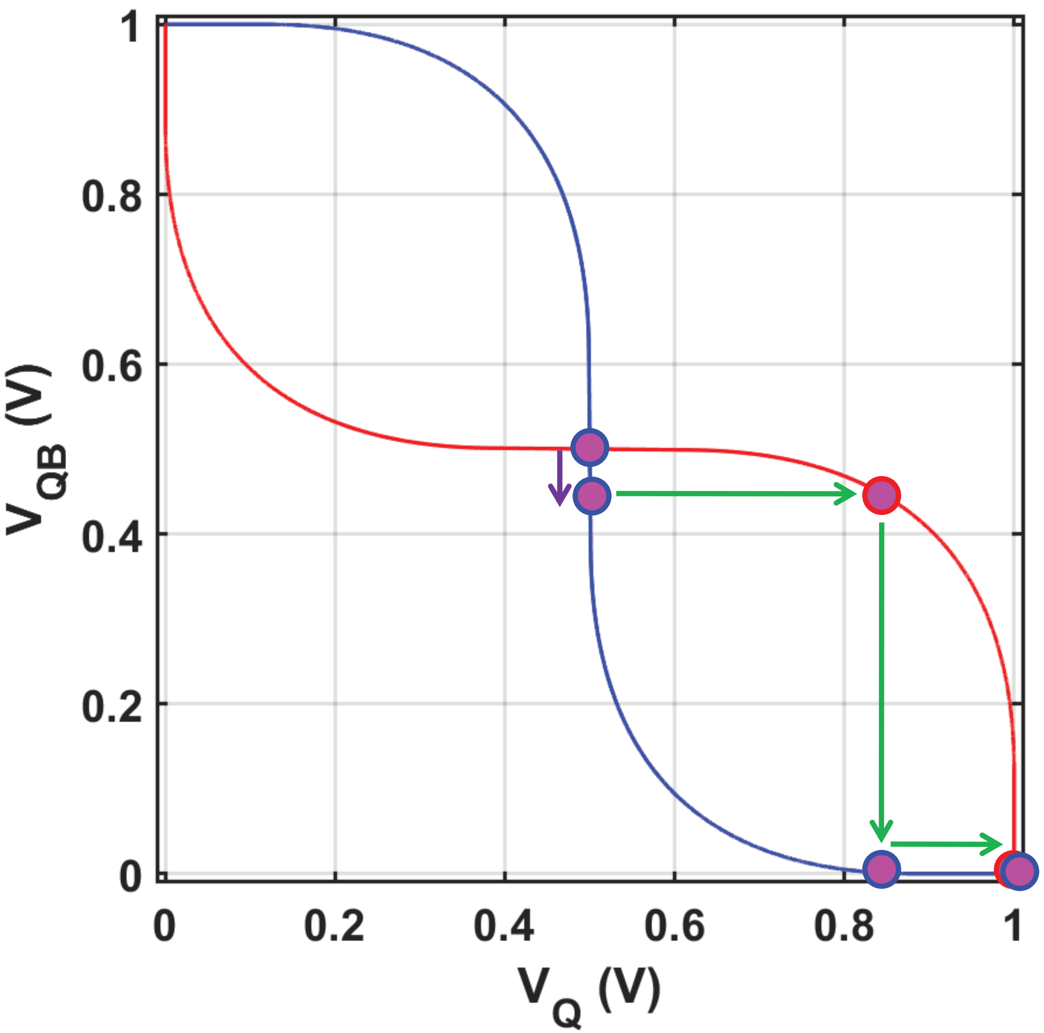
The Bistability Principle



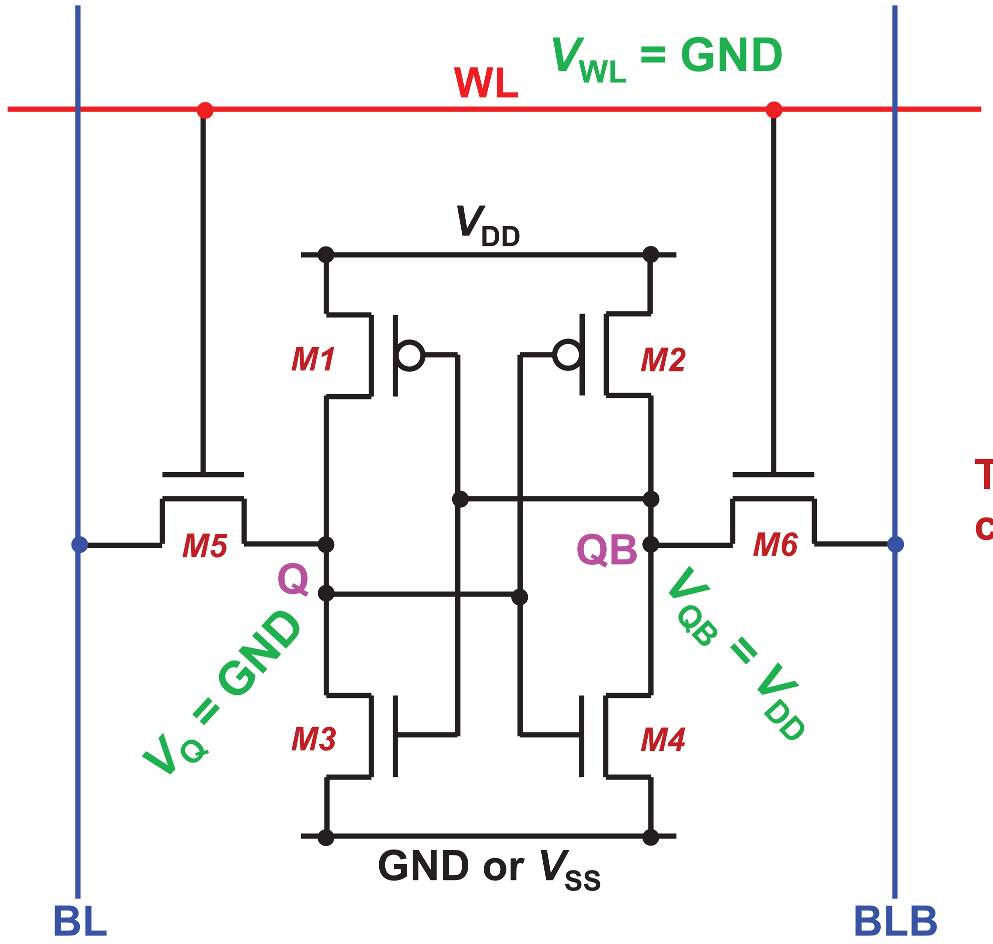


The Bistability Principle



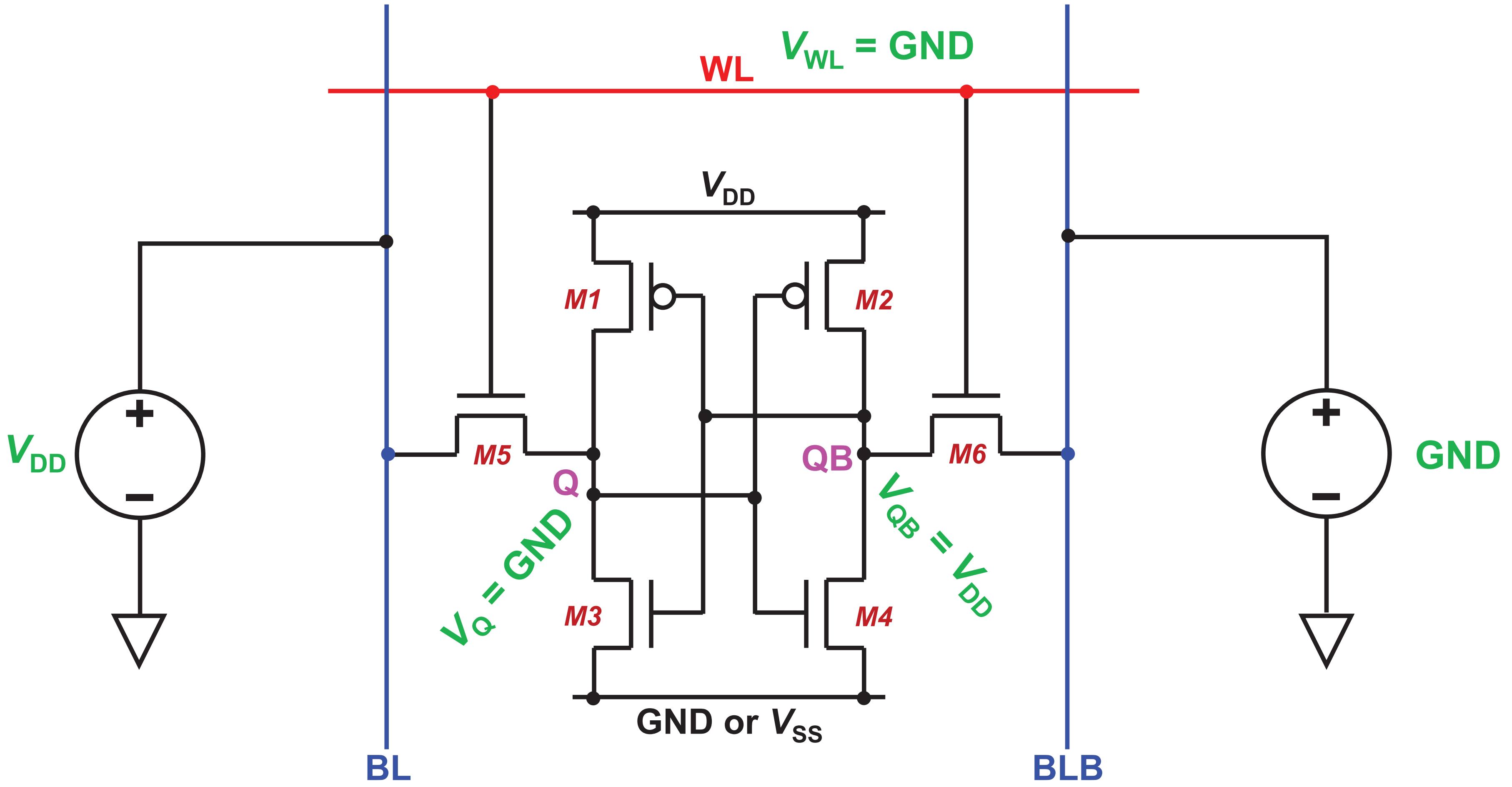


Writing Data into the 6T SRAM

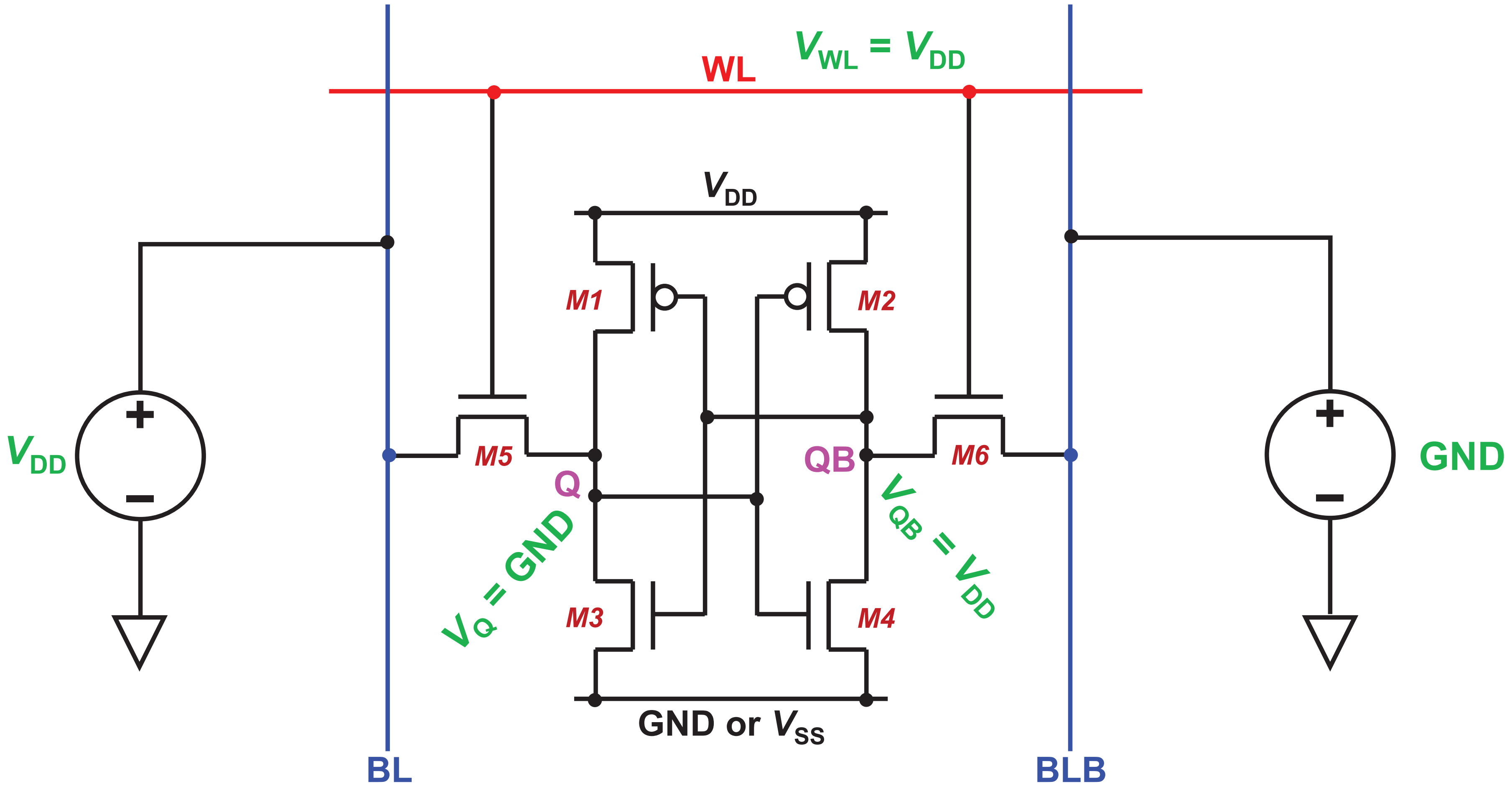


Transistors *M5* and *M6* are called ***access transistors***

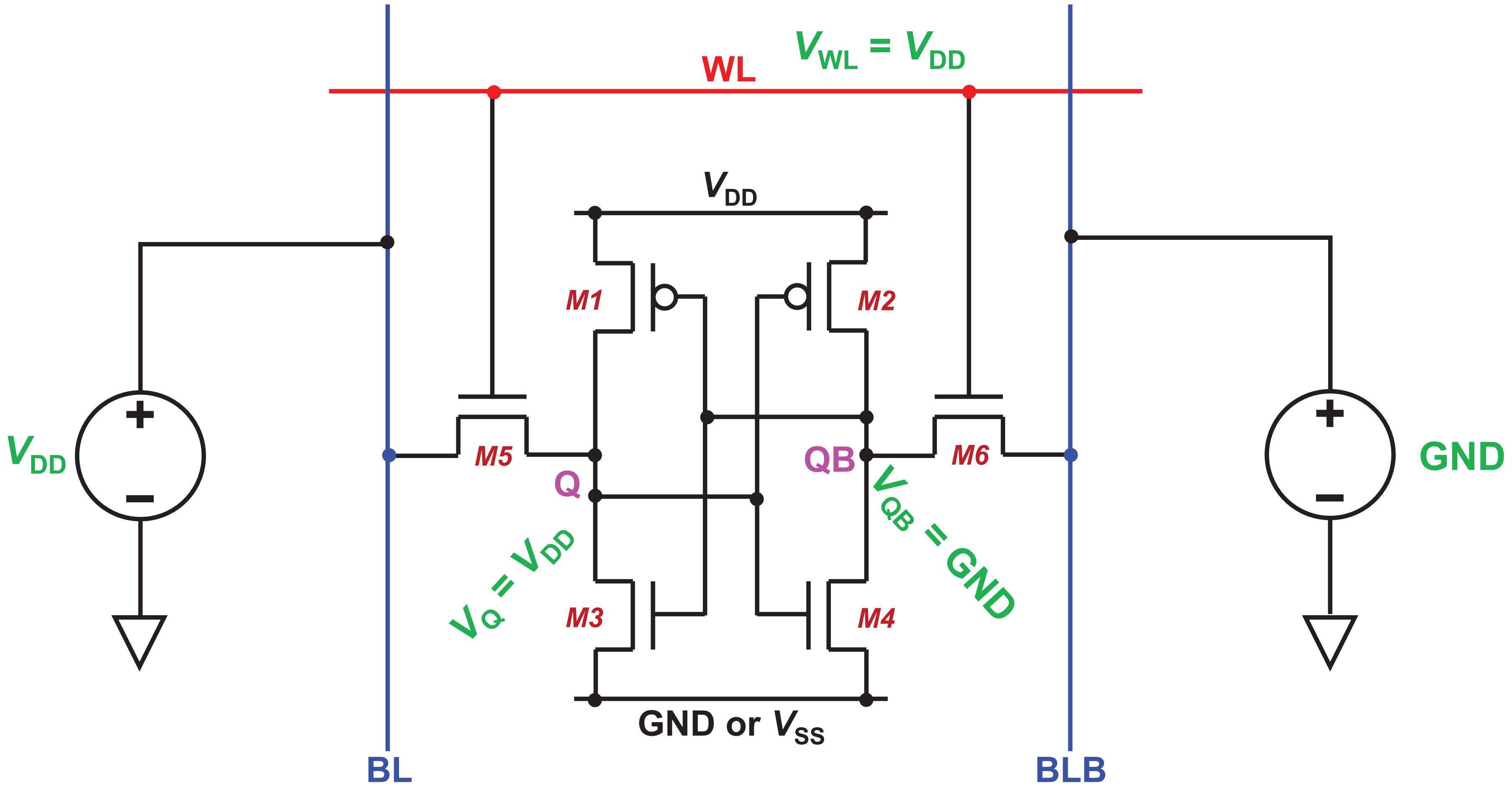
Writing Data into the 6T SRAM – Put data on BL & BLB



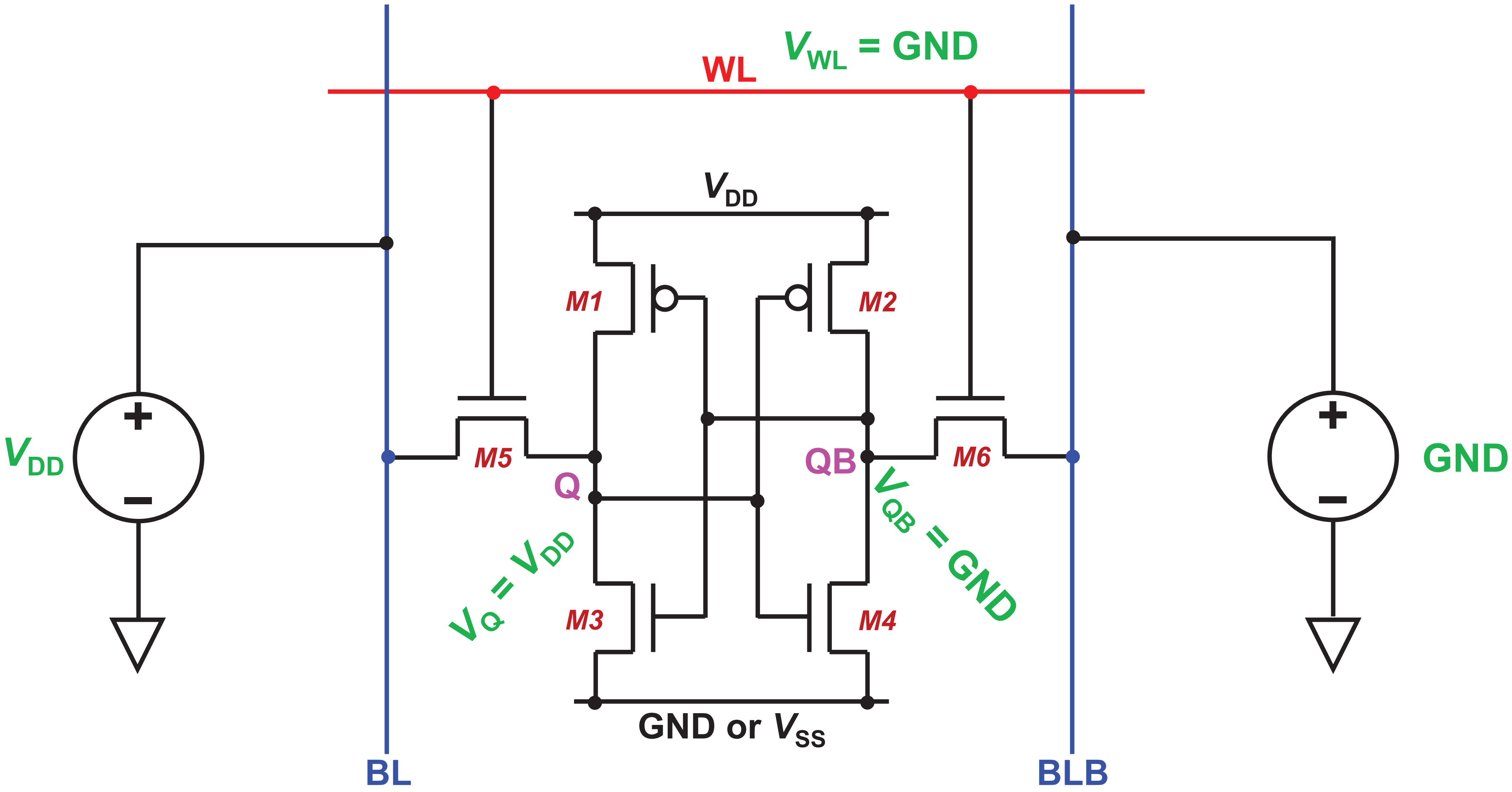
Writing Data into the 6T SRAM – Turn On WL



