



香港中文大學

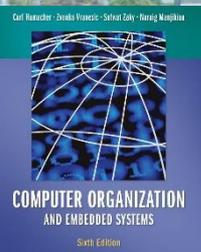
The Chinese University of Hong Kong

CSCI2510 Computer Organization

Lecture 06: Memory Hierarchy

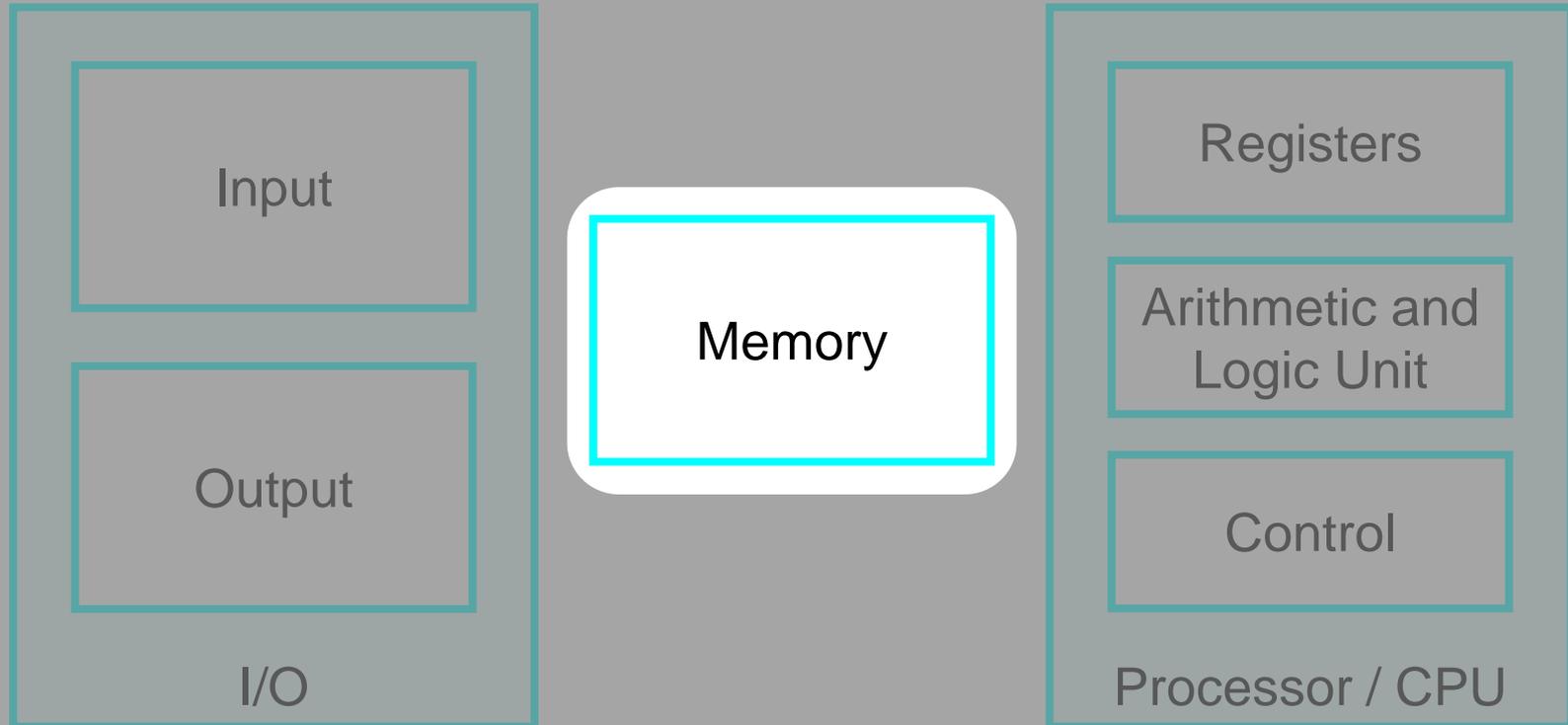
Ming-Chang YANG

mcyang@cse.cuhk.edu.hk



Reading: Chap. 8.1~8.5

Basic Functional Units of a Computer



- **Input:** *accepts* coded information from human operators.
- **Memory:** *stores* the received information for later use.
- **Processor:** *executes* the instructions of a program stored in the memory.
- **Output:** *reacts* to the outside world.
- **Control:** *coordinates* all these actions.

- An Overview of Memory
- Memory Technologies
 - Random Access Memory (RAM)
 - Read-Only Memory (ROM)
 - Non-Volatile Memory (NVM)
- Memory Hierarchy



Why We Need Memory?



- Reason: **Programs** and the **data** must be held in the **memory** of the computer to be executed.

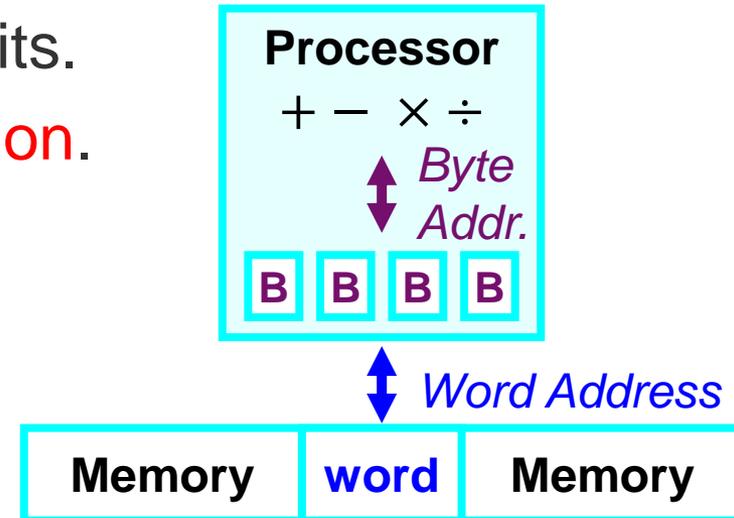
Task Manager Performance tab showing system resource usage. The Memory column is highlighted with a red box, showing 35% usage.

Name	CPU	Memory	Disk	Network
Apps (8) in-use!				
Adobe Acrobat (32 bit)	0%	62.4 MB	0 MB/s	0 Mbps
Google Chrome (2)	0%	147.2 MB	0.1 MB/s	0.1 Mbps
Instant Dictionary (32 bit)	0.2%	21.1 MB	0 MB/s	0 Mbps
Microsoft PowerPoint	0%	282.1 MB	0 MB/s	0 Mbps
Skype (32 bit)	0.1%	67.0 MB	0.1 MB/s	0 Mbps
Snipping Tool	0.5%	4.6 MB	0 MB/s	0 Mbps
Task Manager	0.5%	14.6 MB	0 MB/s	0 Mbps
Windows Explorer	0.4%	60.3 MB	0 MB/s	0 Mbps

Revisit: Memory Basics



- Most machines are **byte-addressable**.
 - Each memory address location refers to a **byte (B)**.
- Memory is designed to store/retrieve in **words**.
 - A **word** is usually of 16, 32 or 64 bits.
 - Reason? **Performance consideration**.

