CS 423
Computer Architecture
Spring 2012

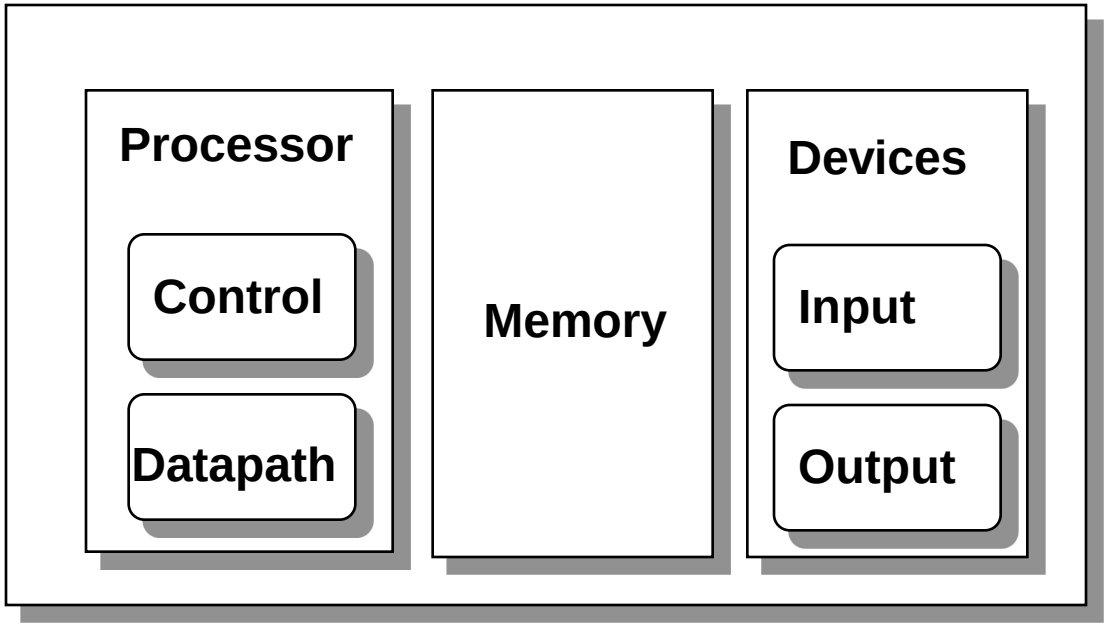
Lecture 07:
Memory

Ozcan Ozturk

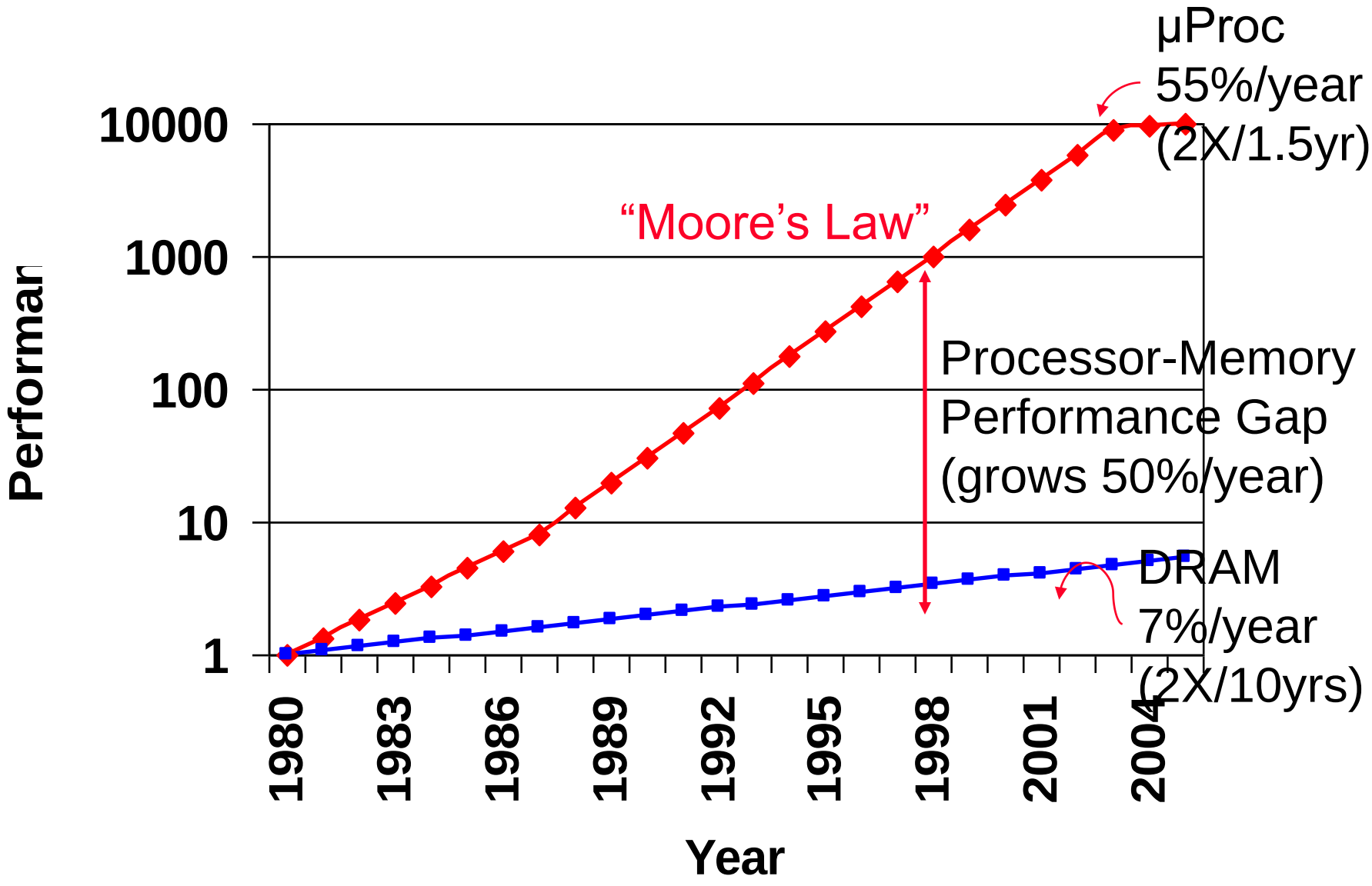
http://www.cs.bilkent.edu.tr/~ozturk/cs423/

[Adapted from *Computer Organization and Design*,
Patterson & Hennessy, © 2005, UCB]