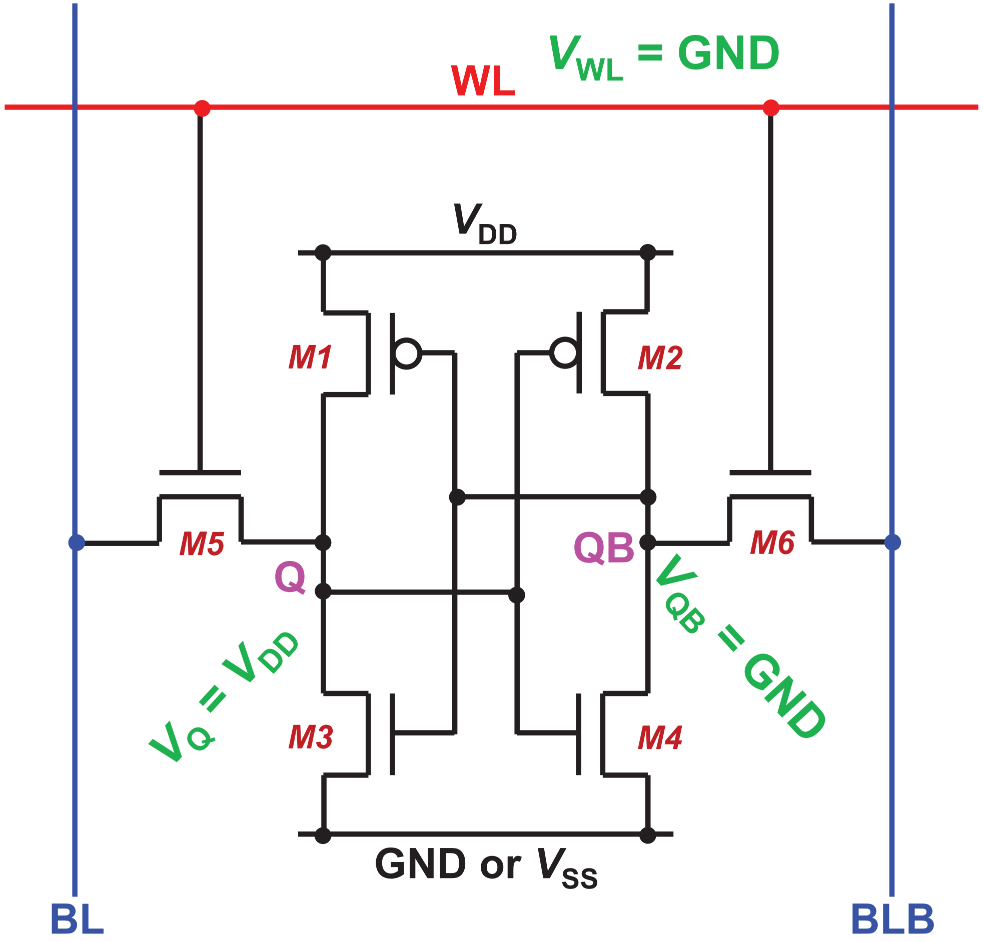
Writing Data into the 6T SRAM – Turn On WL



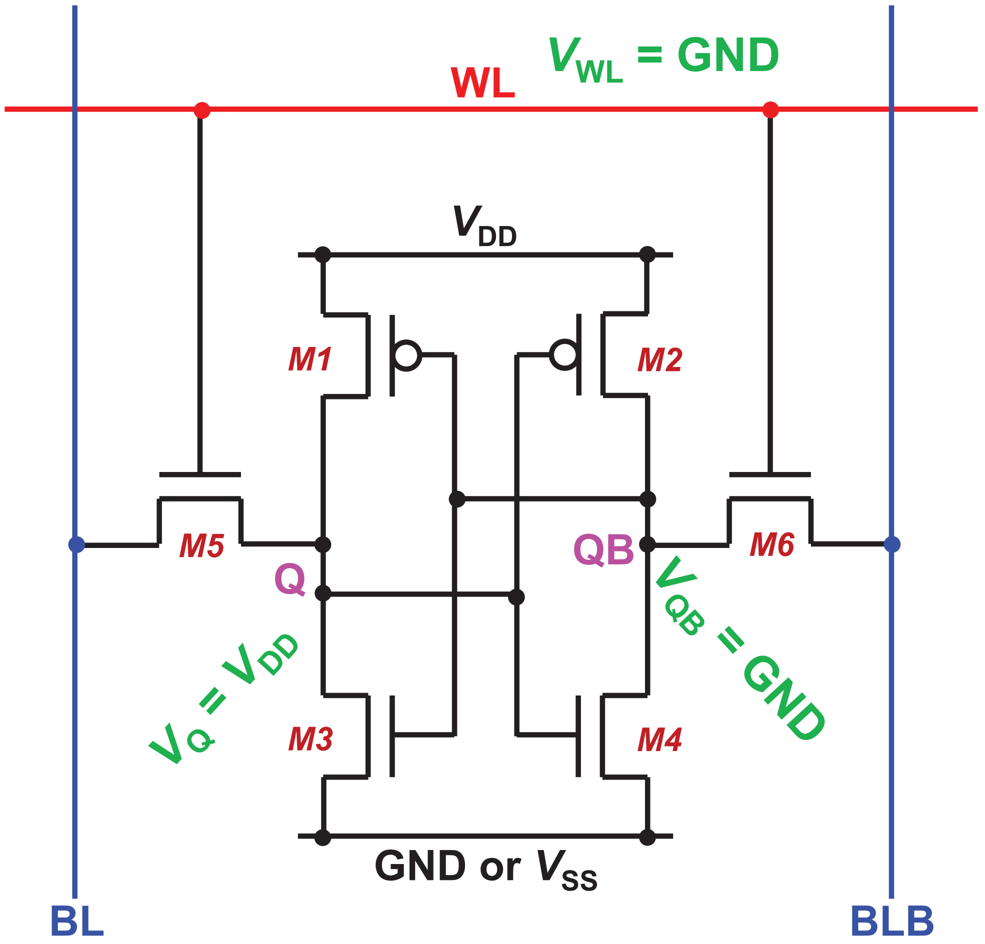
Writing Data into the 6T SRAM – Turn Off WL



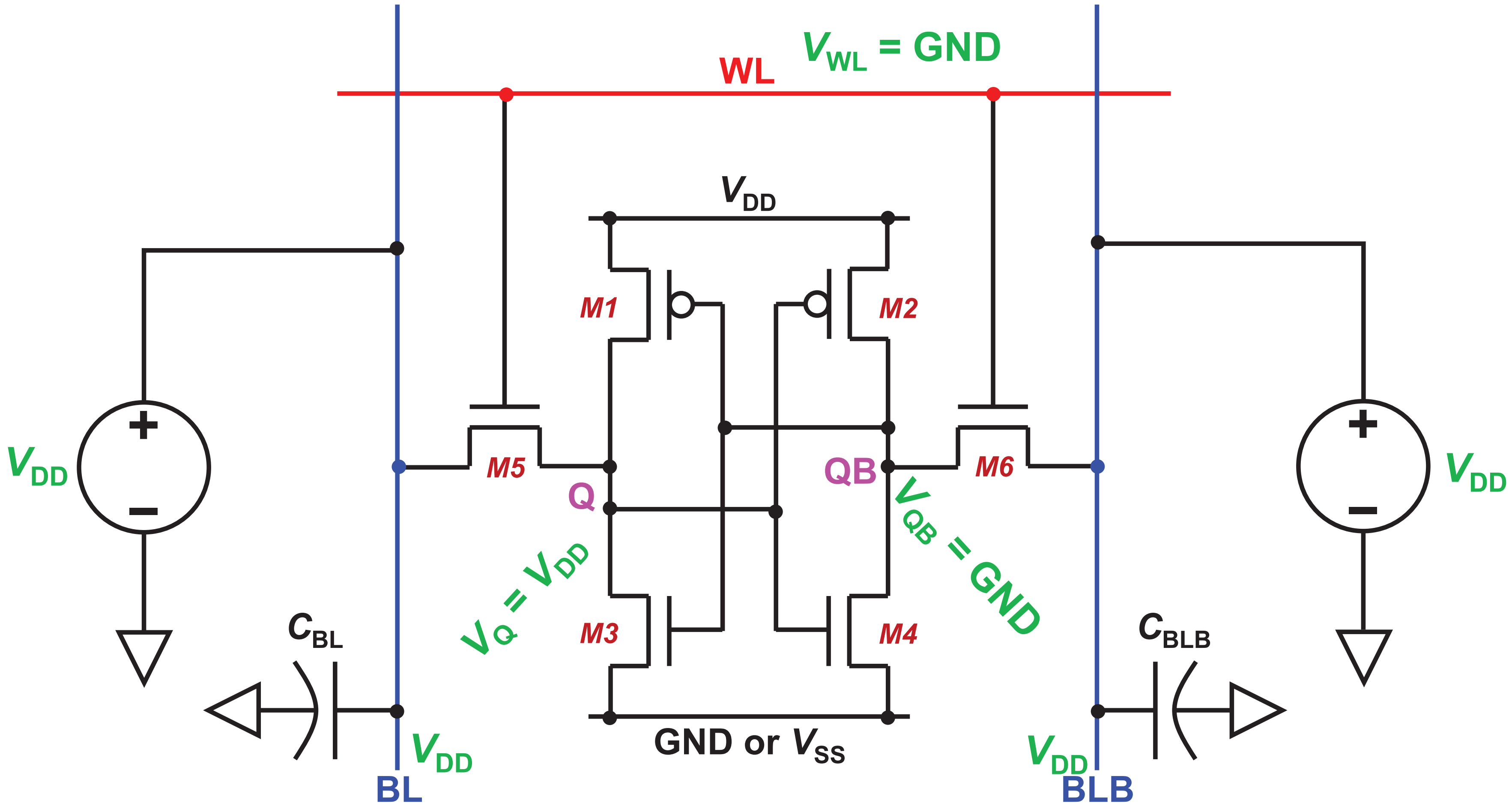
Writing Data into the 6T SRAM – Disconnect BL & BLB



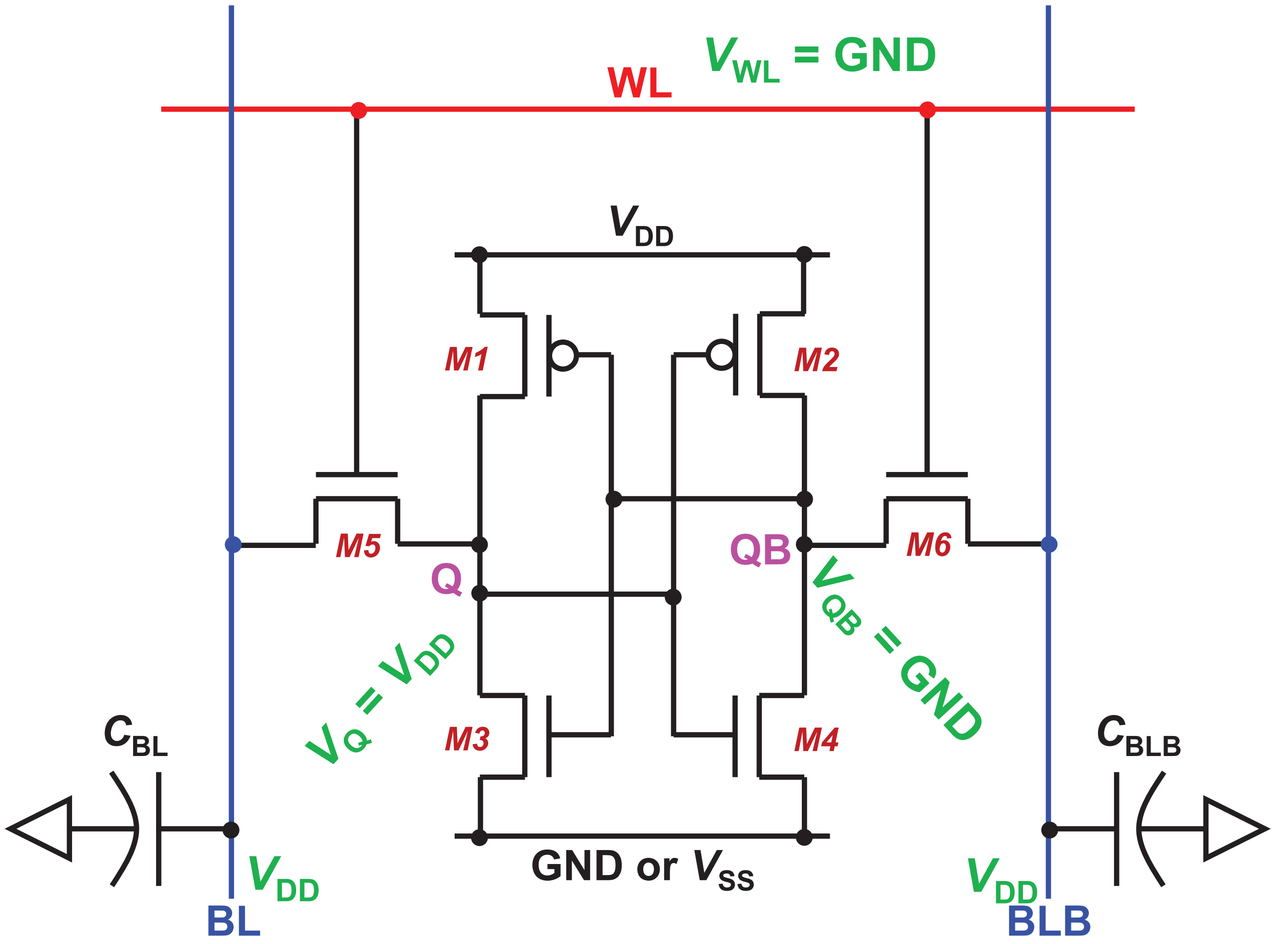
Reading Data from the 6T SRAM



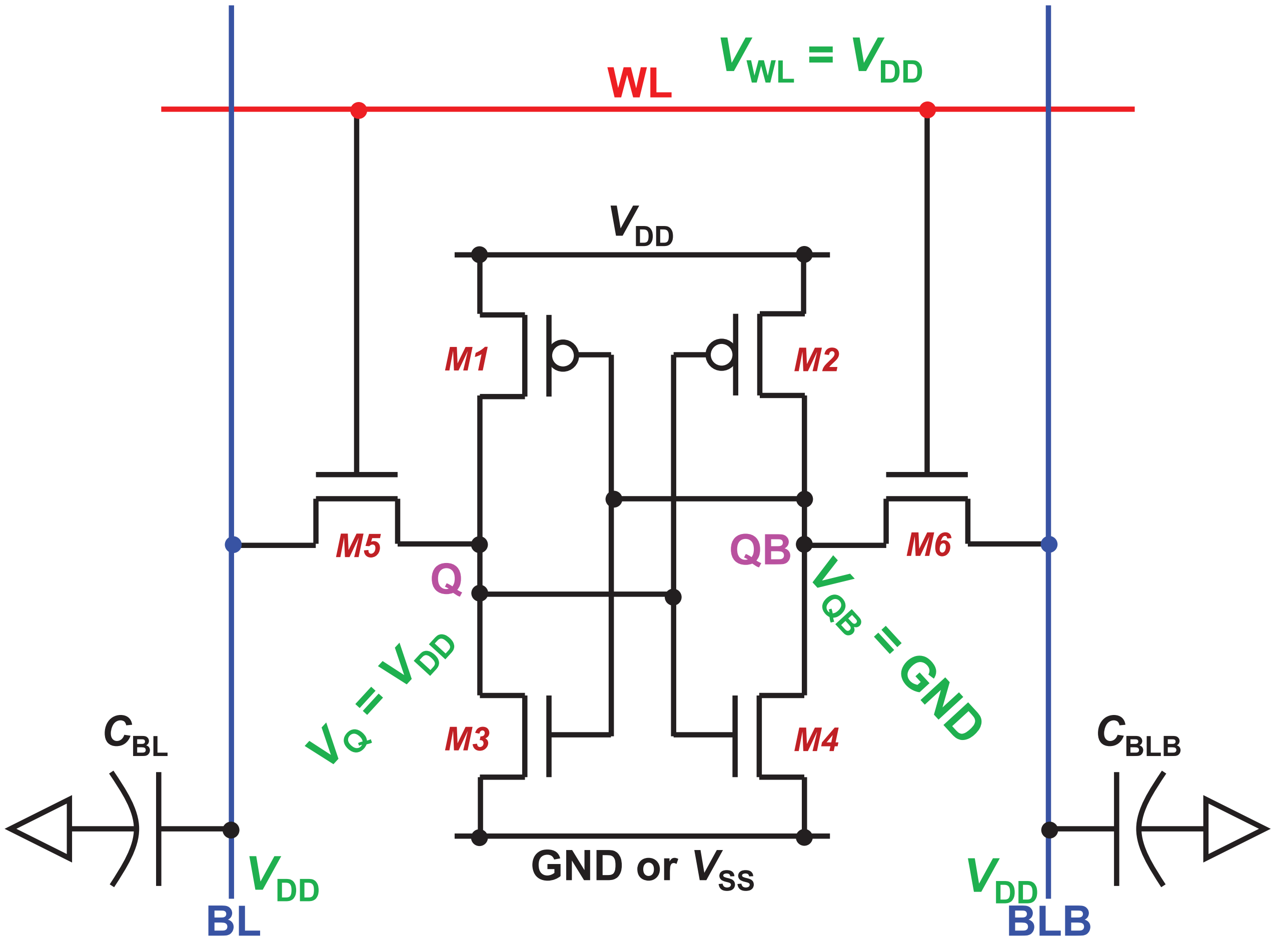
Reading Data from the 6T SRAM – Precharge BL & BLB



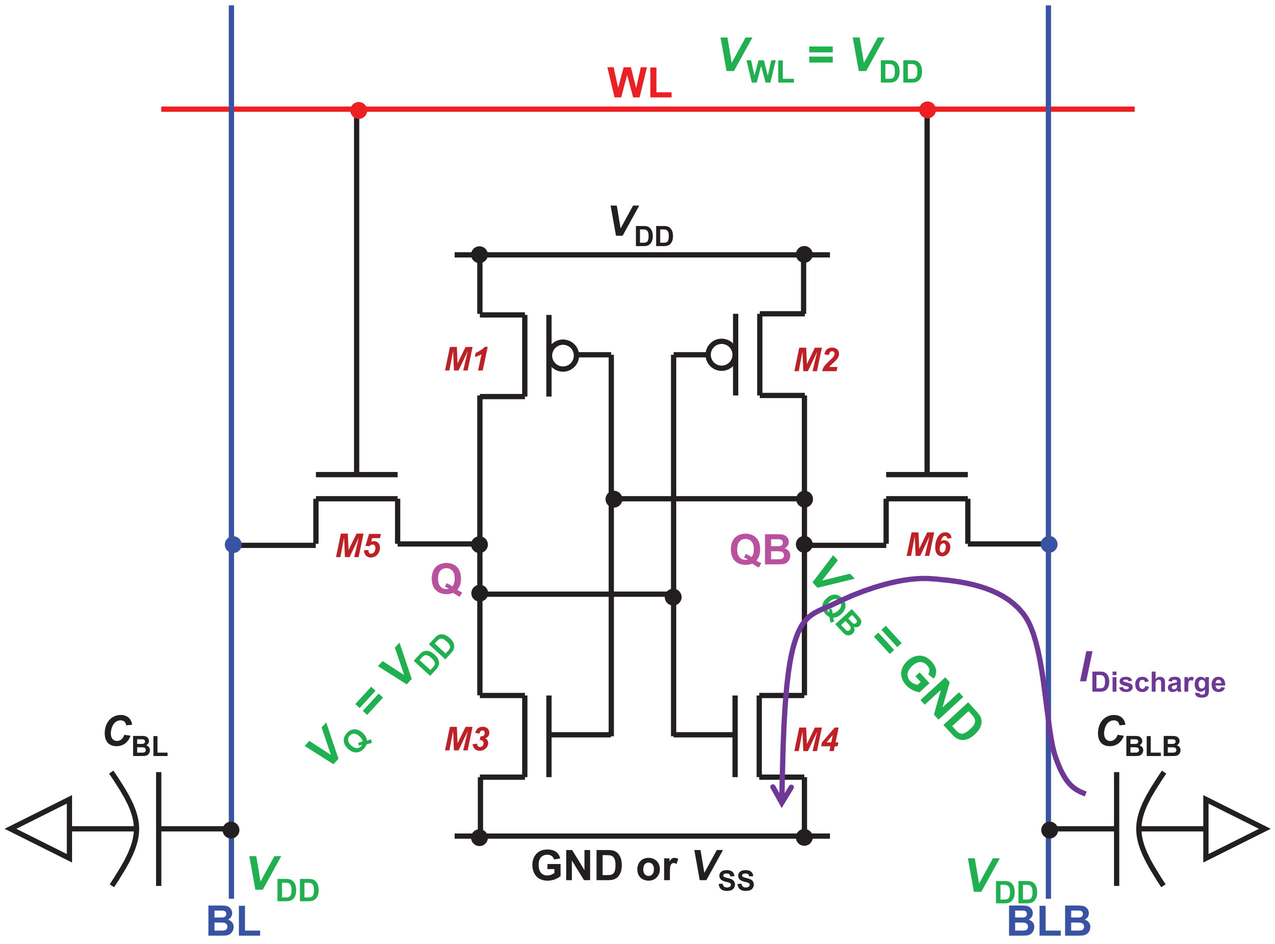
Reading Data from the 6T SRAM – Precharge BL & BLB



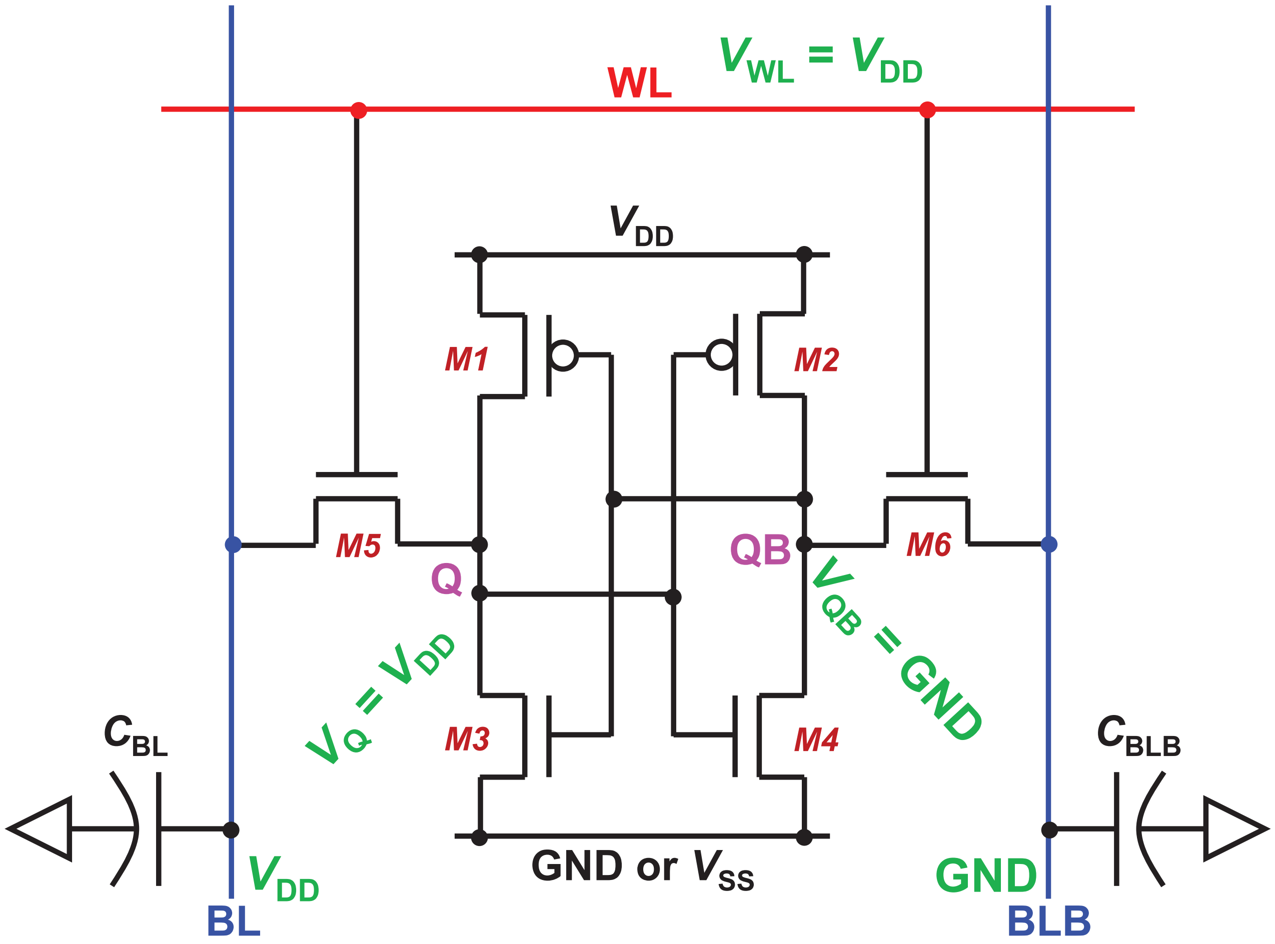
Reading Data from the 6T SRAM – Turn On WL