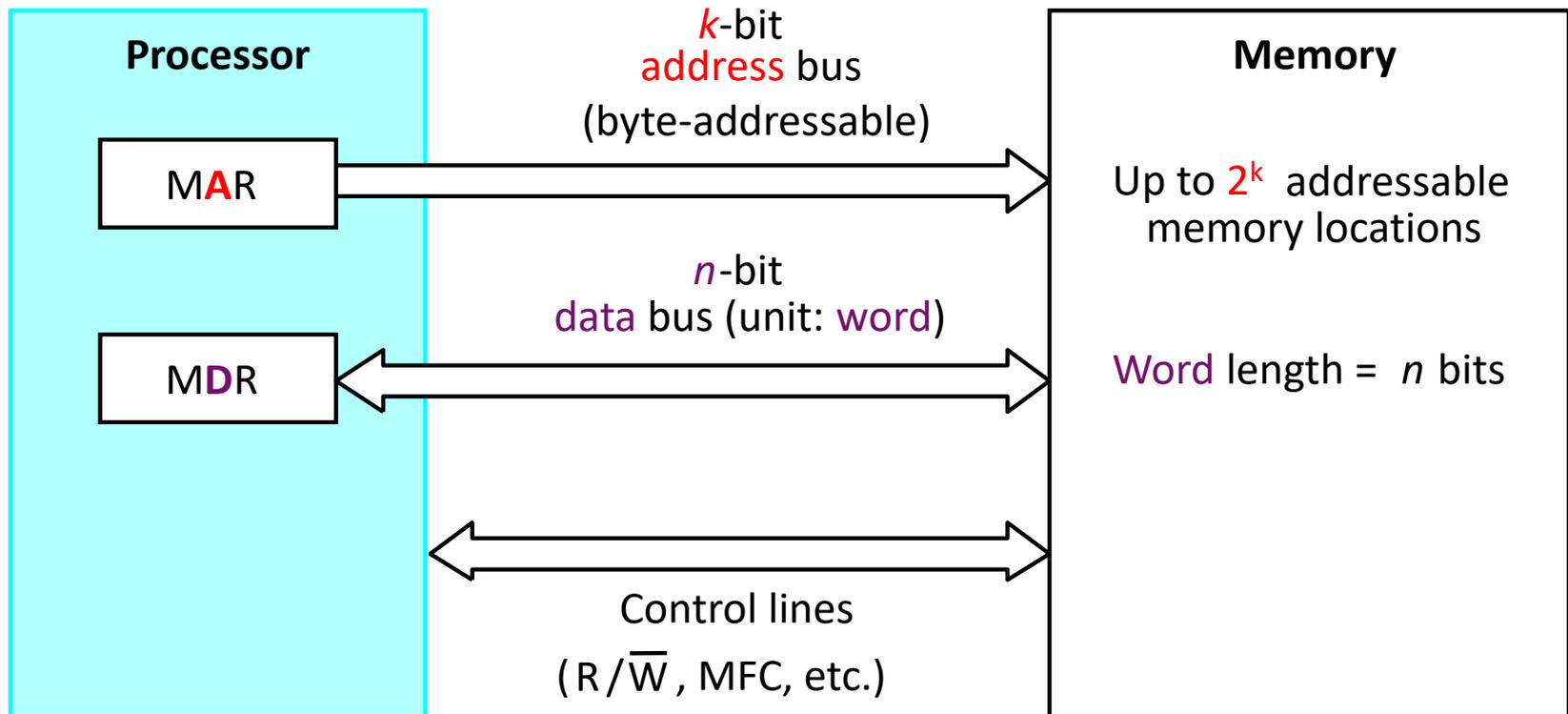


- The maximum size of memory is determined by the **addressing capability**.
 - For example, a 32-bit machine (that uses 32-bit addresses) can utilize a memory that contains up to 2^{32} bytes = 4GB.

Simplified View: Processor-Memory



- Data transferring takes place through MAR and MDR.
 - **MAR**: Memory **A**ddress Register
 - **MDR**: memory **D**ata Register



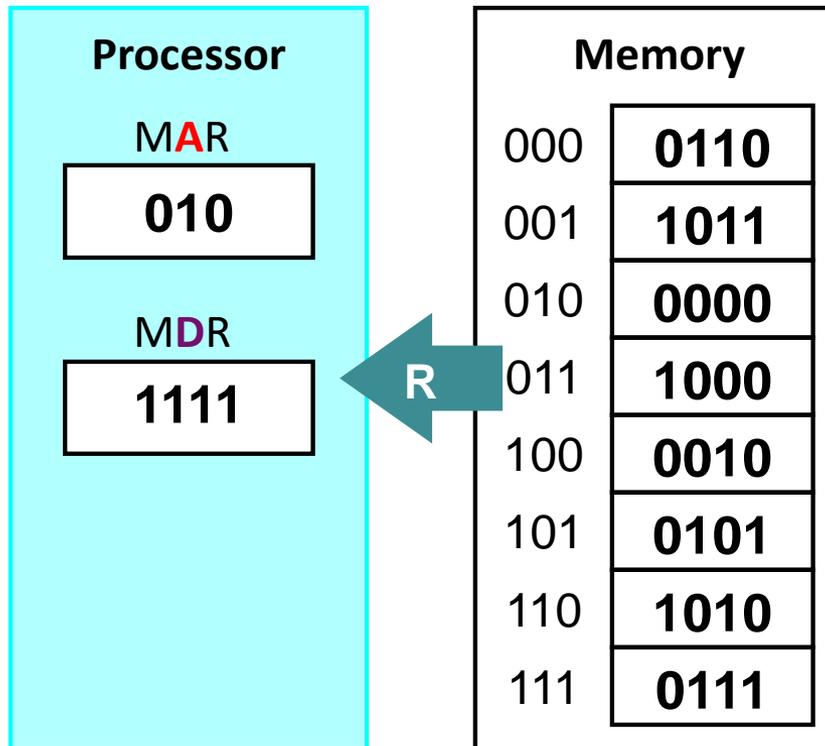
**MFC (Memory Function Completed): Indicating the requested operation has been completed.*

Class Exercise 6.1

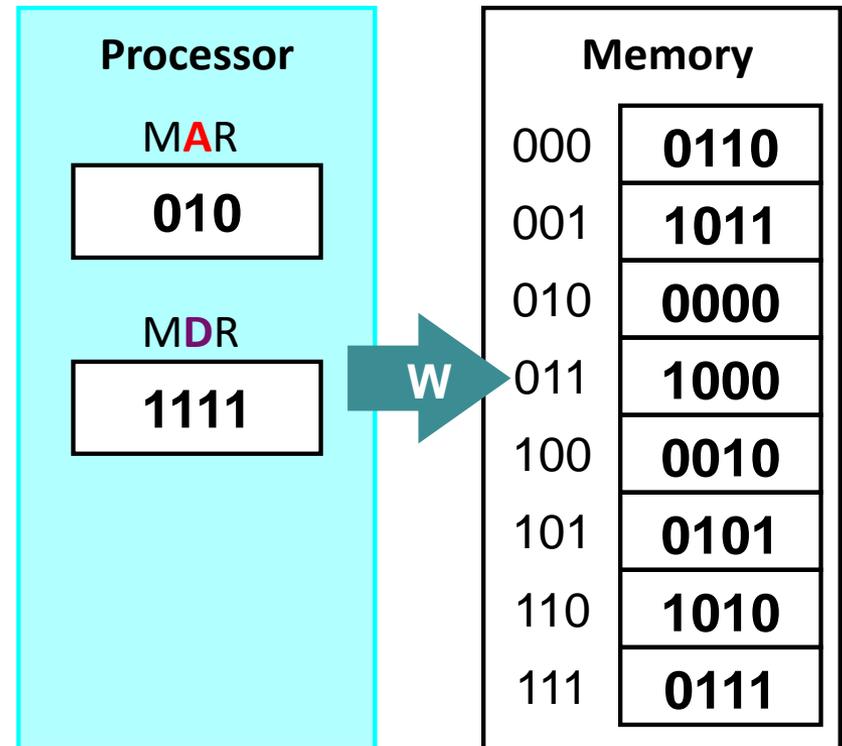
Student ID: _____ Date: _____
Name: _____

- Assume **3-bit address bus** (i.e. $k=3$) and **4-bit data bus** (i.e. $n=4$) are used.
- What will be the contents of MAR, MDR, and the memory after a read or write operation is performed?