Review: Major Components of a Computer

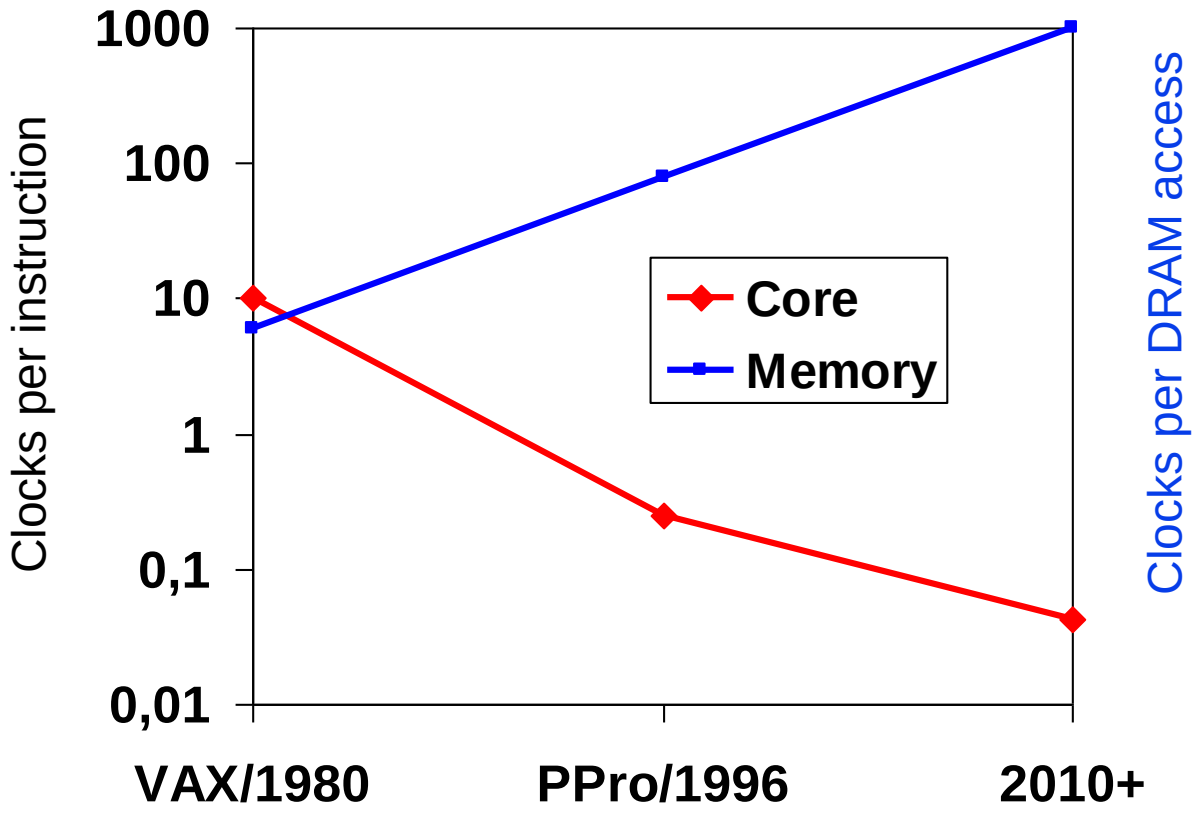


Processor-Memory Performance Gap



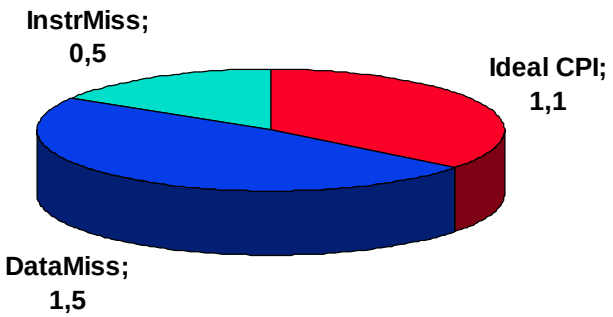
The “Memory Wall”

- Logic vs DRAM speed gap continues to grow



Memory Performance Impact on Performance

- Suppose a processor executes at
  - ideal CPI = 1.1
  - 50% arith/logic, 30% ld/st, 20% control



and that 10% of data memory operations miss with a 50 cycle miss penalty

- CPI = ideal CPI + average stalls per instruction
  = 1.1(cycle) + (0.30 (datamemops/instr)
  x 0.10 (miss/datamemop) x 50 (cycle/miss))
  = 1.1 cycle + 1.5 cycle = 2.6

so 58% of the time the processor is stalled waiting for memory!

- A 1% instruction miss rate would add an *additional* 0.5 to the CPI!