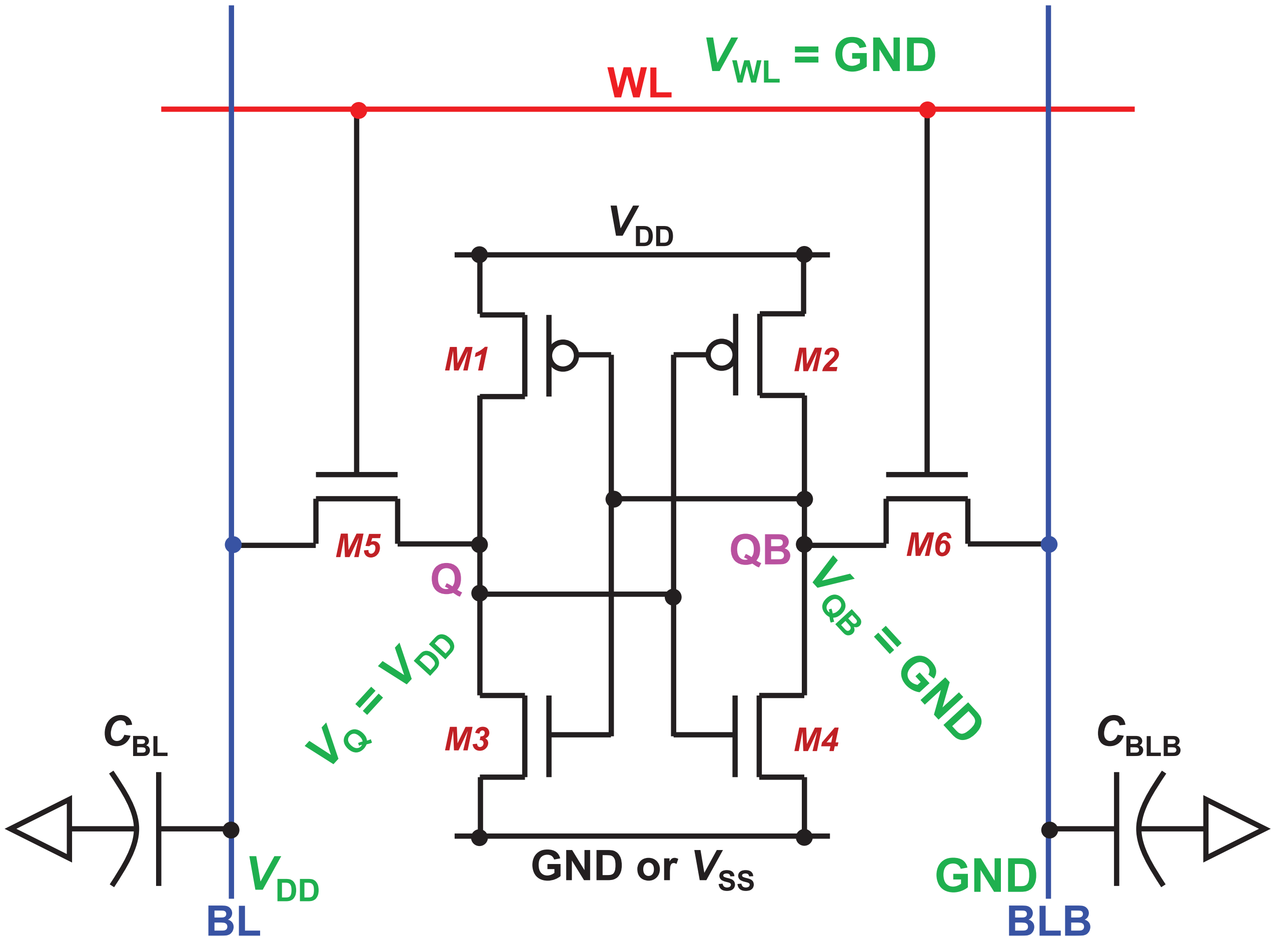
Reading Data from the 6T SRAM – Discharge BL/BLB



Reading Data from the 6T SRAM – Sense BL & BLB



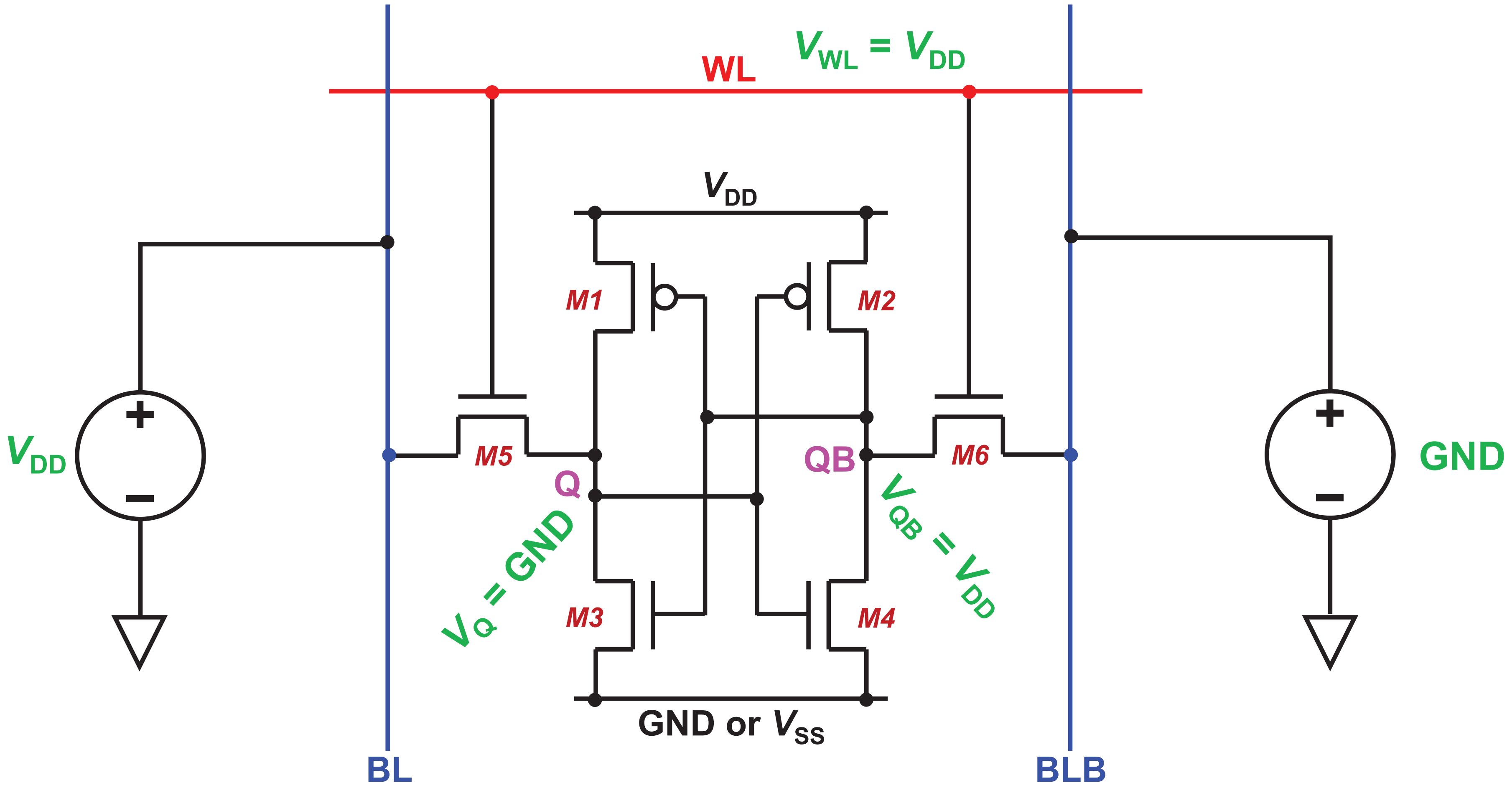
Reading Data from the 6T SRAM – Turn Off WL



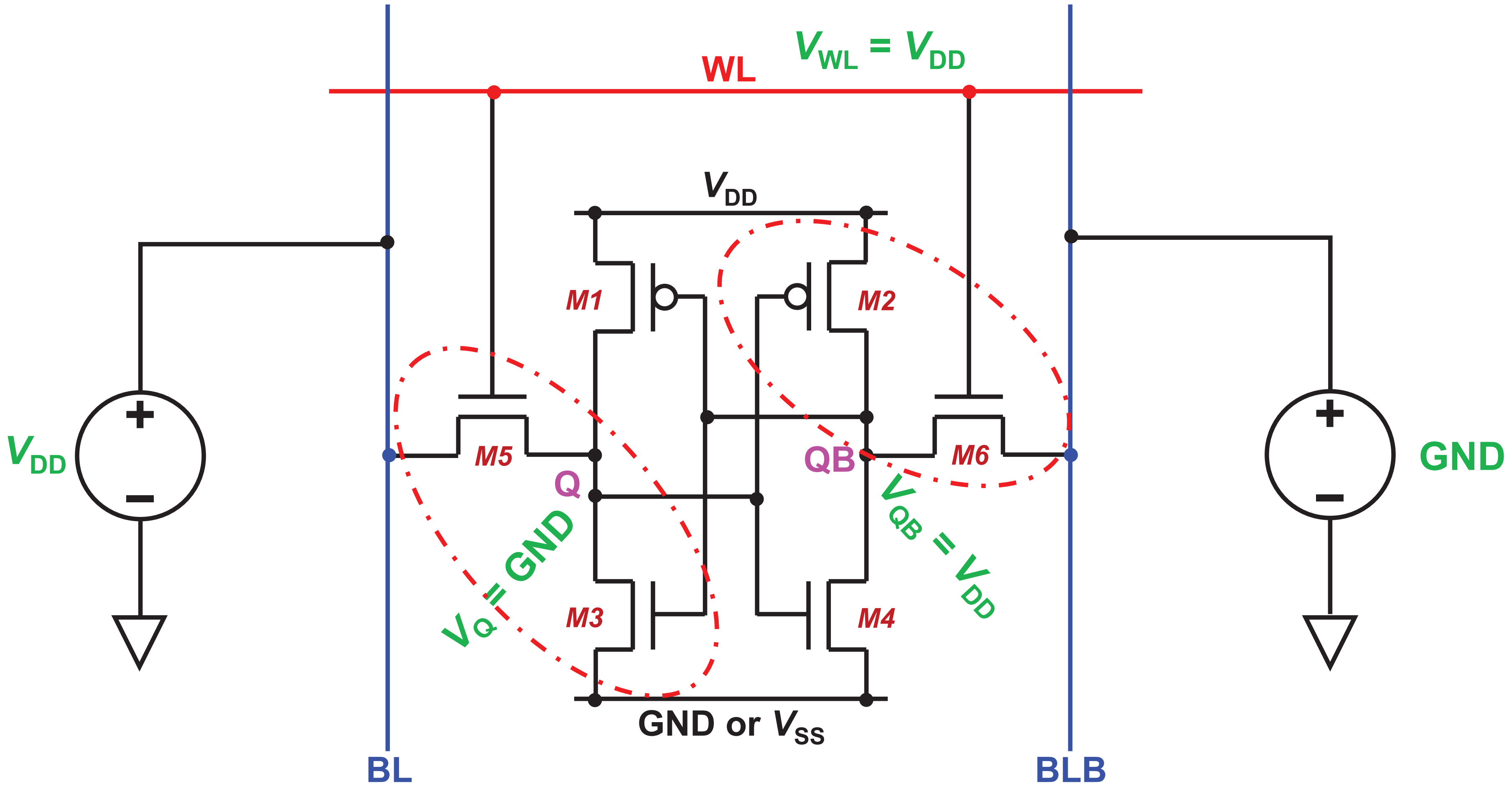
Week 5-4

Sizing Transistors in SRAM Bit-cell

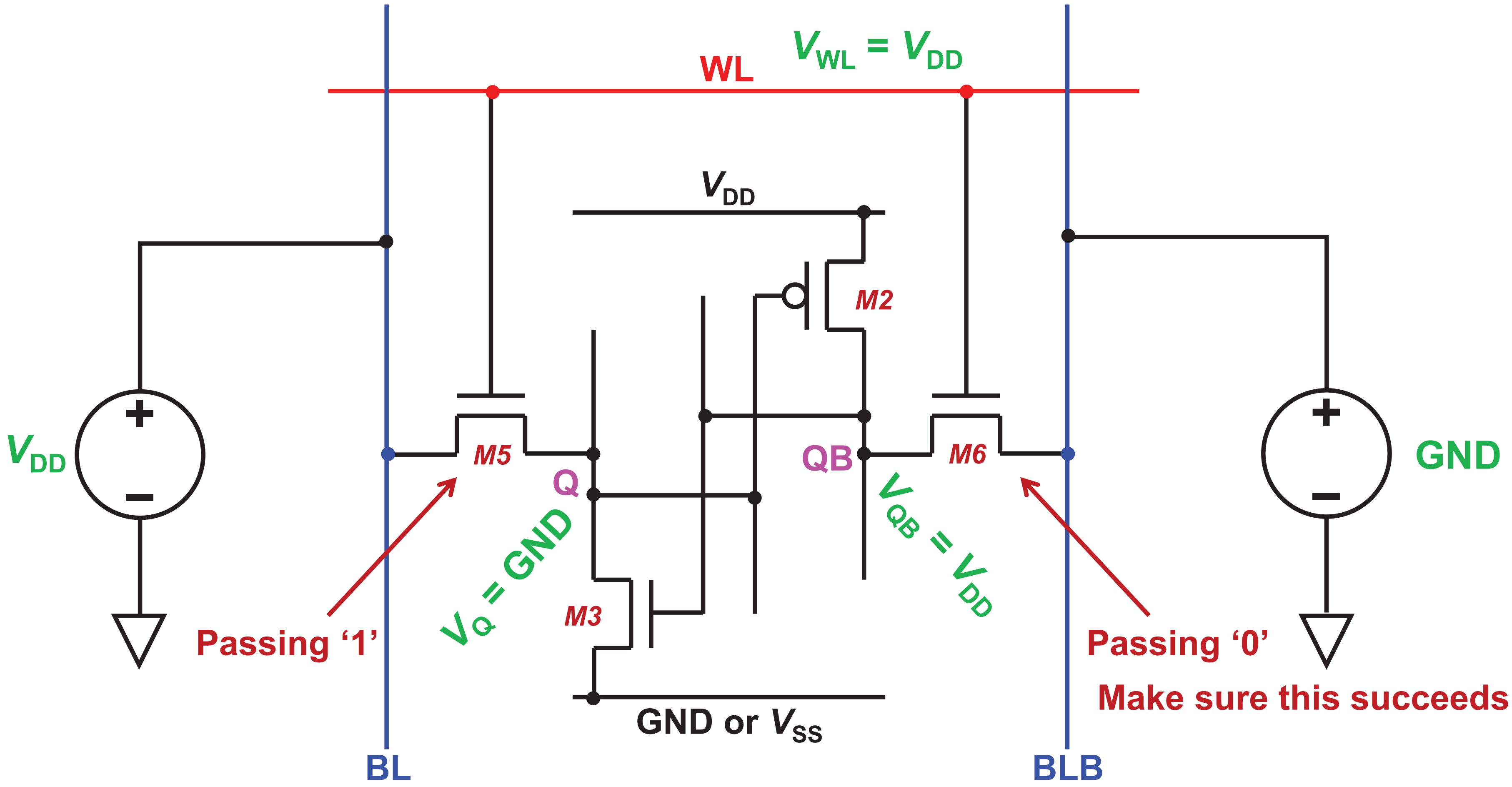
Writing Data into the 6T SRAM



Writing Data into the 6T SRAM

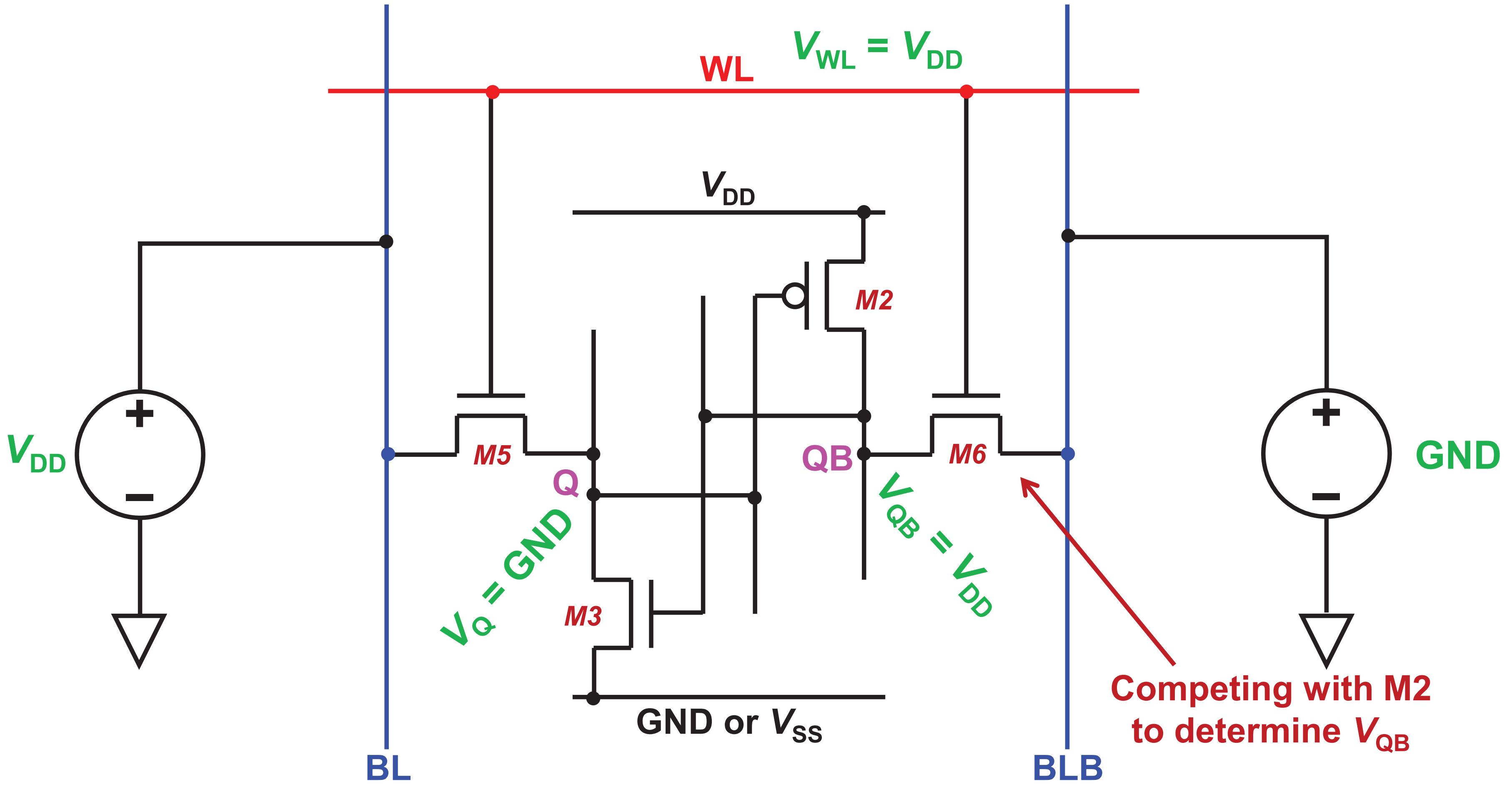


Writing Data into the 6T SRAM – Half-cell Analysis



Make sure this succeeds

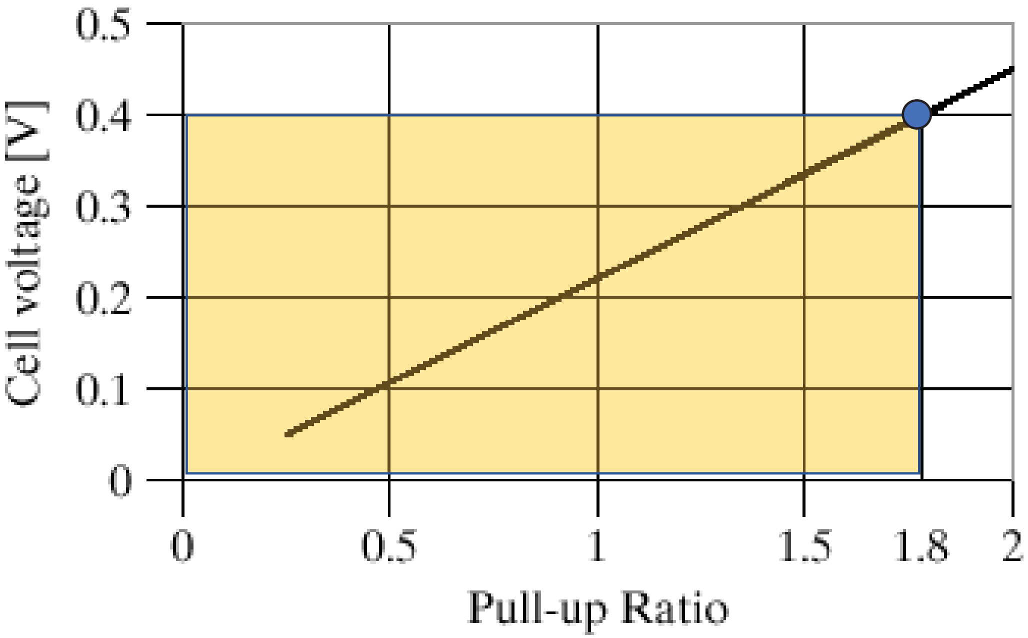
Writing Data into the 6T SRAM – Half-cell Analysis