(a) Read Operation



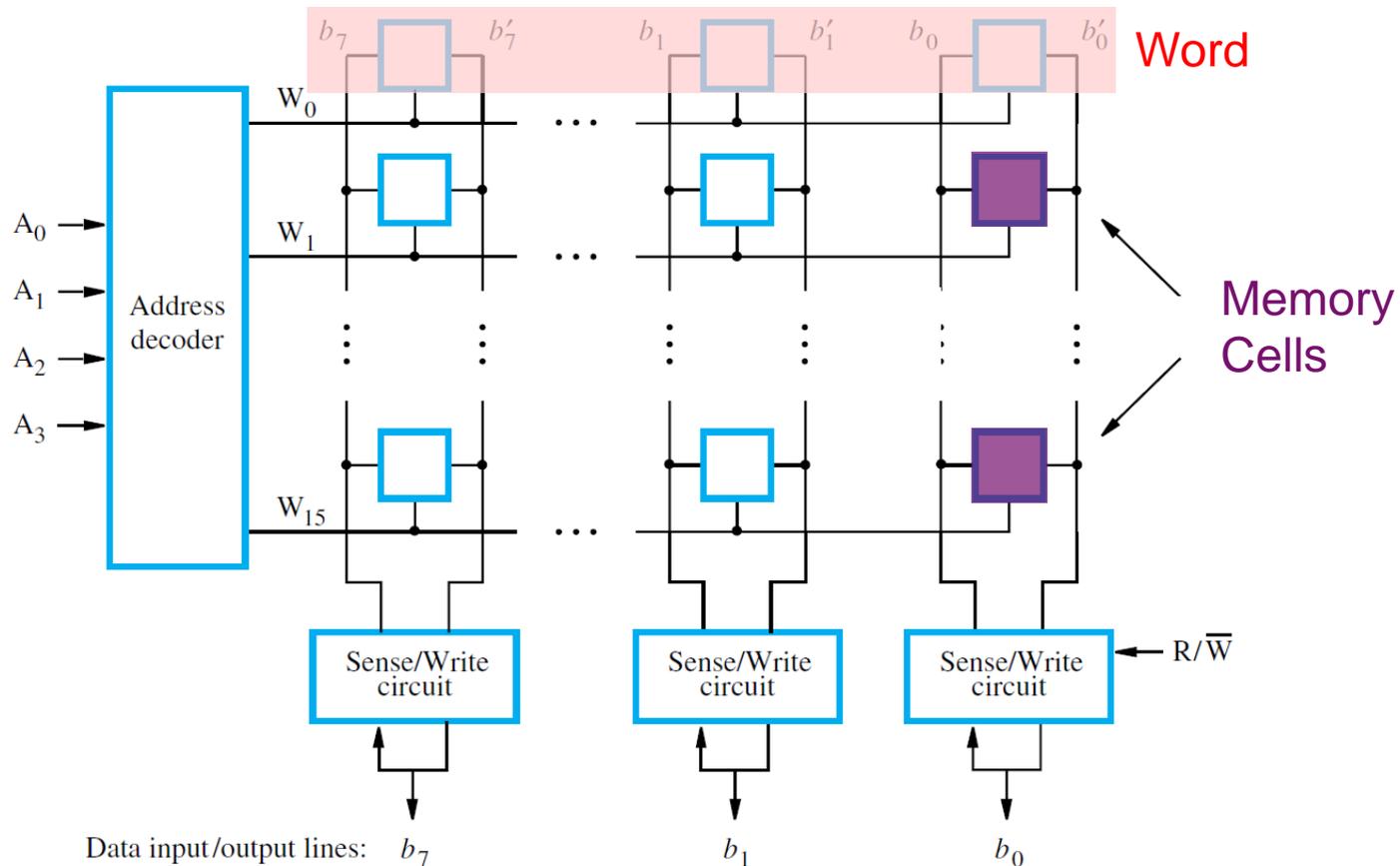
(b) Write Operation



Memory Cell Organization



- Memory cells are usually organized as an **array**:
 - Each **cell** can store one **bit** of information, and
 - Each **row** of cells constitutes a memory **word**.



Class Exercise 6.2



- In the previous example, the small memory circuit contains **16 words**, and each word has **8 bits**.
- How many **bits of data** can be stored in this memory?
- Answer: _____
- How many **bits of address bus** do we needed?
- Answer: _____
- How many **bits of data bus** do we needed?
- Answer: _____
- How many **control lines** do we needed?
- Answer: _____



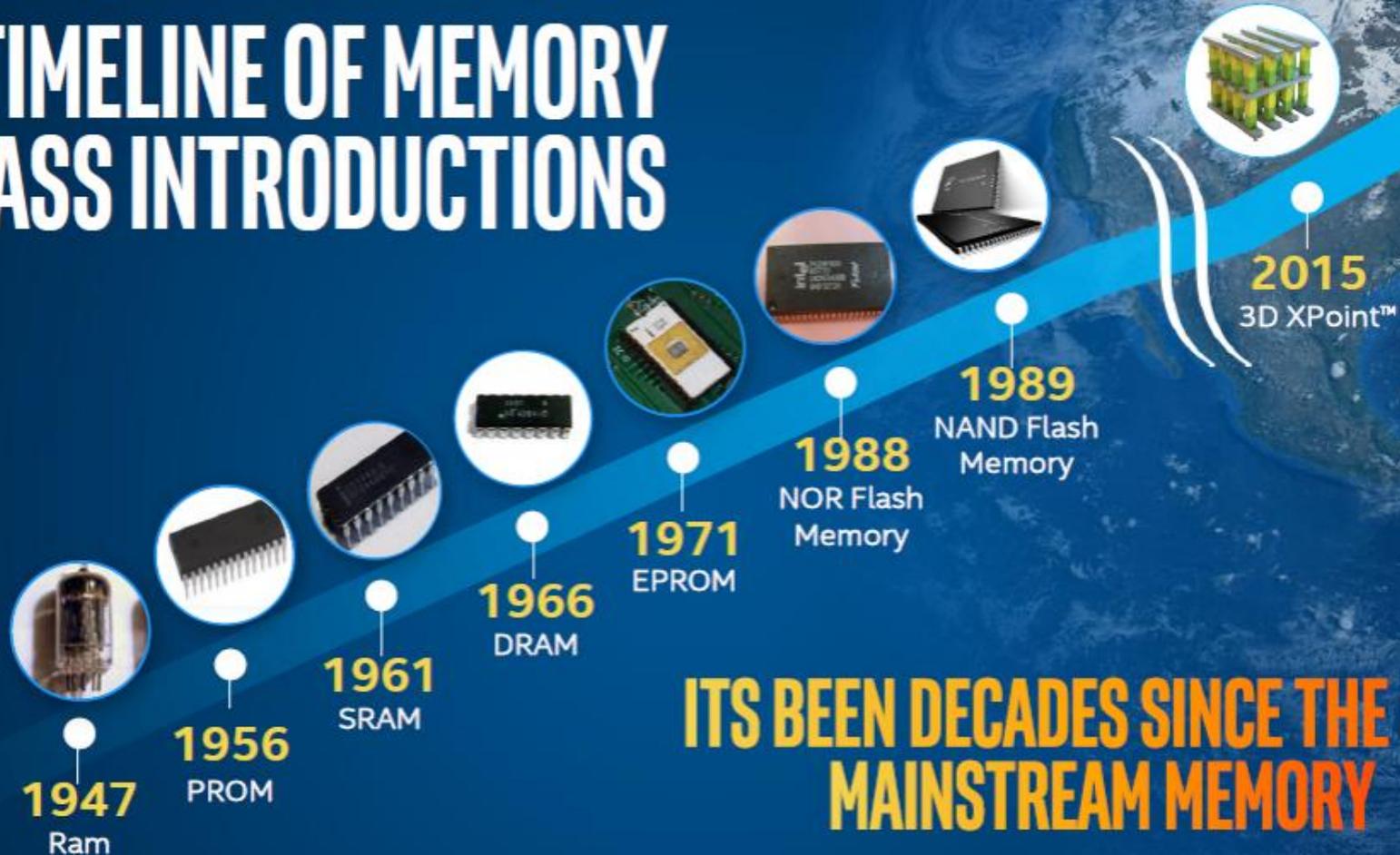
- An Overview of Memory
- **Memory Technologies**
 - Random Access Memory (RAM)
 - Read-Only Memory (ROM)
 - Non-Volatile Memory (NVM)
- Memory Hierarchy