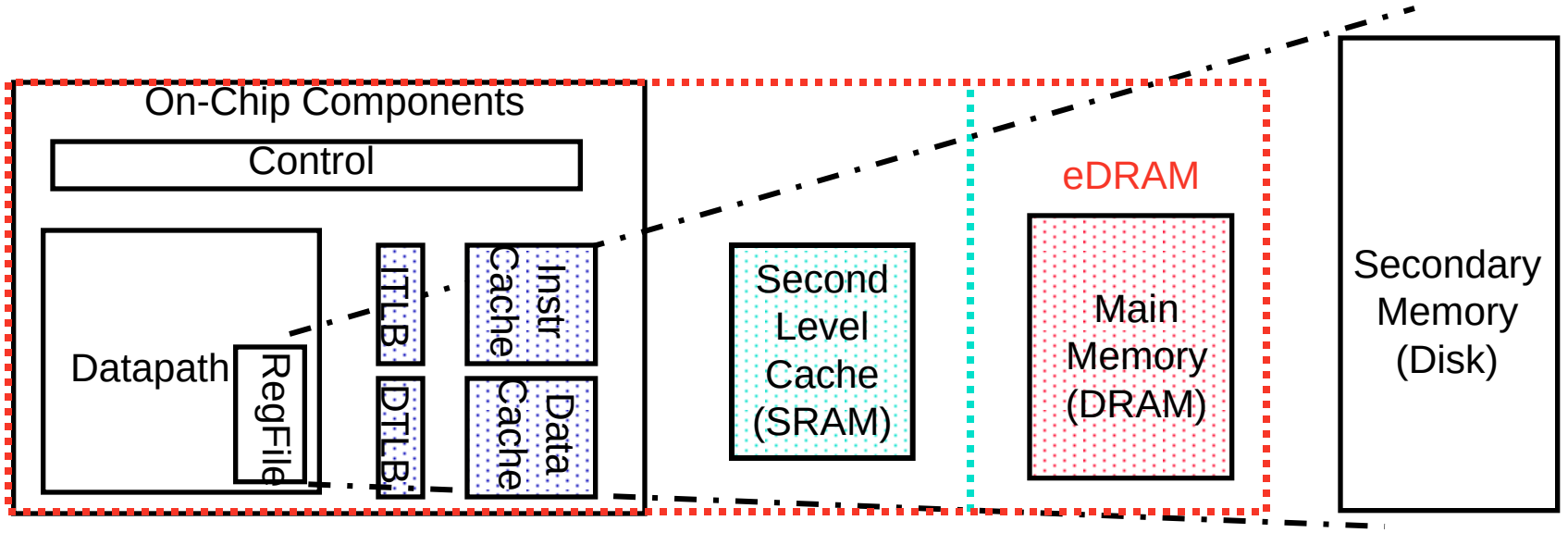
The Memory Hierarchy Goal

- Fact: Large memories are slow and fast memories are small

- How do we create a memory that gives the illusion of being large, cheap and fast (most of the time)?
  - With hierarchy
  - With parallelism