Competing with M2 to determine V_{QB}

Pull-up Ratio **[(W_2/L_2) / (W_6/L_6)] determines if V_{QB} will fall below V_M of inverter (lower is better)**

Impact of Pull-up Ratio on Write Operation of 6T SRAM



$$k_{n,M6}\left((V_{DD}-V_{TN})V_{QB}-\frac{V_{QB}^2}{2}\right)=$$

$$k_{p,M2}\left((V_{TP}-V_{DD})V_{DS,SATp}-\frac{V_{DS,SATp}^2}{2}\right)$$

$$V_{QB}=V_{DD}-V_{TN}$$

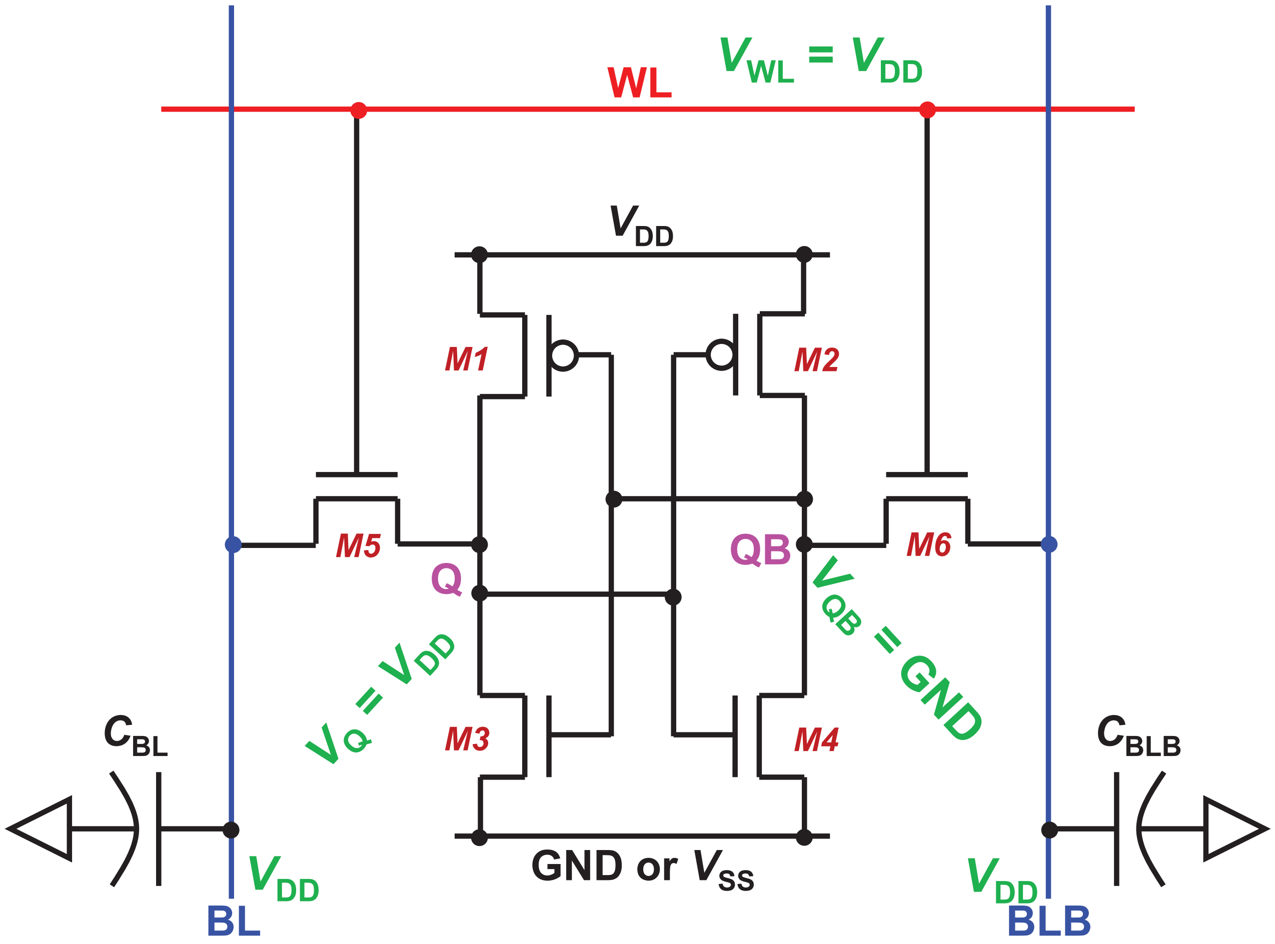
$$-\sqrt{(V_{DD}-V_{TN})^2-2\frac{\mu_p}{\mu_n}\left((V_{TP}-V_{DD})V_{DS,SATp}-\frac{V_{DS,SATp}^2}{2}\right)PR}$$

$$PR=\frac{W_2/L_2}{W_6/L_6}$$

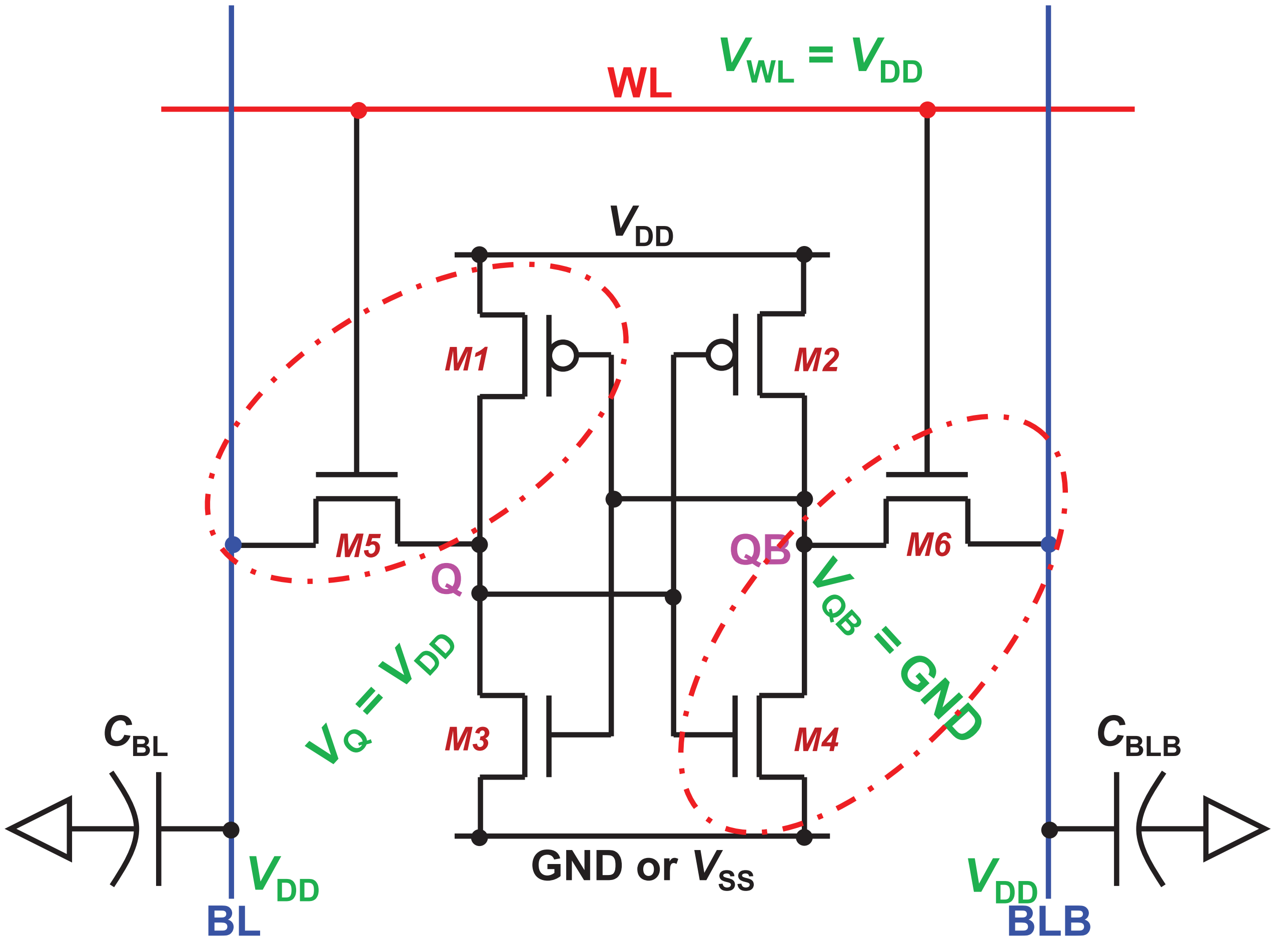
PR < 1.8 to ensure write of “0”
(otherwise, M3 stays on and we cannot write “0” into QB)

Want smaller aspect ratio for *M1* & *M2*
Want larger aspect ratio for *M5* & *M6*

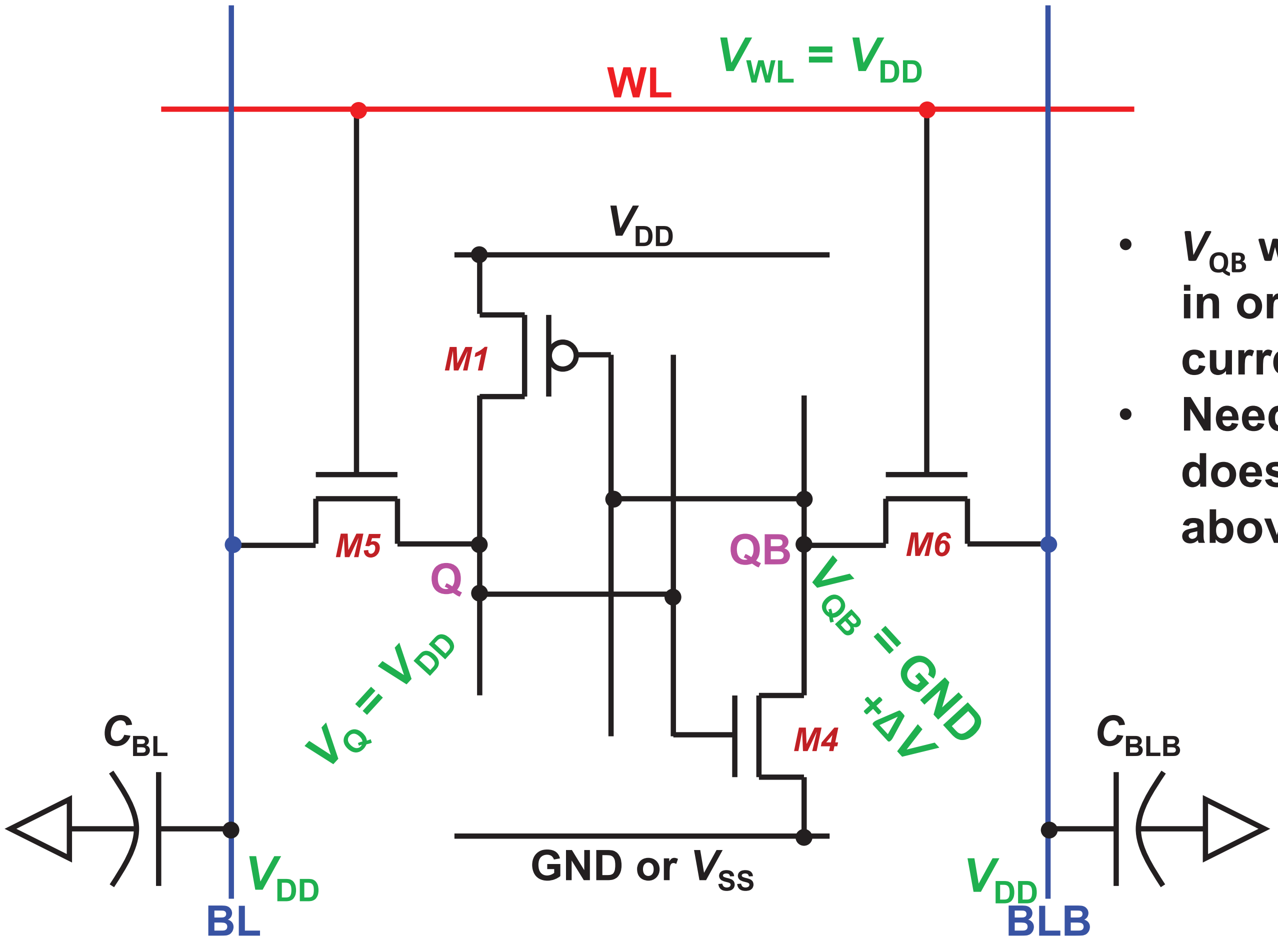
Reading Data from the 6T SRAM



Reading Data from the 6T SRAM



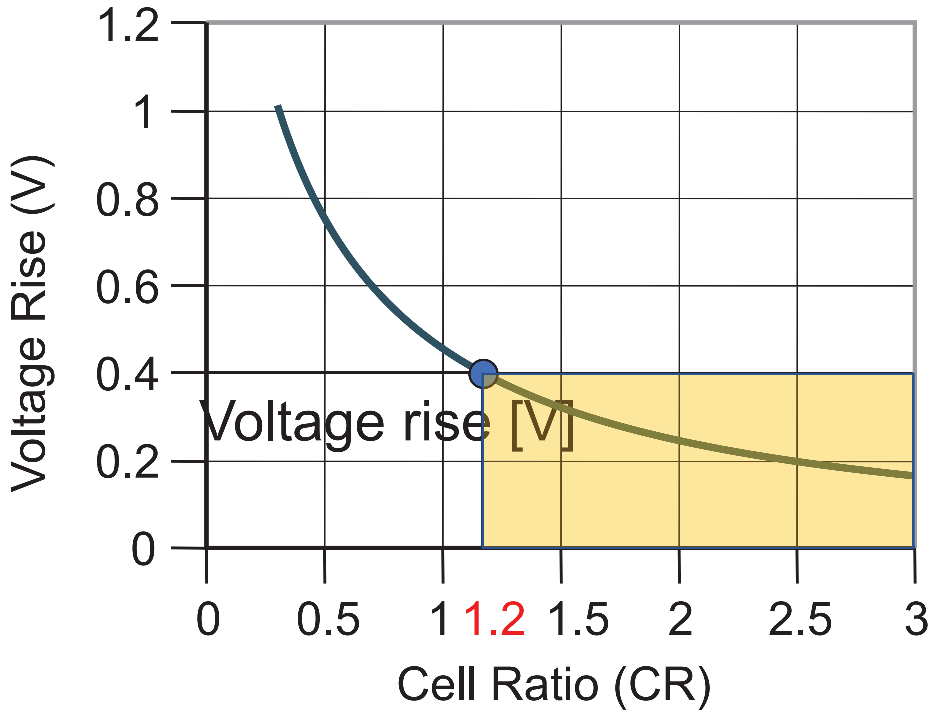
Reading Data from the 6T SRAM – Half-cell Analysis



- V_{QB} will increase by +ΔV in order for *M4* to pass current
- Need to ensure V_{QB} does not increase above V_M of inverter

Cell Ratio **[(W_4/L_4) / (W_6/L_6)] determines +Δ*V* on V_{QB} (larger is better)**

Impact of Cell Ratio on Read Operation of 6T SRAM



$$k_{n,M6}\left((V_{DD}-\Delta V-V_{TN})V_{DS,SATn}-\frac{V_{DS,SATn}^2}{2}\right)=k_{n,M6}\left((V_{DD}-V_{TN})\Delta V-\frac{\Delta V^2}{2}\right)$$

$$\Delta V=\frac{V_{DS,SATn}}{CR}+CR(V_{DD}-V_{TN})-\frac{\sqrt{V_{DS,SATn}^2(1+CR)+CR^2(V_{DD}-V_{TN})^2}}{CR}$$

$$CR=\frac{W_4/L_4}{W_6/L_6}$$

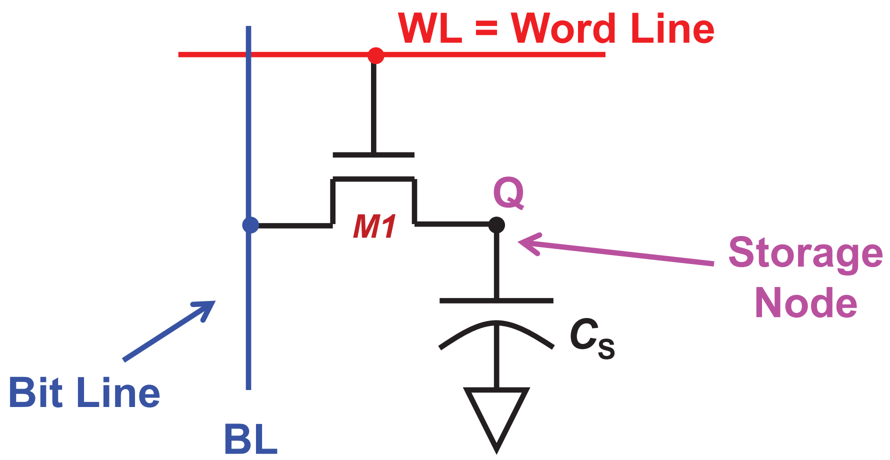
CR > 1.2 to ensure read stability
(otherwise, bit flip occurs while read operation)

Want larger aspect ratio for *M3* & *M4*
Want smaller aspect ratio for *M5* & *M6*

Week 5-5

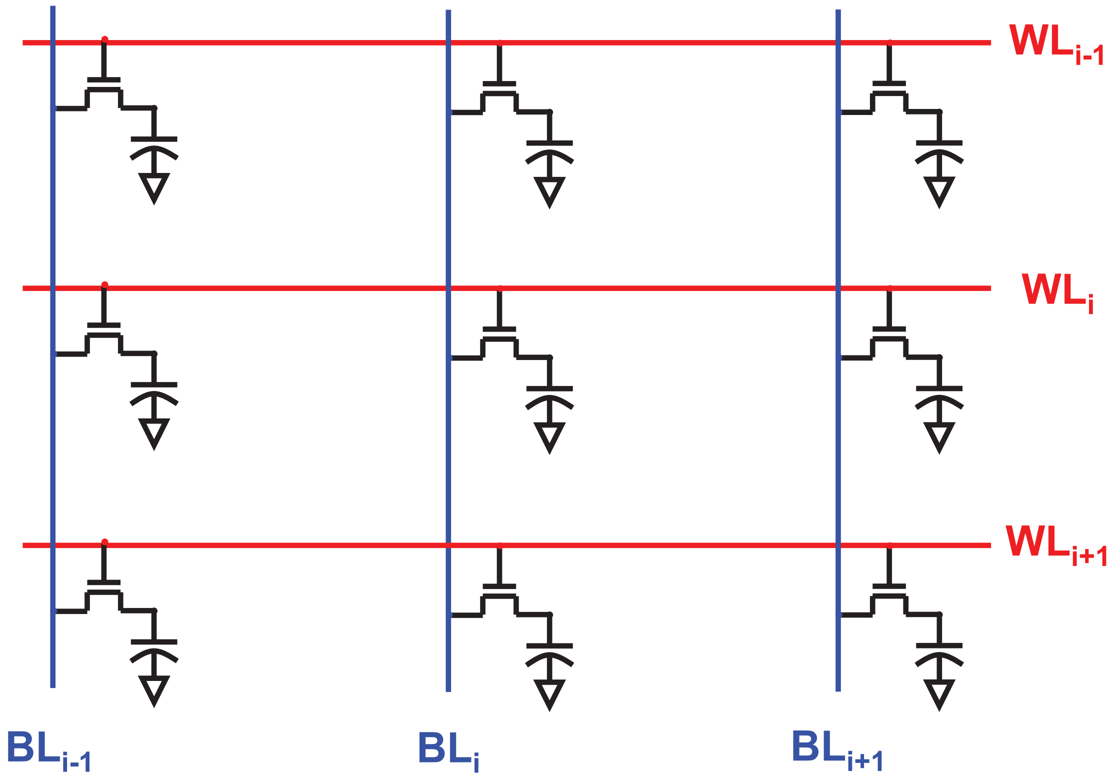
Introduction to the 1-Transistor Dynamic RAM (1T DRAM)

The 1-Transistor Dynamic RAM (1T DRAM) Bitcell



- Unlike SRAM
  - The 1T DRAM cell has only one transistor and one capacitor
  - Single-ended
- Simple structure and compact in size (good for density)
- Data is stored as the amount of charge in the storage capacitor
- Node Q is in high-***Z*** state when $\boldsymbol{V_{WL}}$ = **GND**
  - Stored charge will leak due to transistor leakage
  - Requires dynamic refresh of charge stored to retain data (hence, *dynamic* RAM)

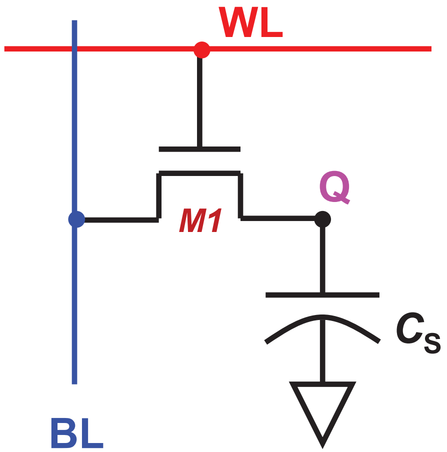
Array of 1T DRAM Bitcells (Bitcell Array)



The 1T DRAM Bitcell Structure

V_{WL} = GND during hold

$V_{WL} = V_{DD}$ during access

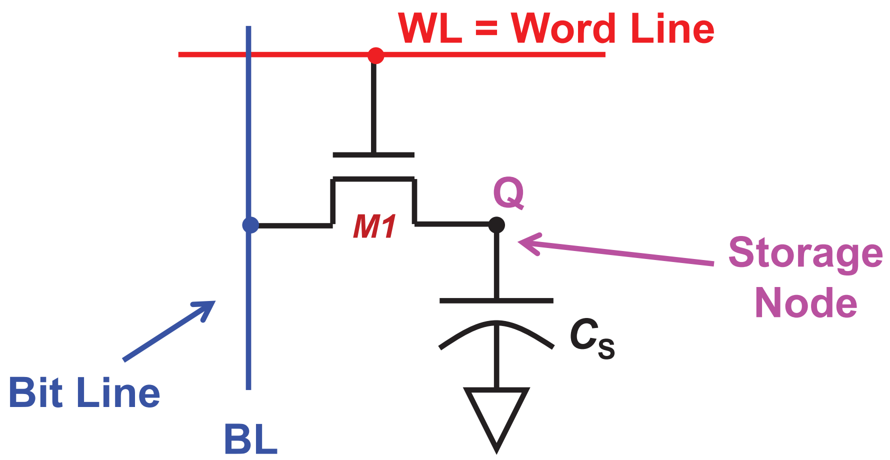


- The bitcell is *accessed* to read data from or write data into the bitcell
  - Electrically connect BL to storage node Q
- Otherwise, bitcell is in hold or hold mode and retains its stored data
  - Affected by leakage
  - Higher C_S allows more charge to be stored
    - Longer data retention

