



Bridge of Life
Education

SOC Design Memory Technology

Jiin Lai

Topics

1. Issues with Memory System - Speed, Capacity, Power
2. **Memory Technology**
3. Cache
4. DRAM
5. FPGA Memory
6. Application

Topics

1. SRAM
2. DRAM
3. Non-volatile Memory - Flash

Memory Technology Basics – SRAM v.s. DRAM

SRAM – Static Random Access Memory

- Static – holds data as long as power is maintained
- Requires multiple transistors to retain one bit and has low density compared to DRAM, thus more expensive
- Faster than DRAM
- Used for caches

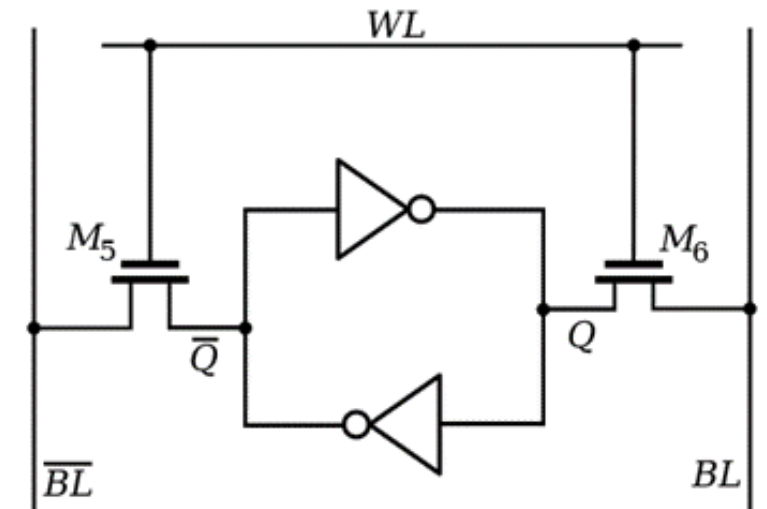
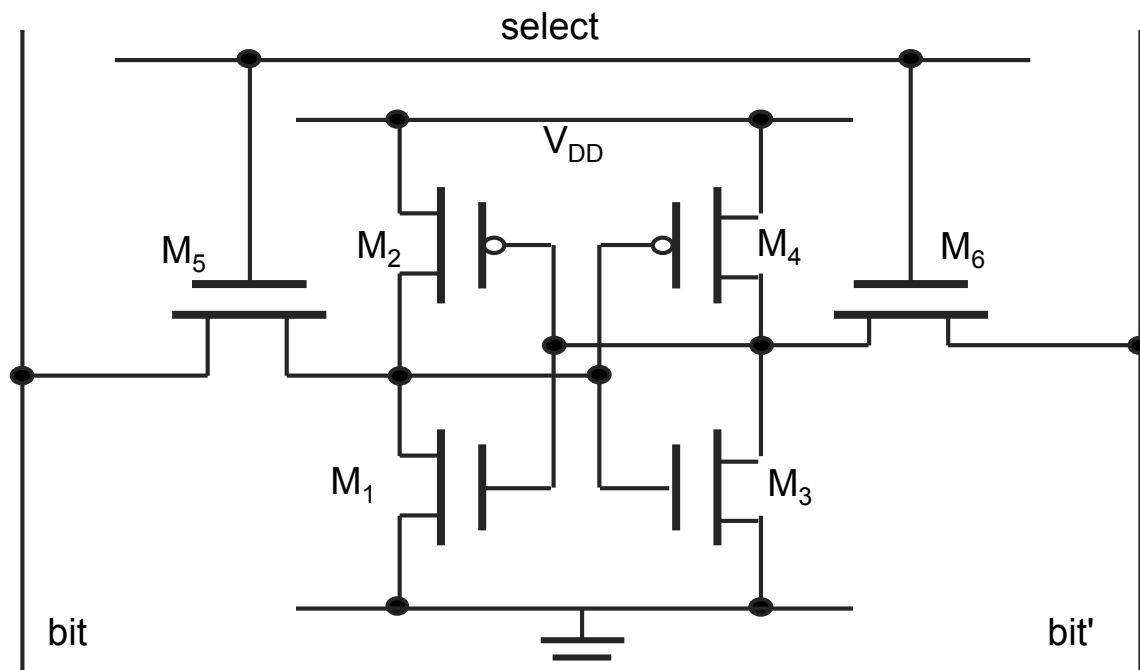
DRAM – Dynamic Random Access Memory

- Dynamic – must be refreshed periodically to hold data
- Requires only one transistor (and one capacitor) to retain one bit of data
- High density, thus cheaper than SRAM
- Used for main memory and sometimes for larger caches

SRAM

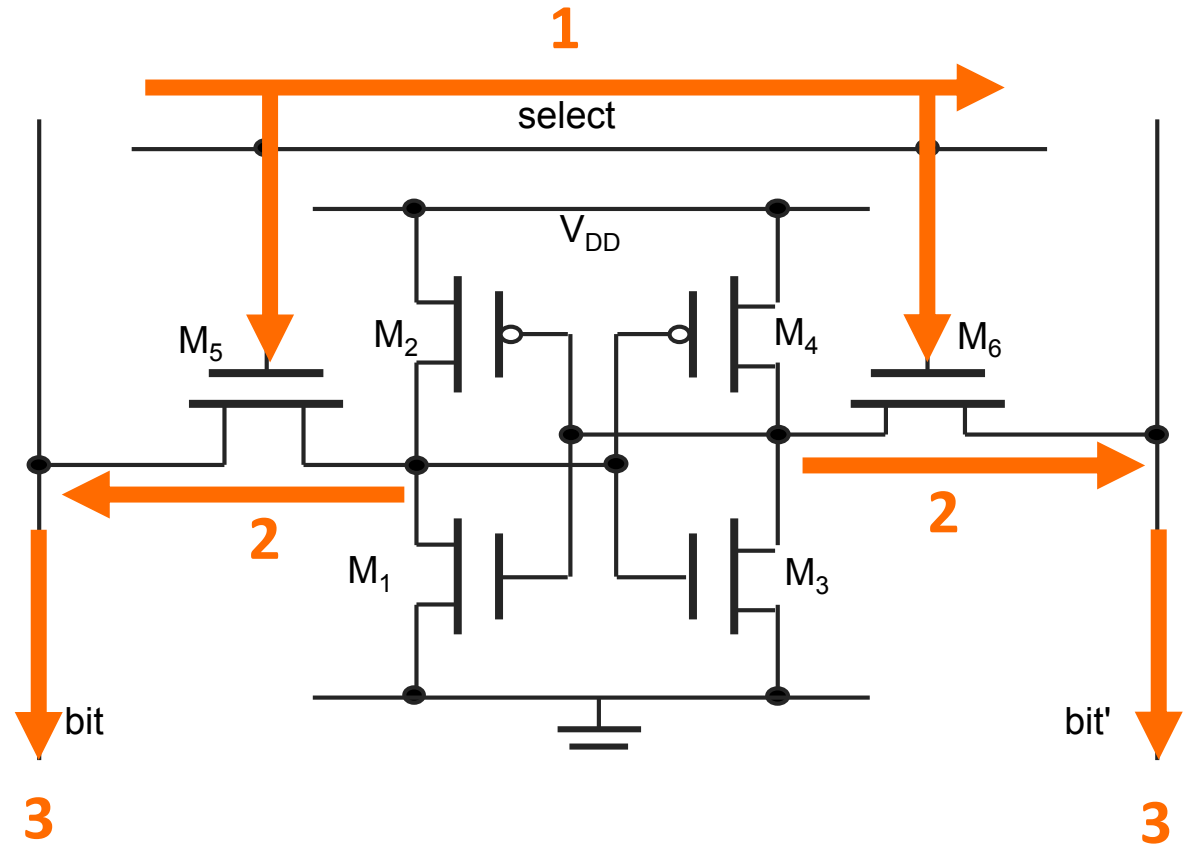
SRAM cell

- An SRAM cell is typically made up of ^{2 inverter} six transistors (MOSFETs).
 - A single bit is stored on 4 transistors (M1-M4), which form two inverters that are cross-coupled.
 - Write: set BL and BL to 0 and VDD or VDD and 0 and then enable WL (i.e., set to VDD)
 - Read: Charge BL and BL to VDD and then enable WL (i.e., set to VDD). Sense a small change in BL or BL



Accessing SRAM – Read operation

- The address is decoded and the desired cell is then selected, in which case the select line is set to one.
- Depending on the value of the 4 transistors (M1-M4), one of the bit lines (bit or bit') will be charged to 1 and the other will be drained to 0.
- The states of the two bit lines are then read out as 1-bit data.