Mainstream Memory Technologies



- There are many types of memory in the market:

A TIMELINE OF MEMORY CLASS INTRODUCTIONS



**ITS BEEN DECADES SINCE THE LAST
MAINSTREAM MEMORY**



- An Overview of Memory
- **Memory Technologies**
 - Random Access Memory (RAM)
 - Read-Only Memory (ROM)
 - Non-Volatile Memory (NVM)
- Memory Hierarchy

Random Access Memory (RAM)

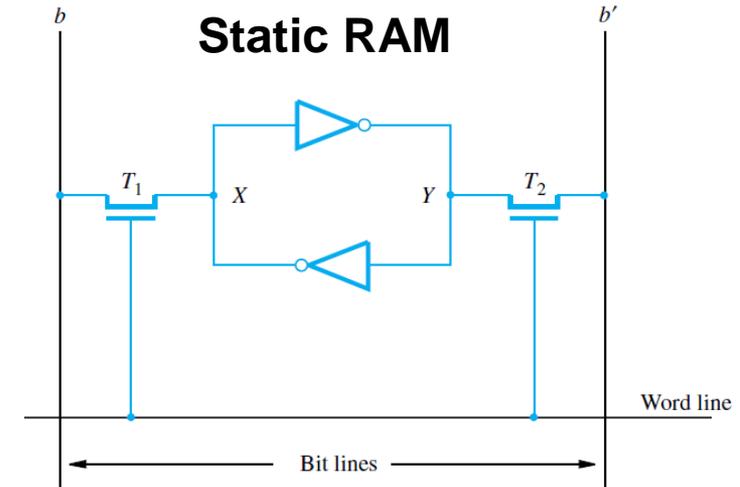


- **Random Access Memory (RAM):** The **access (R/W) time** to any location is the same, independent of the location's address.
 - **Memory Access Time:** The time between start and finish of a memory request.
- That is, we can “**randomly**” access any location of the RAM with **the same memory access time**.
- RAM are available in a wide range of types:
 - 1) Static RAM (SRAM)
 - 2) Dynamic RAM (DRAM)
 - 3) Synchronous DRAM (SDRAM)

Static RAM (SRAM)

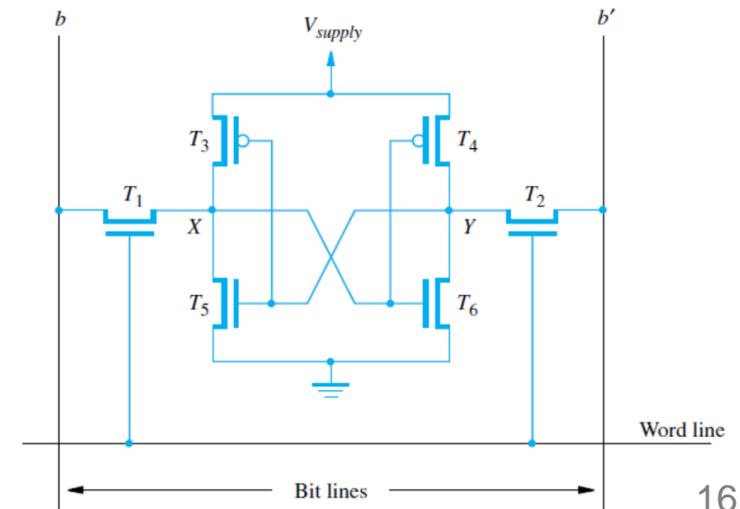


- **Static RAM (SRAM):** Capable of “**statically**” retaining the cell state (i.e. data) as long as power is applied (i.e., **volatile**).
 - Fast: Access times are on the order of a few nanoseconds.
 - Low power:
 - In **SRAM**, **continuous** power is needed for retaining its state; otherwise, the contents are lost.
 - In **CMOS SRAMs**, current flows **only when** accessing the cells.
 - Costly: Several transistors are required.
 - As a result, the capacity is small.



If the logic value at point X is 1 and at point Y is 0, this state is maintained as long as the signal on the word line is at ground level. Assume that this state represents the value 1.

CMOS Static RAM

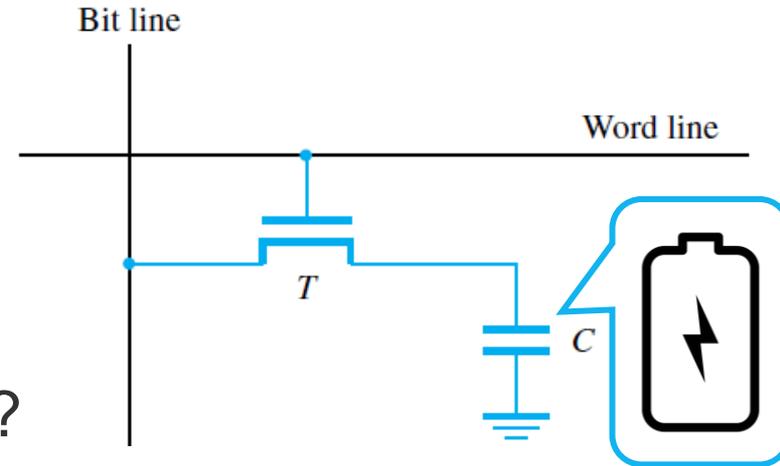


Dynamic RAM (DRAM)



- **Dynamic RAM (DRAM):** Store data in the form of “**dynamical**” charges on a capacitor.

- A DRAM cell is cheaper, simpler, but slower than a SRAM cell.



- Why a DRAM cell is “**dynamical**”?

- Charges can be maintained for only tens of milliseconds.
- That is, the charges will leak away as time goes (i.e., **volatile**).

- The contents of DRAM cells must be **refreshed periodically**.

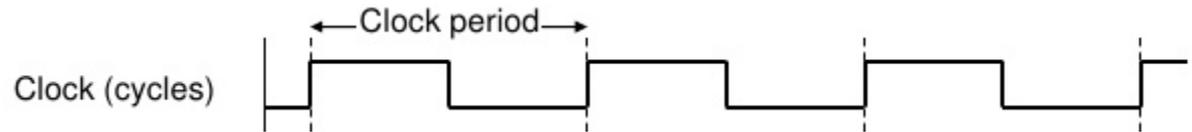
- By recharging the capacitor.

→ A DRAM cell consumes more power than a SRAM cell.

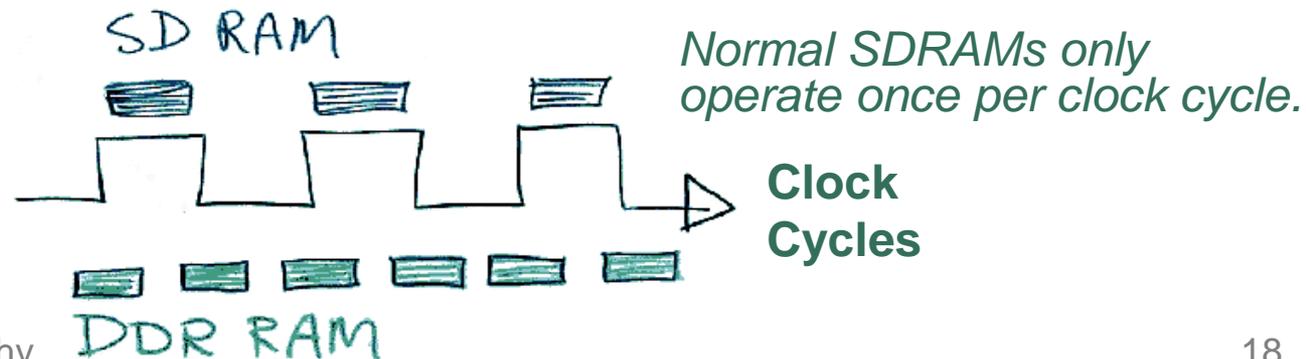
Synchronous DRAM (SDRAM) (1/3)



- **Synchronous DRAM (SDRAM):** Use the same cells as DRAM but use a **clock** to **synchronize** operations.
 - Why to synchronize operations?
 - The refresh operation can be transparent to the users.
 - The data can be transferred at “double data rate” (faster!).
 - Etc.