Week 5-6

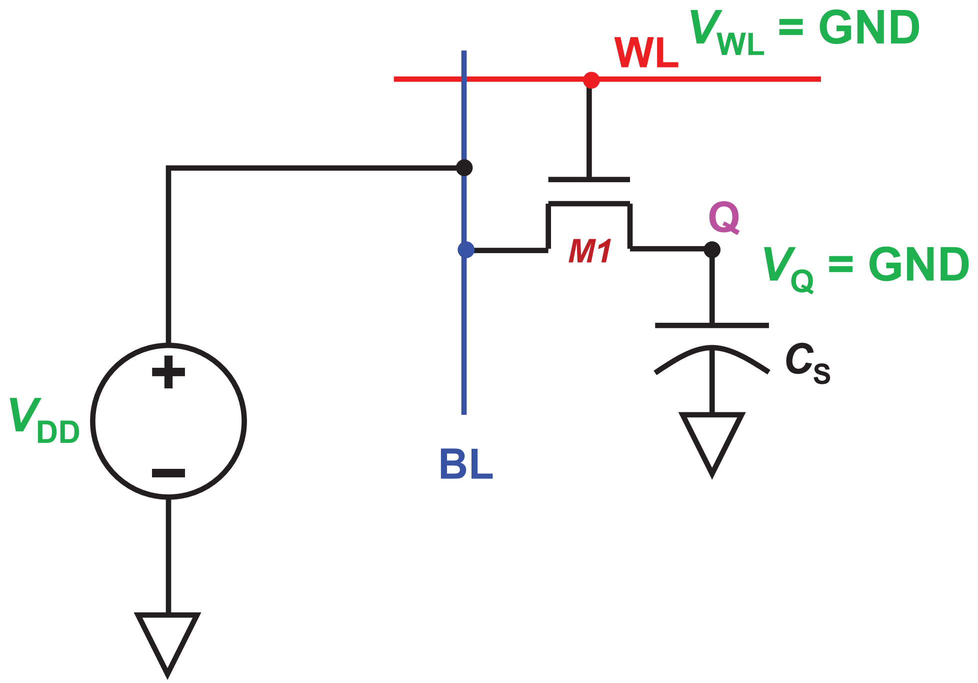
DRAM Read and Write Operations

The 1T DRAM

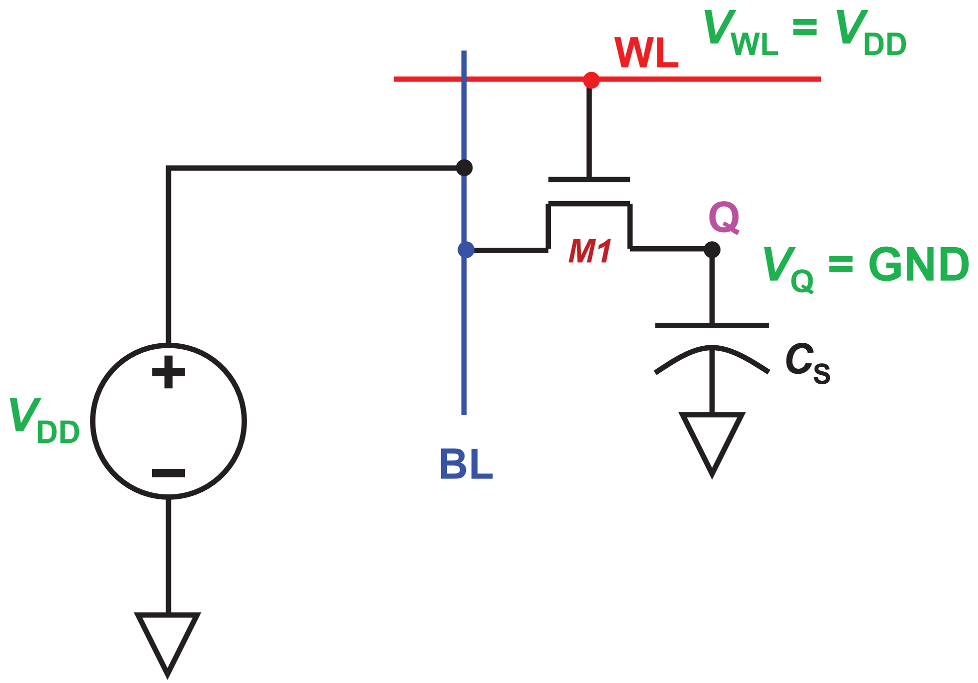


- Unlike SRAM
  - The 1T DRAM cell has only one transistor and one capacitor
  - Single-ended
- Simple structure and compact in size (good for density)
- Data is stored as the amount of charge in the storage capacitor
- Node Q is in high-***Z*** state when $\boldsymbol{V_{WL}}$ = **GND**
  - Stored charge will leak due to transistor leakage
  - Requires dynamic refresh of charge stored to retain data (hence, *dynamic* RAM)

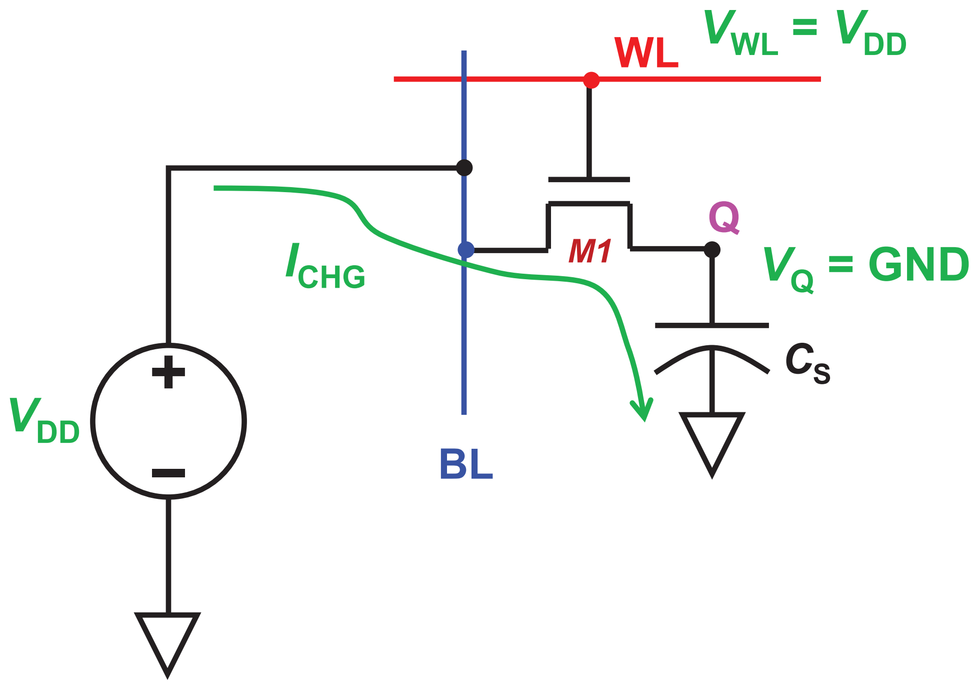
Writing Data into the 1T DRAM – Put data on BL



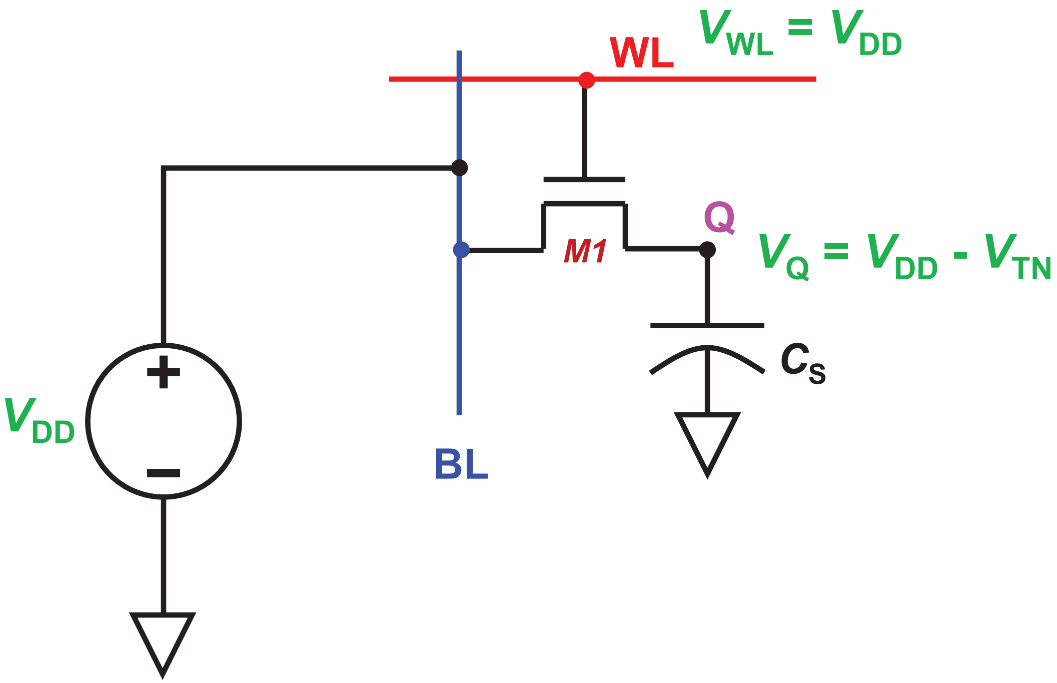
Writing Data into the 1T DRAM – Turn on WL



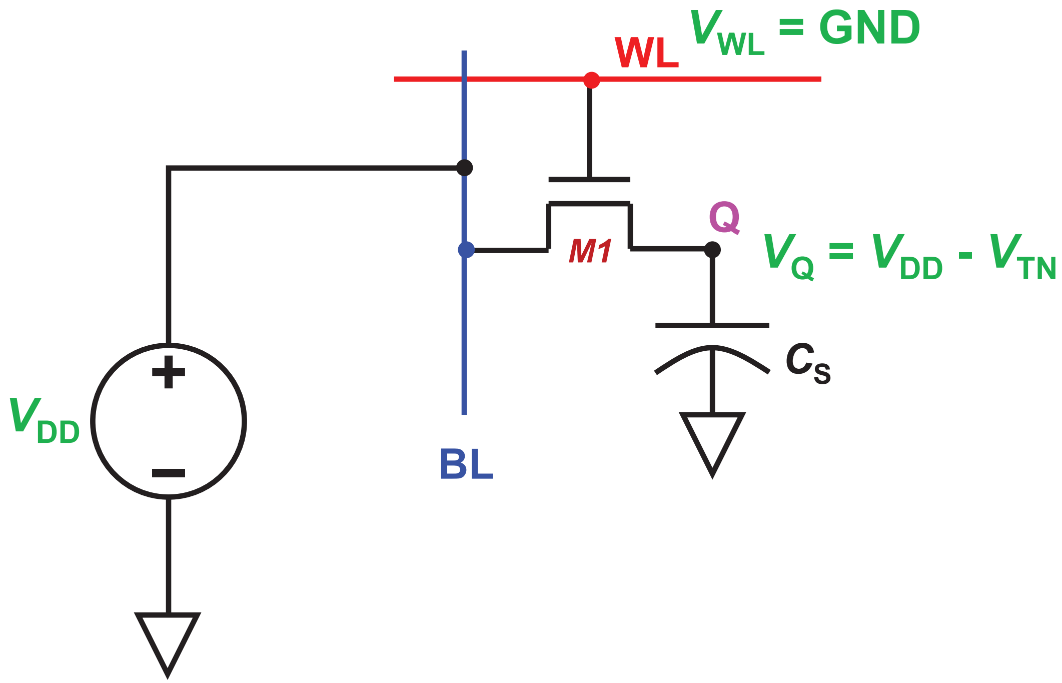
Writing Data into the 1T DRAM – Turn on WL



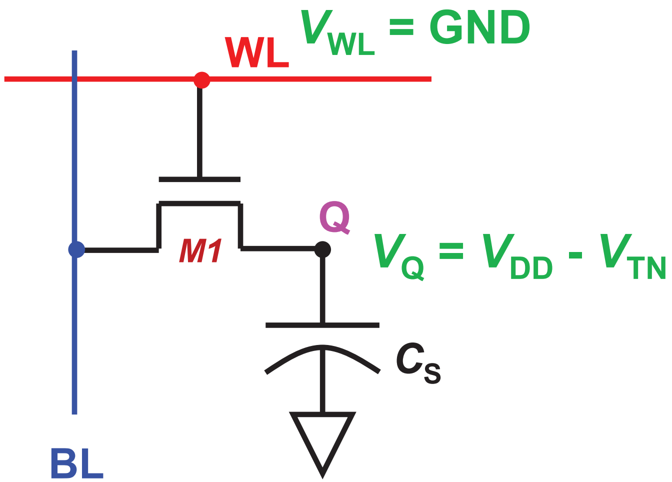
Writing Data into the 1T DRAM – Turn on WL



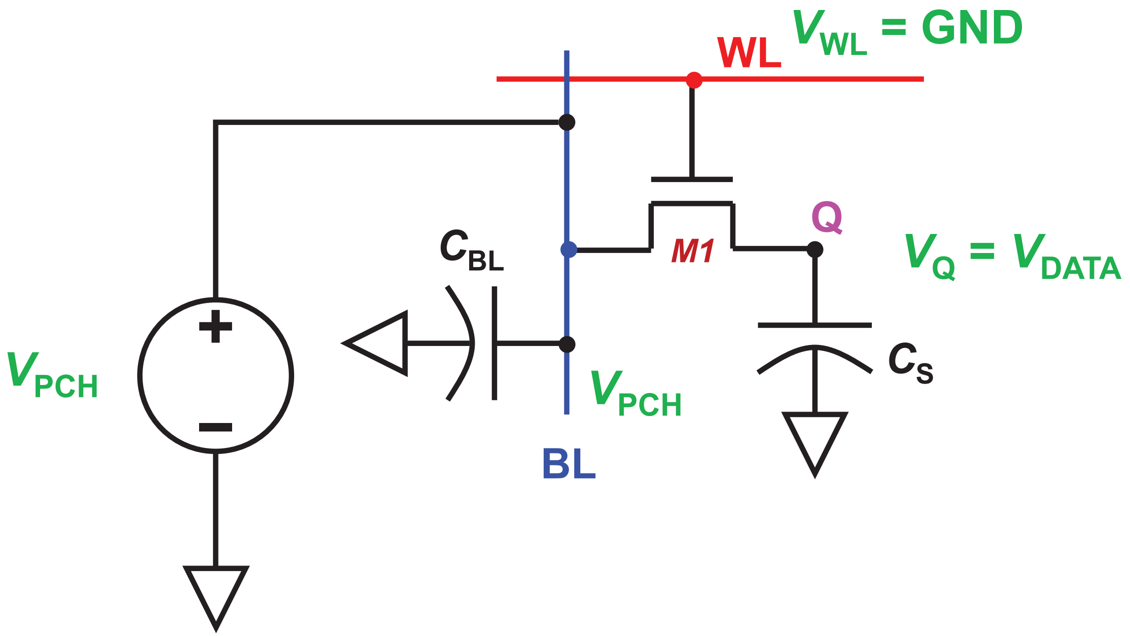
Writing Data into the 1T DRAM – Turn off WL



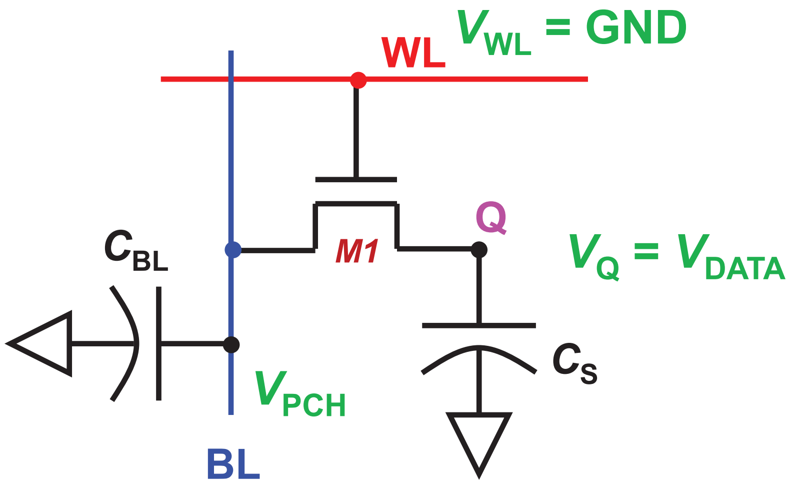
Writing Data into the 1T DRAM – Disconnect BL



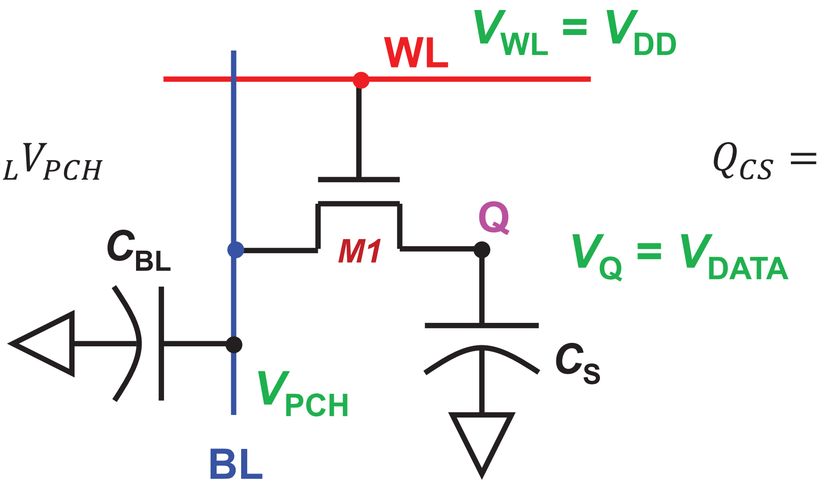
Reading Data from the 1T DRAM – Precharge BL



Reading Data from the 1T DRAM – Precharge BL



Reading Data from the 1T DRAM – Turn on WL



$Q_{BL} = C_{BL}V_{PCH}$

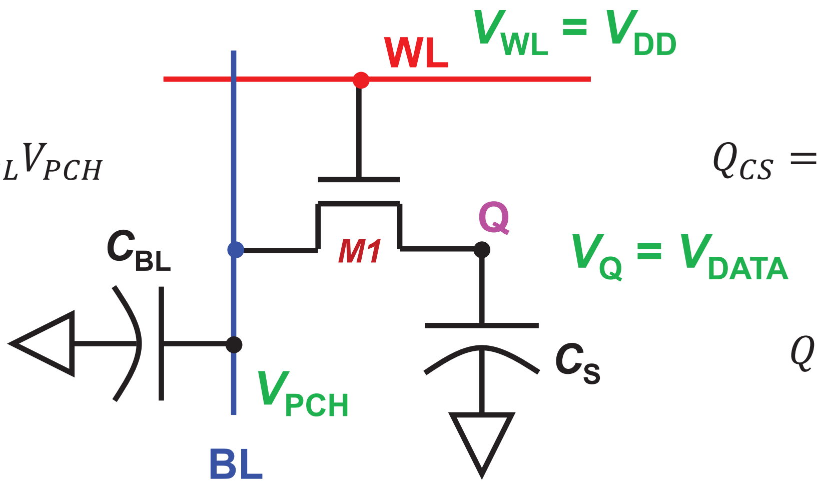
$Q_{CS} = C_S V_{DATA}$

$Q_{TOTAL} = Q_{BL} + Q_{CS}$

- M1 electrically shorts C_{BL} and C_S together
    - Charge sharing between C_{BL} and C_S
    - ΔV_Q and ΔV_{BL} depends on size of C_{BL} relative C_S, and whether M1 gets turned off or not

Reading Data from the 1T DRAM – Turn on WL

$$Q_{BL} = C_{BL}V_{PCH}$$



$$Q_{CS} = C_S V_{DATA}$$

$$Q_{TOTAL} = Q_{BL} + Q_{CS}$$

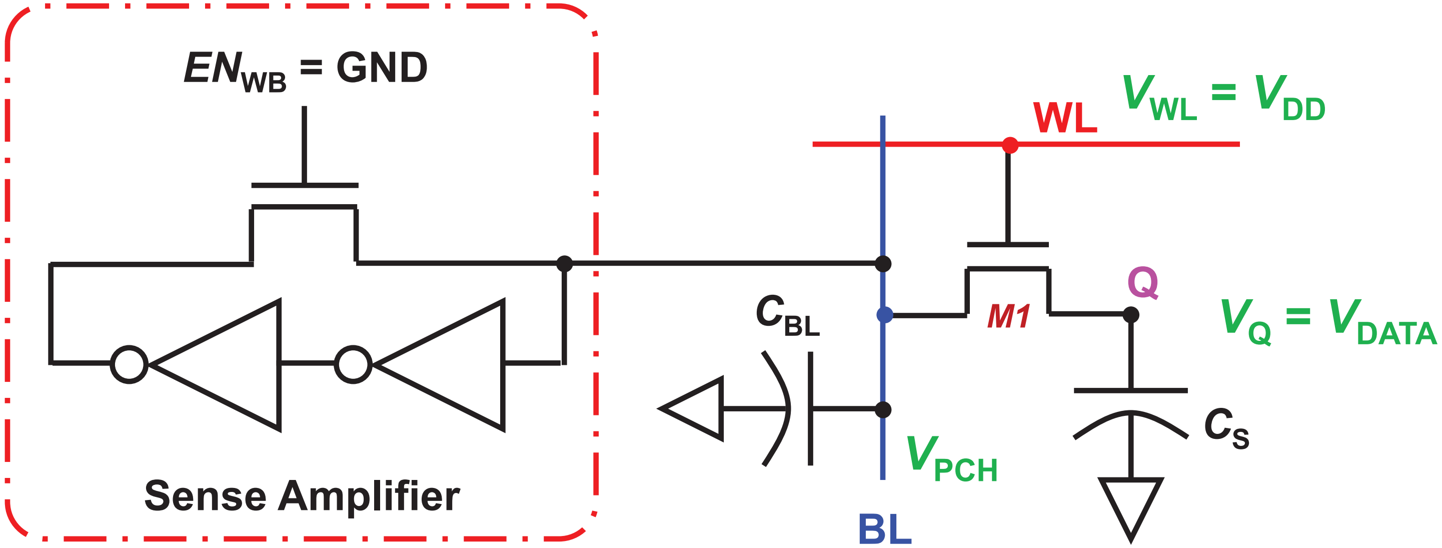
$$Q_{TOTAL} = C_{BL}V_{PCH} + C_S V_{DATA}$$

$$\Delta V_{BL} = \frac{Q_{TOTAL}}{C_{BL} + C_S} - V_{PCH}$$

$$\Delta V_Q = \frac{Q_{TOTAL}}{C_{BL} + C_S} - V_{DATA}$$

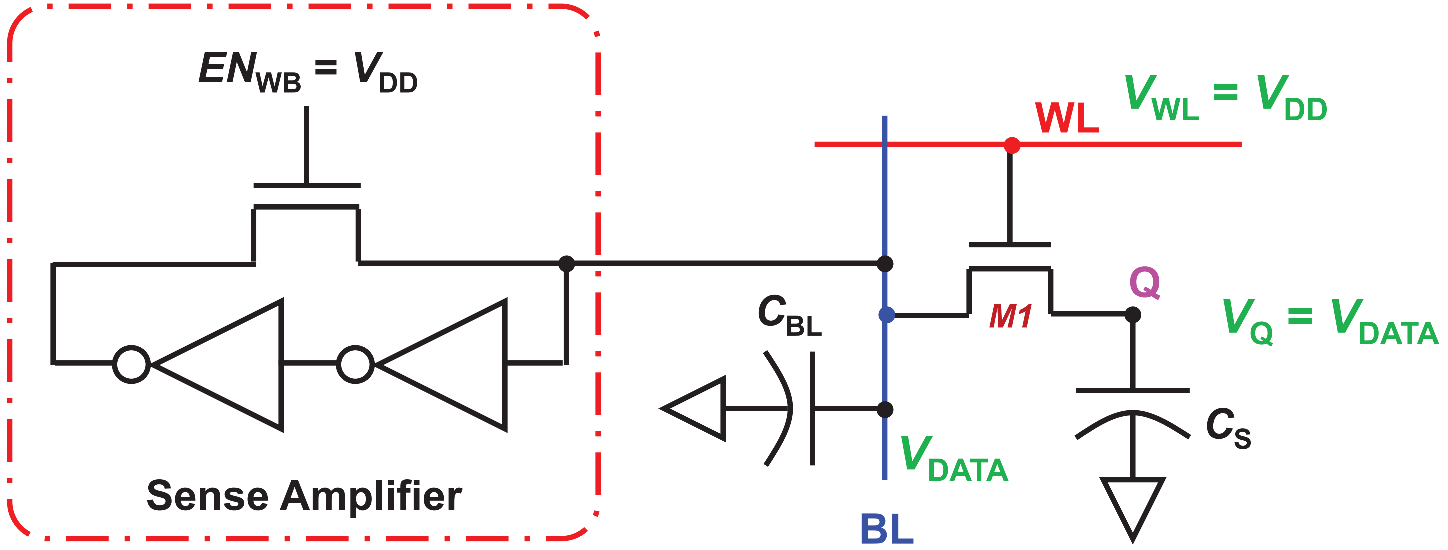
- M1 electrically shorts C_{BL} and C_S together
    - Charge sharing between C_{BL} and C_S
    - ΔV_Q and ΔV_{BL} depends on size of C_{BL} relative C_S, and whether M1 gets turned off or not