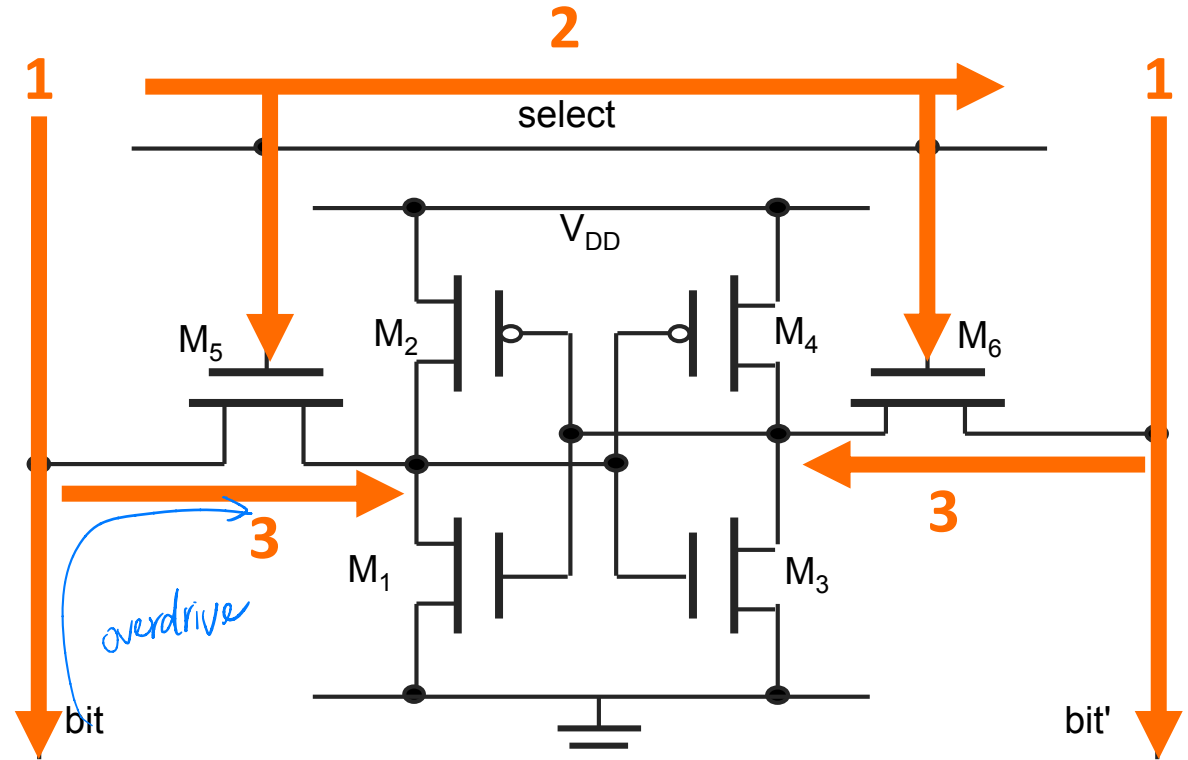


Accessing SRAM – Write operation

The two bit lines are pre-charged to the desired value (e.g., bit = VDD, bit' = VSS).

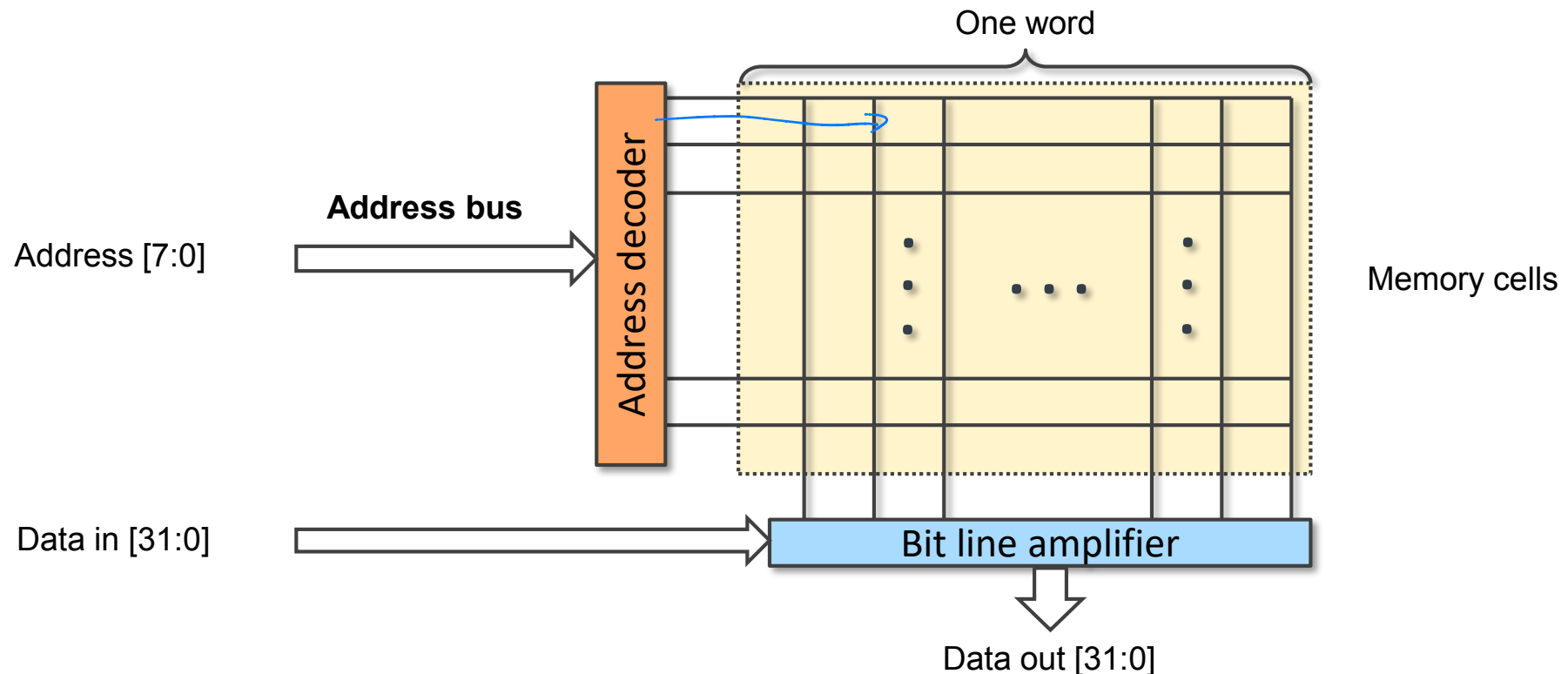
The address is decoded and the desired cell is then selected, in which case the select line is set to one.

The 4 transistors (M1-M4) are then forced to flip their states (either charged or discharged) since the bit lines normally have much higher capacitance than the 4 transistors.



Accessing SRAM

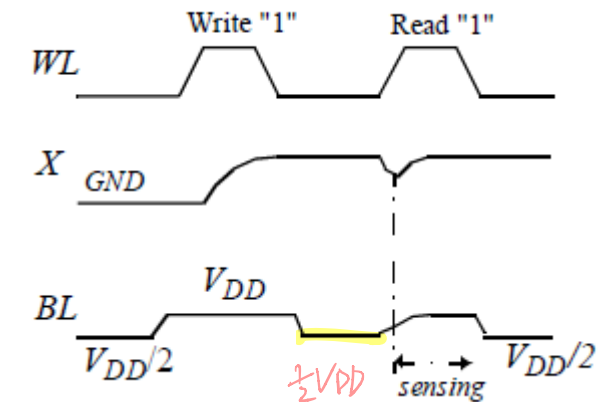
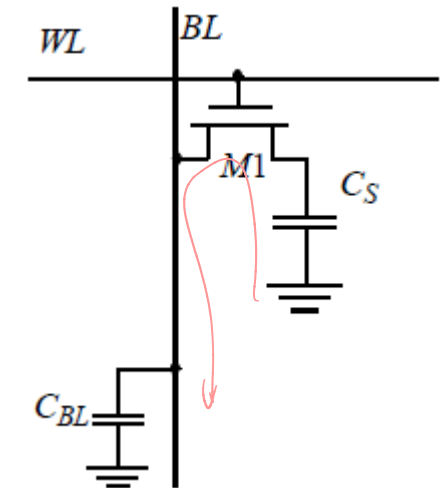
- The SRAM cells are organized into rows, with a whole row accessed at once.
 - For example, a memory architecture with an 8-bit address and 32-bit data is shown below.
- The address decoder uses the address to select a single row, and all its data are read out.