- The most common type used today as the **main memory**.
- **Double Data Rate (DDR) SDRAM:** Transfer data on **both clock edges**.



Synchronous DRAM (SDRAM) (2/3)



- **Memory Modules:** The standard for today's computers to hold multiple SDRAM chips.

SO-DIMM (for laptop)
Small Outline Dual In-line
Memory Module



DIMM (for desktop)
Dual In-line Memory
Module



Synchronous DRAM (SDRAM) (3/3)



- **Enhanced Versions:** DDR-2, DDR-3, and DDR-4
 - They offer larger size, lower power and faster clock rates.
- The table below compares the **theoretical maximum bandwidths** of different SDRAM types.

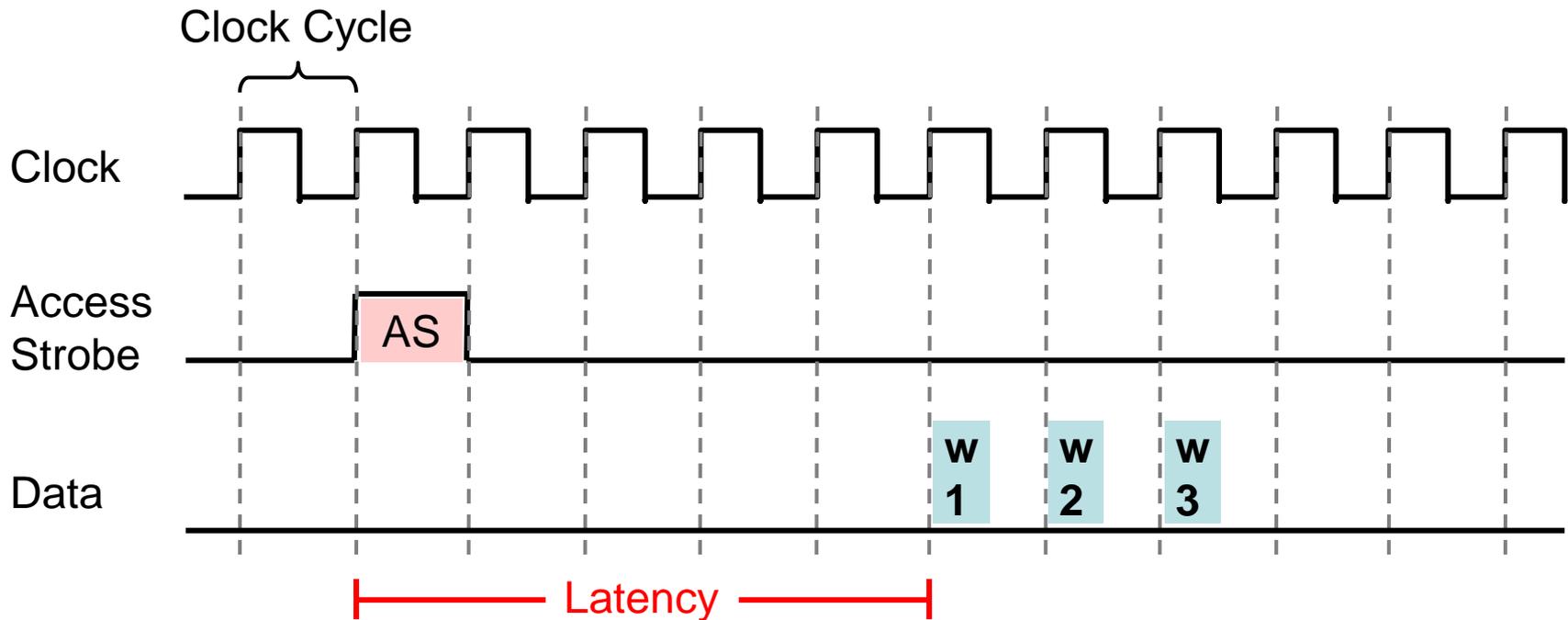
RAM Type	Theoretical Maximum Bandwidth
SDRAM 100 MHz (PC100)	100 MHz X 64 bit/ cycle = 800 MByte/sec
SDRAM 133 MHz (PC133)	133 MHz X 64 bit/ cycle = 1064 MByte/sec
DDR SDRAM 200 MHz (PC1600)	2 X 100 MHz X 64 bit/ cycle \approx 1600 MByte/sec
DDR SDRAM 266 MHz (PC2100)	2 X 133 MHz X 64 bit/ cycle \approx 2100 MByte/sec
DDR SDRAM 333 MHz (PC2600)	2 X 166 MHz X 64 bit/ cycle \approx 2600 MByte/sec
DDR-2 SDRAM 667 MHz (PC2-5400)	2 X 2 X 166 MHz X 64 bit/ cycle \approx 5400 MByte/sec
DDR-2 SDRAM 800 MHz (PC2-6400)	2 X 2 X 200 MHz X 64 bit/ cycle \approx 6400 MByte/sec

- SDRAM does not perform as good as the table shown, due to **latencies**.

Bandwidth vs. Latency



- **Bandwidth:** The maximal number of bits or bytes that can be transferred in one second.
- **Latency:** The amount of time it takes to transfer the first word, after issuing an access (i.e., *access strobe*).



Class Exercise 6.3



- Suppose the clock rate is 500 MHz, and each word (i.e., w_1 , w_2 , w_3) is 16 bits in the previous example.
- What is the **bandwidth** and **latency** on transferring data via the SDRAM?
- Answer:



- An Overview of Memory
- **Memory Technologies**
 - Random Access Memory (RAM)
 - **Read-Only Memory (ROM)**
 - Non-Volatile Memory (NVM)
- Memory Hierarchy

Read-Only Memory (ROM) (2/2)



- Some other ROM designs allow the data to be programmed and erased:
 - **Programmable ROM (PROM):**
 - **Irreversibly** allow the data to be loaded by the user (**write once!**).
 - **Erasable Reprogrammable ROM (EPROM):**
 - Allow the stored data to be **erased** and new data to be **written** into it.
 - Provide flexibility for the development of digital systems.
 - **Electrically EPROM (EEPROM):**
 - An EPROM must be physically removed from the circuit for reprogramming, and the stored data cannot be erased selectively.
 - EEPROM can be erased and reprogrammed **electrically**.
 - Different voltages for erasing/writing/reading increases complexity.
- Nevertheless, ROM is **much slower** than RAM.



- An Overview of Memory
- **Memory Technologies**
 - Random Access Memory (RAM)
 - Read-Only Memory (ROM)
 - **Non-Volatile Memory (NVM)**
- Memory Hierarchy

Non-Volatile Memory (NVM)



- A new approach similar to EEPROM technology.
- **Non-Volatile Memory (NVM)**
 - NVM can be read, written, and erased, and it's **non-volatile**.
 - Features: greater density, higher capacity and lower cost, lower power, shock resistant, but still slower than RAM.
 - The most famous example: **Flash memory**



Smart Phone
(micro SD)



Digital Camera
(SD Card)



Notebook
(SSD)



USB
Drives

- There are many other types of NVM: PCRAM, ReRAM, STTRAM, Racetrack Memory, 3D XPoint etc.

