EECS 151/251A SP2022 Discussion 11

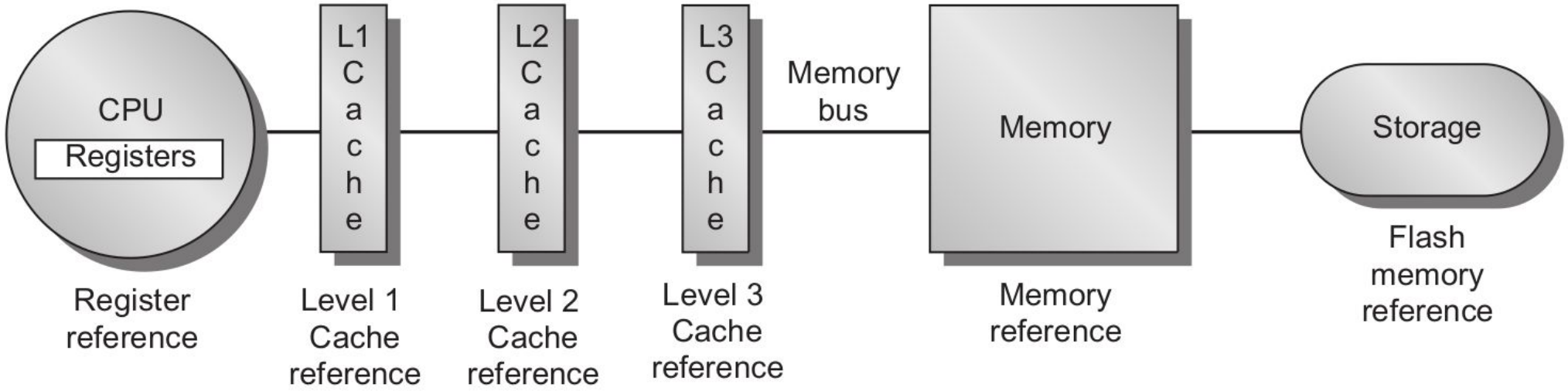
GSI: DIMA NIKIFOROV, YIKUAN CHEN

Agenda

- Caches
- SRAM Decoders
- Other Memories (DRAM, CAM, Flash)

Caches

Memory Hierarchy Overview



		Register reference	Level 1 Cache reference	Level 2 Cache reference	Level 3 Cache reference	Memory reference	Flash memory reference
Laptop	Size:	1000 bytes	64 KB	256 KB	4-8 MB	4–16 GB	256 GB-1 TB
	Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	50-100 uS
Desktop	Size:	2000 bytes	64 KB	256 KB	8-32 MB	8–64 GB	256 GB-2 TB
	Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	50-100 uS

(B) Memory hierarchy for a laptop or a desktop

Hennessy, John L., and David A. Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

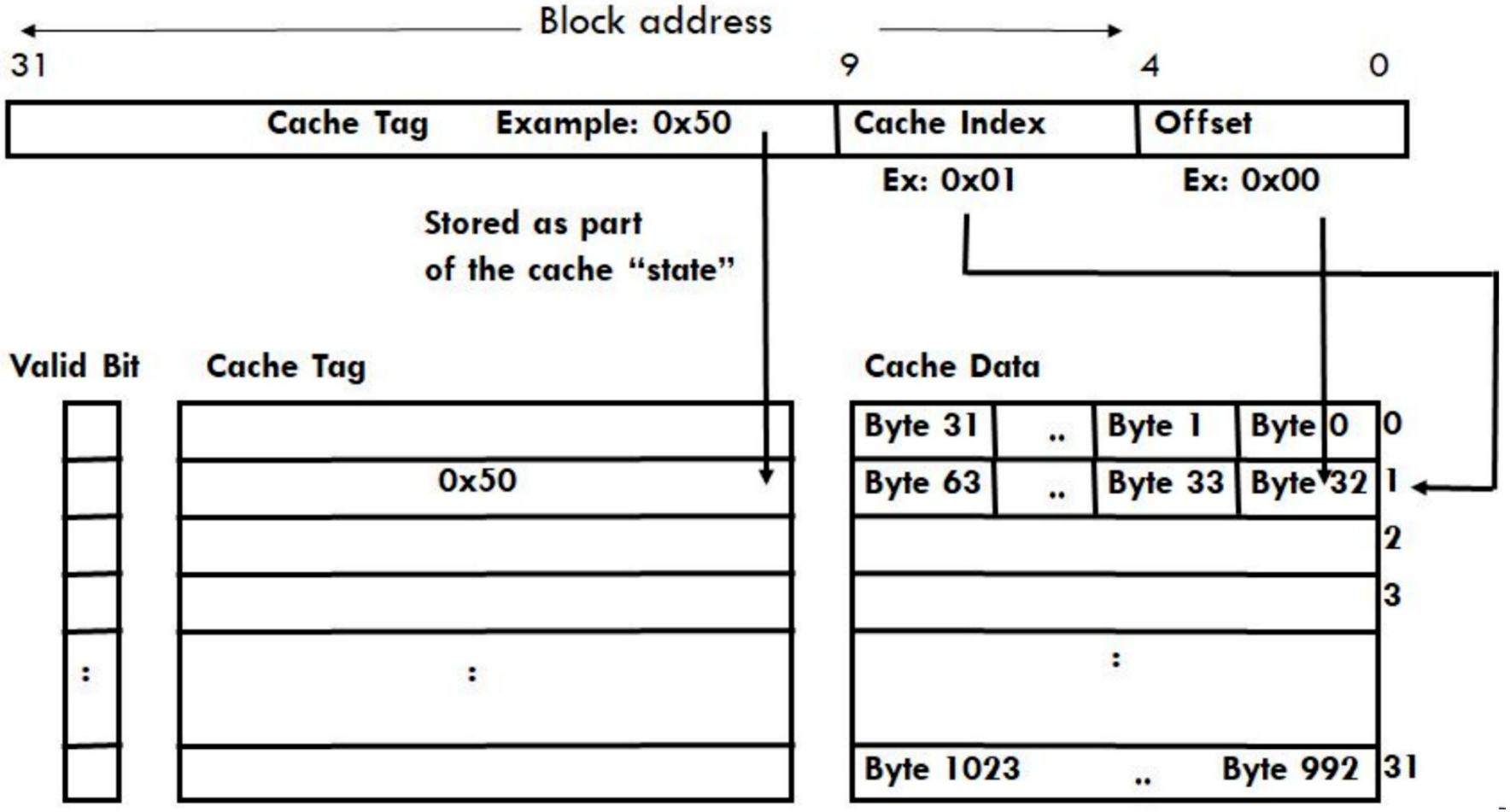
What is a Cache?

- A cache holds **commonly used memory data**.