Reading Data from the 1T DRAM – Sense BL



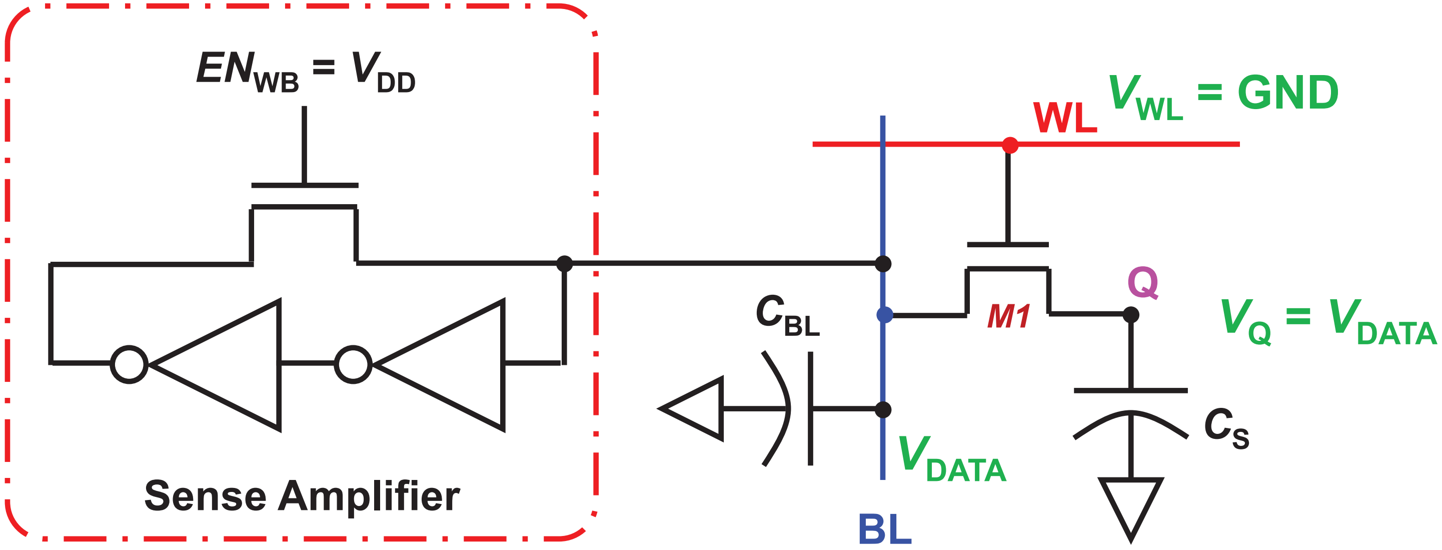
- Use inverter chain to sense logic value on bit-line
  - NMOS pass gate prevents writing back during sensing

Reading Data from the 1T DRAM – Writeback into Cell



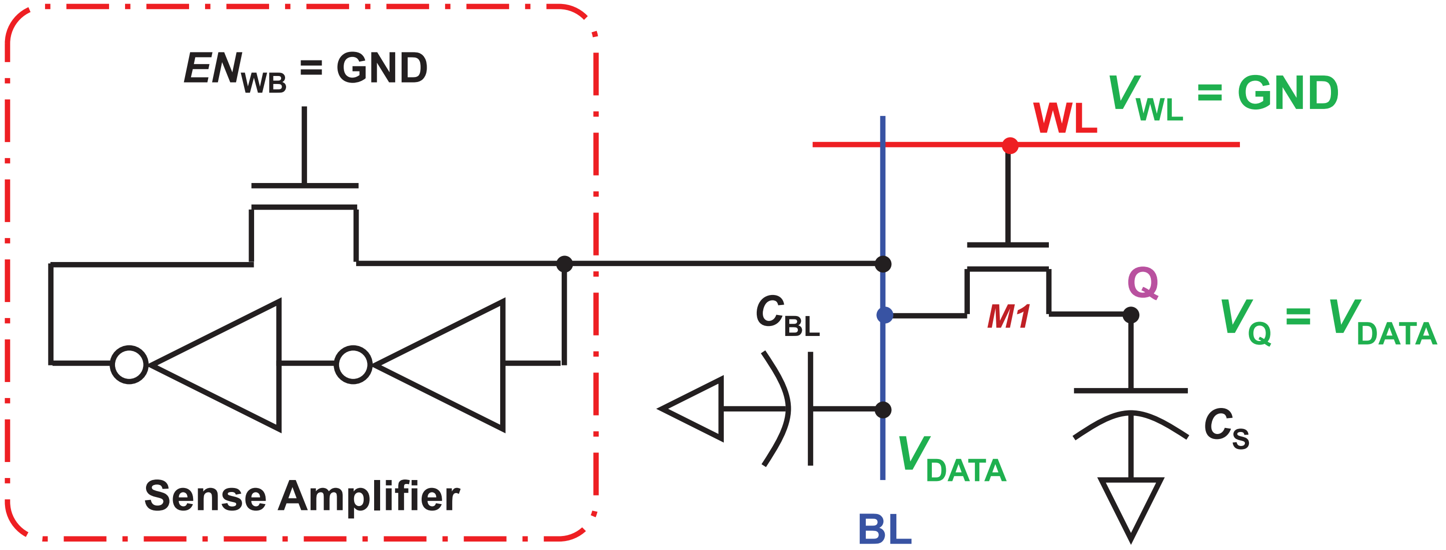
- Use inverter chain to sense logic value on bit-line
  - NMOS pass gate prevents writing back during sensing
  - Pass gate activates after sensing to write sensed data back into the cell

Reading Data from the 1T DRAM – Turn off WL



- Use inverter chain to sense logic value on bit-line
  - NMOS pass gate prevents writing back during sensing
  - Pass gate activates after sensing to write sensed data back into the cell

Reading Data from the 1T DRAM – Turn off Pass Gate

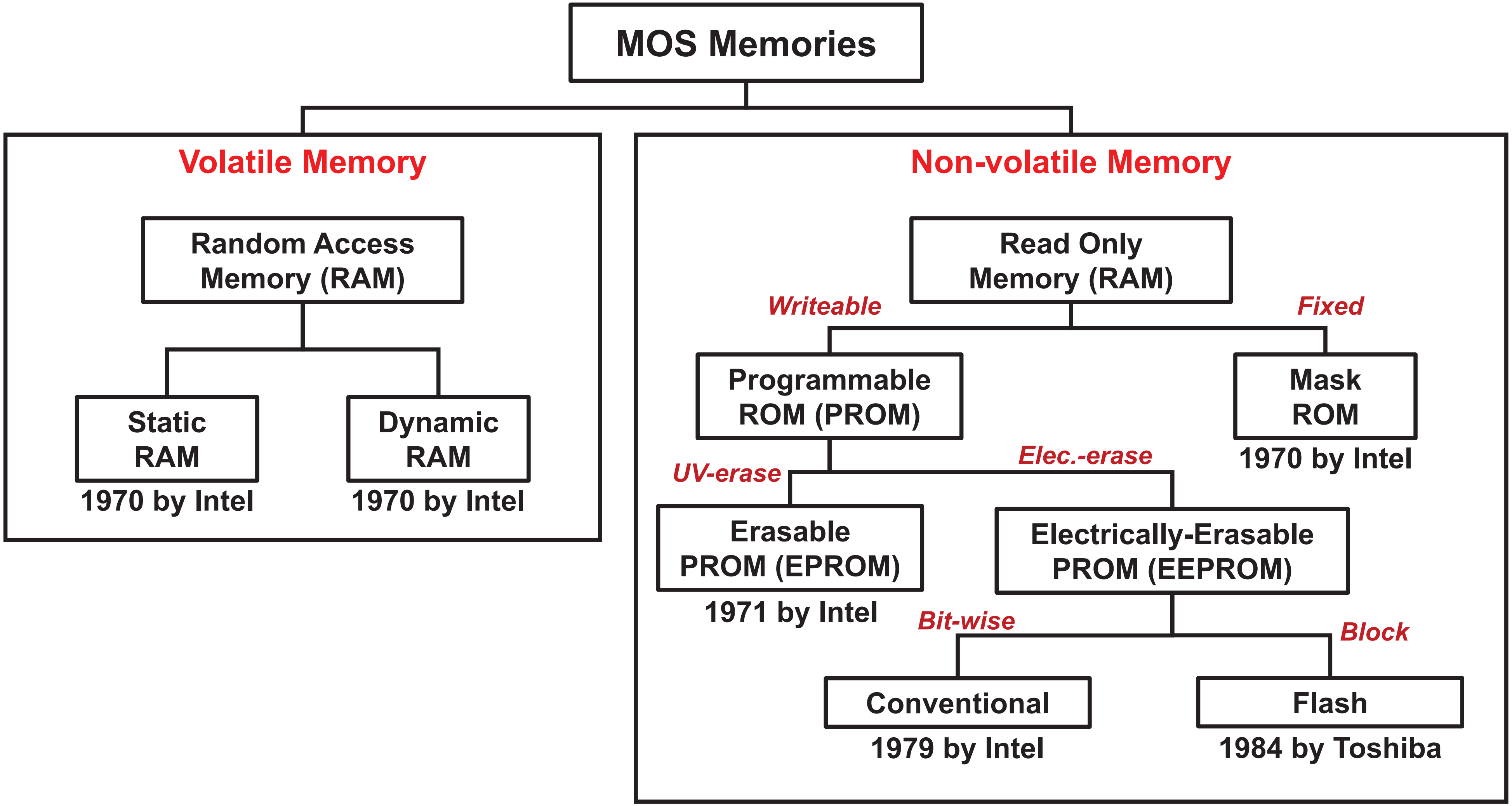


- Use inverter chain to sense logic value on bit-line
  - NMOS pass gate prevents writing back during sensing
  - Pass gate activates after sensing to write sensed data back into the cell

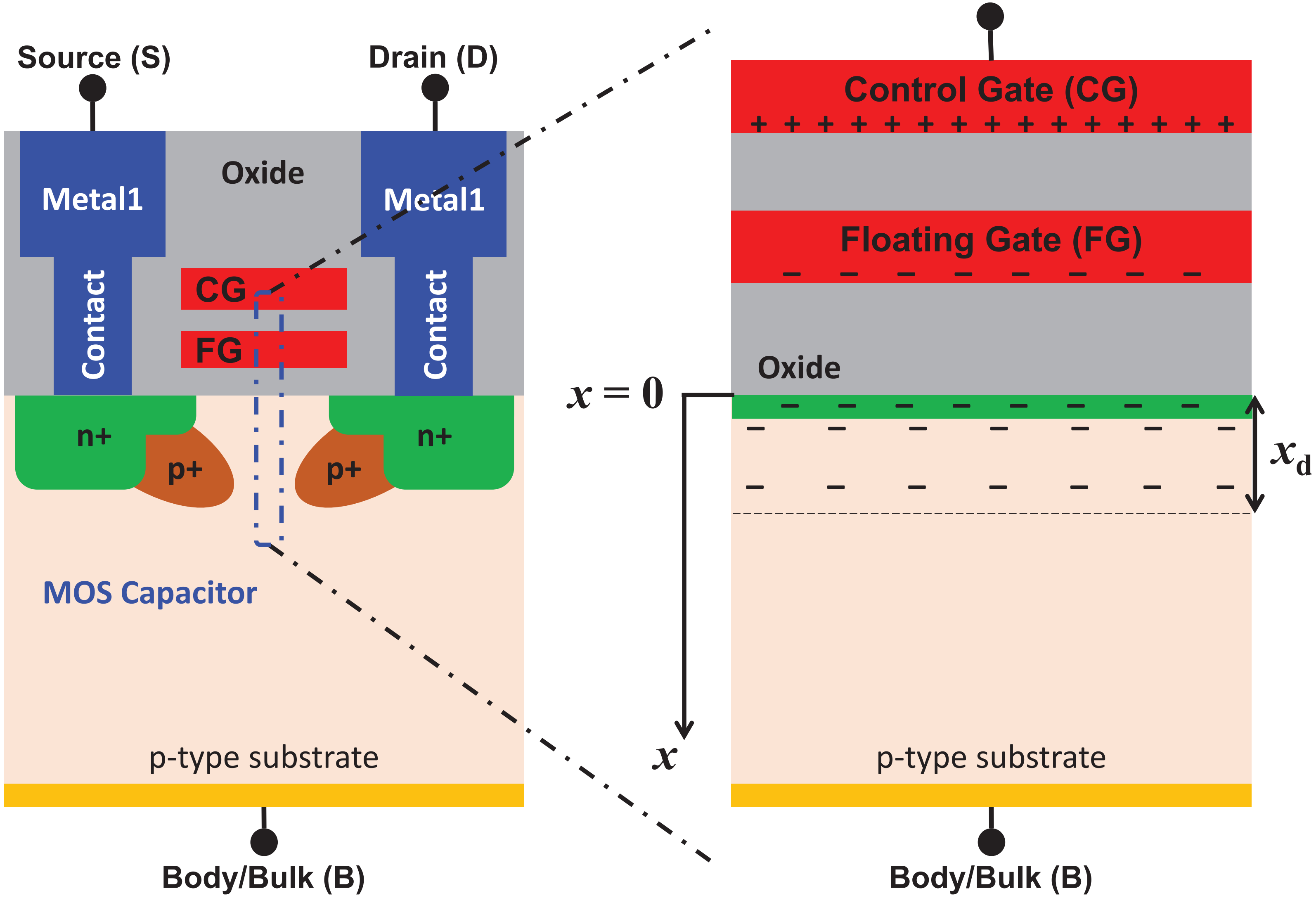
Week 5-7

Introduction to Flash Memory

MOS Memory Classification

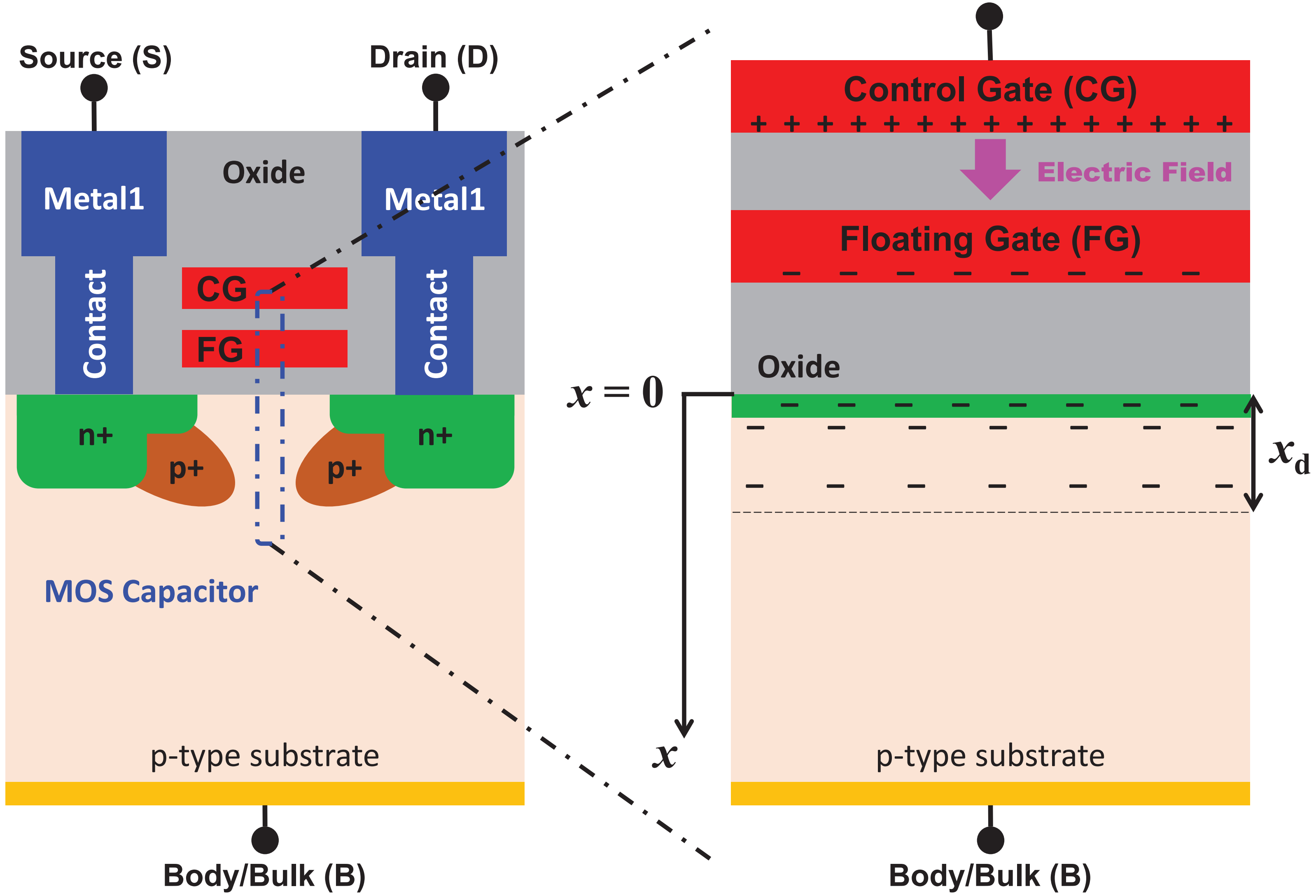


The Floating Gate Field-effect Transistor (FET)



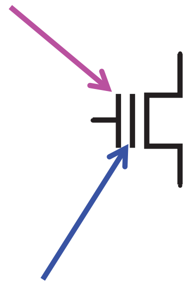
- Floating gate is embedded in oxide between control gate and channel region
- Control gate, source, drain and body terminals are connected to rest of circuit
- Control gate used like gate of an FET
- Charge can be stored in floating gate and sensed as change in threshold voltage, V_{TH}, of the FET

The Floating Gate Field-effect Transistor (FET)



- Charges added to control gate must be balanced by charges in substrate AND floating gate so that overall charge is zero (charge neutrality)
- Charges on floating gate shields substrate from electric field charges on control gate, which then changes the threshold voltage, V_{TH}, of the transistor

Flash Memory Based on Floating Gate FET



Control Gate (CG)

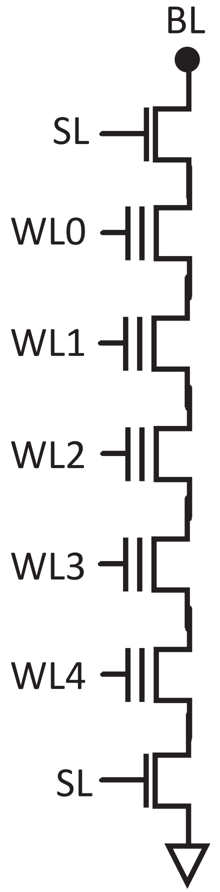
Memory cell or *bitcell*

Floating Gate (FG)

Flash Memory Based on Floating Gate FET

Architecture of NAND Flash

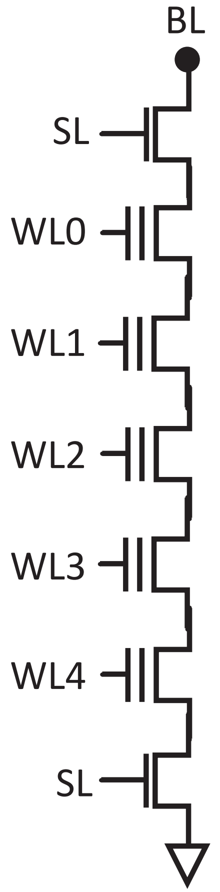
BL: Bit Line
SL: Select Line
WL: Word Line