NAND Flash Memory

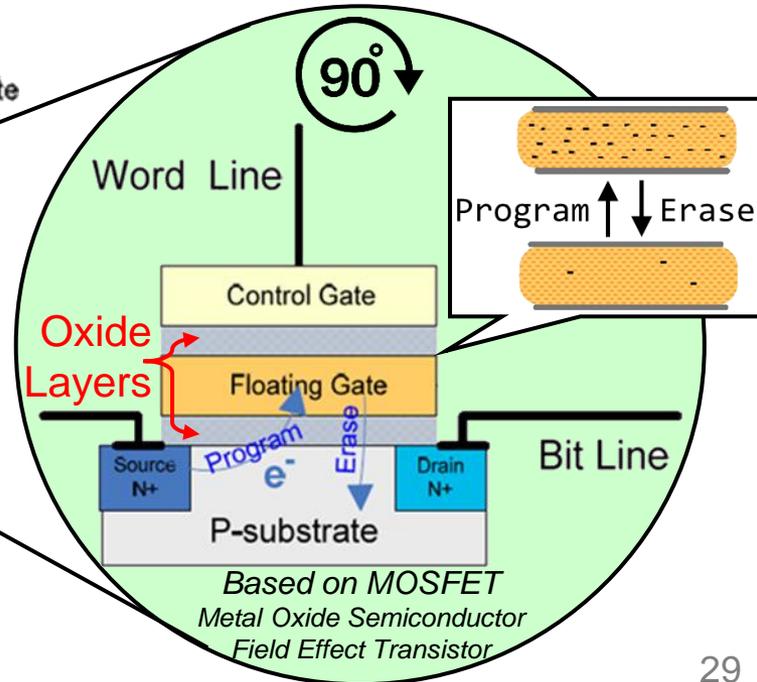
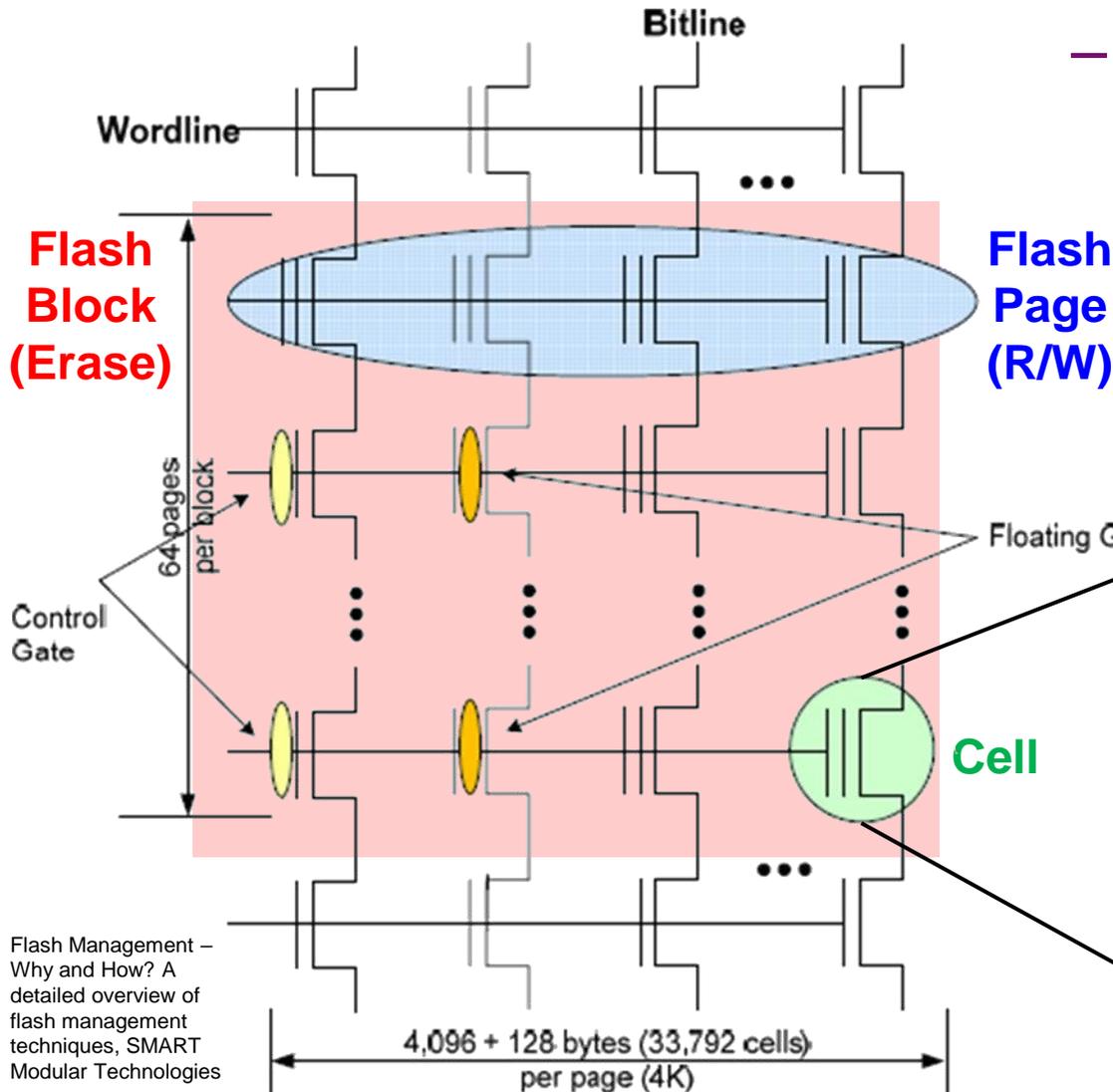


• NAND Flash Array

• NAND Flash Cell

– Floating-Gate Transistor

- **Program:** Inject electrons into FG to raise voltage
- **Erase:** Remove electrons from FG to lower voltage
- **Read:** Sense voltage of FG

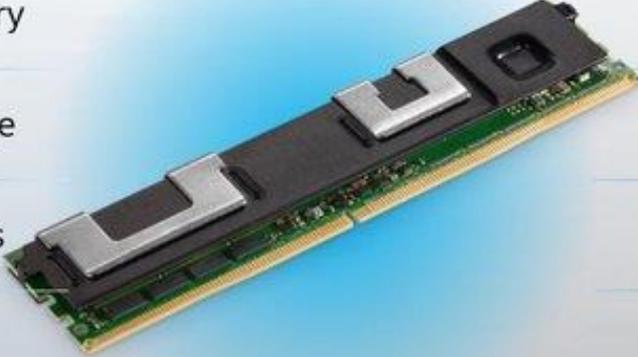


Flash Management – Why and How? A detailed overview of flash management techniques, SMART Modular Technologies

3D XPoint (1/2)



- Intel® Optane™ DC persistent memory is the latest, innovative memory technology.
 - It delivers affordable large space and data persistence.
 - 10X higher density than DRAM.
 - It adopts **3D XPoint** as the memory media.



