

CPS104
Computer Organization and Programming
Lecture 16: Memory Systems

Dietolf (Dee) Ramm

Oct 25, 1999

<http://www.cs.duke.edu/~dr/cps104.html>

Today's Lecture

- **Memory**

Outline

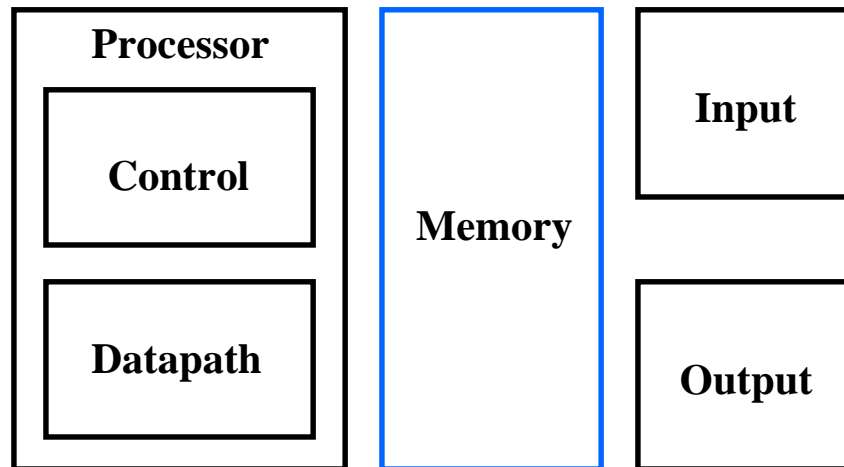
- **Review**
- **Big Picture of Memory**
- **Memory Technology**
 - SRAM
 - DRAM