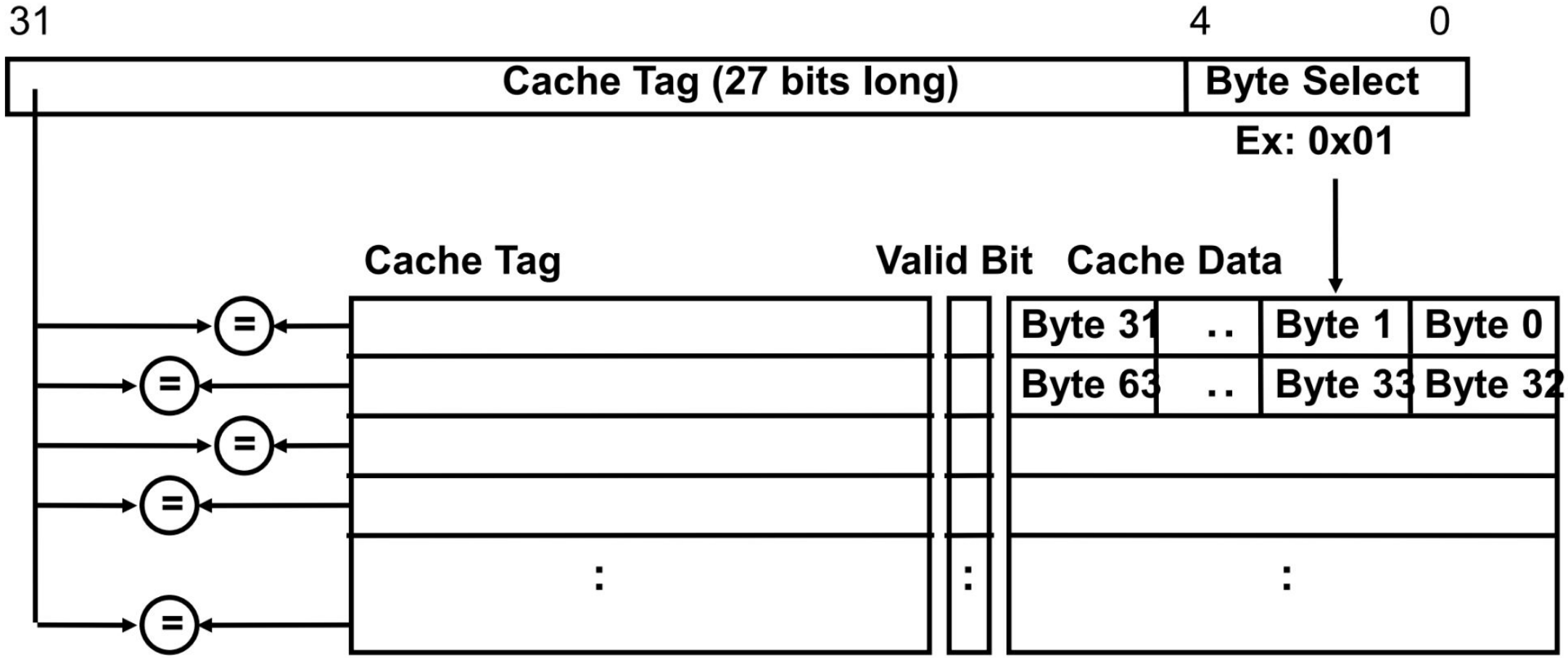
- An **ideal cache** would anticipate all of the data needed by CPU, and fetch it from main memory ahead of time, so that it has zero miss rate.

- Caches are specified by:
  - **C:** capacity
  - **b:** block size: Granularity of memory loaded into cache
  - **B:** number of blocks (B = C / b)
  - **S:** number of sets
  - **N:** degree of associativity