Big and Affordable Memory	128, 256, 512GB
High Performance Storage	DDR4 Pin Compatible
Direct Load/Store Access	Hardware Encryption
Native Persistence	High Reliability

<https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html>

<https://www.slideshare.net/Syntech/intel-micron-unveil-breakthrough-3d-xpoint-memory-tech-a-revolutionary-breakthrough-in-memory-technology>



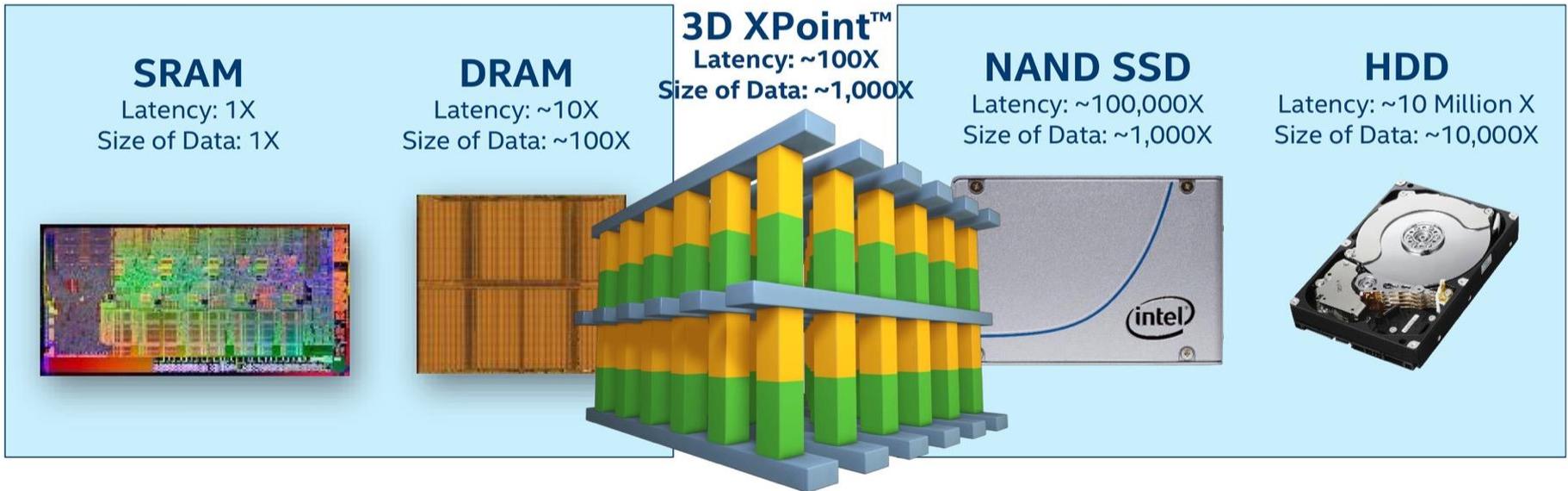
3D XPOINT™ MEMORY MEDIA

Breaks the memory/storage barrier

MEMORY

+

STORAGE



Technology claims are based on comparisons of latency, density and write cycling metrics amongst memory technologies recorded on published specifications of in-market memory products against internal Intel specifications.

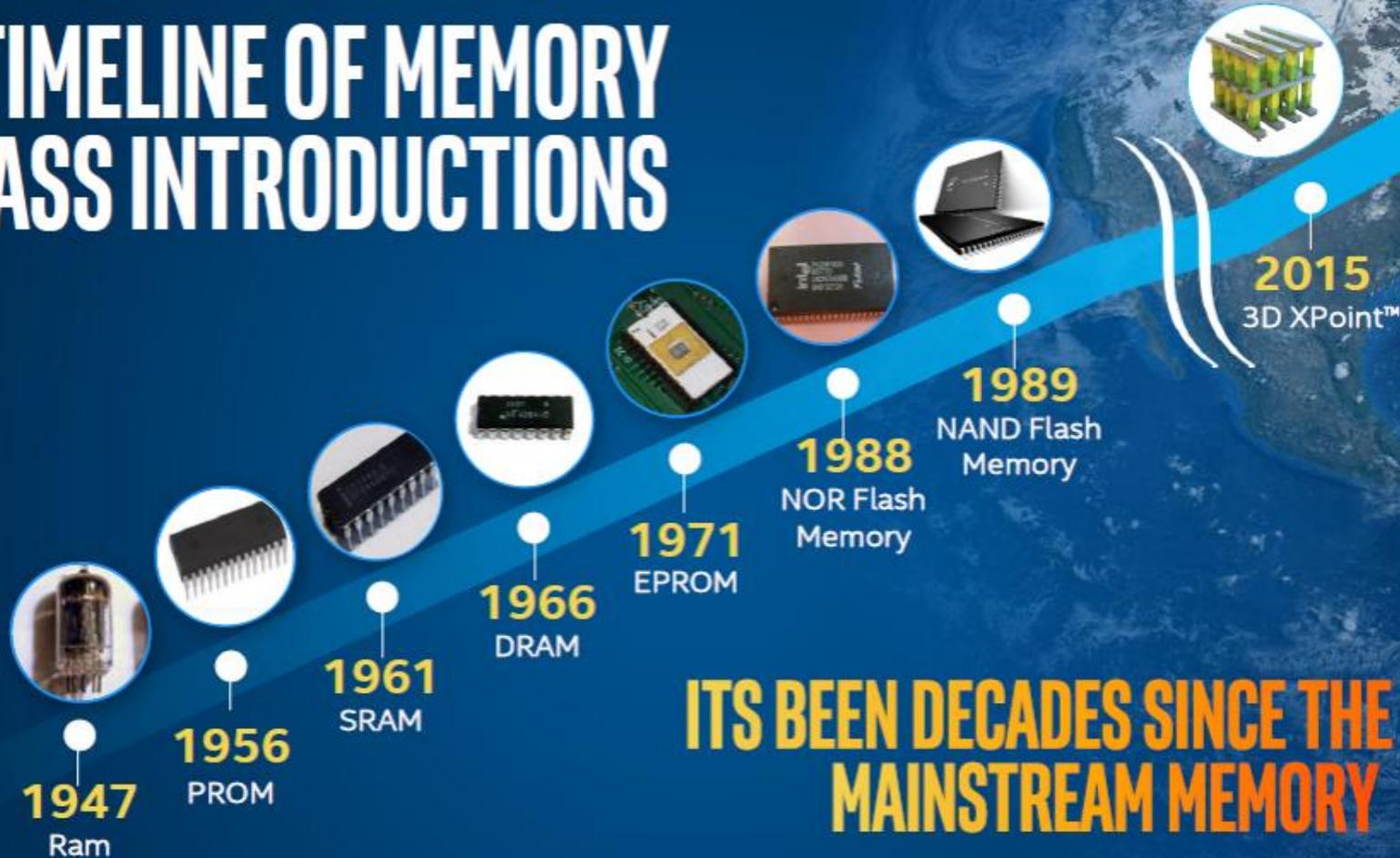


Revisit: Memory Technologies



- What is the “best” choice for the computer memory?

A TIMELINE OF MEMORY CLASS INTRODUCTIONS



**ITS BEEN DECADES SINCE THE LAST
MAINSTREAM MEMORY**



- An Overview of Memory
- Memory Technologies
 - Random Access Memory (RAM)
 - Read-Only Memory (ROM)
 - Non-Volatile Memory (NVM)
- **Memory Hierarchy**

Mix-and-Match: Best of ALL



- An ideal memory would be fast, large, and cheap.
- The fact is different memories have its pros and cons.
- **SRAM** is **fast**, but **expensive** and **not very dense**:
 - Good choice for providing the user the **fastest access time**
 - Good for **registers, L1 and L2 cache** in the processor
- **SDRAM** is **slower**, but **cheap** and **dense**:
 - Good choice for providing the user a **big memory space**
 - Good for **main memory**