DRAM

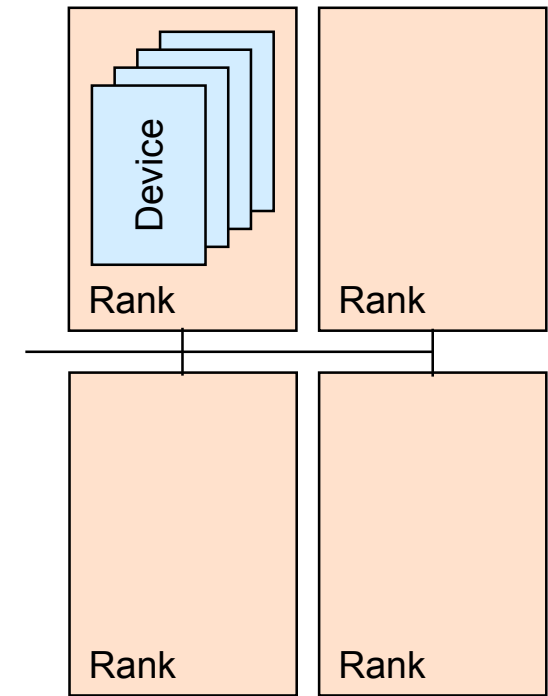
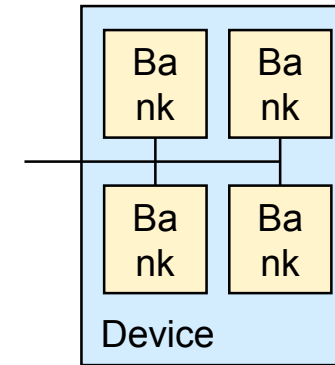
DRAM

- A DRAM cell is composed of a transistor and a capacitor
- A charged capacitor represents a logic high '1', a discharged capacitor is a logic low '0'
 - Charging/discharge is done via the wordline and bitline
- To Write: set Bit Line (BL) to 0 or VDD & enable Word Line (WL) (i.e., set to VDD)
- complex ← To Read: set Bit Line (BL) to VDD / 2 & enable Word Line (i.e., set it to VDD)
- The cell needs to be refreshed (or recharged) periodically, e.g. every 7.8 ms
 - The capacitor is drained on a read and charged (if storing 1) on a write.
 - The capacitor leaks its charge.
- DRAM is higher density than SRAM.
- DRAM can be categorized according to its synchronization and data rate.
 - Most DRAM is now synchronous (SDRAM), so it has a clock, rather than asynchronous.
 - Double data rate (DDR) DRAM transfers data on both the rising and falling clock edges.



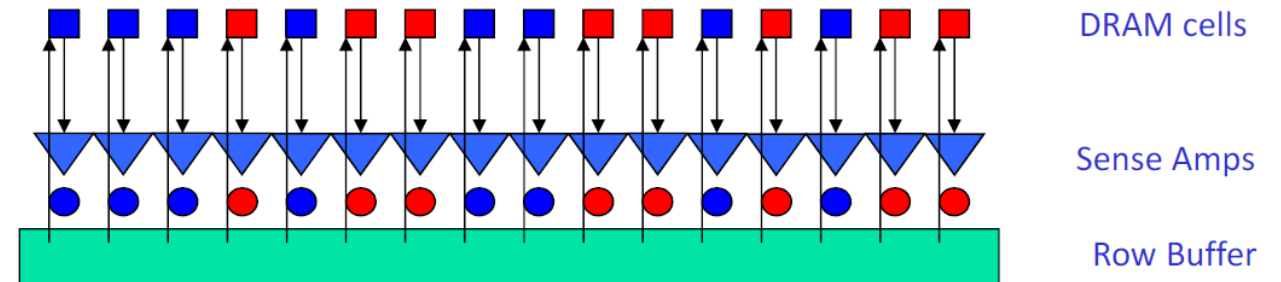
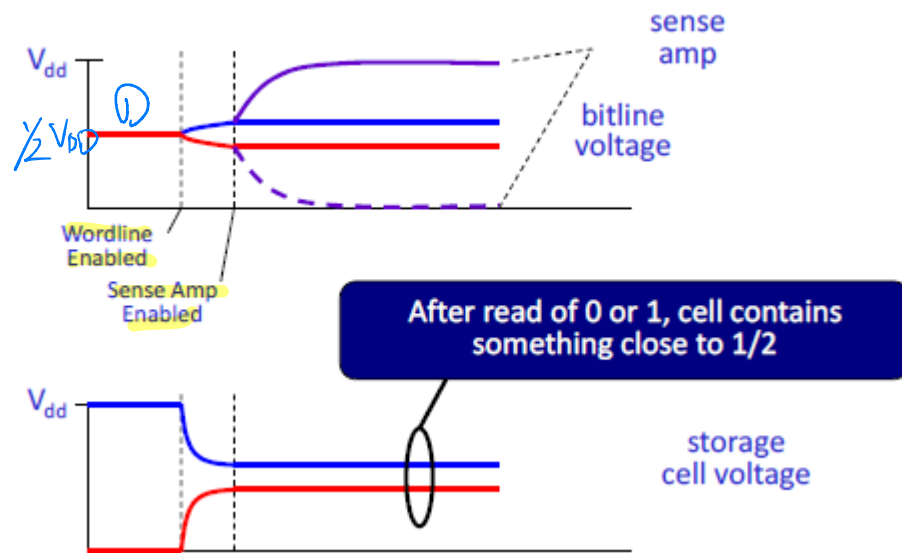
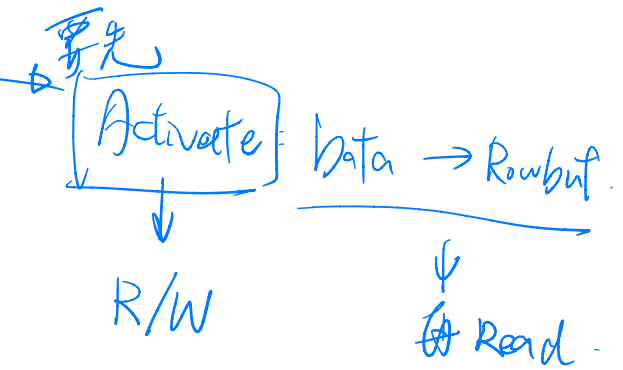
DRAM Organization

- **DRAM chip** has multiple banks (4 or 8 Banks)
- Each **device bank** operates independently.
 - This allows multiple accesses to occur concurrently.
- **Rank:** Devices may be grouped into ranks.
 - All devices in a rank are accessed together.
 - “2Rx8” – 2 ranks, 8 bits from each DRAM chip
- **DIMM** (Dual Inline Memory Module)
 - Contain one or two ranks



Read is destructive – Row Buffer

- Read is destructive – After read, cell contains 1/2 charge
 - So after a read, the content of DRAM cell are gone
 - The value are stored in the row buffer
 - Write them back into the cells for the next read
- DRAM cell will lose its contents (leak) even if it's not accessed
- DRAM rows need to be regularly read and re-written (Refresh)



Techniques for Improving DRAM Performance

Row Buffer (Page)

- To buffer recently accessed data without having to make another access
- Read and refresh several words and in parallel.
- The row buffer is essentially the sense amps at the bottom of each array.

DRAM Banking

- Increase the number of parallel banks to improve bandwidth with simultaneous accesses.