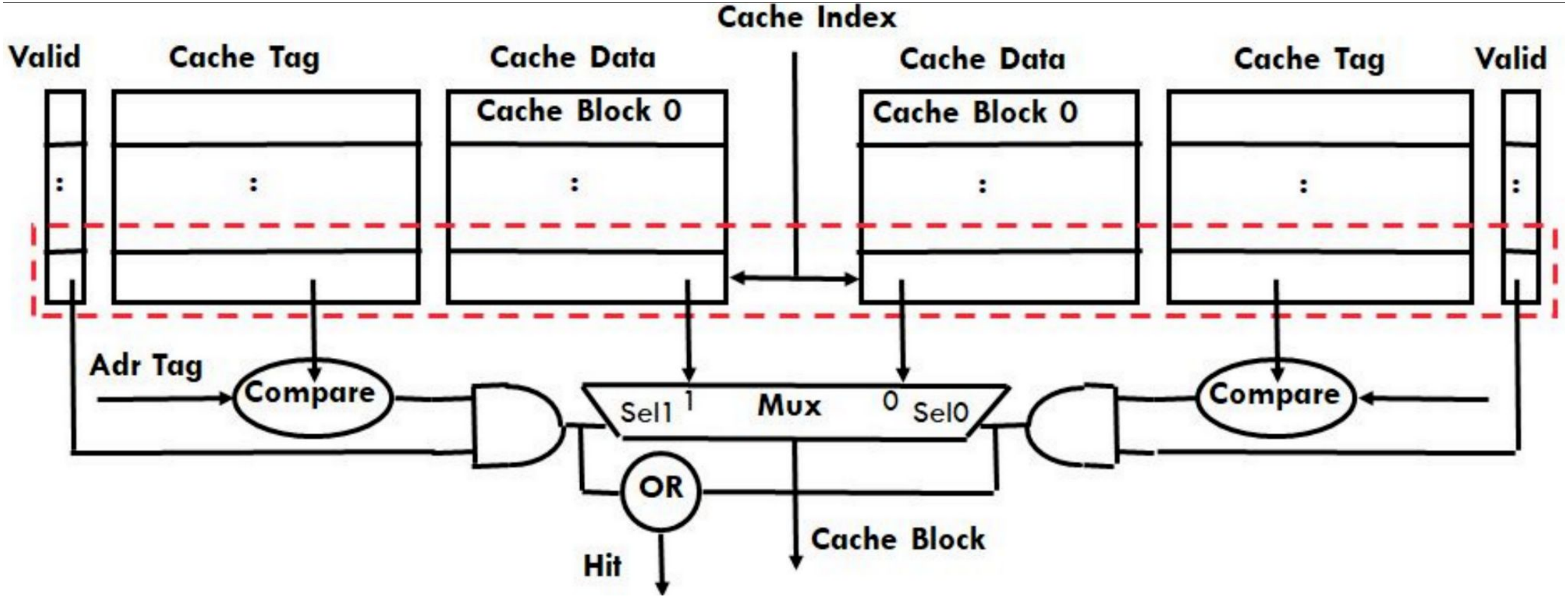
Direct Mapped Cache



Fully Associative Cache

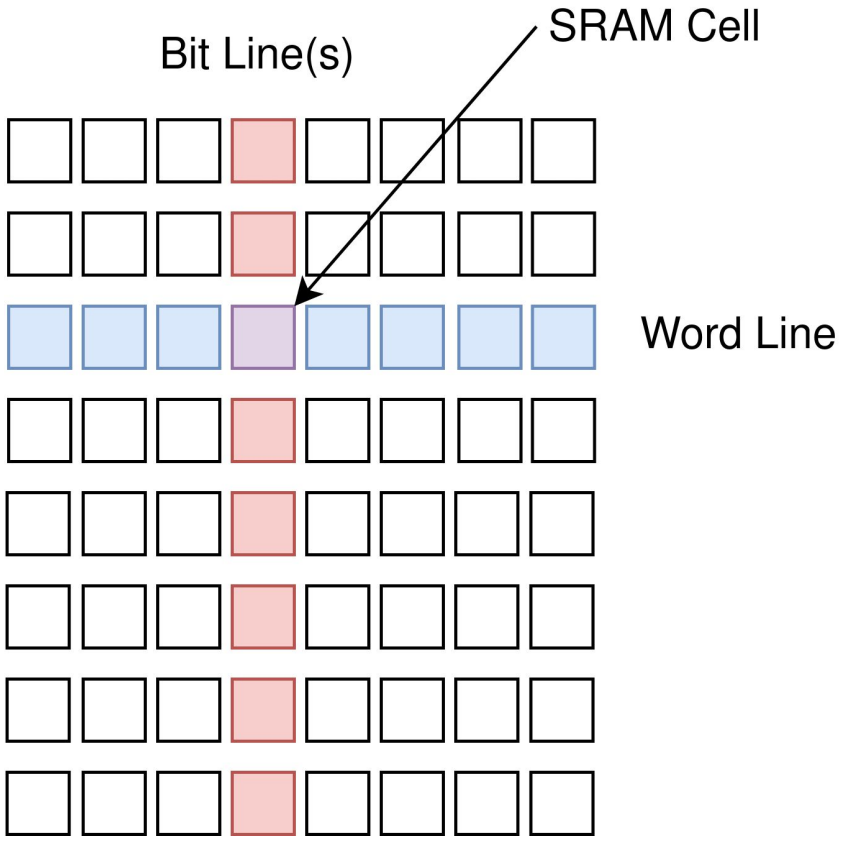


N-Way Set Associative Cache