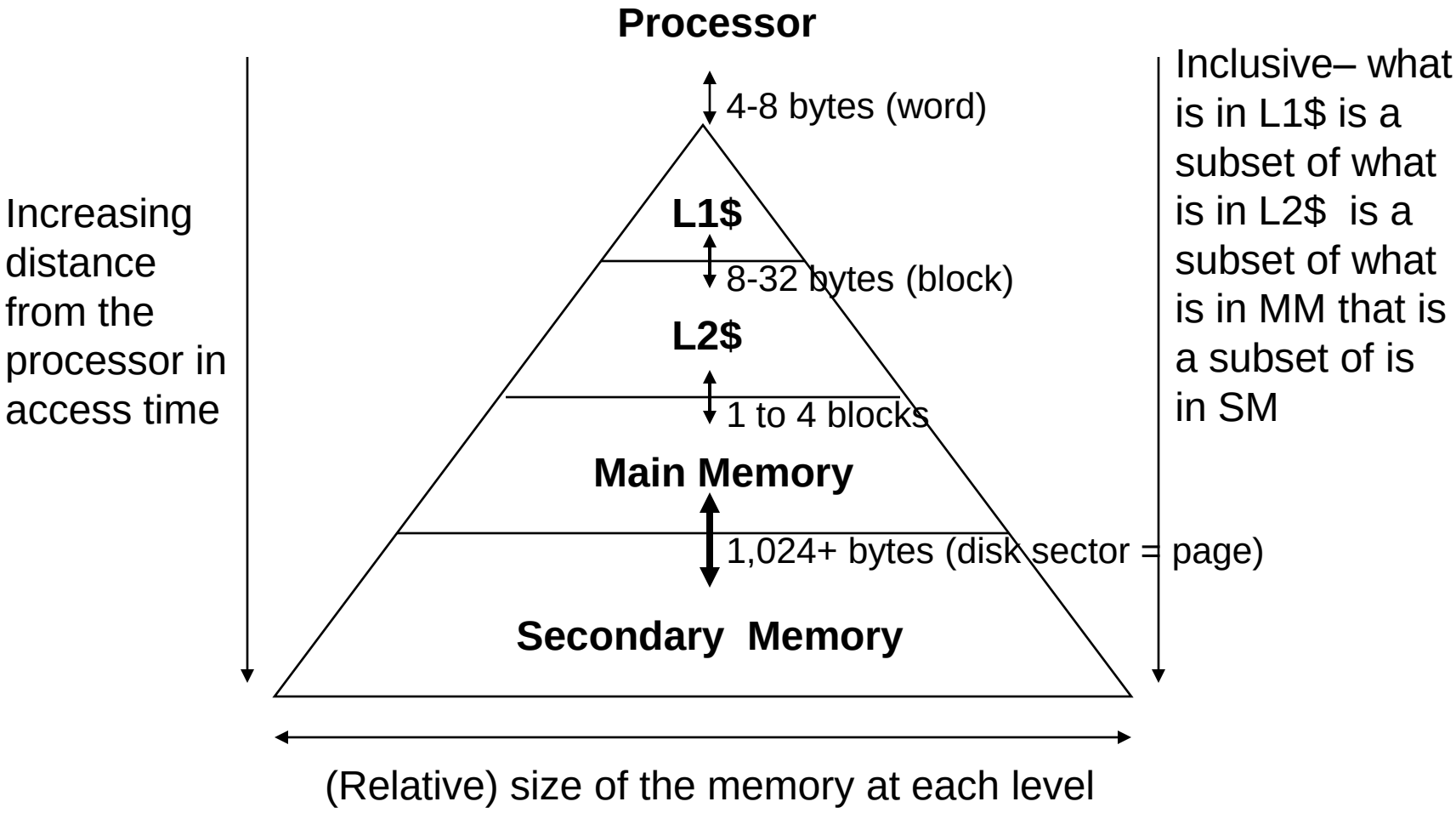
A Typical Memory Hierarchy

- By taking advantage of the principle of locality
  - Can present the user with as much memory as is available in the cheapest technology
  - at the speed offered by the fastest technology



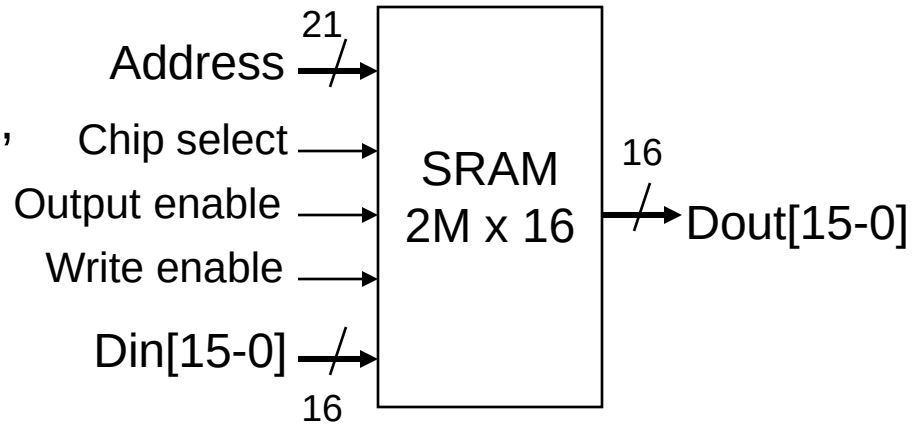
Speed (%cycles):	½'s	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	G's to T's
Cost:	highest				lowest