Double Data Rate

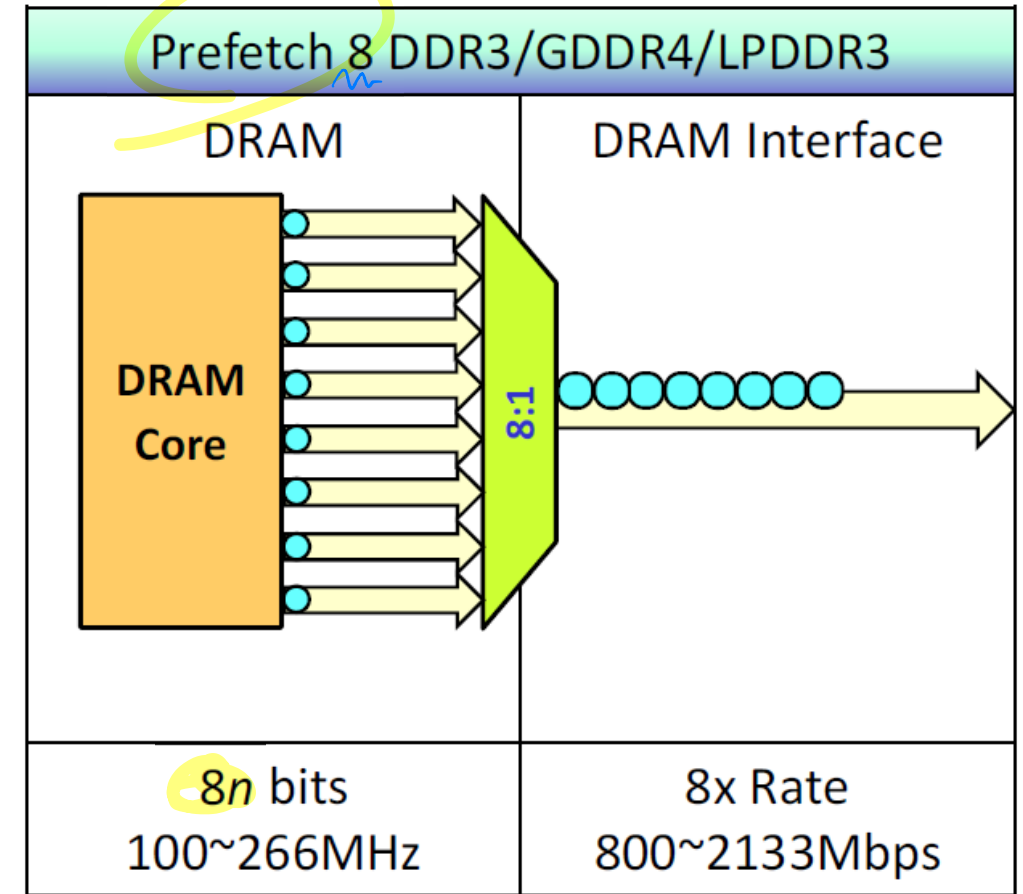
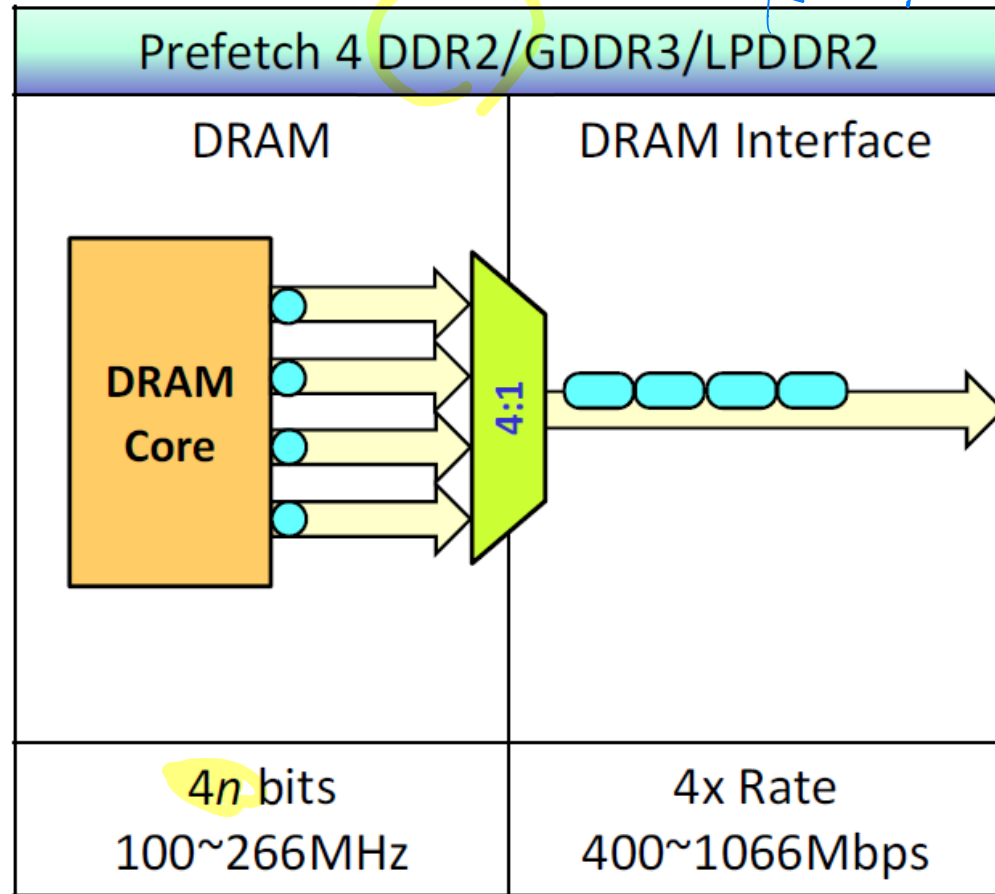
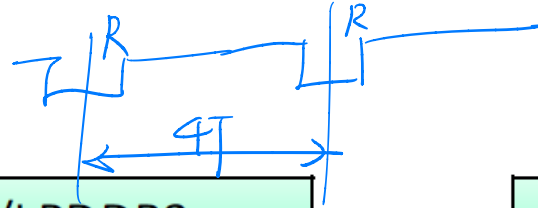
- Transfer data on the rising and falling clock edges to double the bandwidth.

Prefetch Architecture - Bursting

- Transfer data on the rising and falling clock edges to double the bandwidth.

DRAM Prefetch Architecture

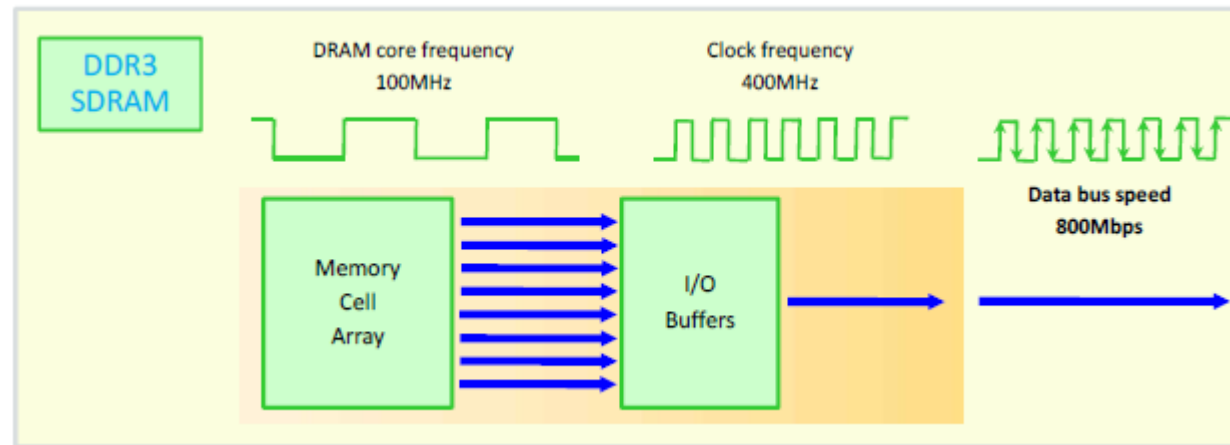
$4T < \text{prefetch}$



Cacheline Size = $60R$ burst length = 8
8x64

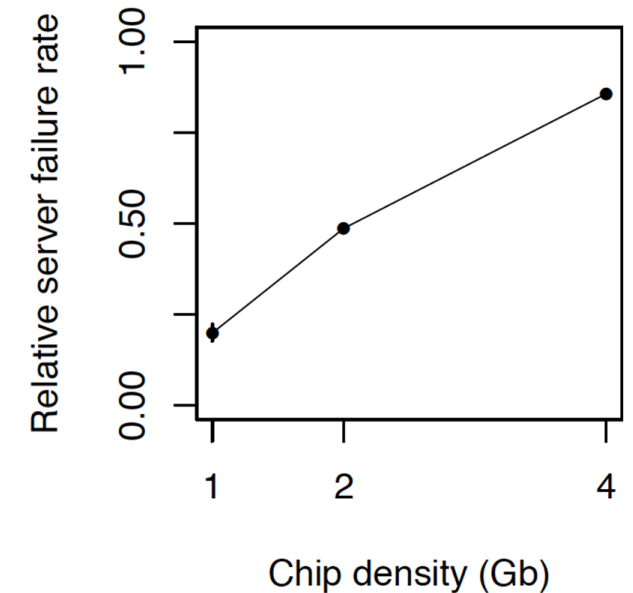
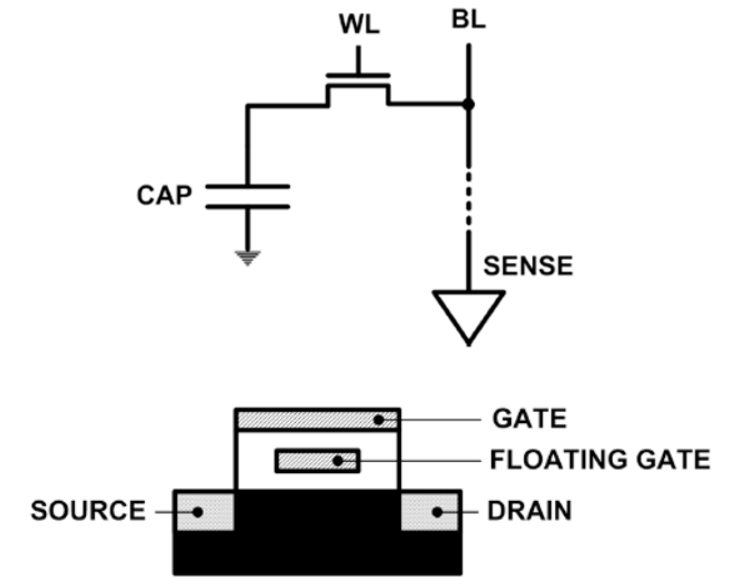
Double Data Rate (DDR) *periods < skew*