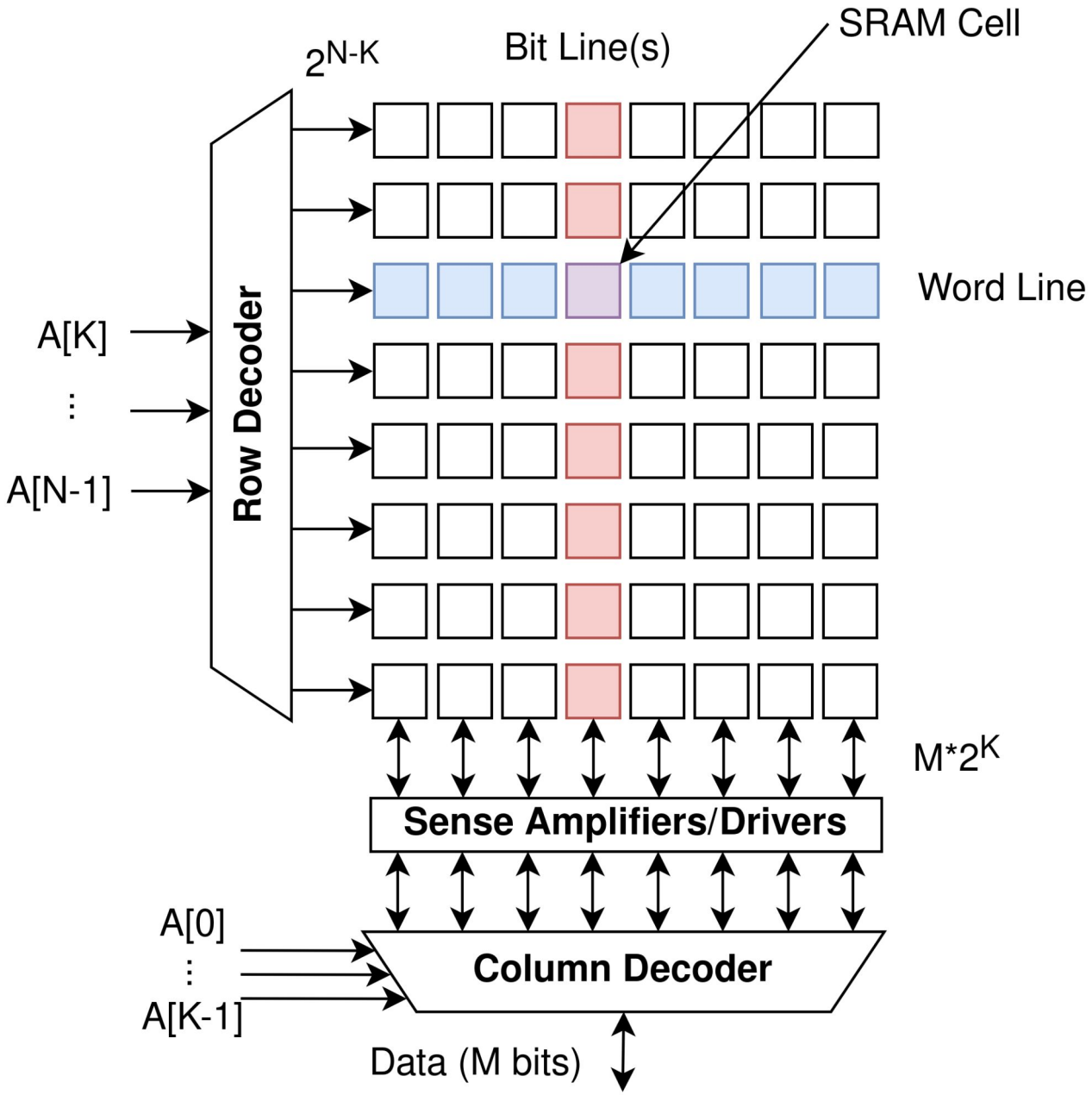


SRAM Decoders

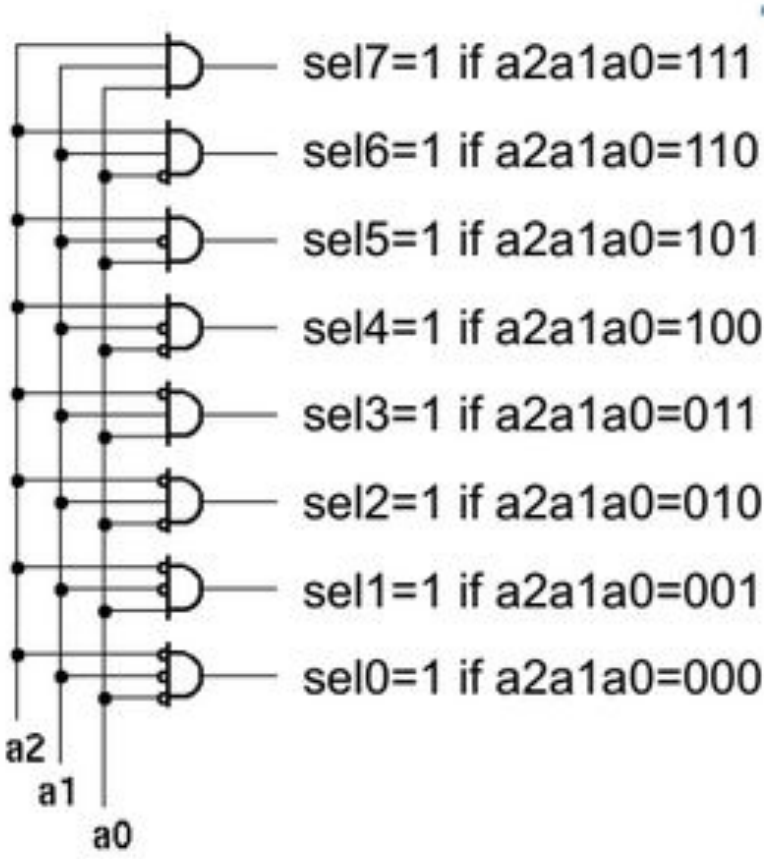
SRAM Structure:

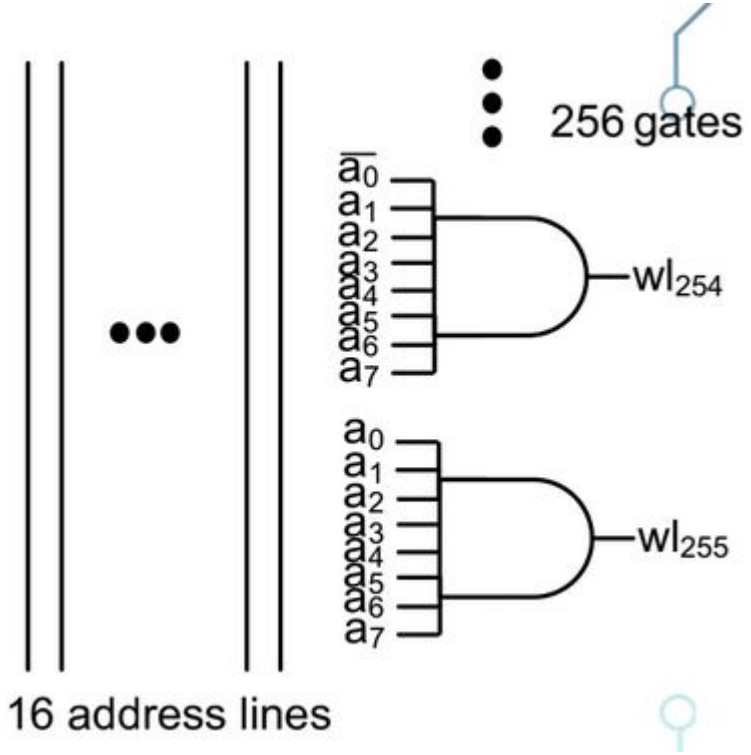


SRAM Structure:

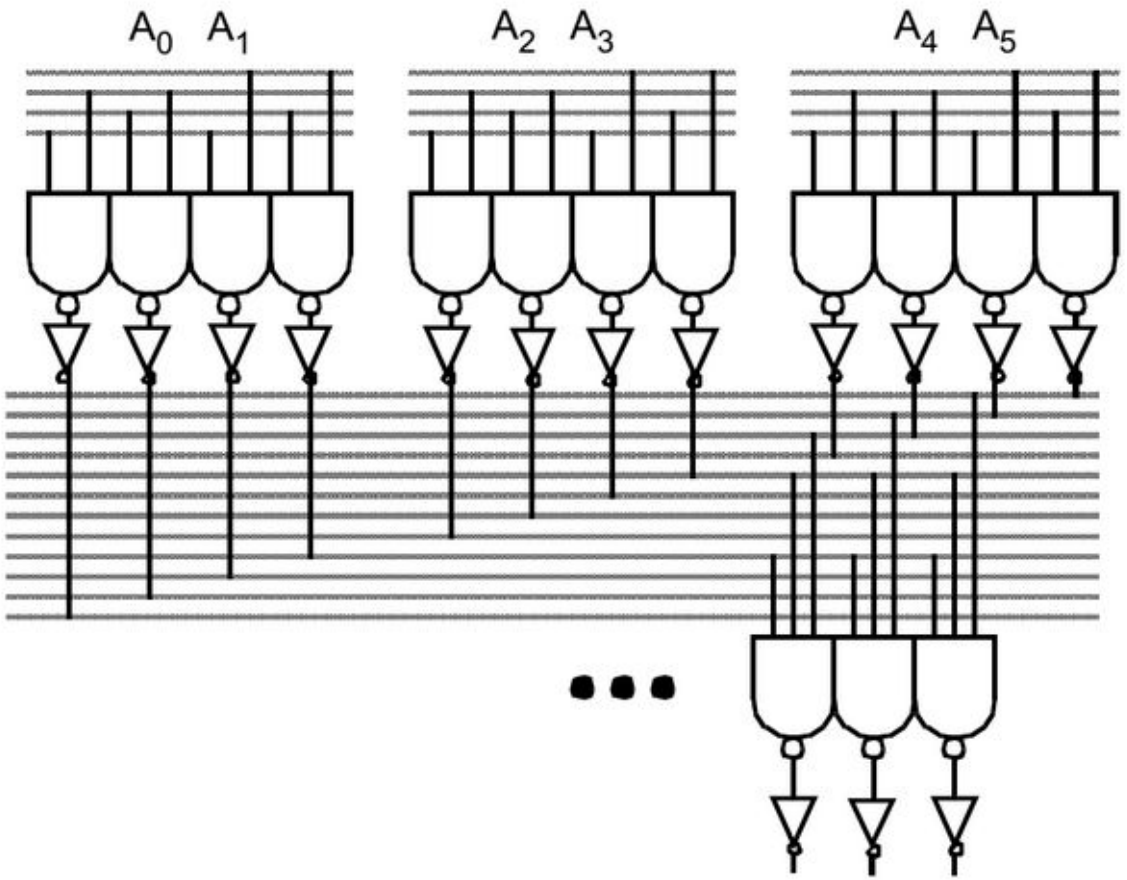


Row Decoder: Naive Implementation



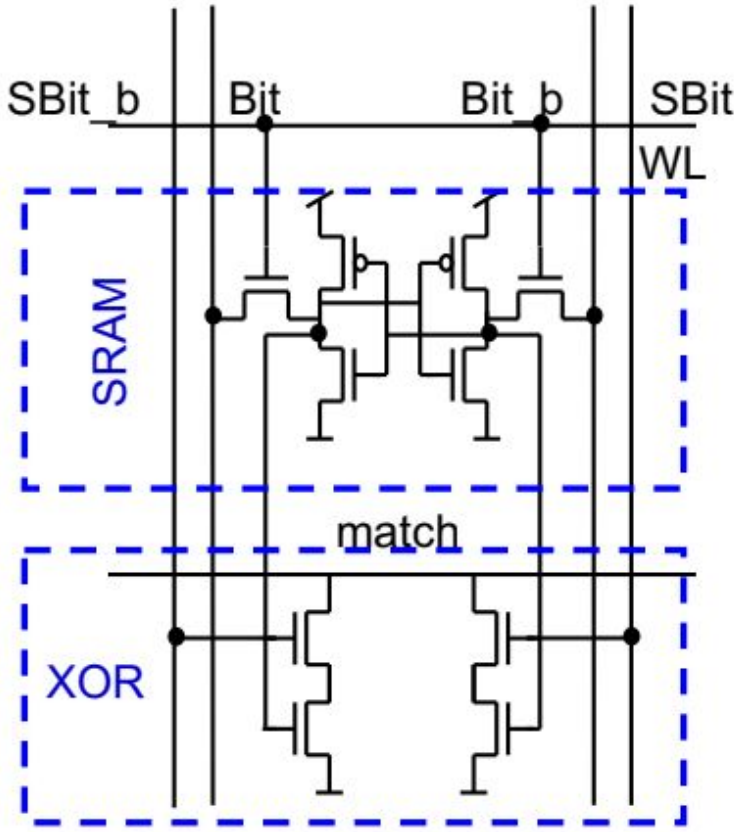


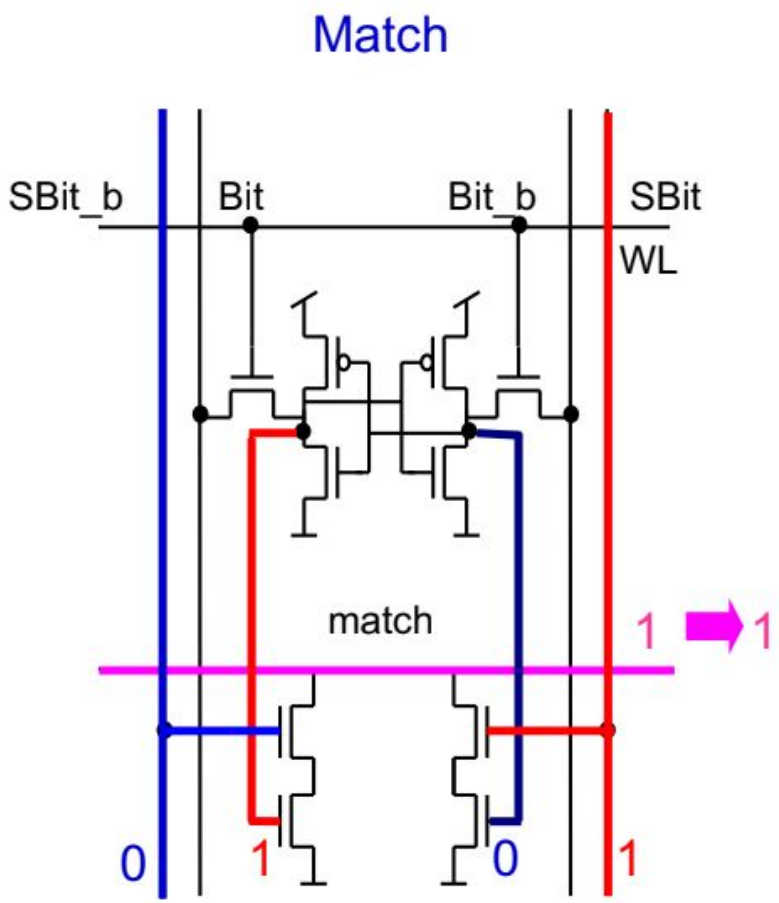