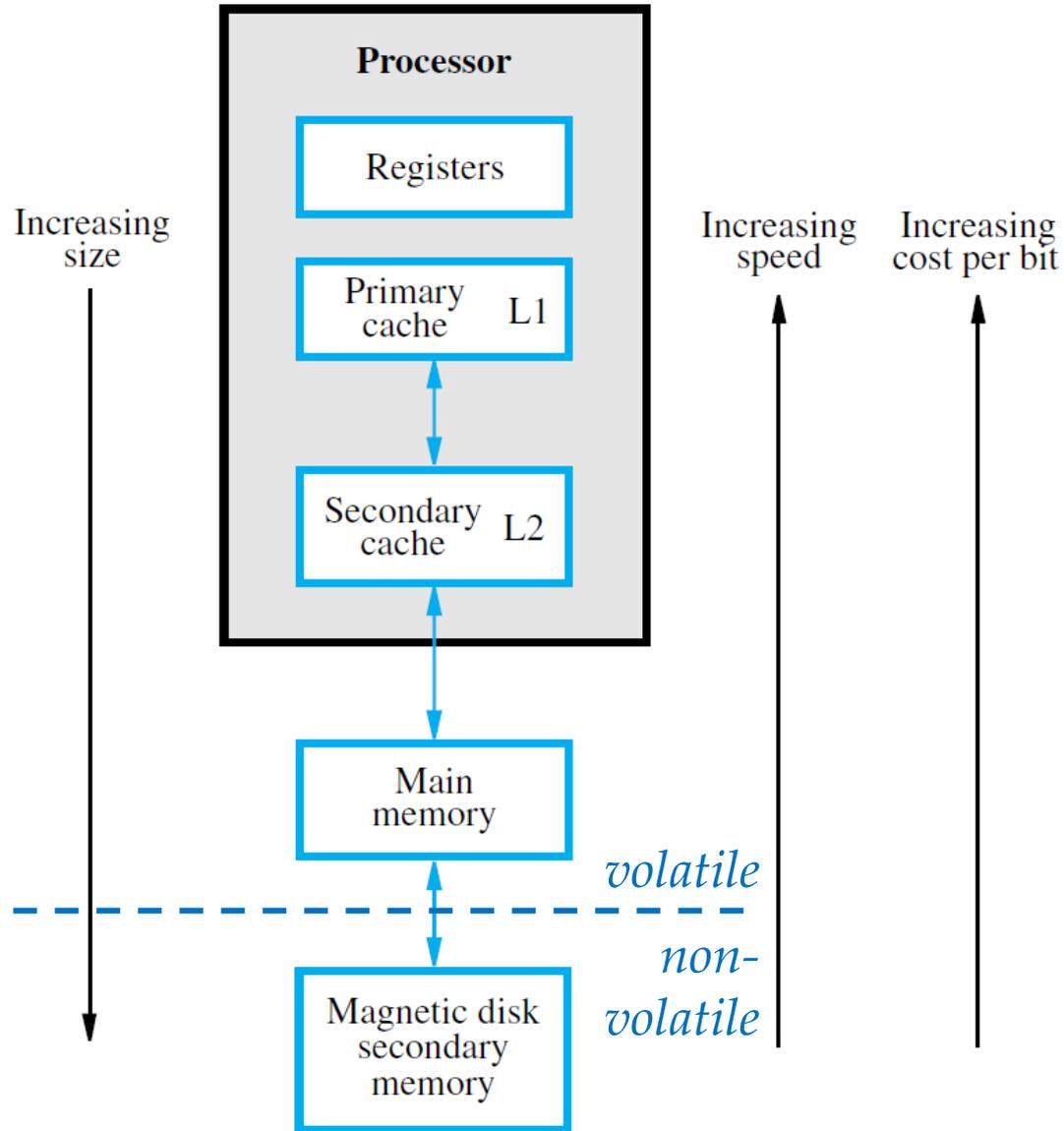
- **NVM/SSD/Disk** are **even slower**, but *volatile*
even cheaper, denser and *non-volatile* **non-volatile**:
 - Good choice for cost-effective and **non-volatile data storage**
 - Good for **secondary storage**

Solution: Memory Hierarchy (1/2)



Processor

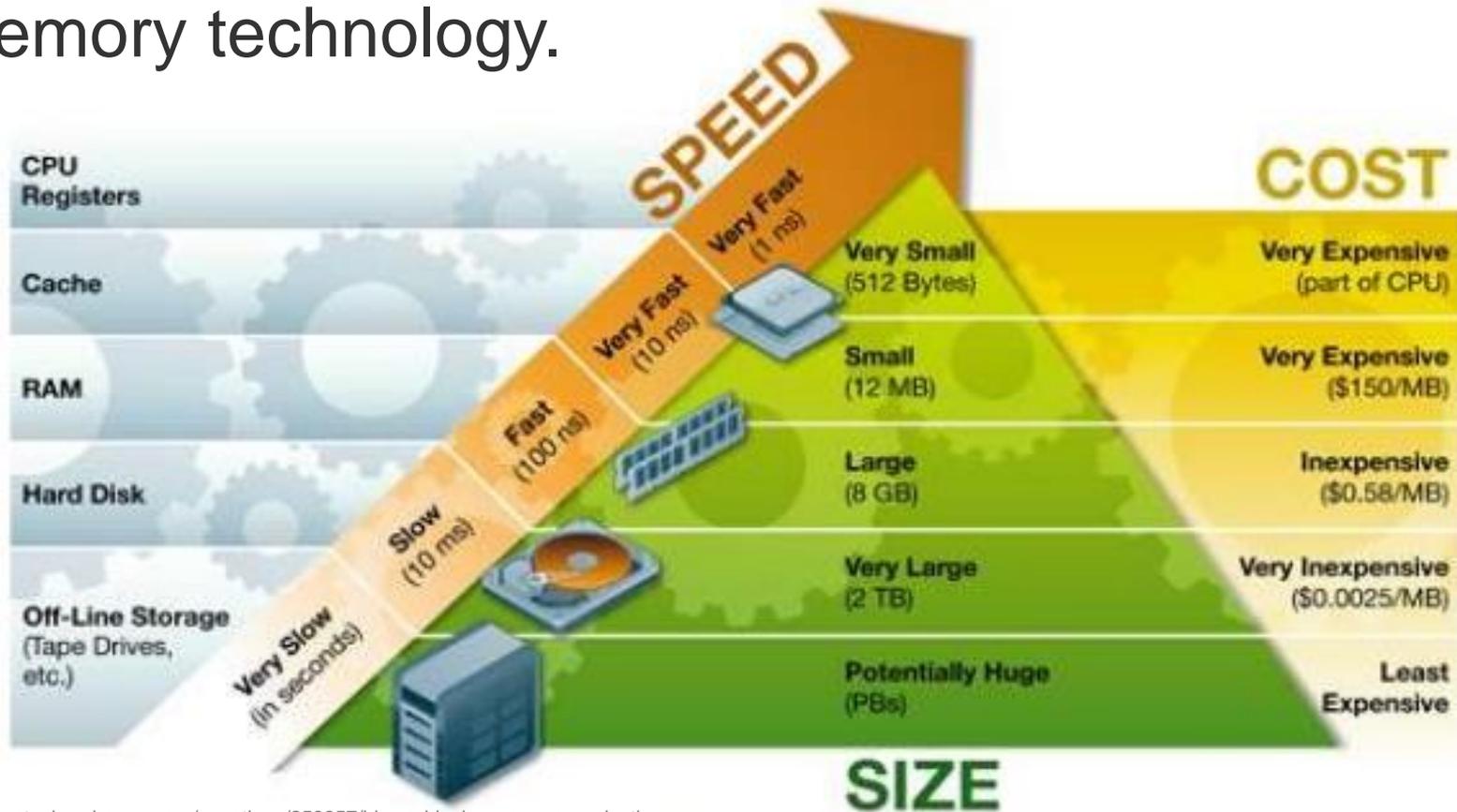
- Register: SRAM
- L1, L2 cache: SRAM
- Main memory: SDRAM
- Secondary storage: Hard disks or NVM



Solution: Memory Hierarchy (2/2)



- Provide the user with as much memory as is available in the **cheapest** memory technology.
- Provide access at the speed offered by the **fastest** memory technology.





- An Overview of Memory
- Memory Technologies
 - Random Access Memory (RAM)
 - Read-Only Memory (ROM)
 - Non-Volatile Memory (NVM)
- Memory Hierarchy