- Achieve high-speed data transfer
 - Data words are transferred on the rising and falling edges of CLK
 - Command & address inputs are still registered on the rising edge of CLK
- DDR1: 2n-prefetch
- DDR2: 4n-prefetch
- DDR3: 8n-prefetch



DRAM Scaling Problem

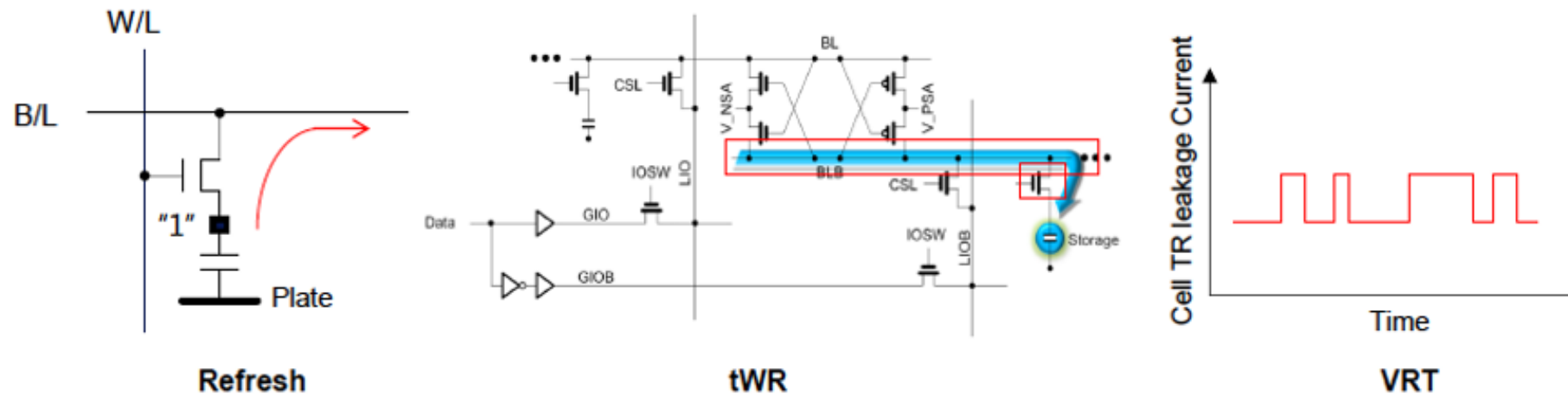
- DRAM stores charge in a **capacitor** (charge-based memory)
 - Capacitor must be large enough for **reliable sensing**
 - Access transistor should be large enough for low leakage and high retention time
- Difficult charge placement and control
 - Flash: floating gate charge
 - DRAM: capacitor charge, **transistor leakage**
- Data retention and ^{refresh} reliable sensing becomes difficult as **charge** storage unit size reduces
- As Memory Scales, It Becomes Unreliable
 - **Alpha Particle-Induced Soft Errors**, or simply soft errors – random and non-recurring.



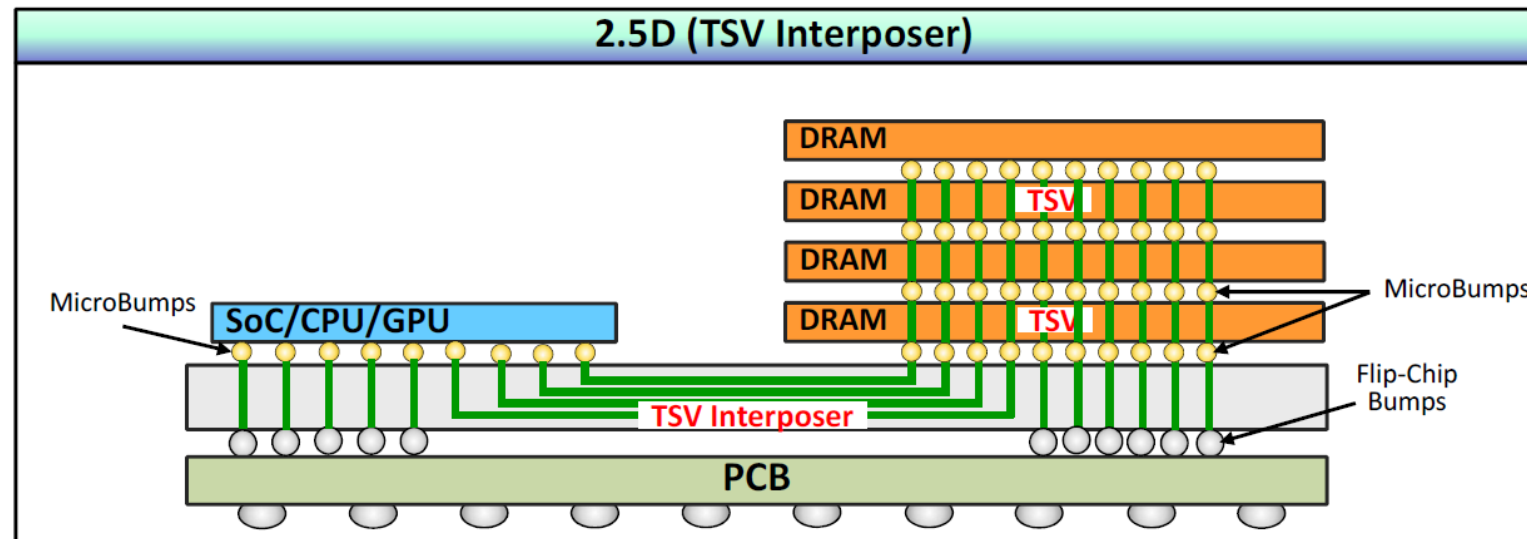
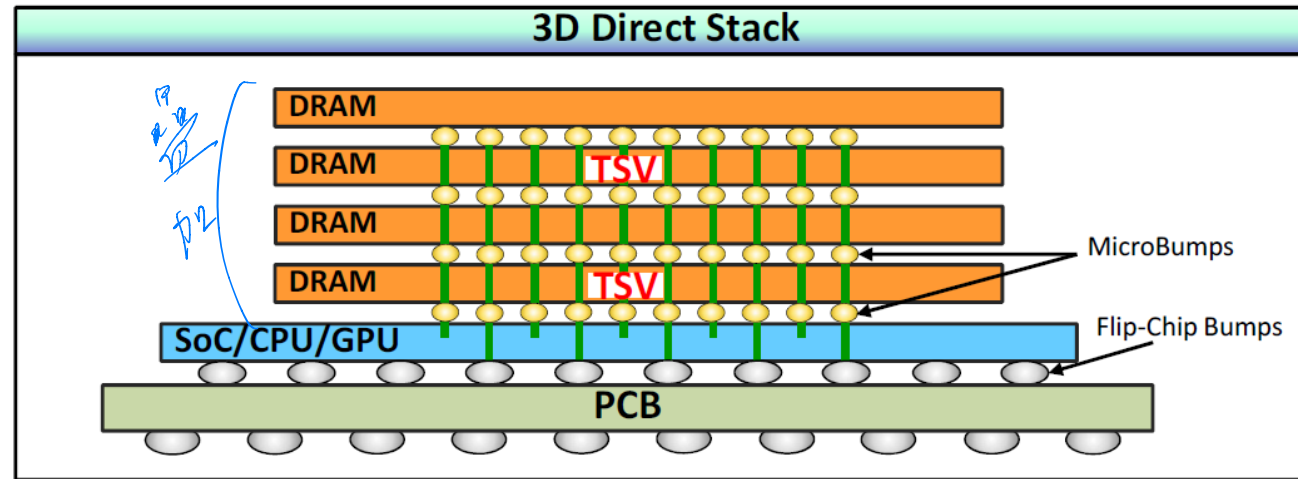
When Process Scaling down



- Frequent Refresh due to
 - Cell capacitance decreases
 - Leakage current of cell access transistors increases
- Write Recovery increases (t_{WR})
 - completion of a valid write operation, before an active bank can be precharged
 - Contact resistance between cell capacitor and access transistor increase
 - On-current of cell access transistor decreases
 - Bitline resistance increases
- Variable retention time (VRT) - Need VRT-aware refresh