Predecoder + Decoder

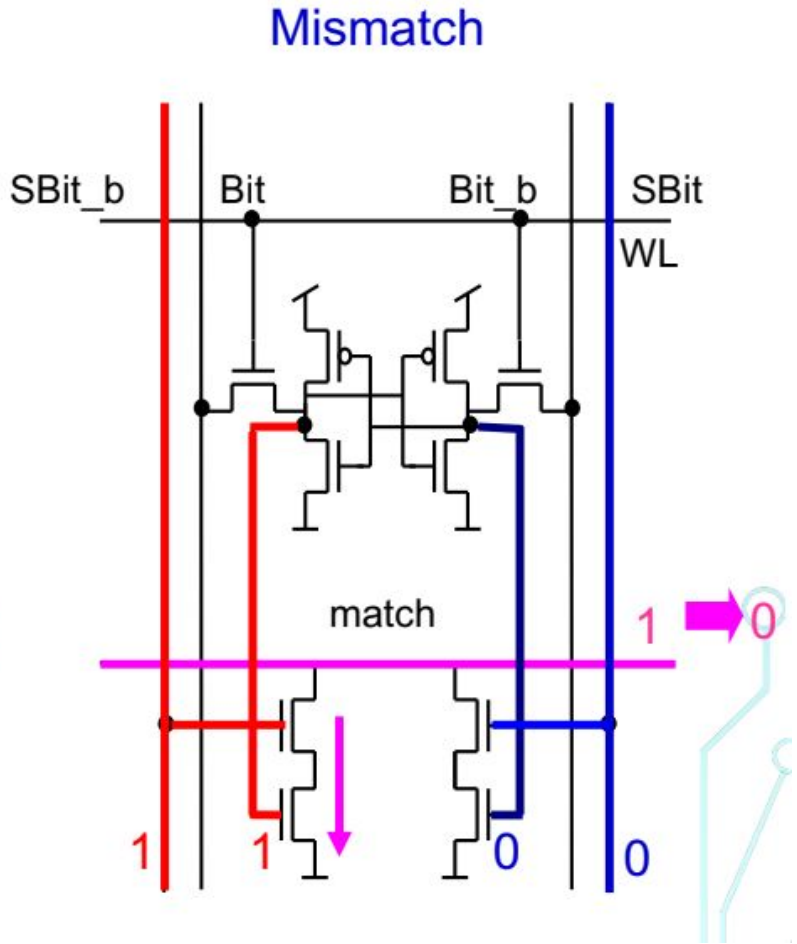


Other Memories

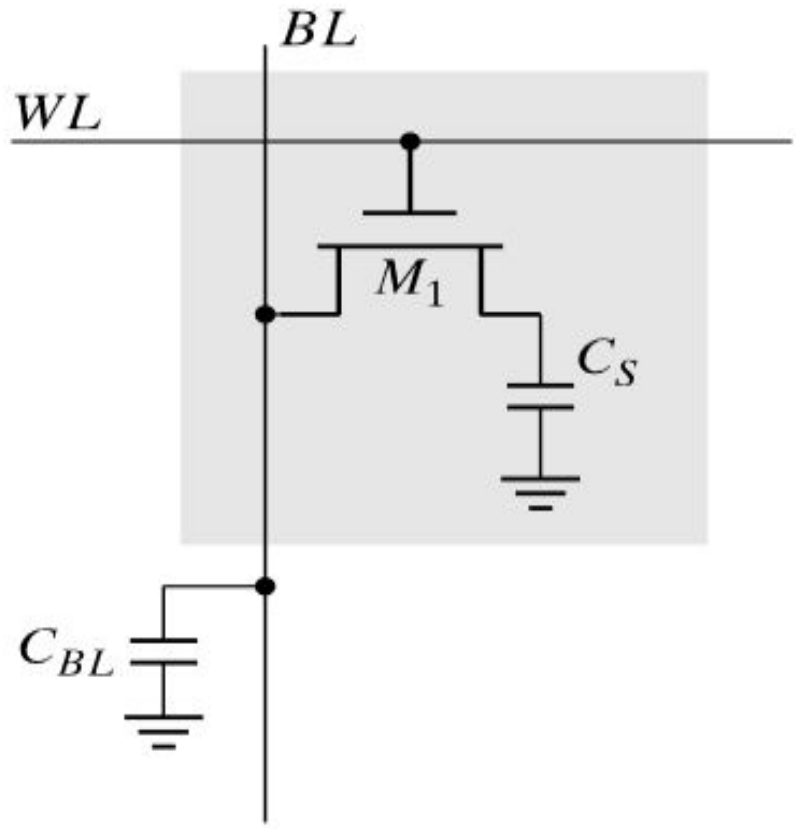
Content Addressable Memory (CAM)