Reading

Chapter 7

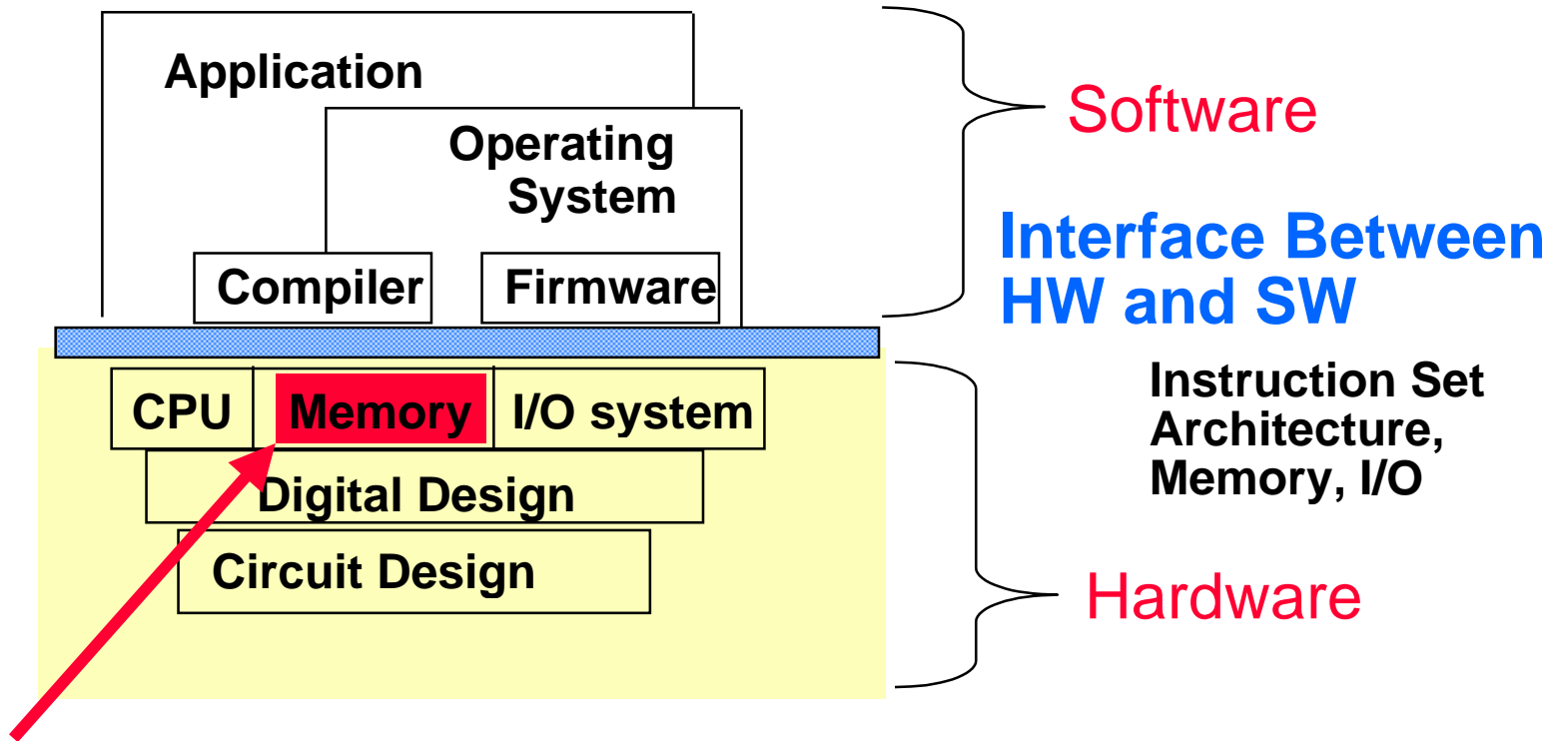
Big Picture

- **The Five Classic Components of a Computer**



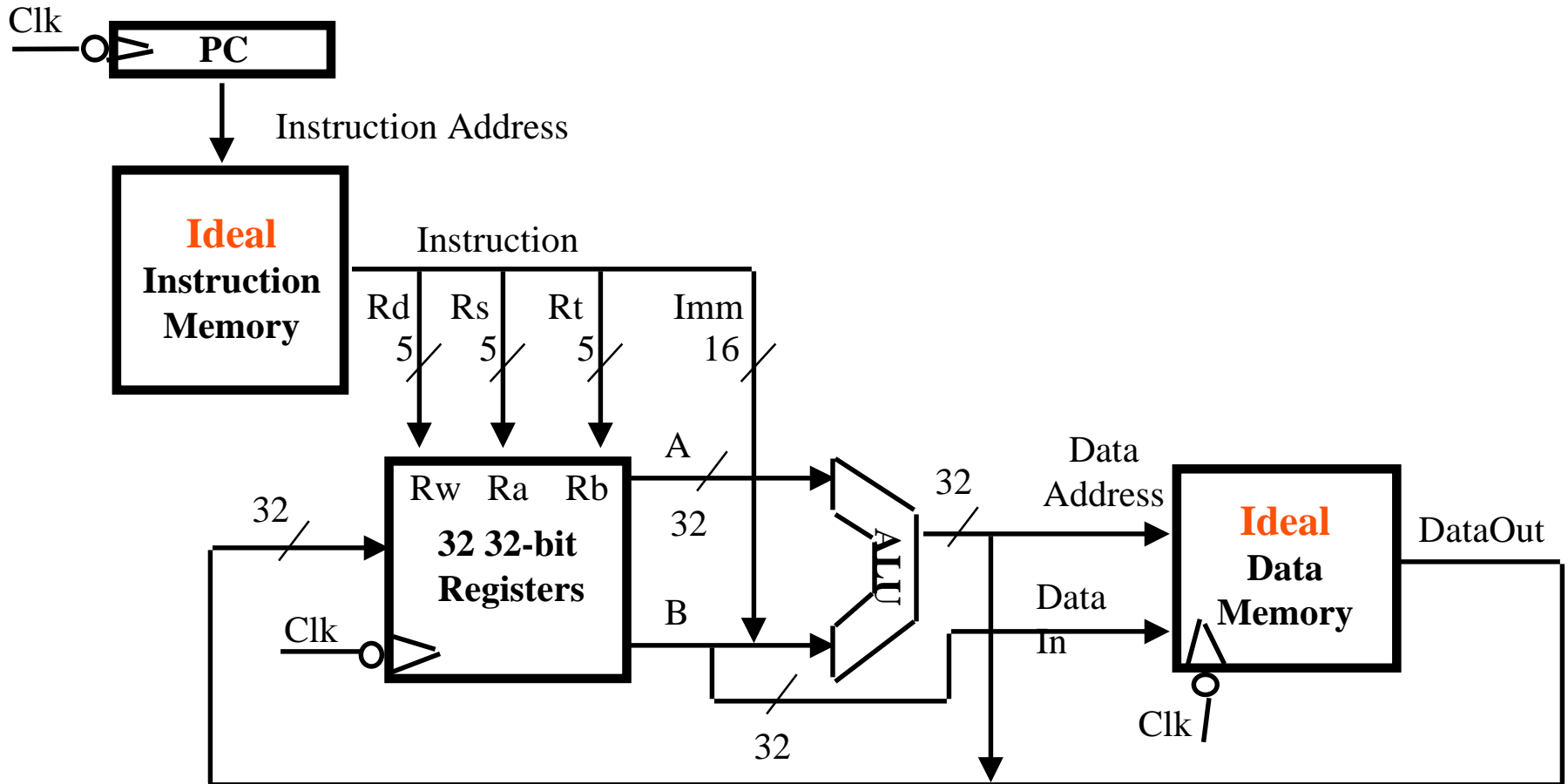
- **Today's Topic: Memory System**

Where Are We?



You are here.