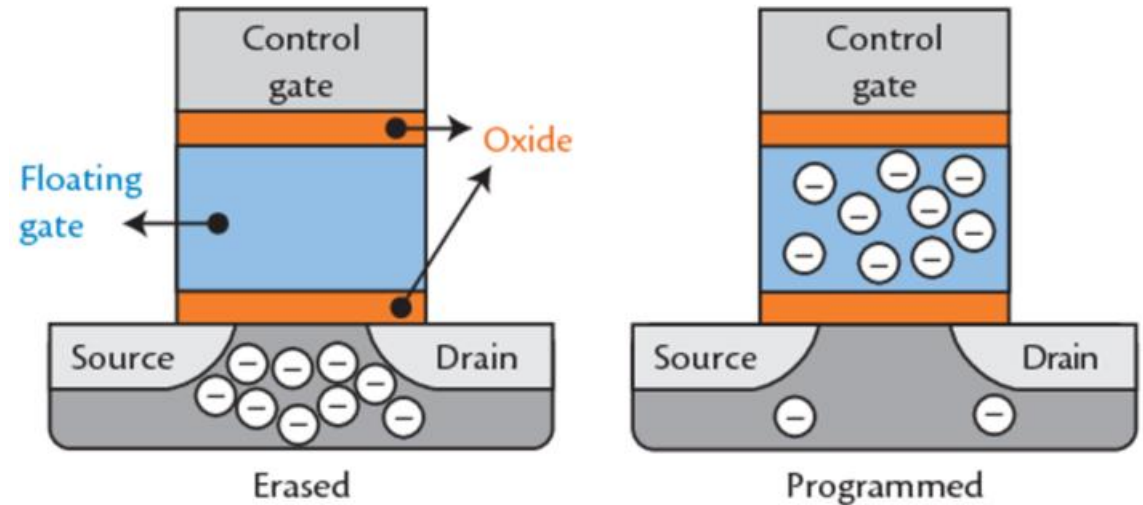
Memory Stacking – 3D/2.5D TSV



Non-volatile Memory

Flash Programming Technology

- Pro
 - Nonvolatile
 - Reprogrammable like SRAM FPGA
 - Area efficient (single floating gate)
 - Lower power
- Con
 - Limited reprogramming -> wearing
 - Non-standard CMOS process



Flash Memory - Floating Gate Device

A type of EEPROM

Non-volatile, solid state technology

Information is stored in an array of memory cells made from floating-gate (FG) transistor

Stacked Gate NMOS Transistor

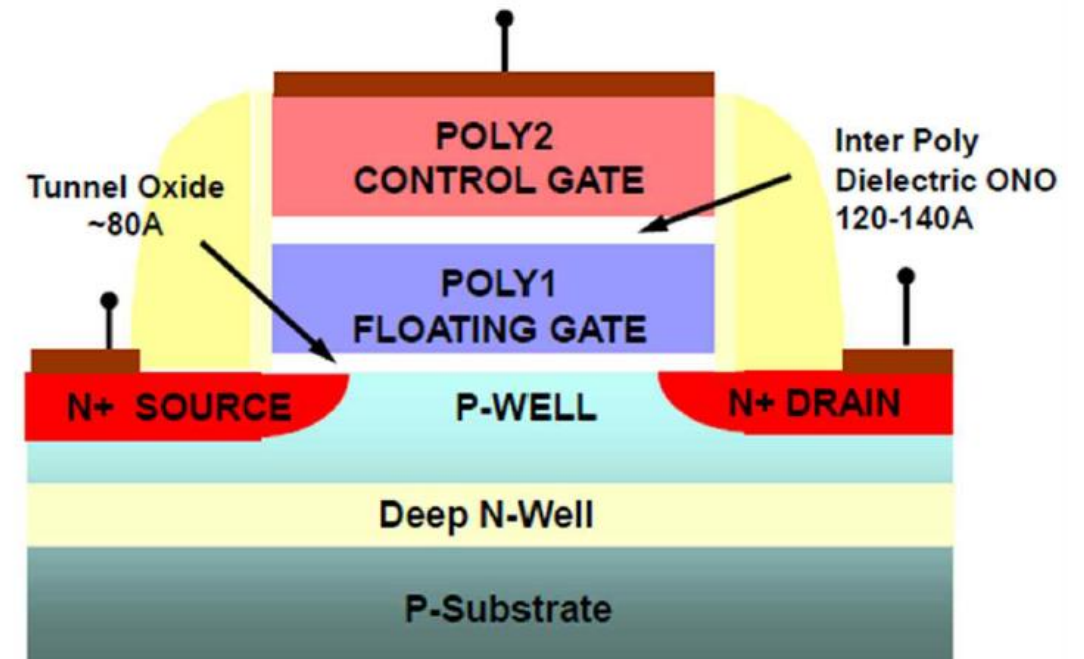
Poly1 Floating Gate for charge storage

Poly2 Control Gate for accessing the transistor

Silicon dioxide for Gate oxide (Tunnel oxide)

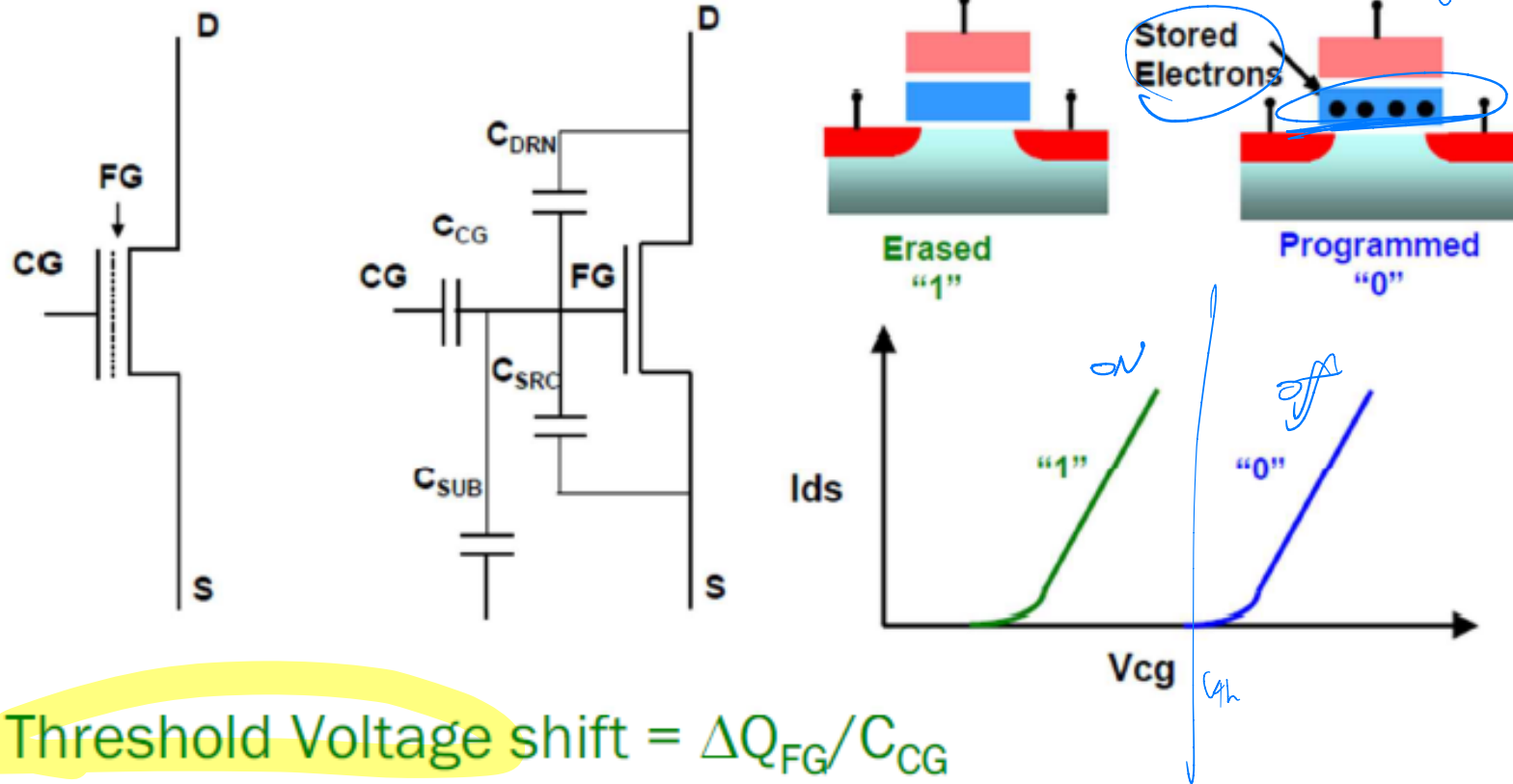
Oxide-Nitride-Oxide (ONO) for the inter Poly Dielectric