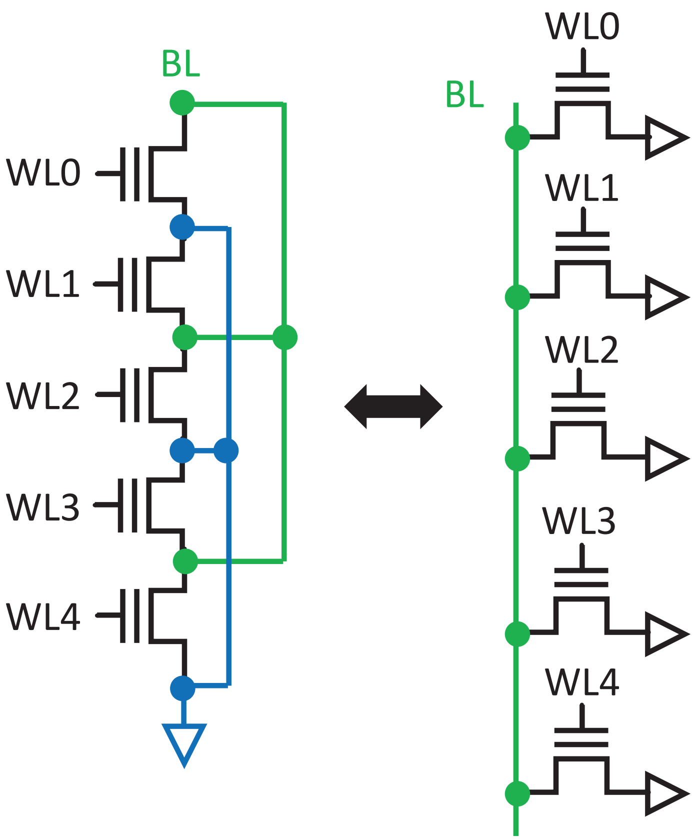
Flash Memory Based on Floating Gate FET

Architecture of NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line

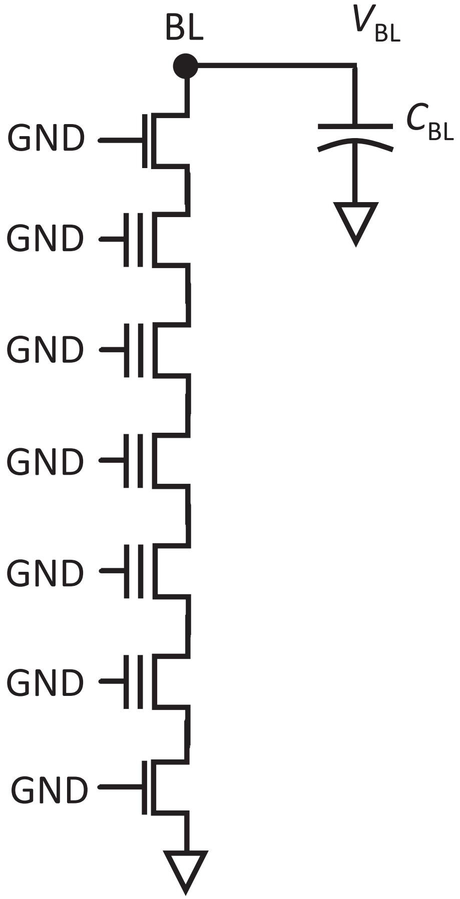


Architecture of NOR Flash



Read Operations in NAND Flash

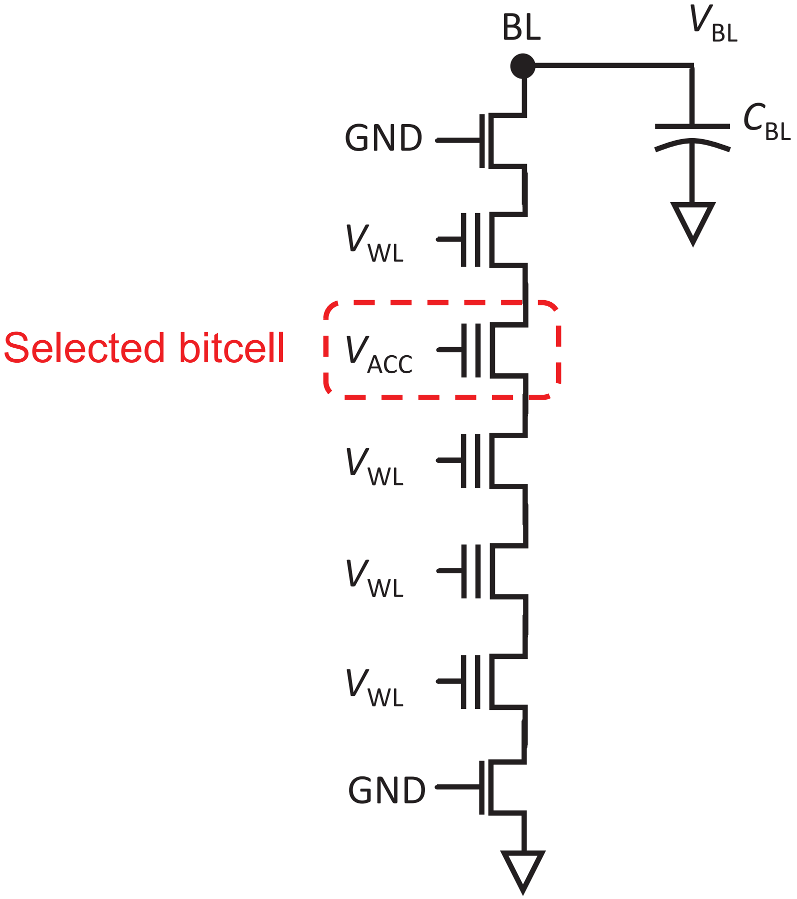
BL: Bit Line
SL: Select Line
WL: Word Line



Step #1: Charge BL to voltage V_{BL}

Read Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line



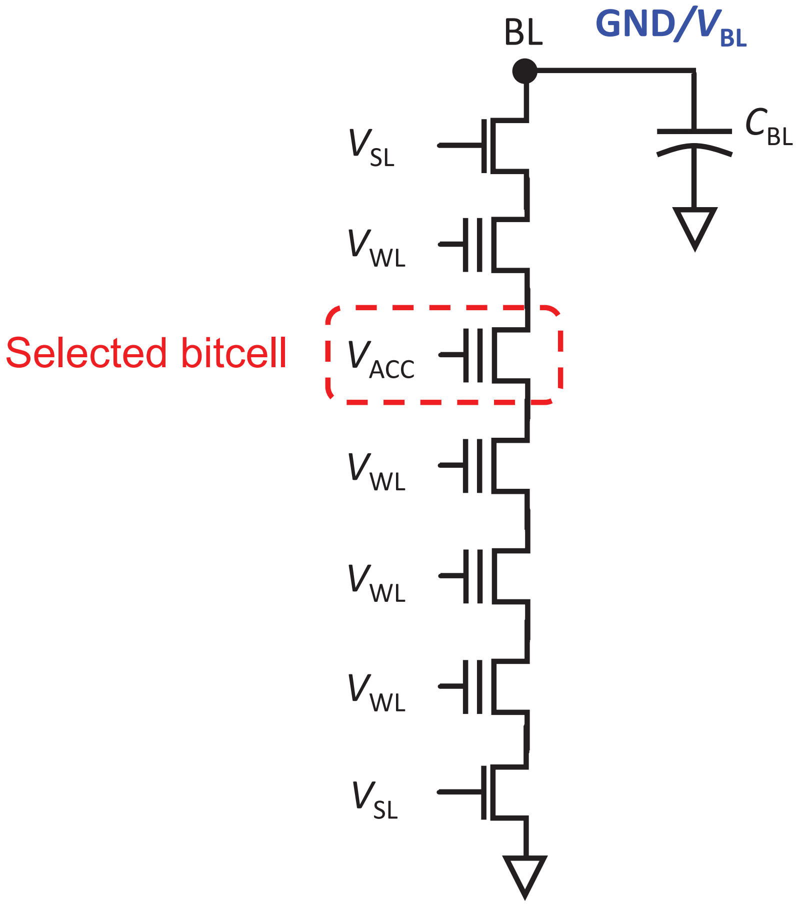
Step #1: Charge BL to voltage V_{BL}

Step #2: Apply voltage V_{WL} to all WLs except the WL being connected to the selected bitcell

- FETs with high V_{TH} are ON
- State of selected FET depends on V_{TH} due to stored charge

Read Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line



Step #1: Charge BL to voltage V_{BL}

Step #2: Apply voltage V_{WL} to all WLs except the WL being connected to the selected bitcell

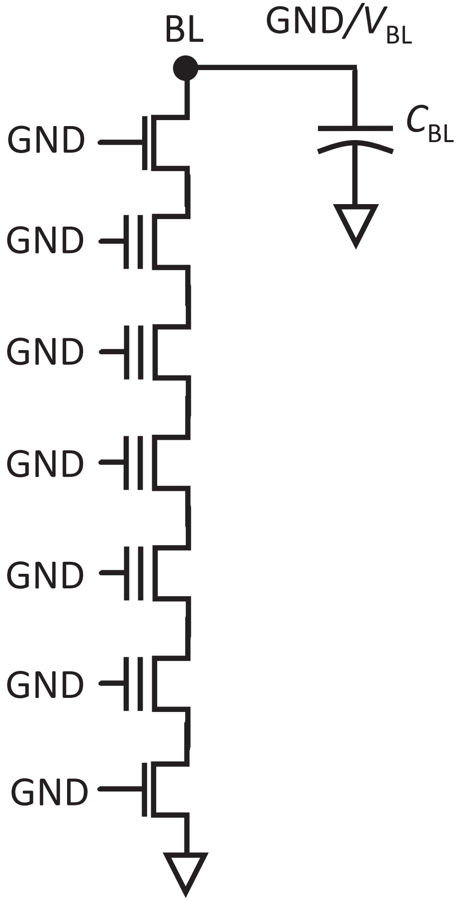
- FETs with high V_{TH} are ON
- State of selected FET depends on V_{TH} due to stored charge

Step #3: Apply voltage, V_{SL}, to SL

- Allow possible path to discharge C_{BL} through the NAND string
- Depends on whether selected FET is ON or OFF
  - ON: V_{BL} → GND

Read Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line



Step #1: Charge BL to voltage V_{BL}

Step #2: Apply voltage V_{WL} to all WLs except the WL being connected to the selected bitcell

- FETs with high V_{TH} are ON
- State of selected FET depends on V_{TH} due to stored charge

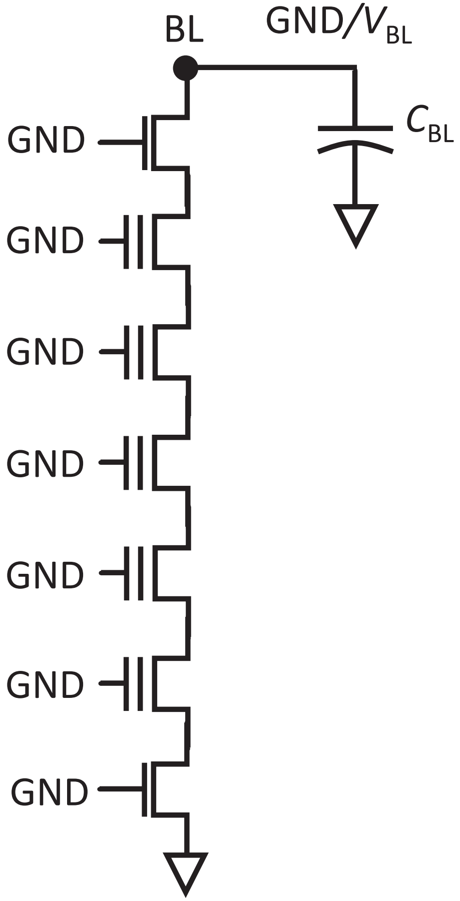
Step #3: Apply voltage, V_{SL}, to SL

- Allow possible path to discharge C_{BL} through the NAND string
- Depends on whether selected FET is ON or OFF
  - ON: V_{BL} → GND

Step #4: Ground all WLs and SLs

Read Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line



Step #1: Charge BL to voltage V_{BL}

Step #2: Apply voltage V_{WL} to all WLs except the WL being connected to the selected bitcell

- FETs with high V_{TH} are ON
- State of selected FET depends on V_{TH} due to stored charge

Step #3: Apply voltage, V_{SL}, to SL

- Allow possible path to discharge C_{BL} through the NAND string
- Depends on whether selected FET is ON or OFF
  - ON: V_{BL} → GND

Step #4: Ground all WLs and SLs

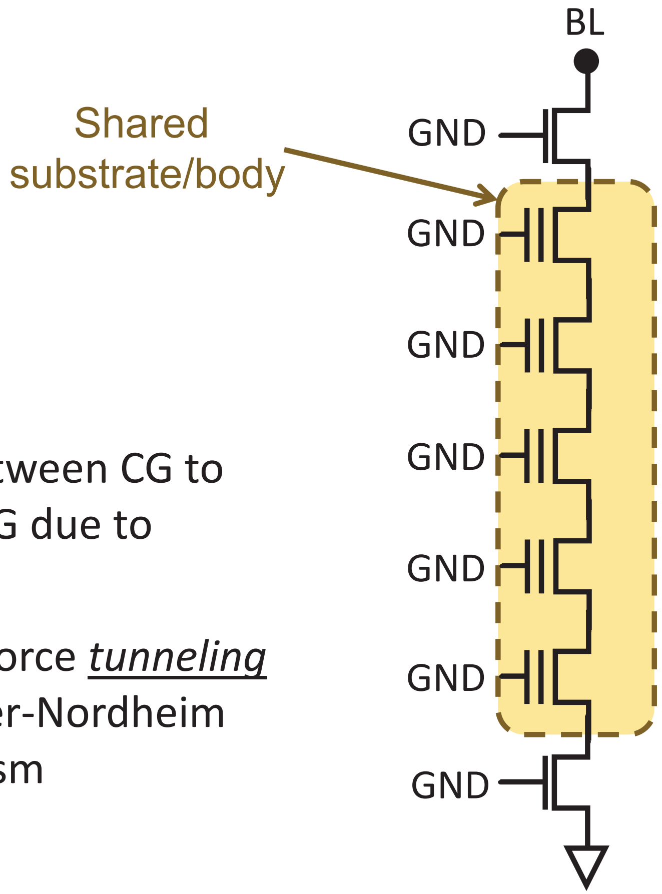
Step #5: Sense voltage of BL

Write Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line

High resistance path between CG to substrate through the FG due to insulating oxide

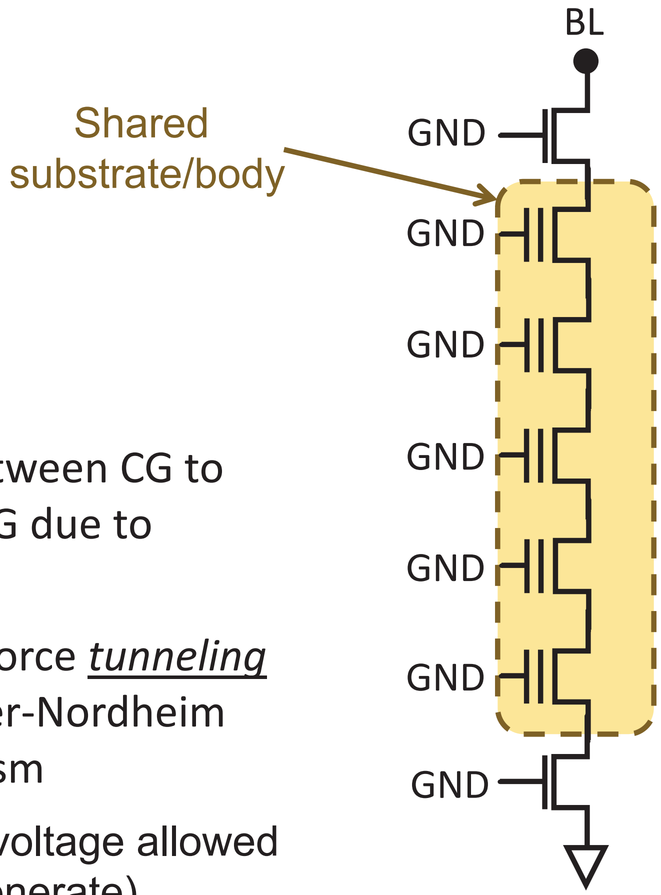
Large electric field can force ***tunneling current*** to flow by Fowler-Nordheim (FN) tunneling mechanism



Main idea: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

Write Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line



Main idea: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

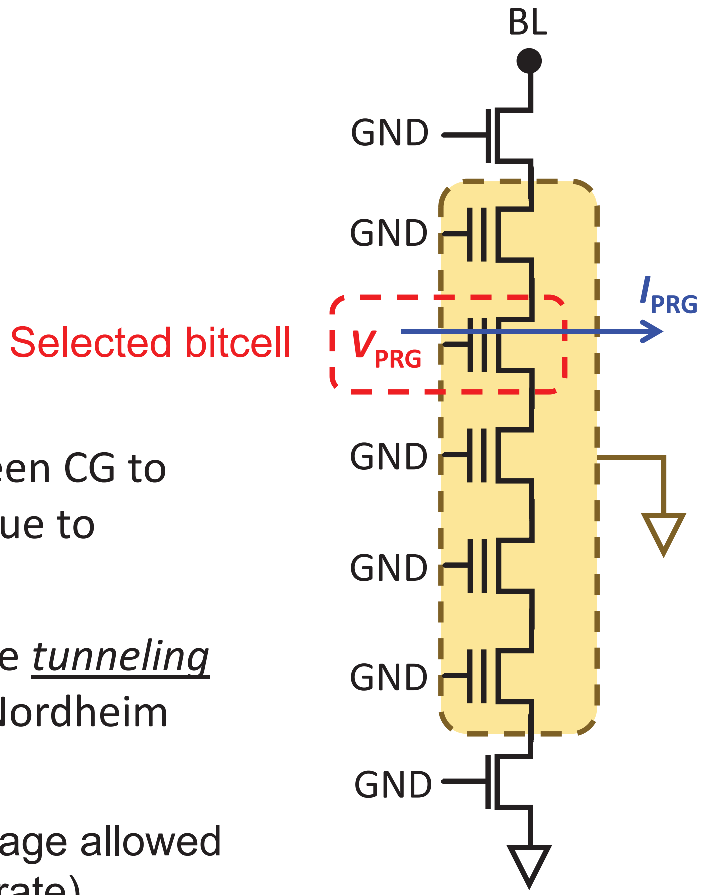
High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force ***tunneling current*** to flow by Fowler-Nordheim (FN) tunneling mechanism

Constraint: no negative voltage allowed
(difficult to generate)

Write Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line



High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force ***tunneling current*** to flow by Fowler-Nordheim (FN) tunneling mechanism

Constraint: no negative voltage allowed (difficult to generate)

Main idea: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

PROGRAM operation:

- Set voltage of substrate voltage to *GND*
- Apply programming voltage (V_{PRG}) to selected bitcell to drive current, I_{PRG}

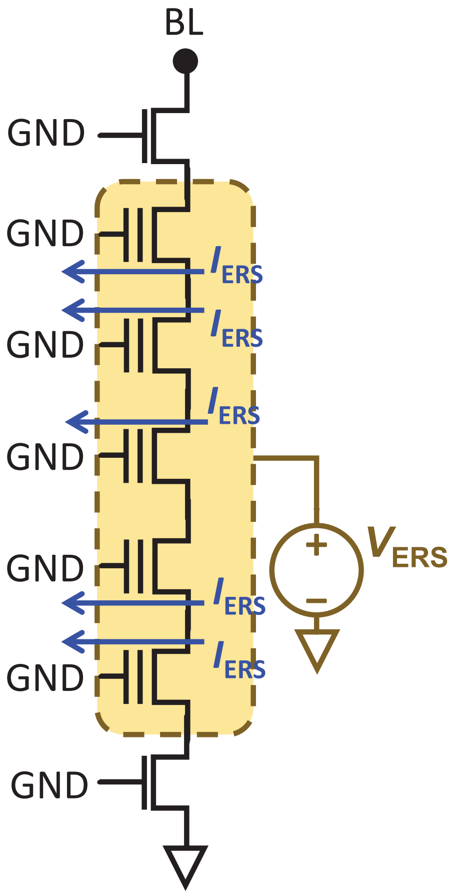
Write Operations in NAND Flash

BL: Bit Line
SL: Select Line
WL: Word Line

High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force *tunneling current* to flow by Fowler-Nordheim (FN) tunneling mechanism

Constraint: no negative voltage allowed (difficult to generate)



Main idea: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

PROGRAM operation:

- Set voltage of substrate voltage to *GND*
- Apply programming voltage (V_{PRG}) to selected bitcell to drive current, I_{PRG}

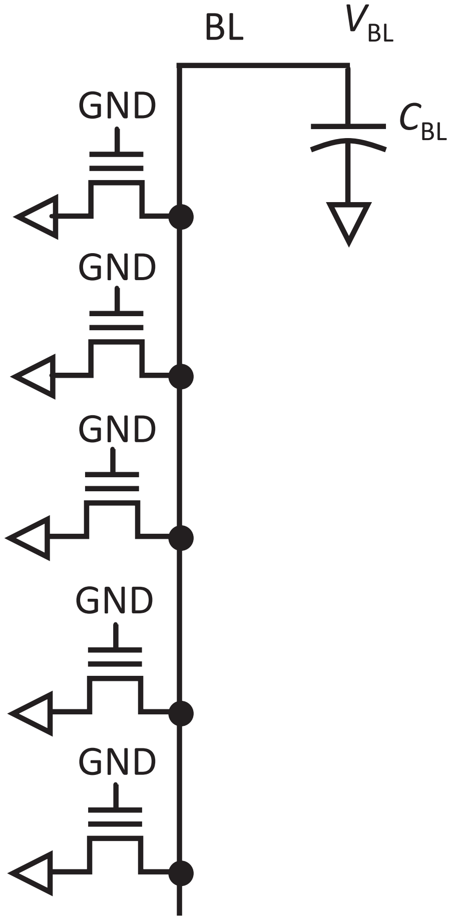
(block) ERASE operation:

- Set all WL voltages to GND
- Apply erase voltage (V_{ERS}) to common substrate to drive current, I_{ERS}, through all bitcells in the NAND string