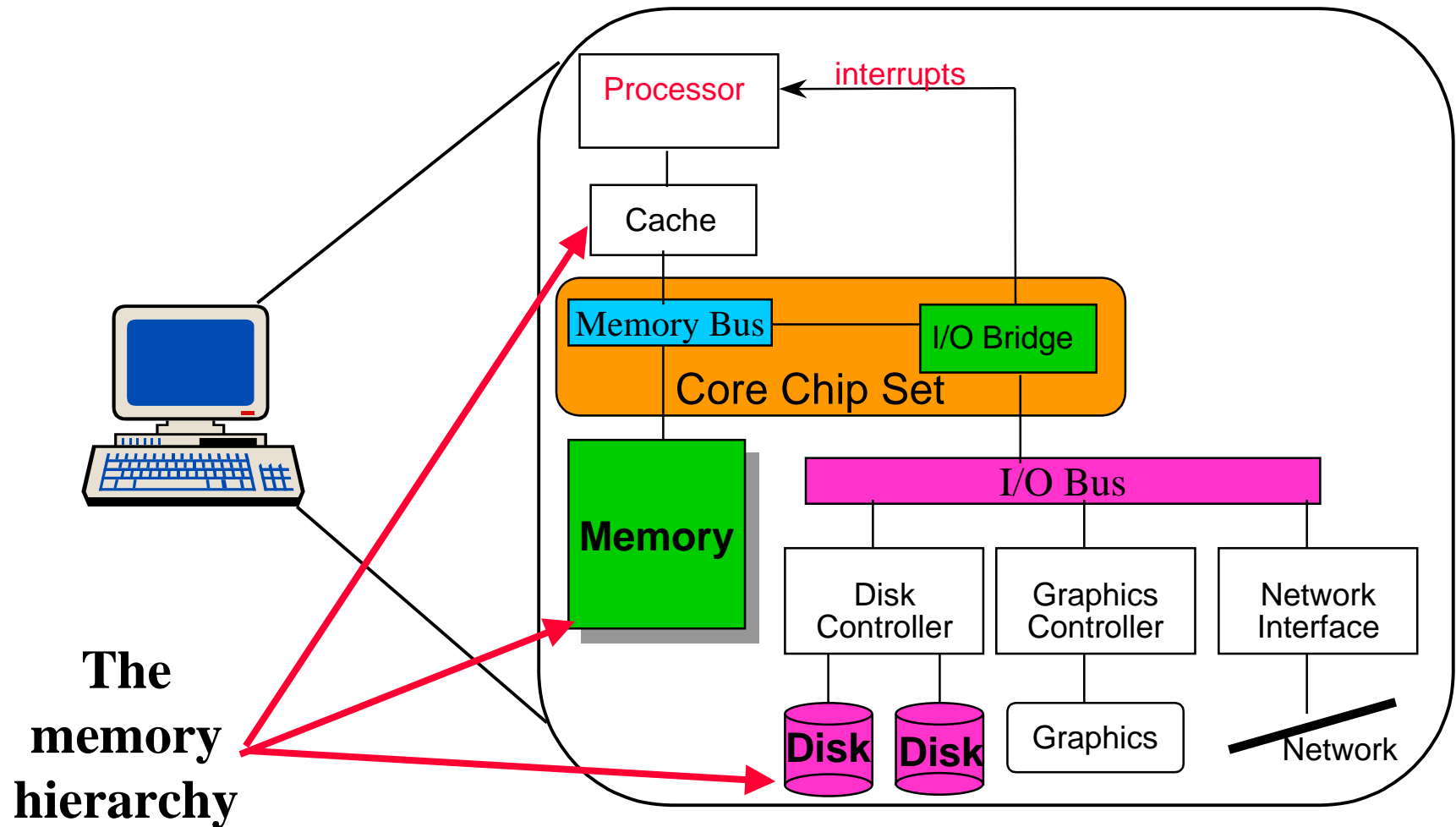
Our Naïve View of Memory



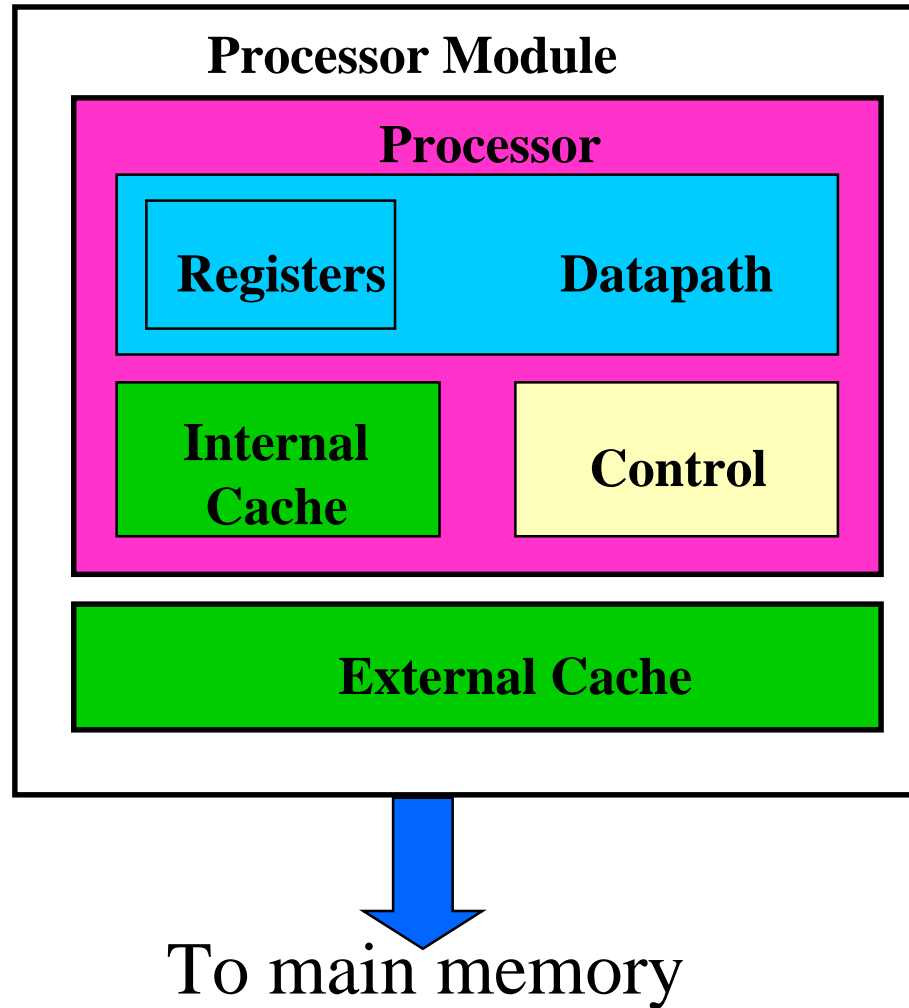
Question

- **What issues do we need to worry about in implementing the memory system?**

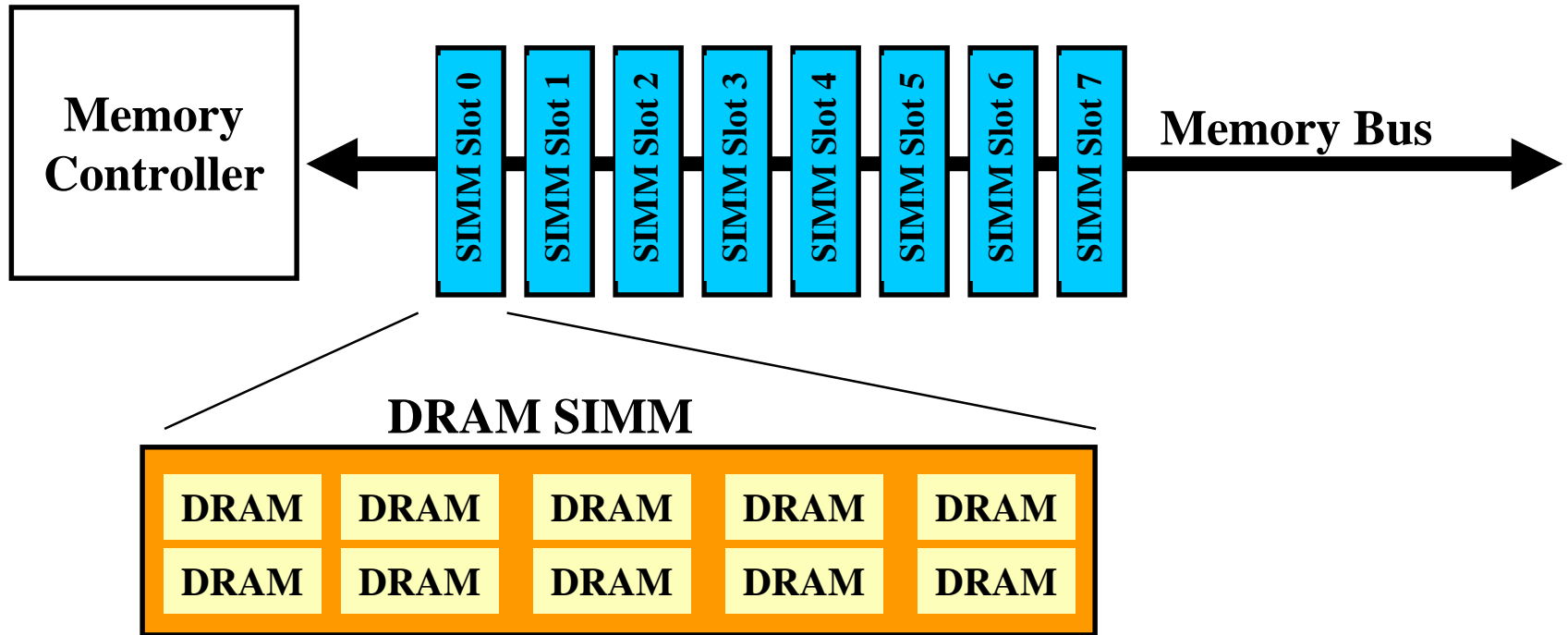
System Organization



Processor and Caches



Memory



Why is it called DRAM?

Memory Technology

- **Random Access:**
 - “Random” is good: access time is the same for all locations
 - **DRAM:** Dynamic Random Access Memory
 - » High density, low power, cheap, slow
 - » Dynamic: needs to be “refreshed” regularly
 - » Main memory
 - **SRAM:** Static Random Access Memory
 - » Low density, high power, expensive, fast
 - » Static: content will last “forever”(until lose power)
 - » Caches
- **“Not-so-random” or “Direct” Access Technology:**
 - Access time varies from location to location and from time to time
 - Examples: Disk, CDROM
- **Sequential Access Technology: access time linear in location (e.g., Tape)**

Random Access Memory (RAM) Technology

- **Why do computer professionals need to know about RAM technology?**
 - Processor performance is usually limited by memory **latency** and **bandwidth**.
 - **Latency**: The time it takes to access a word in memory.
 - **Bandwidth**: The average speed of access to memory (Words/Sec).
 - As IC densities increase, lots of memory will fit on processor chip
 - » **Tailor on-chip memory to specific needs.**
 - Instruction cache
 - Data cache
 - Write buffer
- **What makes RAM different from a bunch of flip-flops?**
 - **Density**: RAM is much more dense
 - **Speed**: RAM access is slower than flip-flop (register) access.

Technology Trends

	Capacity	Speed
Logic:	2x in 3 years	2x in 3 years
DRAM:	4x in 3 years	1.4x in 10 years
Disk:	2x in 3 years	1.4x in 10 years

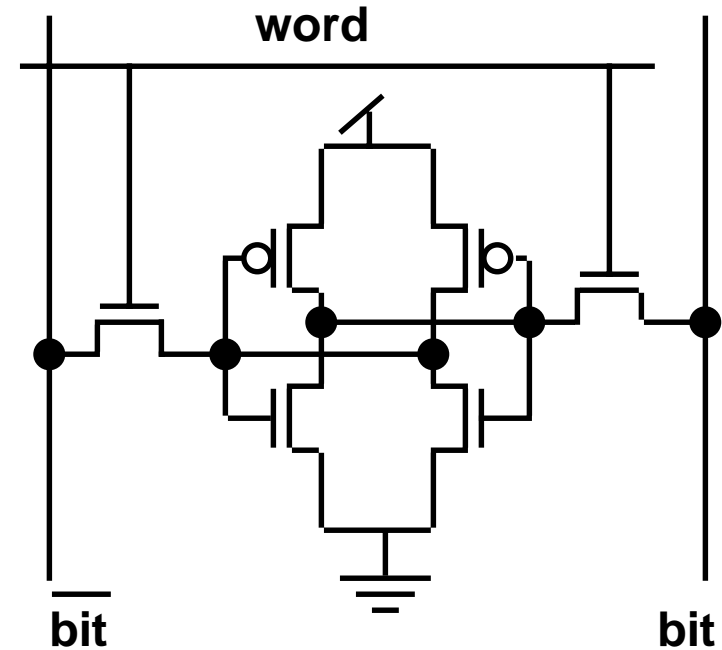
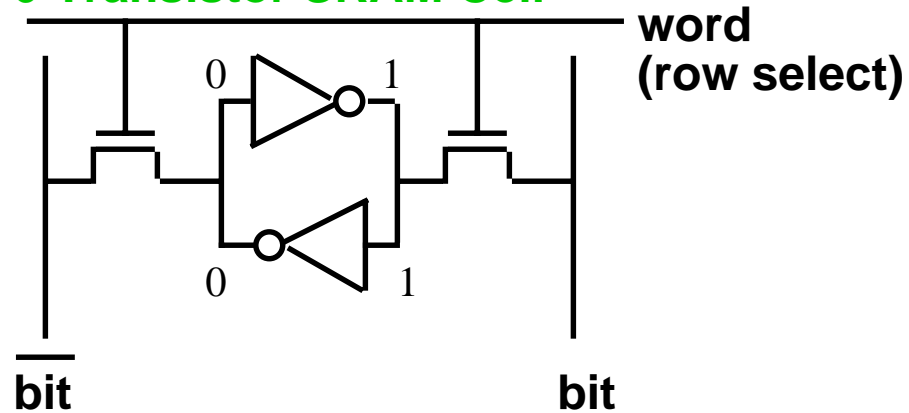
DRAM		
<u>Year</u>	<u>Size</u>	<u>Cycle Time</u>
1980	64 Kb	250 ns
1983	256 Kb	220 ns
1986	1 Mb	190 ns
1989	4 Mb	165 ns
1992	16 Mb	145 ns
1995	64 Mb	120 ns