Source/Drain Junction optimized for Program/Erase



Floating Gate Flash Memory Device

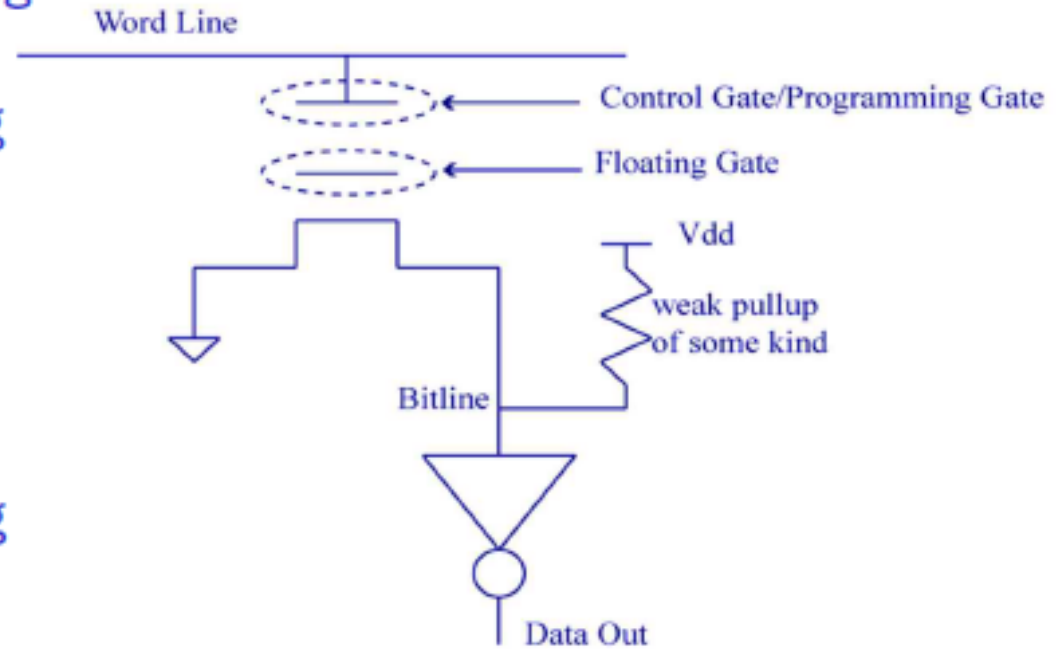
apply diff V_{th}



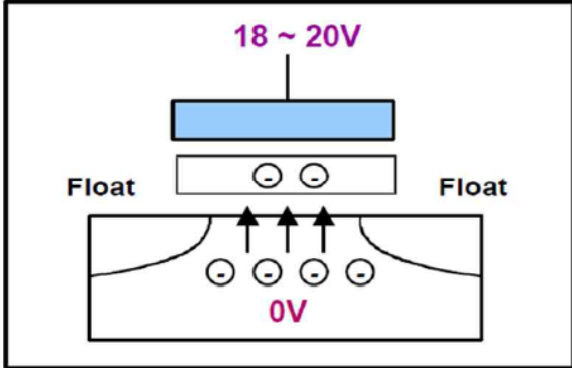
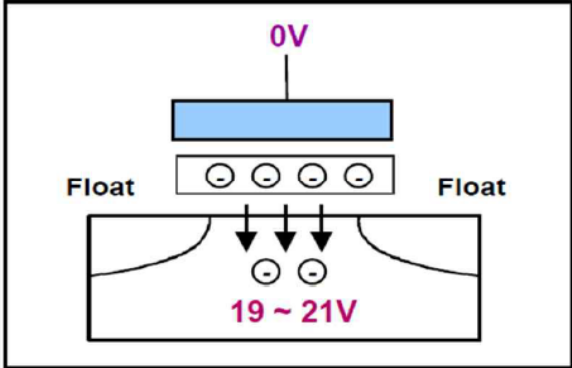
- Threshold Voltage shift = $\Delta Q_{FG}/C_{CG}$
- Programming = Electrons Stored on the FG = High V_{th}
- Erasing = Remove electrons from the FG = Low V_{th}

Floating Gate Operation

- **Not programmed** *V_{th} low \rightarrow on*
 - No electronics trapped on floating gate
 - $WL=1$ turns on transistor, pulling Bitline low
 - Data out = 1
 - As floating gate has no effect
- **Programmed** *V_{th} high*
 - Electronics is trapped on floating gate
 - Increase threshold voltage
 - Transistor remains off when $WL=1$
 - Bitline=1 and Data out=0

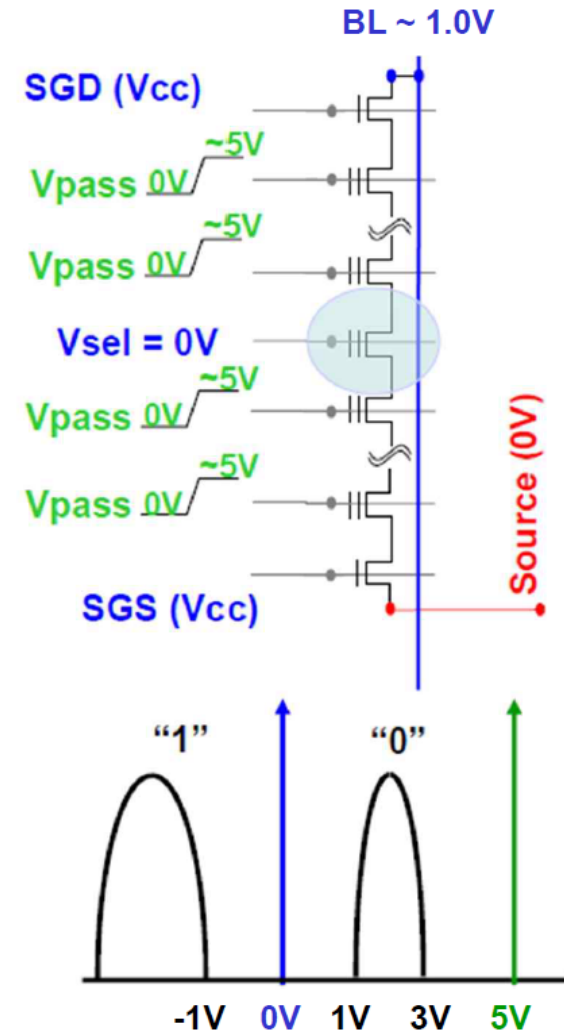


NAND Program / Erase

| Program | Erase |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  <p>The diagram shows a cross-section of a NAND cell during programming. A blue gate is at the top with a voltage of 18 ~ 20V. Below it is a white floating gate with two circles representing electrons. The word 'Float' is on either side. Below the floating gate is the channel region with four circles representing electrons and a voltage of 0V. Three upward-pointing arrows indicate electron tunneling from the channel to the floating gate.</p> |  <p>The diagram shows a cross-section of a NAND cell during erasing. A blue gate is at the top with a voltage of 0V. Below it is a white floating gate with four circles representing electrons. The word 'Float' is on either side. Below the floating gate is the channel region with two circles representing electrons and a voltage of 19 ~ 21V. Three downward-pointing arrows indicate electron tunneling from the floating gate to the channel.</p> |
| <ul style="list-style-type: none">● Use F-N Tunneling● Channel Inversion | <ul style="list-style-type: none">● Use F-N Tunneling● Channel Accumulation |
| <ul style="list-style-type: none">-> No DD Source (Easy Device Scaling)-> No BTBT Current (Easy Voltage Scaling) | |

NAND Flash Cell Read

- Erased Cell V_{th} : $< -1V$
- Programmed Cell V_{th} : $1-3V$
- To Read a Cell:
 - Bitline is pre-charged to $\sim 1.0V$
 - V_{cc} is applied to the select gates of the string (block) to be selected
 - All the deselected WL on this string (block) are biased at V_{pass} ($\sim 5V$) which has to be higher than the highest program V_{th}
 - Selected WL is held at $0V$
 - If the selected cell $V_{th} < 0V$, the string will conduct and the bitline is discharged and Sense Amp reads the data as “1”
 - If the selected cell $V_{th} > 0V$, the string will not conduct and the bitline stays at $\sim 1.0V$ and Sense Amp reads the data as “0”