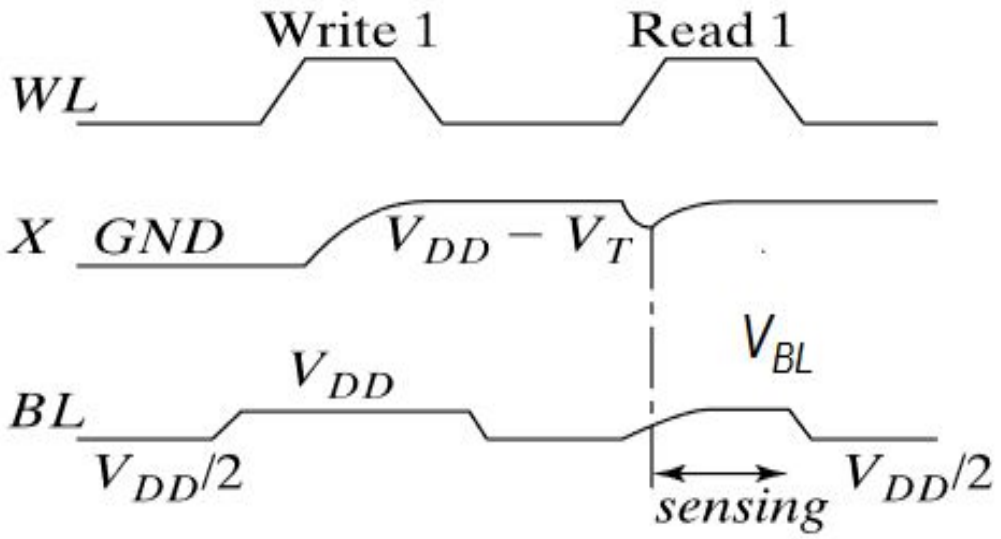






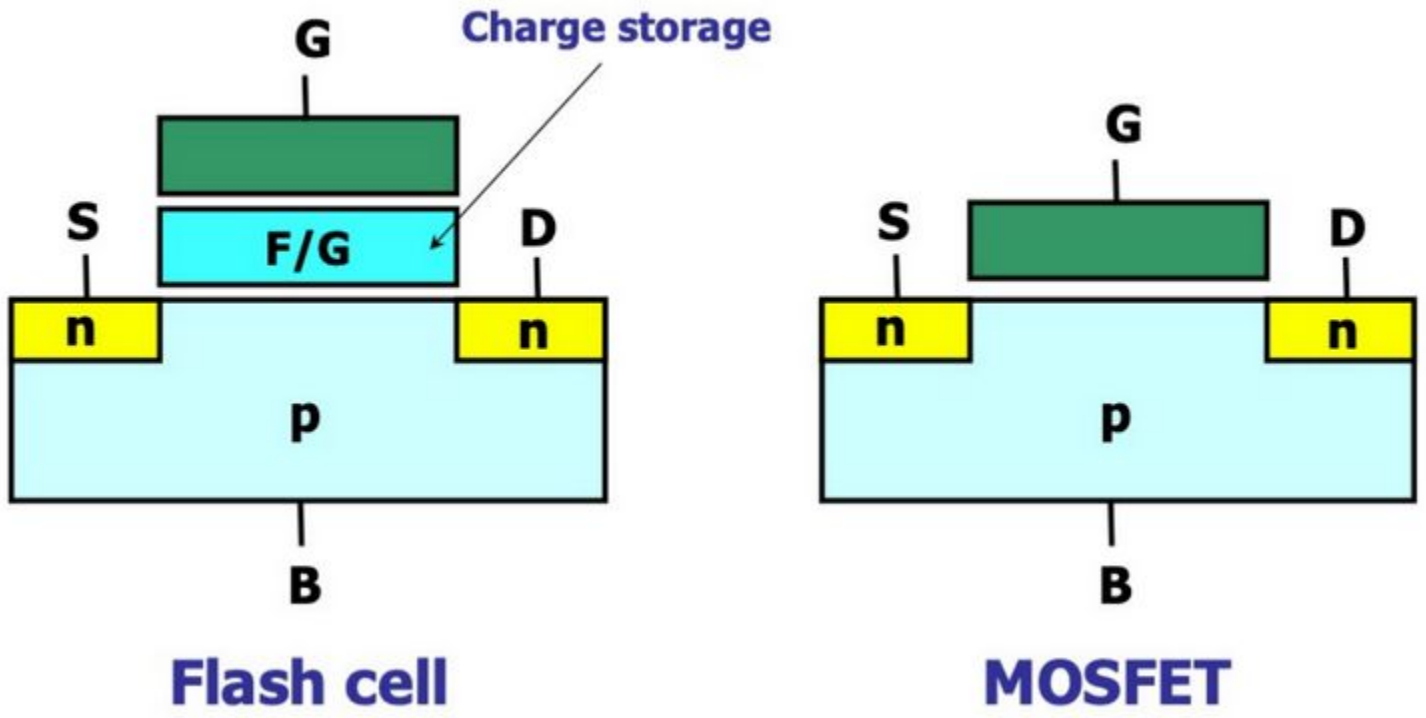
DRAM



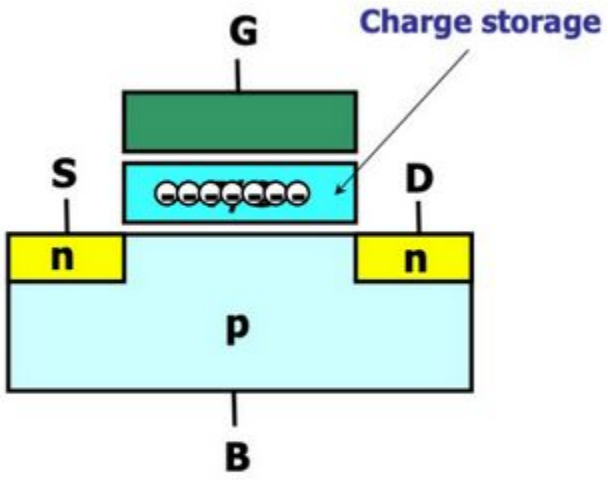


$V_{BIT} = 0$ or $(V_{DD} - V_T)$

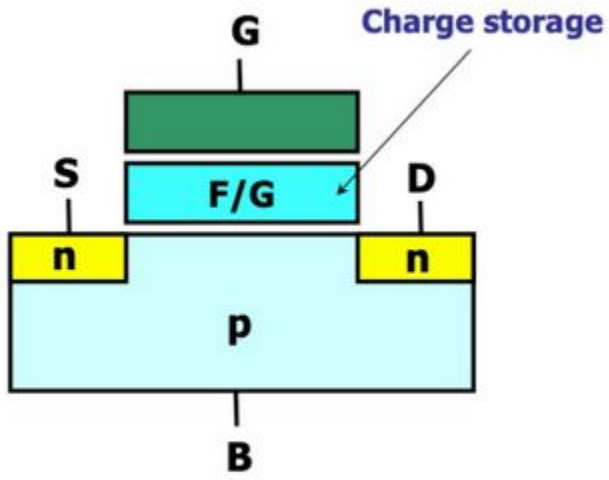
Flash Overview



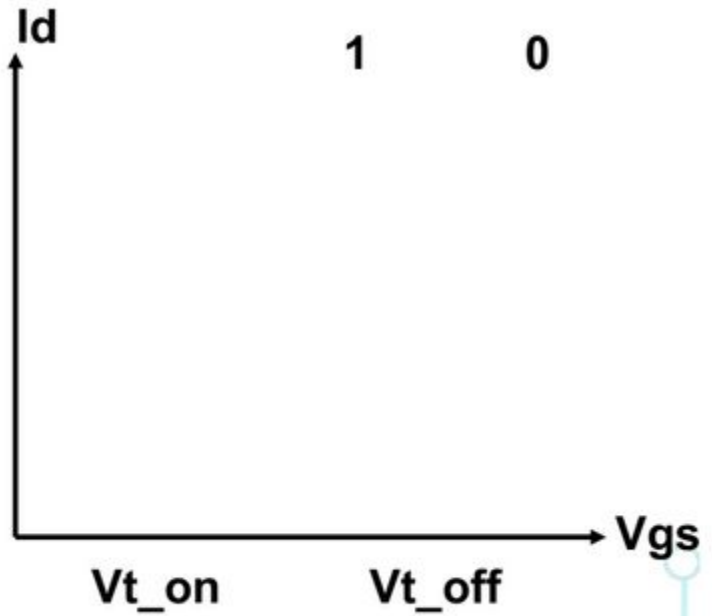
Flash Overview



Storing 0
OFF State

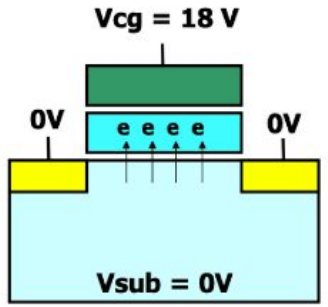


Storing 1
ON State



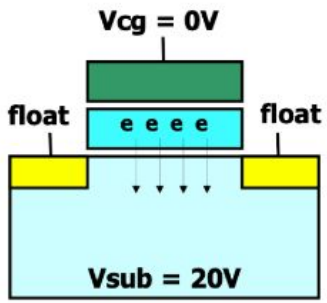
Flash Write

- Step 1: Erasing.
  - Erase all the FG transistors to set them to 1
  - Apply a negative voltage to the gate -> Electrons flow from the floating gate to the substrate.
- Step 2: Programming
  - Reprogram the appropriate FG transistors to set them to 0
  - Apply a high voltage to the gate -> Electrons are tunneled onto the floating gate.



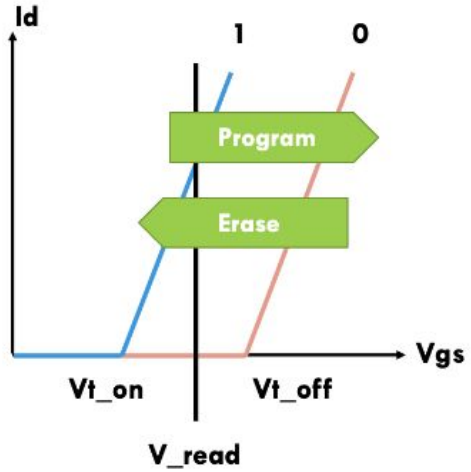
Program
F-N Tunneling

Off cell
(Solid-0)



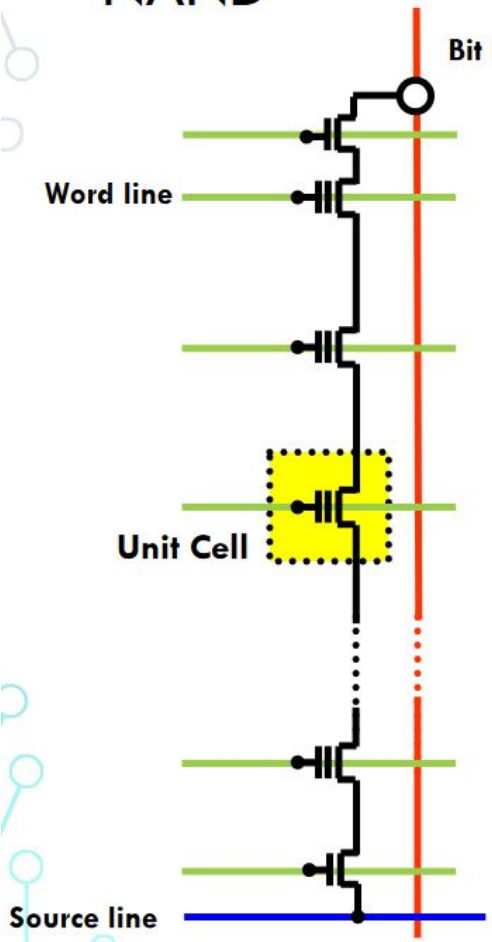
Erase
F-N Tunneling

On cell
(Solid-1)