1000:1! (Size trend from 1980 to 1995)

2:1! (Cycle Time trend from 1980 to 1995)

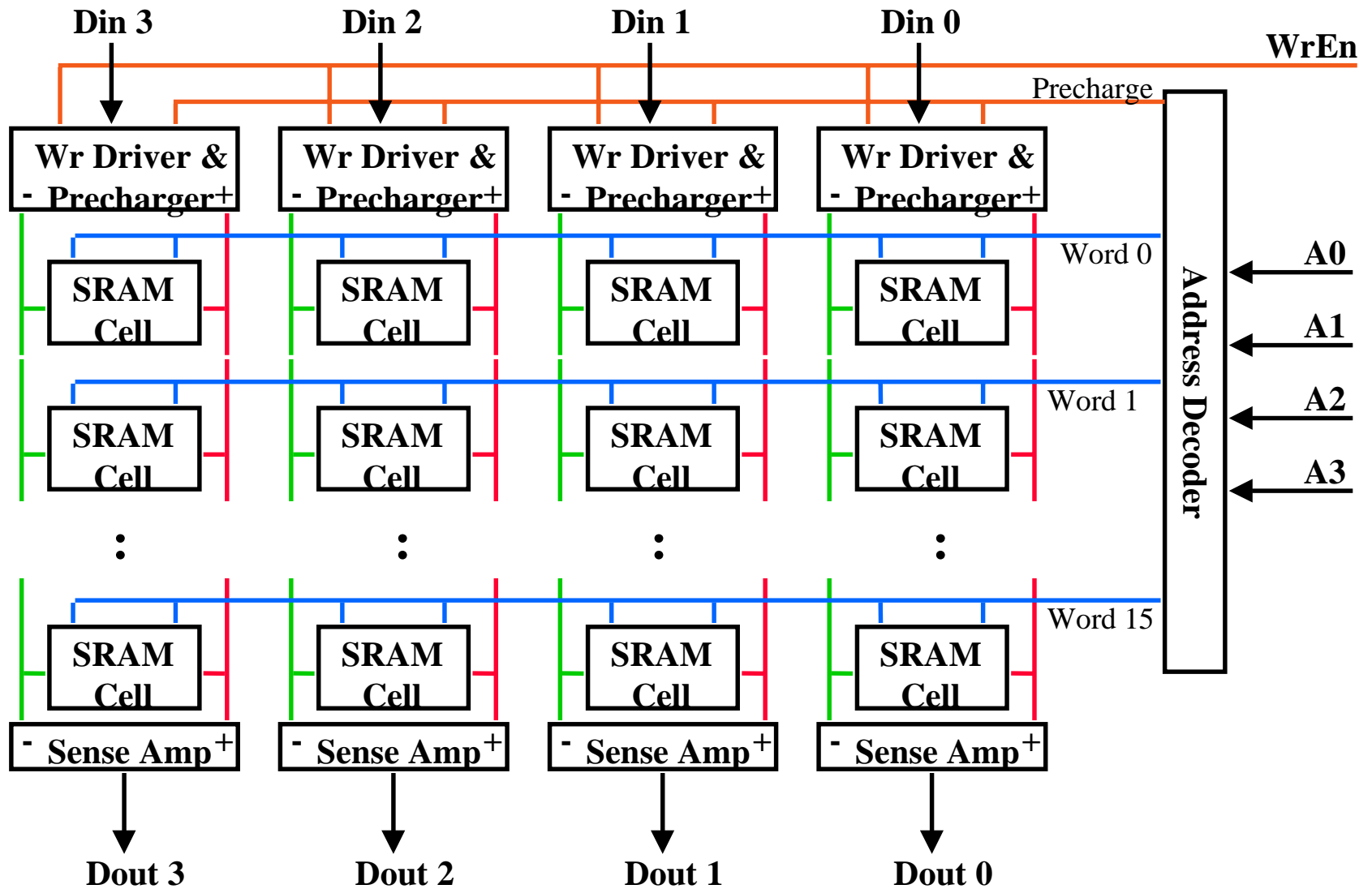
Static RAM Cell

6-Transistor SRAM Cell

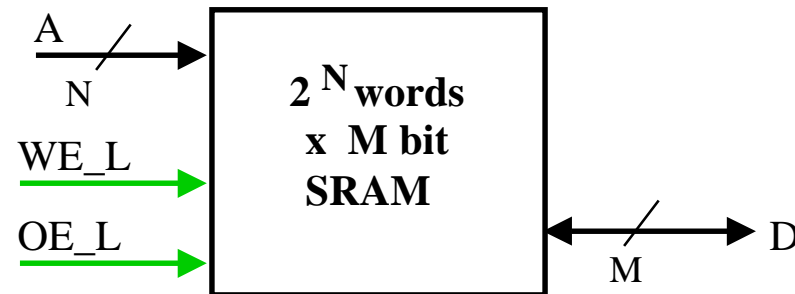


- **Write:**
 1. Drive bit lines ($\text{bit}=1$, $\overline{\text{bit}}=0$)
 2. Select row
- **Read:**
 1. Precharge bit and $\overline{\text{bit}}$ to Vdd (set to 1)
 2. Select row
 3. Cell pulls one line low (pulls to 0)
 4. Sense amp on column detects difference between bit and $\overline{\text{bit}}$