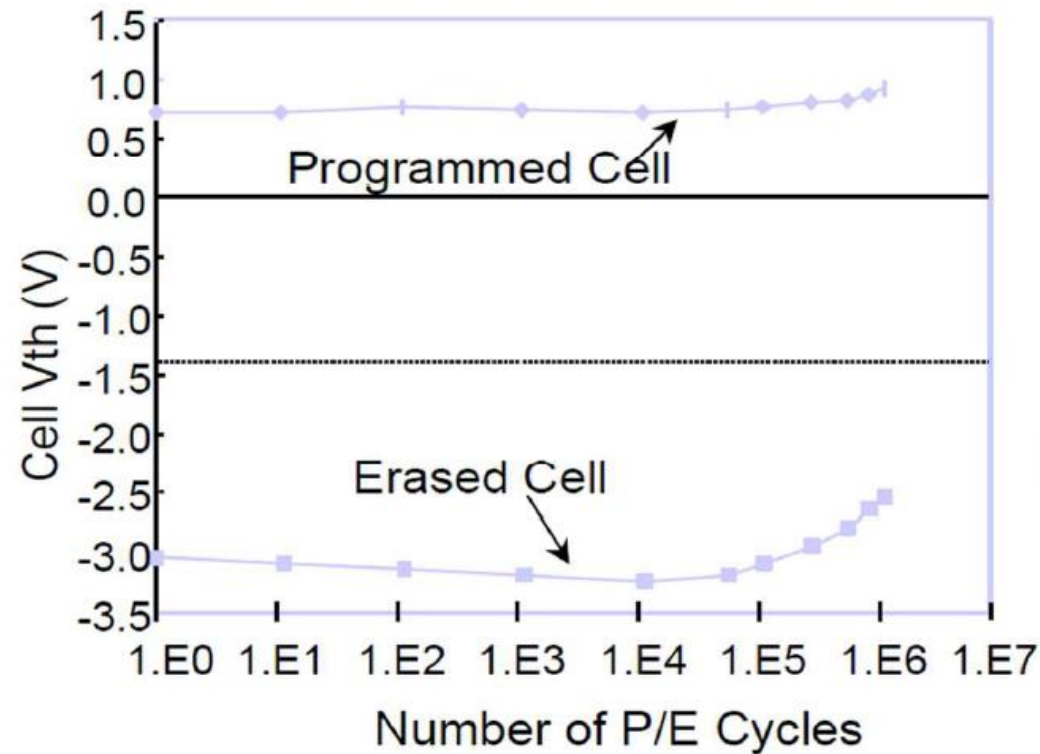


NAND Flash Lifetime

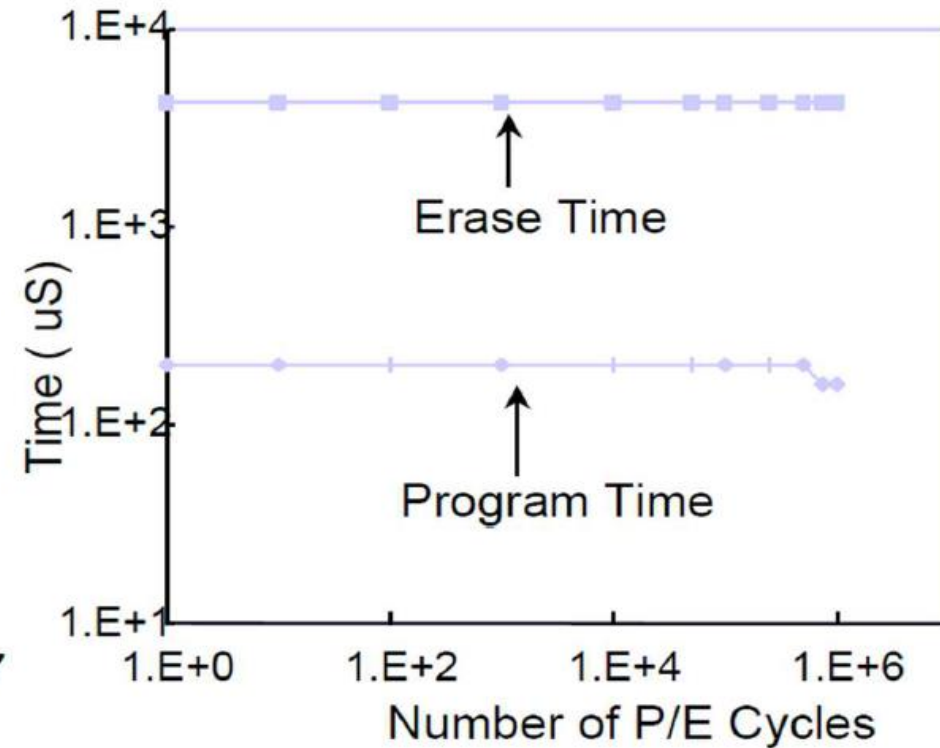
of erase operation is limited due to degradation

- Cell Vth Shift
- Program/Erase Time Variation
- Wear leveling & ECC

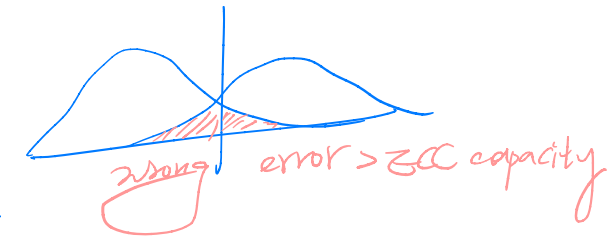
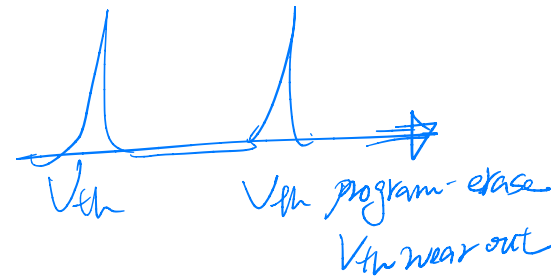
Cell Vth Shift

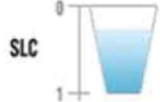
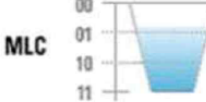
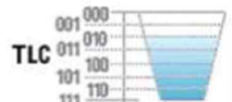


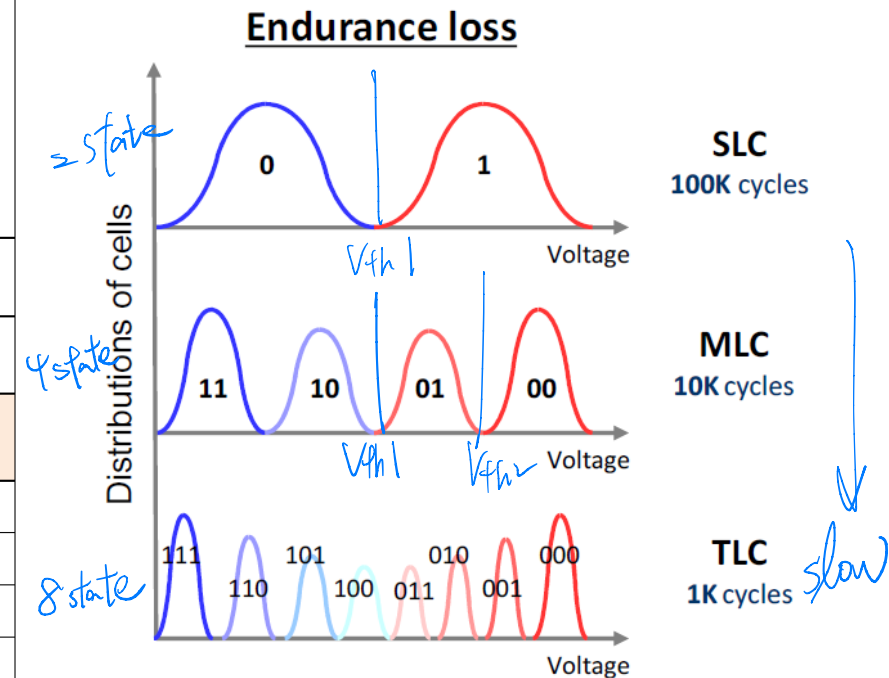
Program/ Erase Time Variation



Multi-Level NAND Flash



| Item | SLC Single Level Cell | MLC Multi Level Cell | TLC Triple Level Cell |
|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Architecture | SLC Flash has only two states: erased (empty) or programmed (full).  | MLC Flash has four states: erased (empty), 1/3, 2/3, and programmed (full).  | TLC Flash has eight states: erased (empty), 1/7, 2/7, 3/7, 4/7, 5/7, 6/7 and programmed (full).  |
| Read/Write Speed | Faster | Slow | Slower |
| Power Consumption | Lower | High | Higher |
| Endurance (P/E Cycles) | 100K | 1~3K | Under 1K |
| Data Retention | 10 Years | 10 Years | 5 Years |
| Density | TLC > MLC > SLC | | |
| Cost | \$\$\$\$+ | \$ | \$- |
| Application | 1. IPC, embedded, automation 2. Commercial application | 1. POS, Kiosk system 2. Commercial application | 1. Consumer application 2. MP3, Pen-Drive, Storage gift |



Enhance Endurance (& Effective Capacity)

redundancy parity bit \rightarrow capacity \downarrow

- ECC (error correction code)

- In data transmission, the sender adds carefully selected **redundant data** to its message
- Allow the receiver to detect and correct errors (within some bound) without the need to ask the sender for additional data

- **Wear leveling** ↗ 每次 W/R 不同 block \rightarrow 均匀使用
P/E ↑ File block 不可固定 \rightarrow (需 W/R)

- Wear leveling attempts to work around these limitations by arranging data so that erase and re-writes are distributed evenly across the medium
- In this way, no single erase block prematurely fails due to a high concentration of write cycles