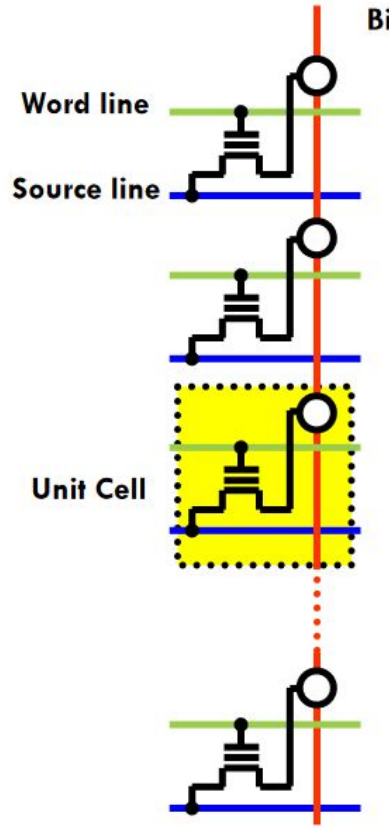
NAND vs NOR Flash

- NAND



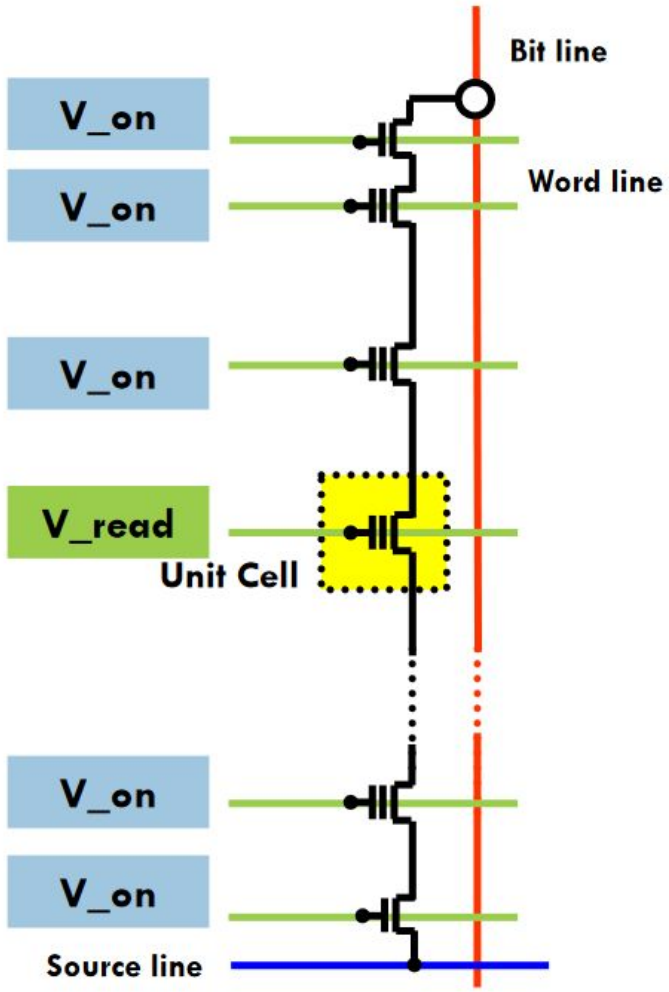
NAND:
- High Density
- Used for data storage
  - USB drives
  - Memory cards
  - SSD

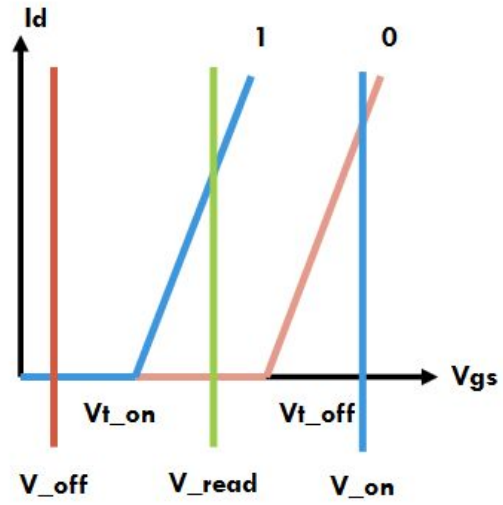
- NOR



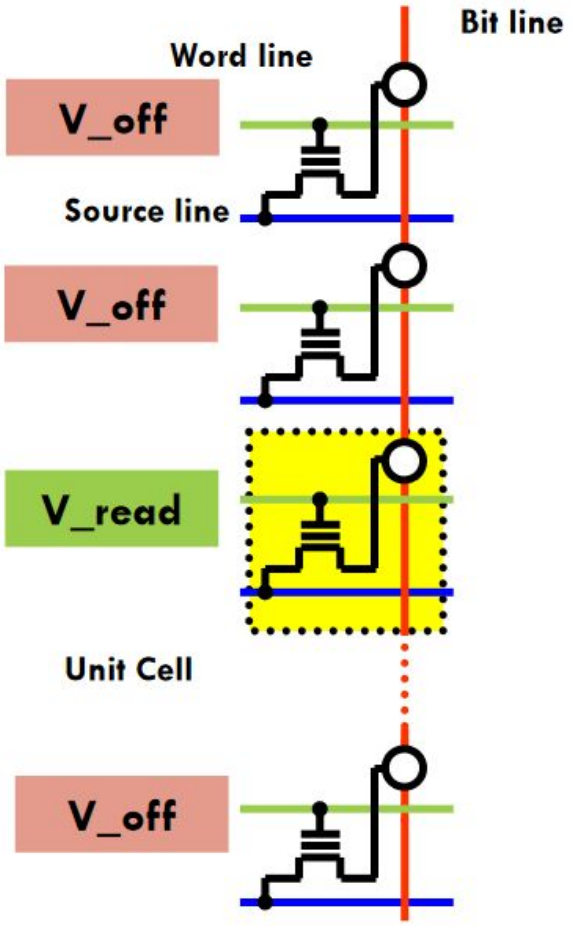
NOR:
- Lower Latency
- Used for code storage
  - Embedded systems

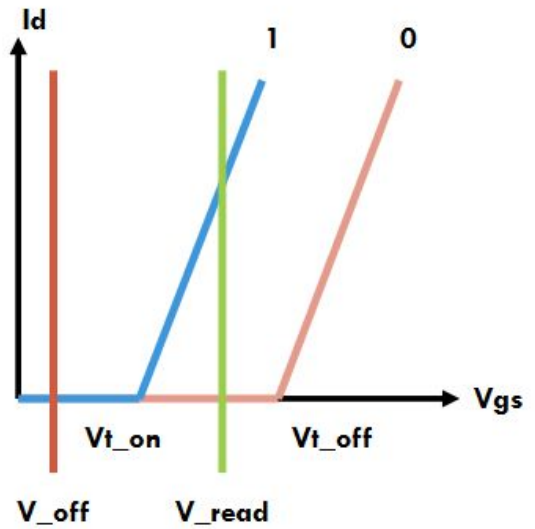
NAND Flash Read





NOR Flash Read





Questions?