Typical SRAM Organization: 16-word x 4-bit

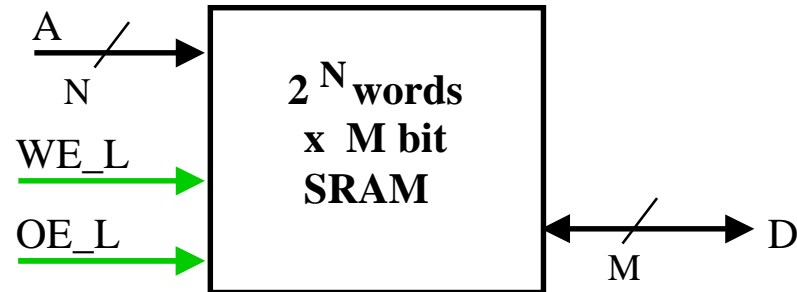


Logic Diagram of a Typical SRAM



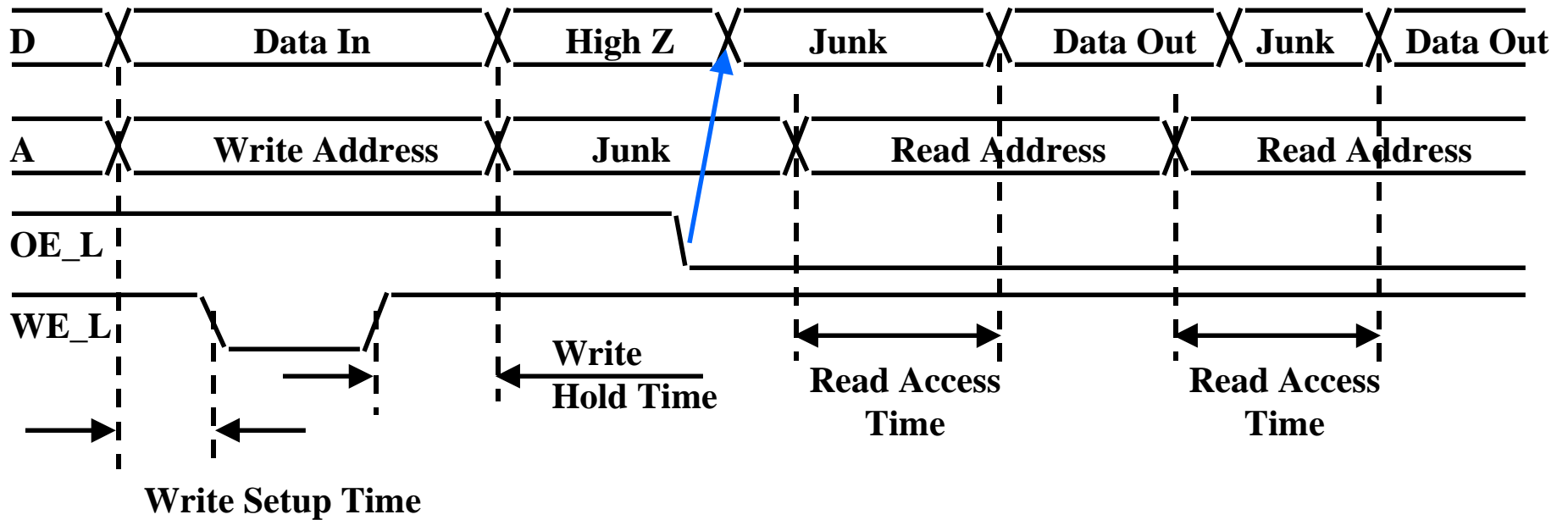
- **Write Enable is usually active low (WE_L)**
- **Din and Dout are combined to save pins:**
 - A new control signal, output enable (OE_L) is needed
 - WE_L is asserted (Low), OE_L is disasserted (High)
 - » D serves as the data input pin
 - WE_L is disasserted (High), OE_L is asserted (Low)
 - » D is the data output pin
 - Both WE_L and OE_L are asserted:
 - » Result is unknown. **Don't do that!!!**

Typical SRAM Timing



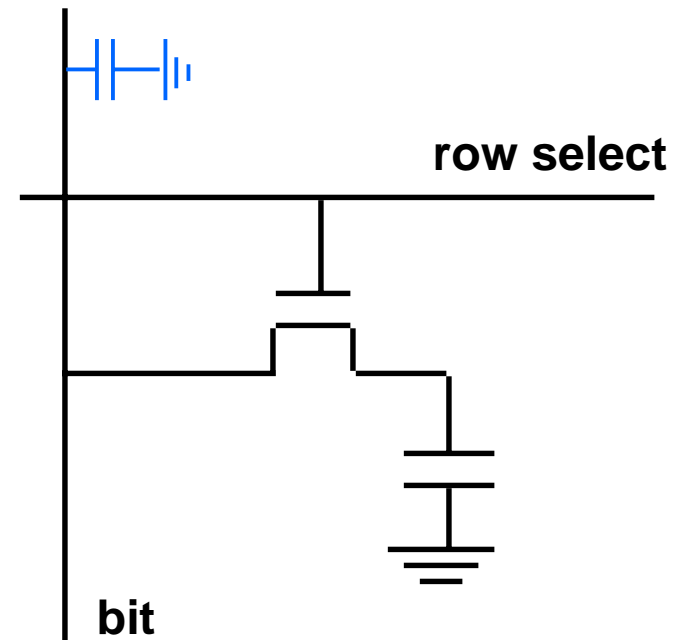
Write Timing:

Read Timing:



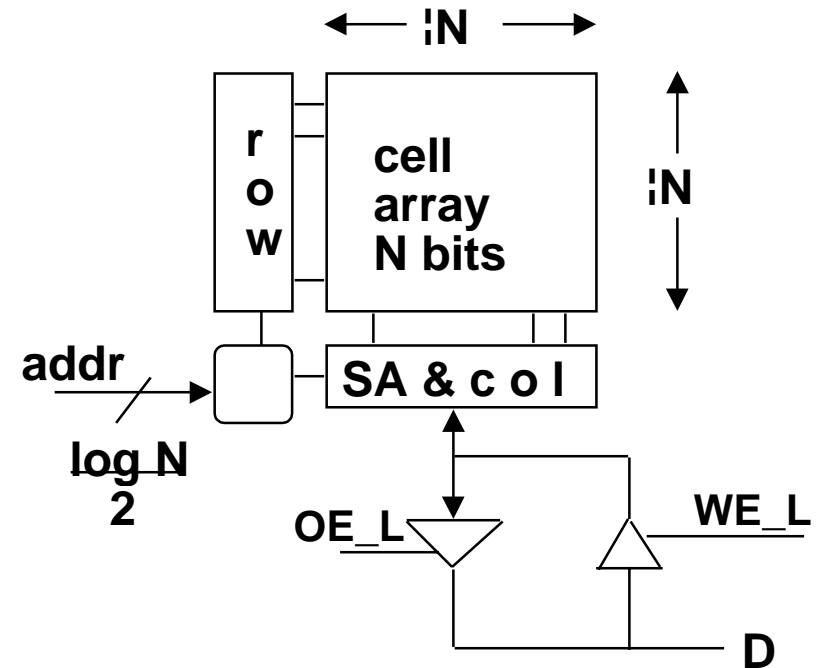
1-Transistor Memory Cell (DRAM)

- **Write:**
 - 1. Drive bit line
 - 2. Select row
- **Read:**
 - 1. Precharge bit line to Vdd (1)
 - 2. Select row
 - 3. Cell and bit line share charges
 - » Very small voltage changes on the bit line
 - 4. Sense (fancy sense amp)
 - » Can detect changes of ~1 million electrons
 - 5. Write: restore the value
- **Refresh**
 - 1. Just do a dummy read to every cell.