Read Operations in NOR Flash

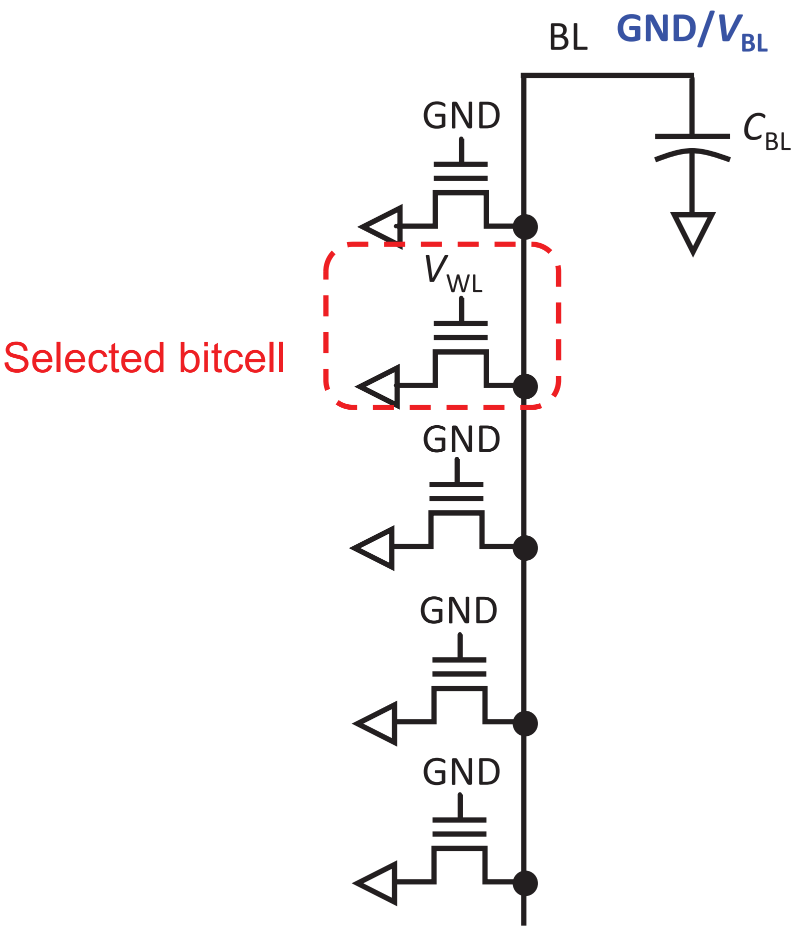
Step #1: Charge BL to voltage V_{BL}

BL: Bit Line
SL: Select Line
WL: Word Line



Read Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line



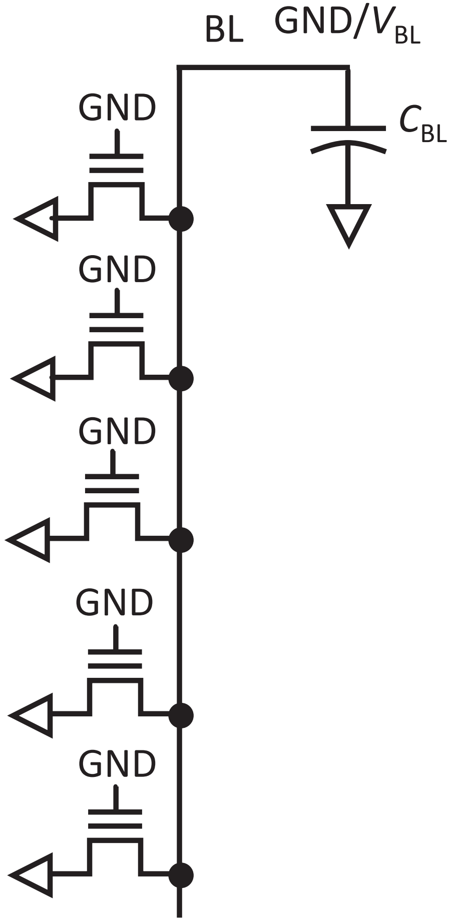
Step #1: Charge BL to voltage V_{BL}

Step #2: Apply voltage V_{WL} the WL being connected to the selected bitcell

- Selected FETs with low V_{TH} are ON
- Allow possible path to discharge C_{BL} through the selected bitcell
- Whether selected FET is ON or OFF depends on stored charge
  - ON: V_{BL} → GND

Read Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line



Step #1: Charge BL to voltage V_{BL}

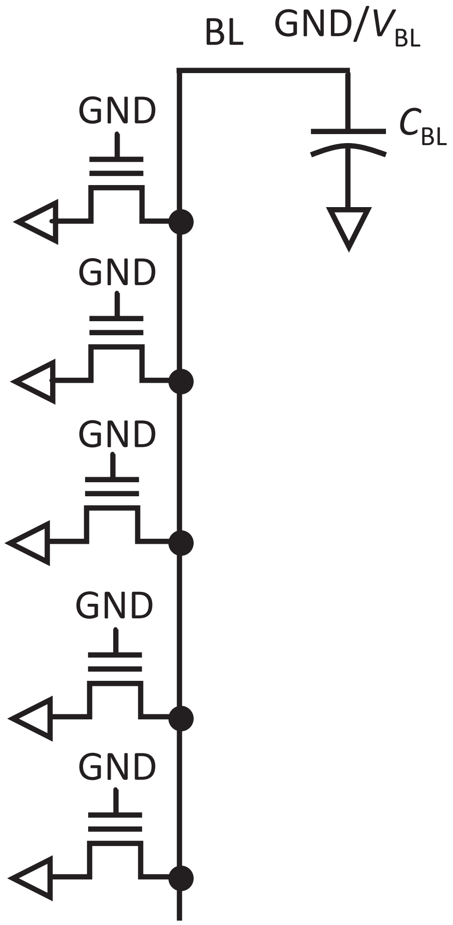
Step #2: Apply voltage V_{WL} the WL being connected to the selected bitcell

- Selected FETs with low V_{TH} are ON
- Allow possible path to discharge C_{BL} through the selected bitcell
- Whether selected FET is ON or OFF depends on stored charge
  - ON: V_{BL} → GND

Step #3: Ground all WLs

Read Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line



Step #1: Charge BL to voltage V_{BL}

Step #2: Apply voltage V_{WL} the WL being connected to the selected bitcell

- Selected FETs with low V_{TH} are ON
- Allow possible path to discharge C_{BL} through the selected bitcell
- Whether selected FET is ON or OFF depends on stored charge
  - ON: V_{BL} → GND

Step #3: Ground all WLs

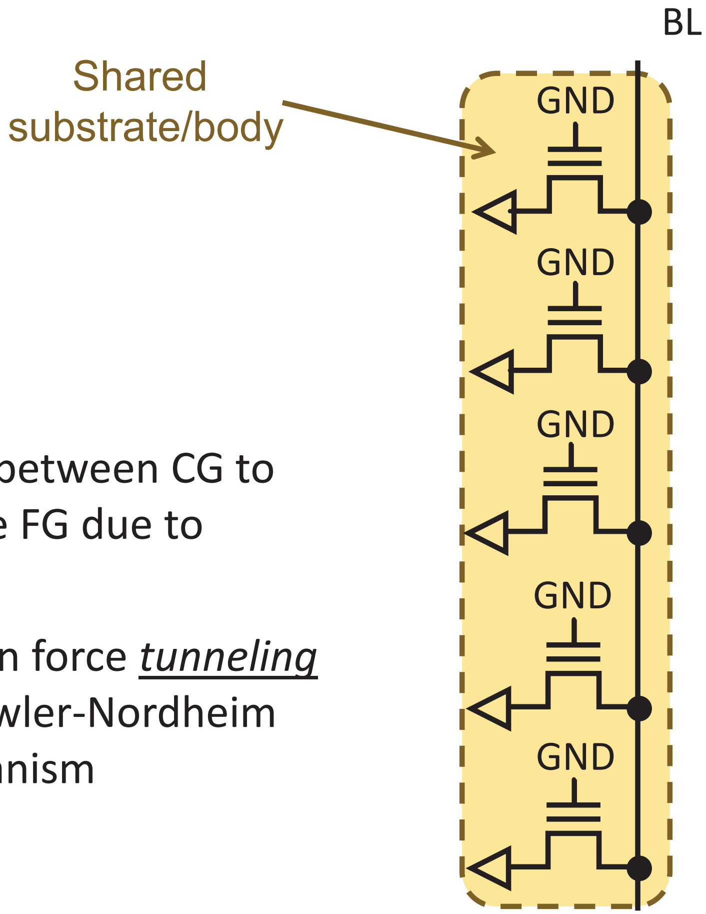
Step #4: Sense voltage of BL

Write Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line

High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force ***tunneling current*** to flow by Fowler-Nordheim (FN) tunneling mechanism



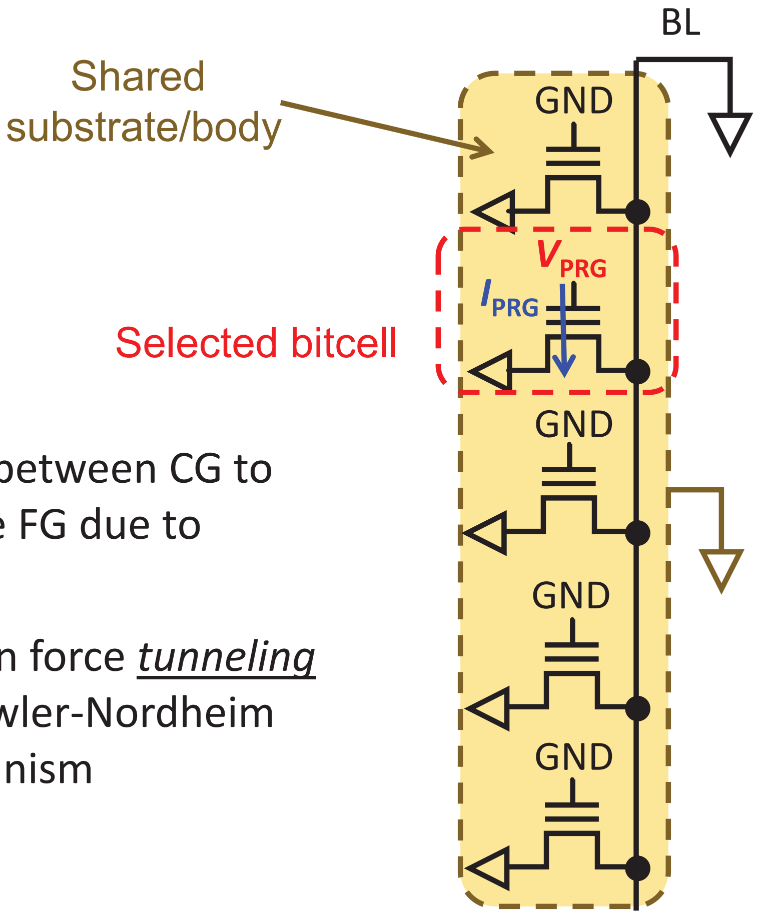
Main idea is the same as in NAND Flash: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

PROGRAM operation:

- Set voltage of BL and substrate to GND
- Apply V_{PRG} to WL of selected bitcell

Write Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line



High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force *tunneling current* to flow by Fowler-Nordheim (FN) tunneling mechanism

Main idea is the same as in NAND Flash: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

PROGRAM operation:

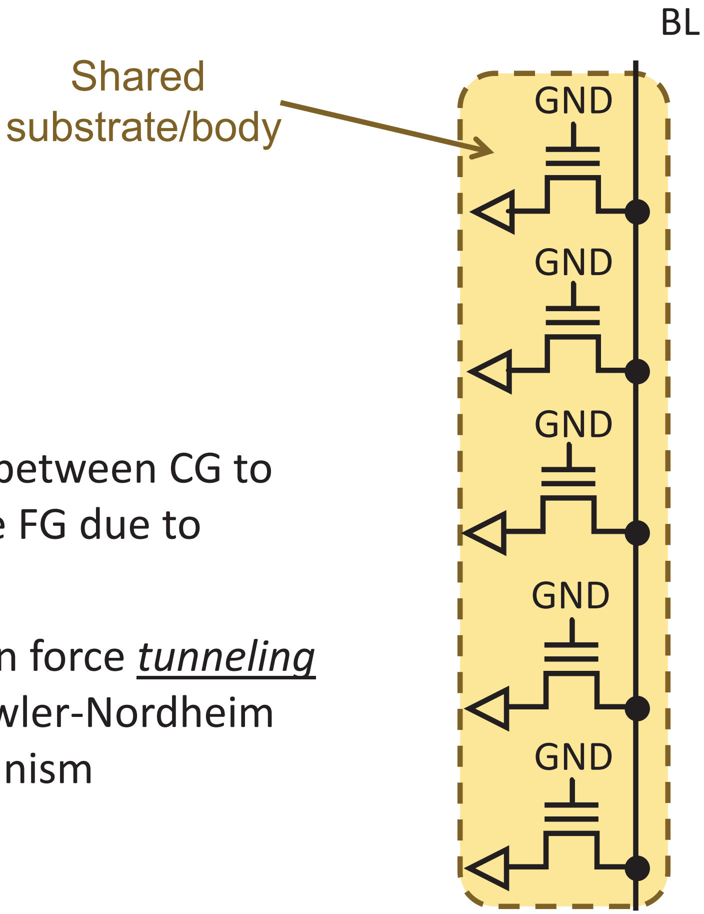
- Set voltage of BL and substrate to GND
- Apply V_{PRG} to WL of selected bitcell

Write Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line

High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force *tunneling current* to flow by Fowler-Nordheim (FN) tunneling mechanism



Main idea is the same as in NAND Flash: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

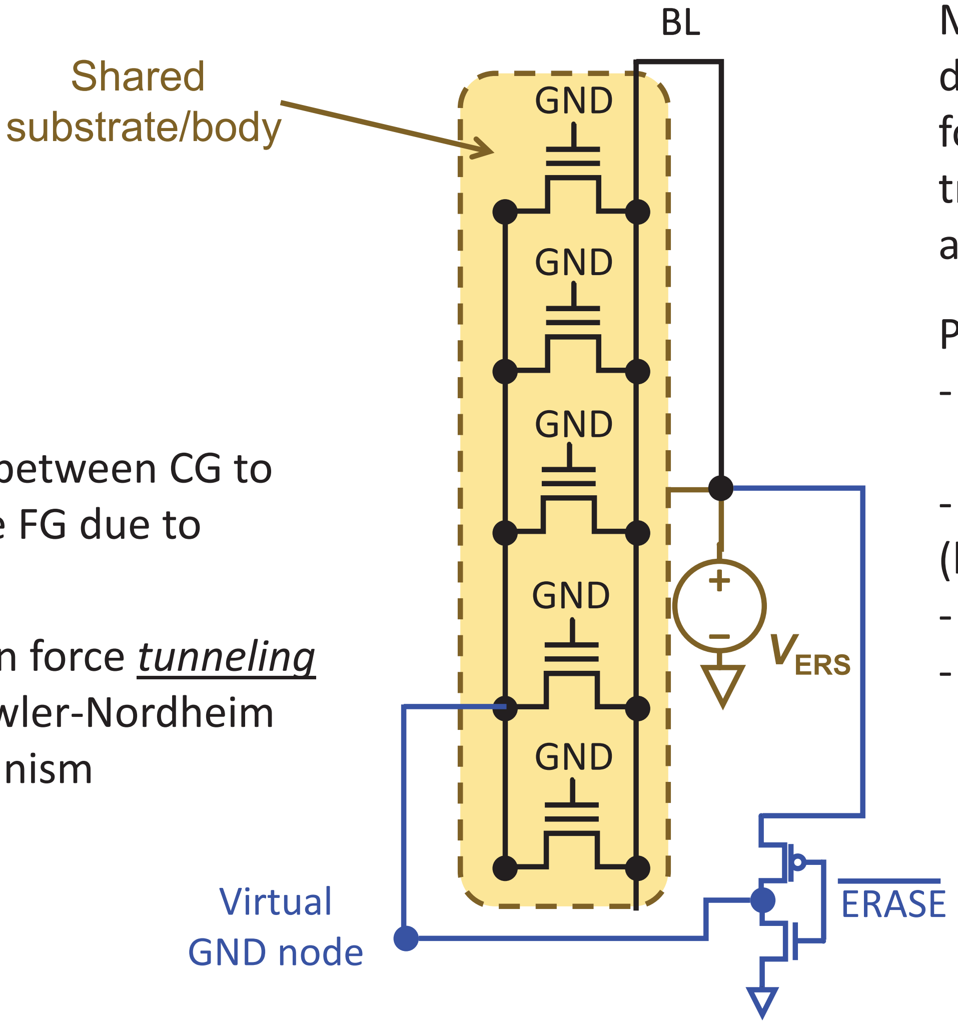
PROGRAM operation:

- Set voltage of BL and substrate to GND
- Apply V_{PRG} to WL of selected bitcell

(block) ERASE operation:

- Set voltage of all WLs to GND
- Apply V_{ERS} to BL, substrate, and virtual GND

Write Operations in NOR Flash



BL: Bit Line
SL: Select Line
WL: Word Line

High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force *__tunneling current__* to flow by Fowler-Nordheim (FN) tunneling mechanism

Main idea is the same as in NAND Flash: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

PROGRAM operation:

- Set voltage of BL and substrate to GND
- Apply V_{PRG} to WL of selected bitcell

(block) ERASE operation:

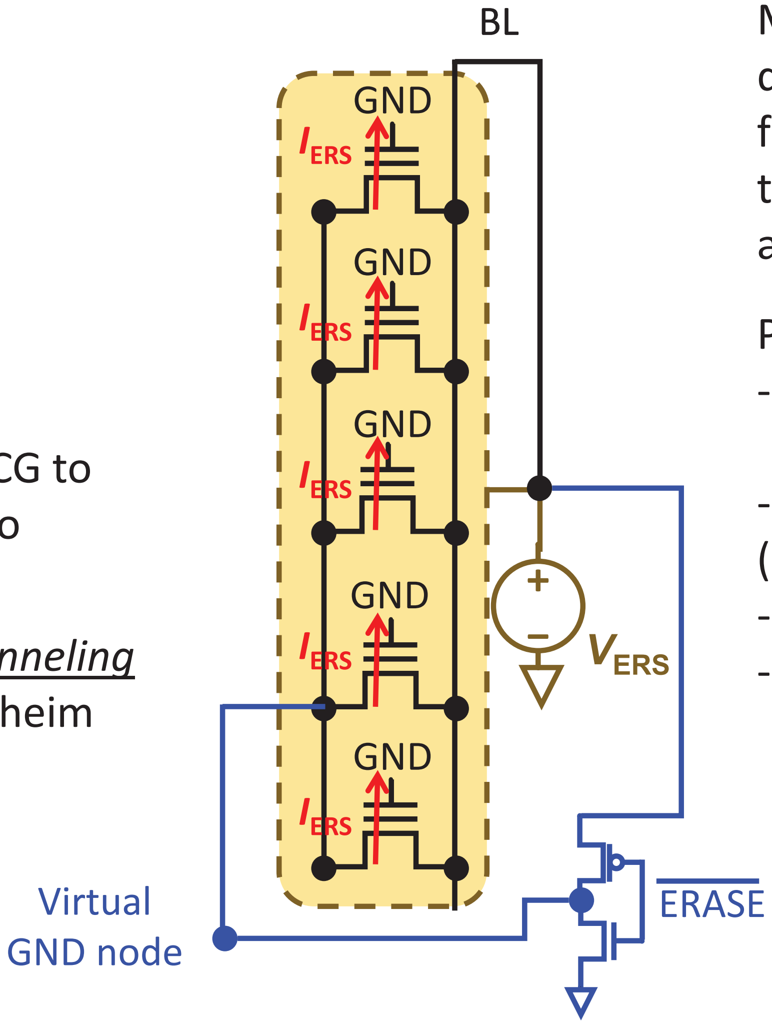
- Set voltage of all WLs to GND
- Apply V_{ERS} to BL, substrate, and virtual GND

Write Operations in NOR Flash

BL: Bit Line
SL: Select Line
WL: Word Line

High resistance path between CG to substrate through the FG due to insulating oxide

Large electric field can force *tunneling current* to flow by Fowler-Nordheim (FN) tunneling mechanism



Main idea is the same as in NAND Flash: drive current between CG and substrate for a short period of time. Charges get trapped on the FG when current is abruptly turned off

PROGRAM operation:

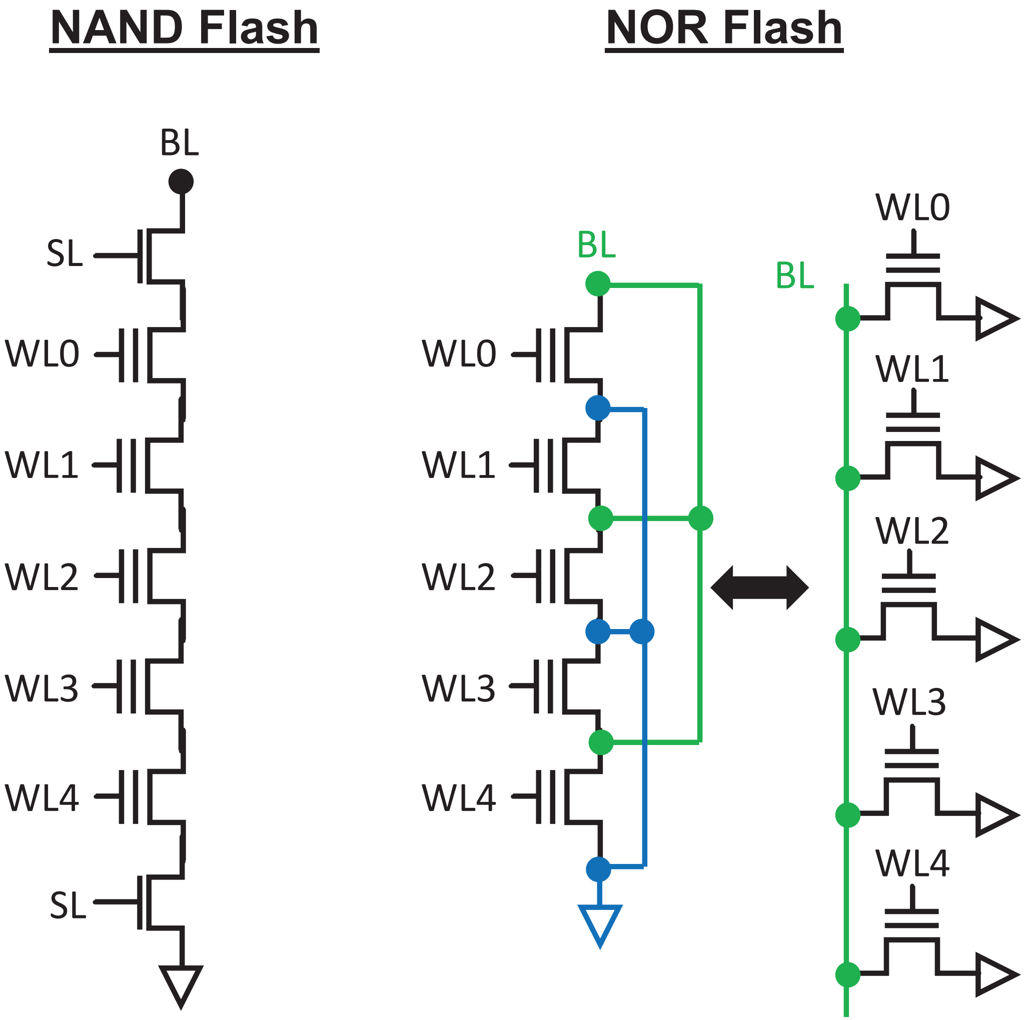
- Set voltage of BL and substrate to GND
- Apply V_{PRG} to WL of selected bitcell

(block) ERASE operation:

- Set voltage of all WLs to GND
- Apply V_{ERS} to BL, substrate, and virtual GND
  - BL and virtual GND voltage cannot be GND or the pn-junctions with substrate will conduct leakage current

Characteristics of Flash Memory

BL: Bit Line
SL: Select Line
WL: Word Line



	NAND	NOR
Speed	X	O
Density	O	X
Program	Cell	
Erase	Block	

Memory management needed to:

- Map memory addresses to physical blocks of memory
  - Redundant blocks are needed
  - When erase operation is needed for a cell, entire block is marked as ‘dirty’ and copied to a ‘clean’ block
- Wear-leveling
  - Each cell can only support < ~10^6 PROGRAM and ERASE ops.