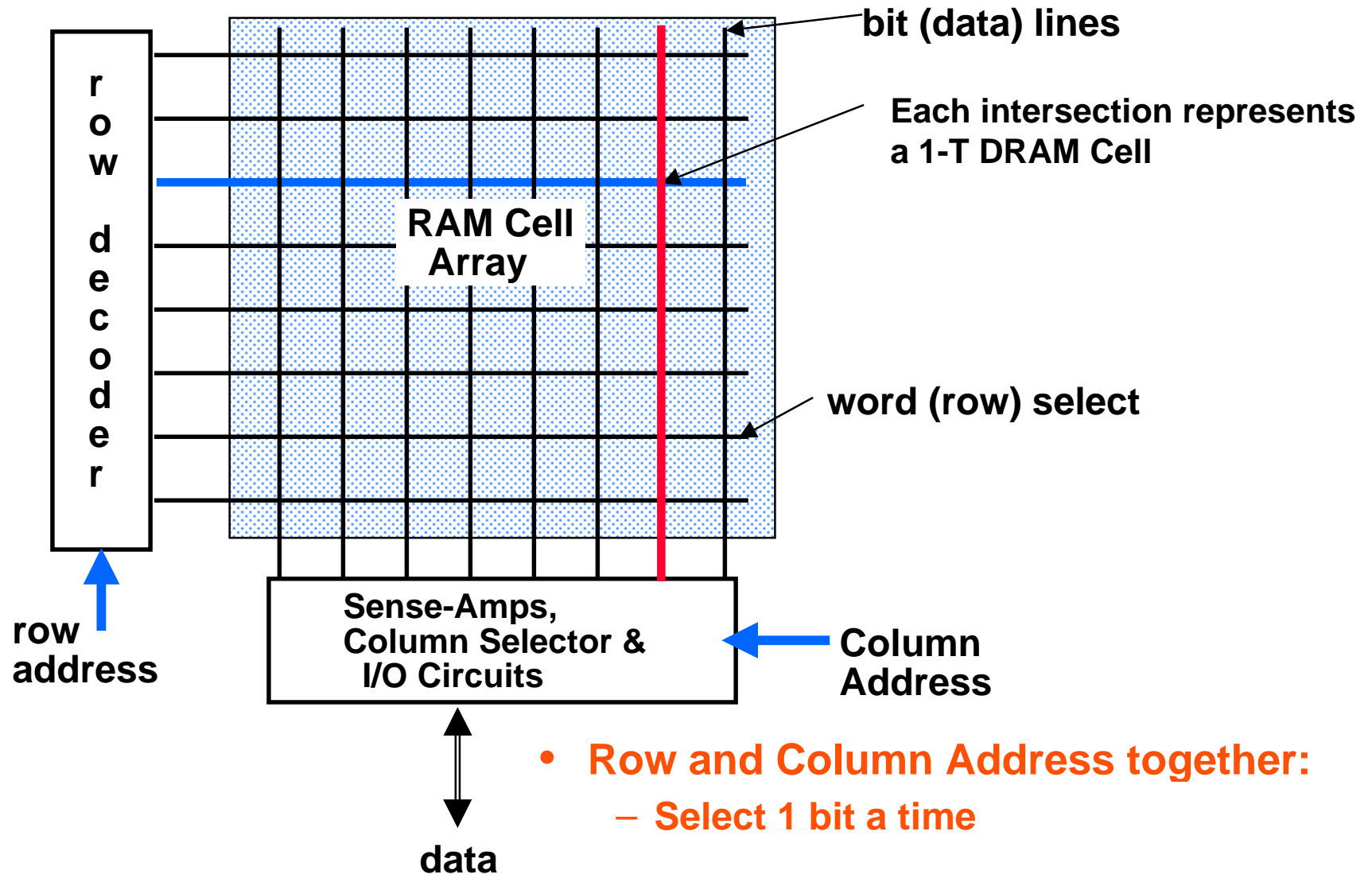


Introduction to DRAM

- **Dynamic RAM (DRAM):**
 - Refresh required
 - Very high density
 - Low power (.1 - .5 W active, .25 - 10 mW standby)
 - Low cost per bit
 - Pin sensitive (few pins):
 - » Output Enable (OE_L)
 - » Write Enable (WE_L)
 - » Row address strobe (ras)
 - » Col address strobe (cas)

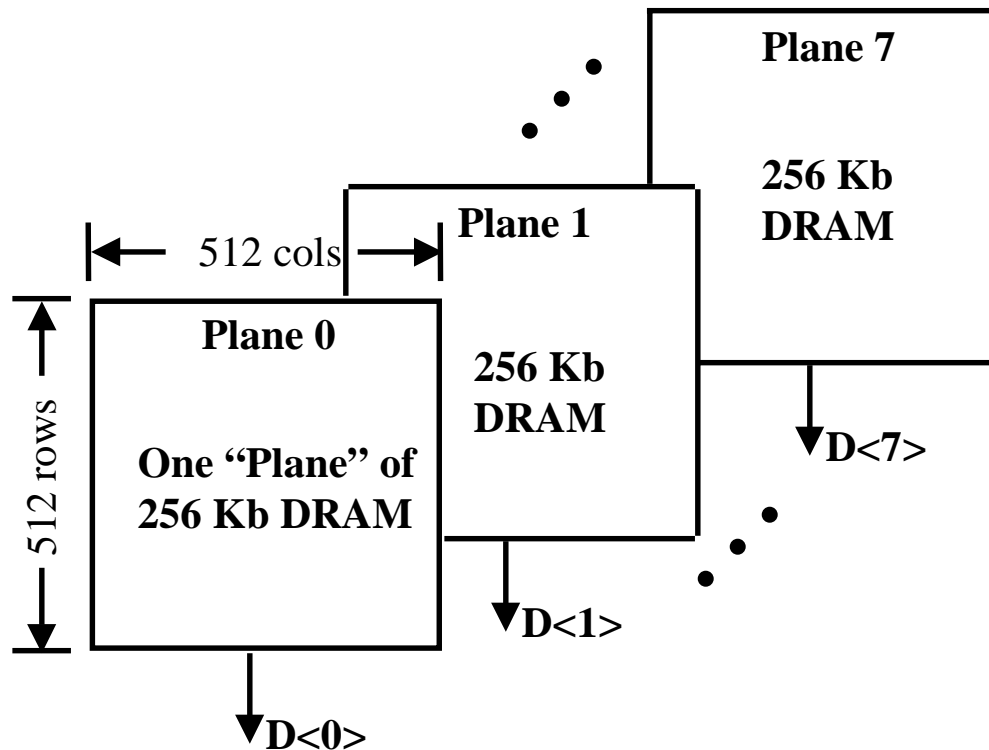


Classical DRAM Organization (square)

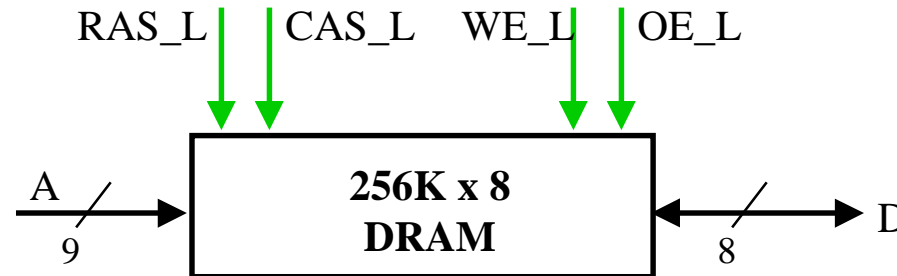


Typical DRAM Organization

- **Typical DRAMs: access multiple bits in parallel**
 - Example: 2 Mb DRAM = 256K x 8 = 512 rows x 512 cols x 8 bits
 - Row and column addresses are applied to all 8 planes in parallel



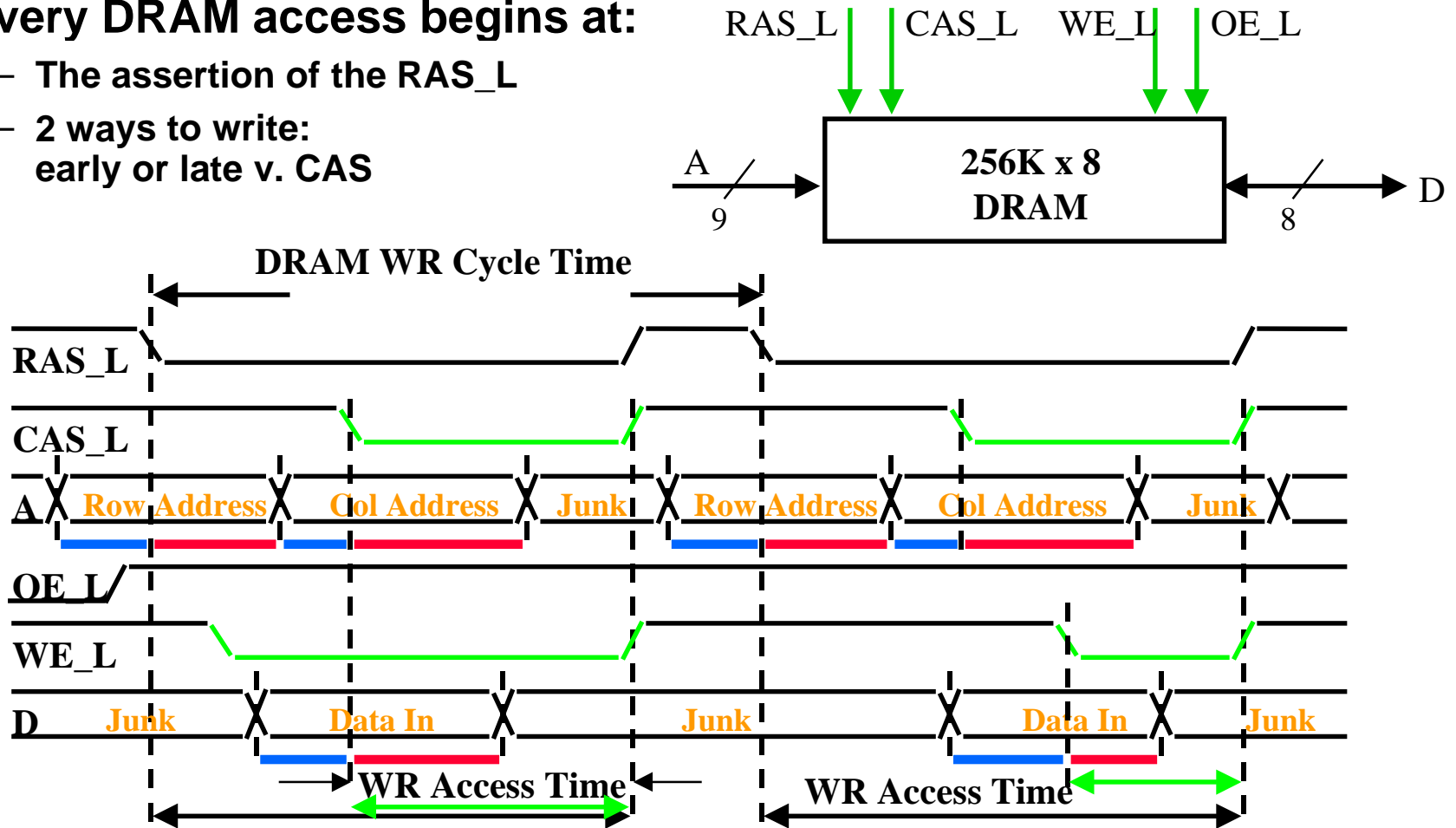
Logic Diagram of a Typical DRAM



- **Control Signals (RAS_L, CAS_L, WE_L, OE_L) are all active low**
- **Din and Dout are combined (D):**
 - WE_L is asserted (Low), OE_L is disasserted (High)
 - » D serves as the data input pin
 - WE_L is disasserted (High), OE_L is asserted (Low)
 - » D is the data output pin
- **Row and column addresses share the same pins (A)**
 - RAS_L goes low: Pins A are latched in as row address
 - CAS_L goes low: Pins A are latched in as column address
 - RAS/CAS edge-sensitive

DRAM Write Timing

- Every DRAM access begins at:
 - The assertion of the RAS_L
 - 2 ways to write: early or late v. CAS

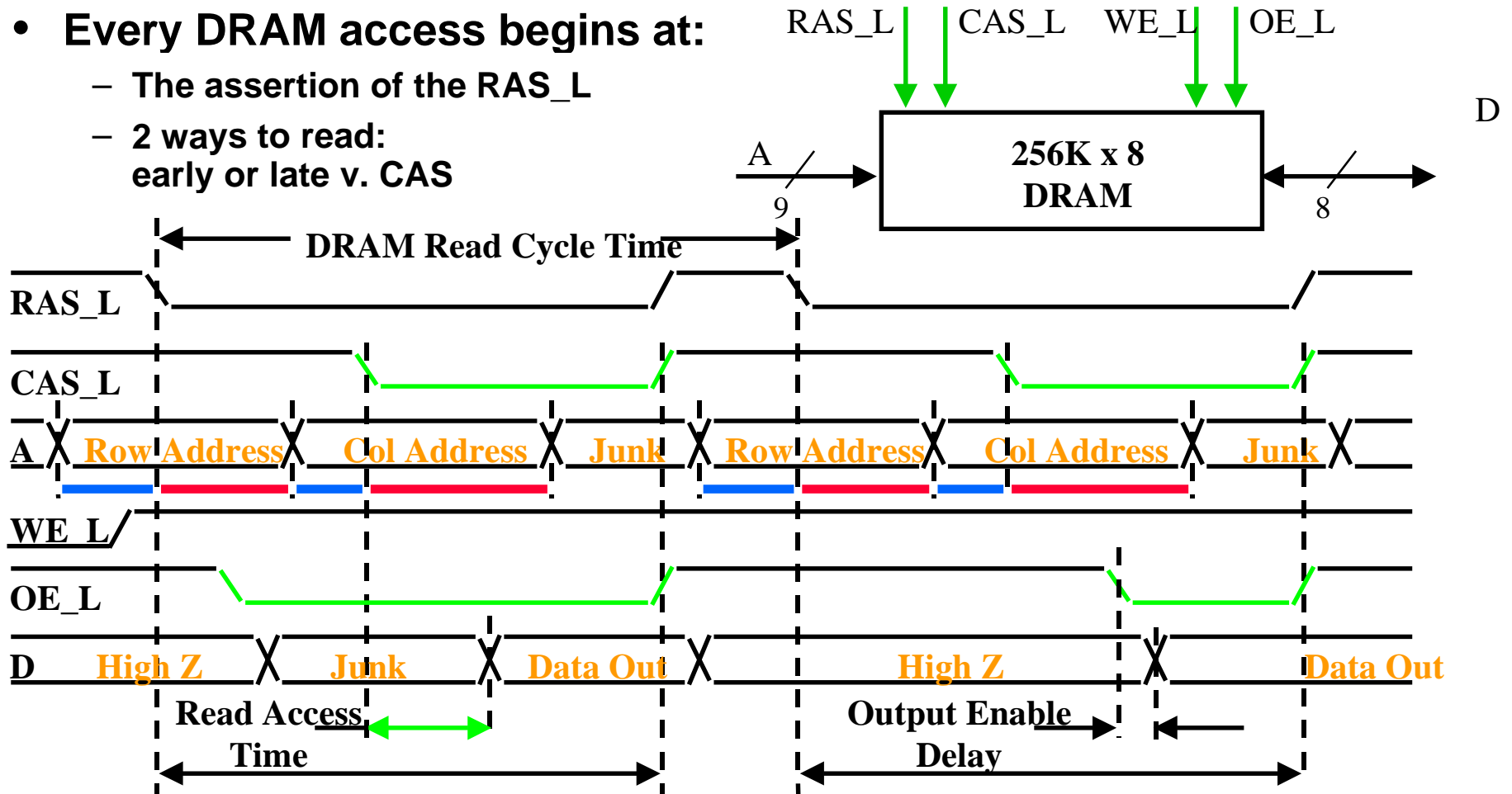


Early Wr Cycle: WE_L asserted before CAS_L Late Wr Cycle: WE_L asserted after CAS_L

DRAM Read Timing

- Every DRAM access begins at:

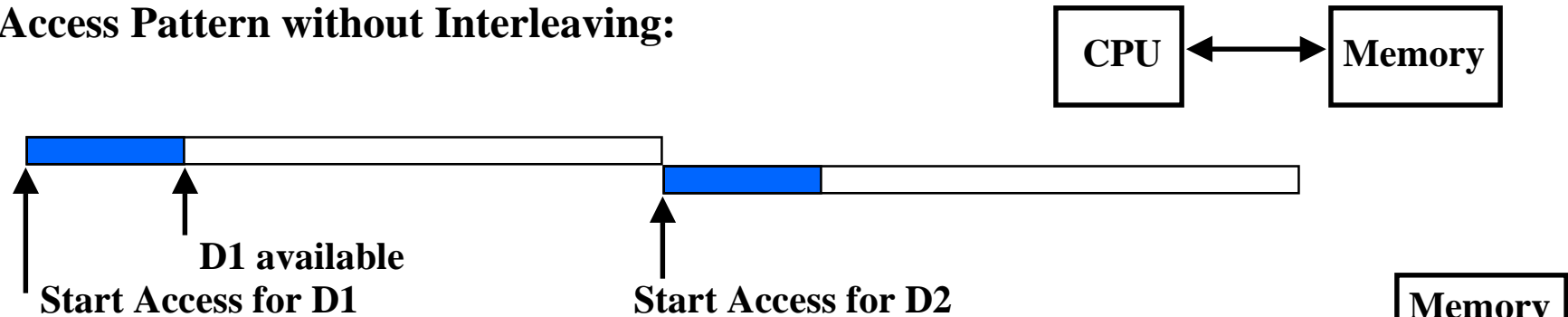
- The assertion of the RAS_L
- 2 ways to read:
early or late v. CAS



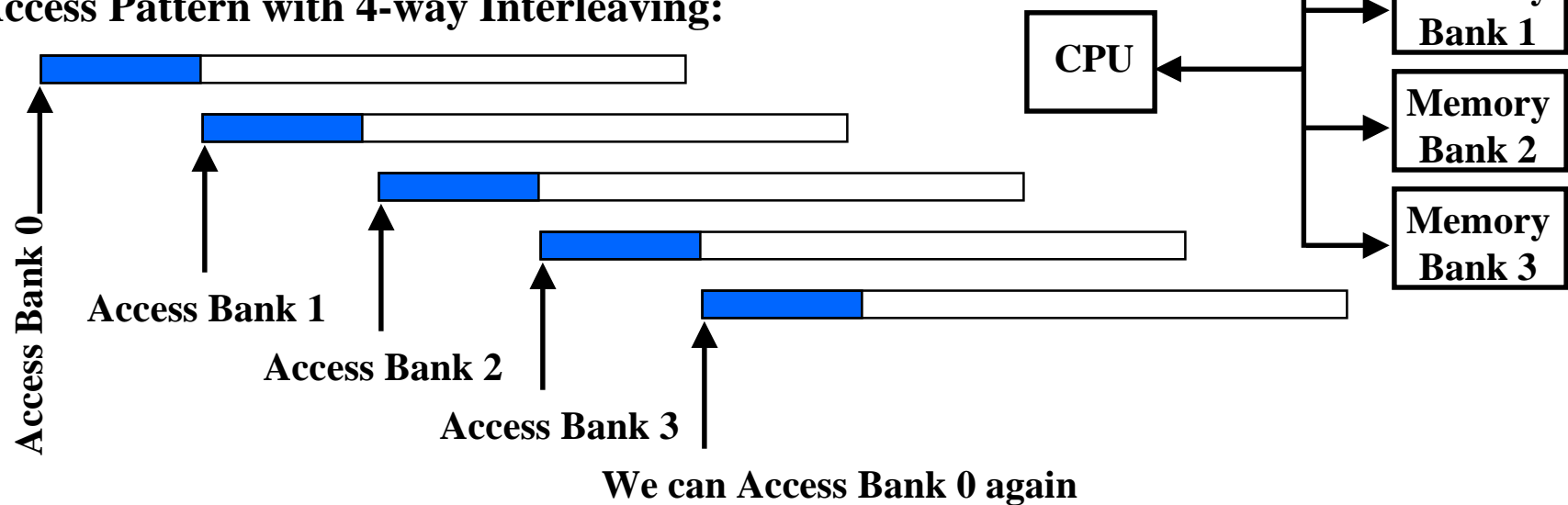
Early Read Cycle: OE_L asserted before CAS_L Late Read Cycle: OE_L asserted after CAS_L

Increasing Bandwidth - Interleaving

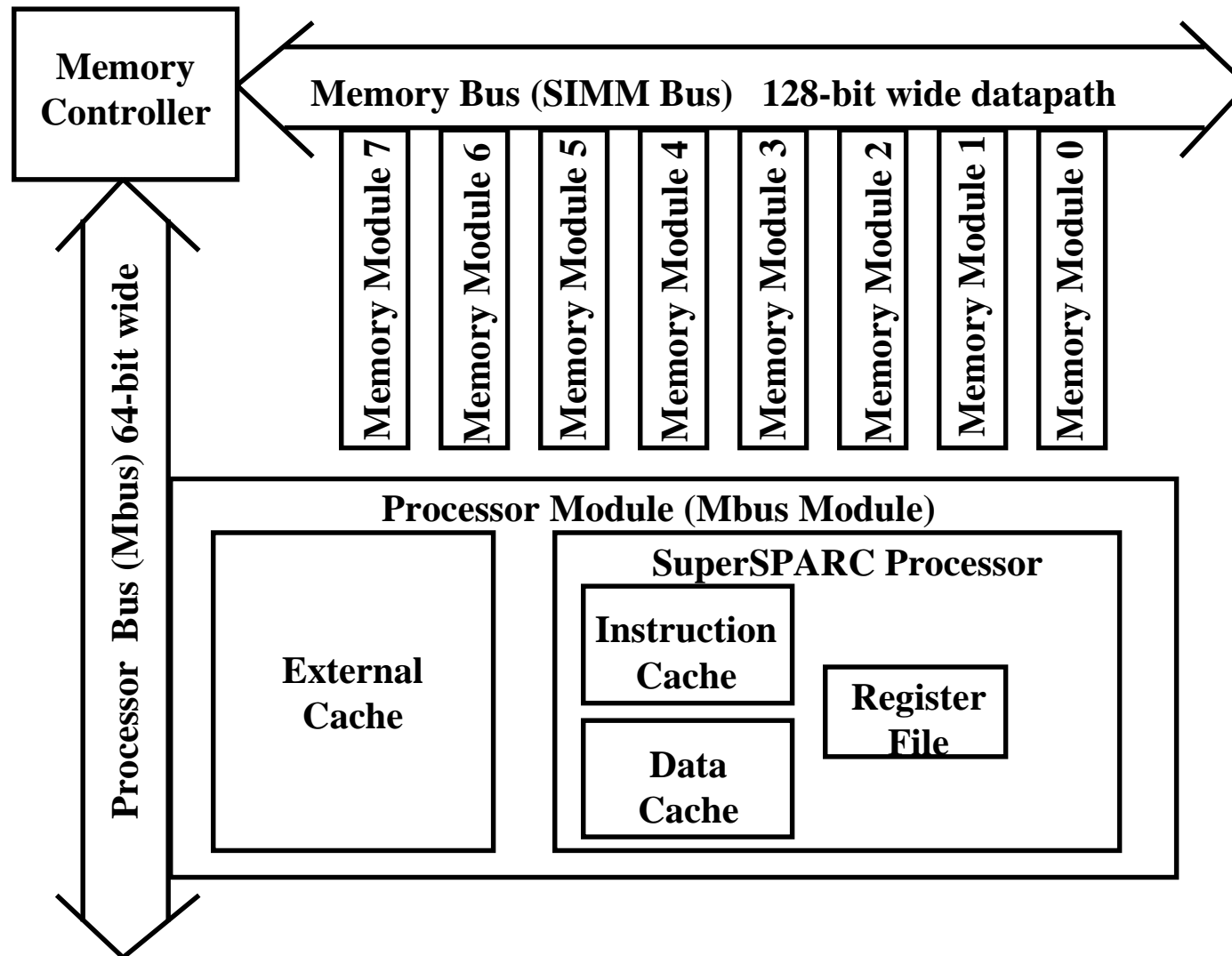
Access Pattern without Interleaving:



Access Pattern with 4-way Interleaving:



SPARCstation 20's Memory System Overview



Fast Memory Systems: DRAM specific

- **Multiple RAS accesses: several names (page mode)**
 - 64 Mbit DRAM: cycle time = 100 ns, page mode = 20 ns
- **New DRAMs?**
 - **Synchronous DRAM**: Provide a clock signal to DRAM, transfer synchronous to system clock
 - **RAMBUS**: reinvent DRAM interface (**Intel will use it**)
 - » Each Chip a module vs. slice of memory
 - » Short bus between CPU and chips
 - » Does own refresh
 - » Variable amount of data returned
 - » 1 byte / 2 ns (500 MB/s per chip)
 - **Cached DRAM (CDRAM)**: Keep entire row in SRAM, Gershon Kedem

Summary of Memory Technology

- **DRAM is slow but cheap and dense:**
 - Good choice for presenting the user with a BIG memory system
 - Uses one transistor, must be refreshed.
- **SRAM is fast but expensive and not very dense:**
 - Good choice for providing the user FAST access time.
 - Uses six transistors, holds state as long as power is supplied.
- **GOAL:**
 - Present the user with large amounts of memory using the cheapest technology.
 - Provide access at the speed offered by the fastest technology.
- **Next Time: Caches**