Characteristics of the Memory Hierarchy



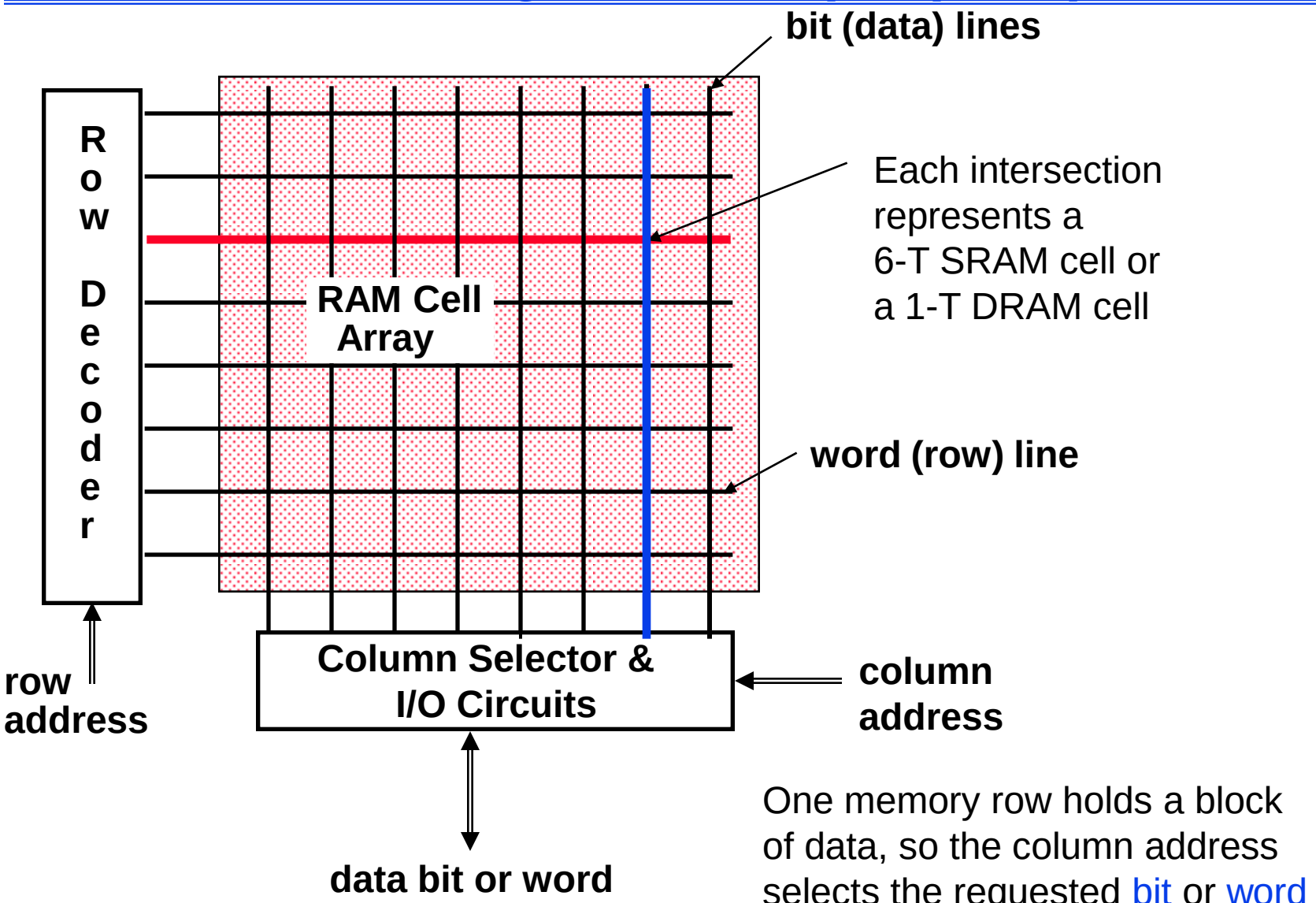
Memory Hierarchy Technologies

- Caches use *SRAM* for speed and technology compatibility
  - Low density (6 transistor cells), high power, expensive, fast
  - Static: content will last "forever" (until power turned off)



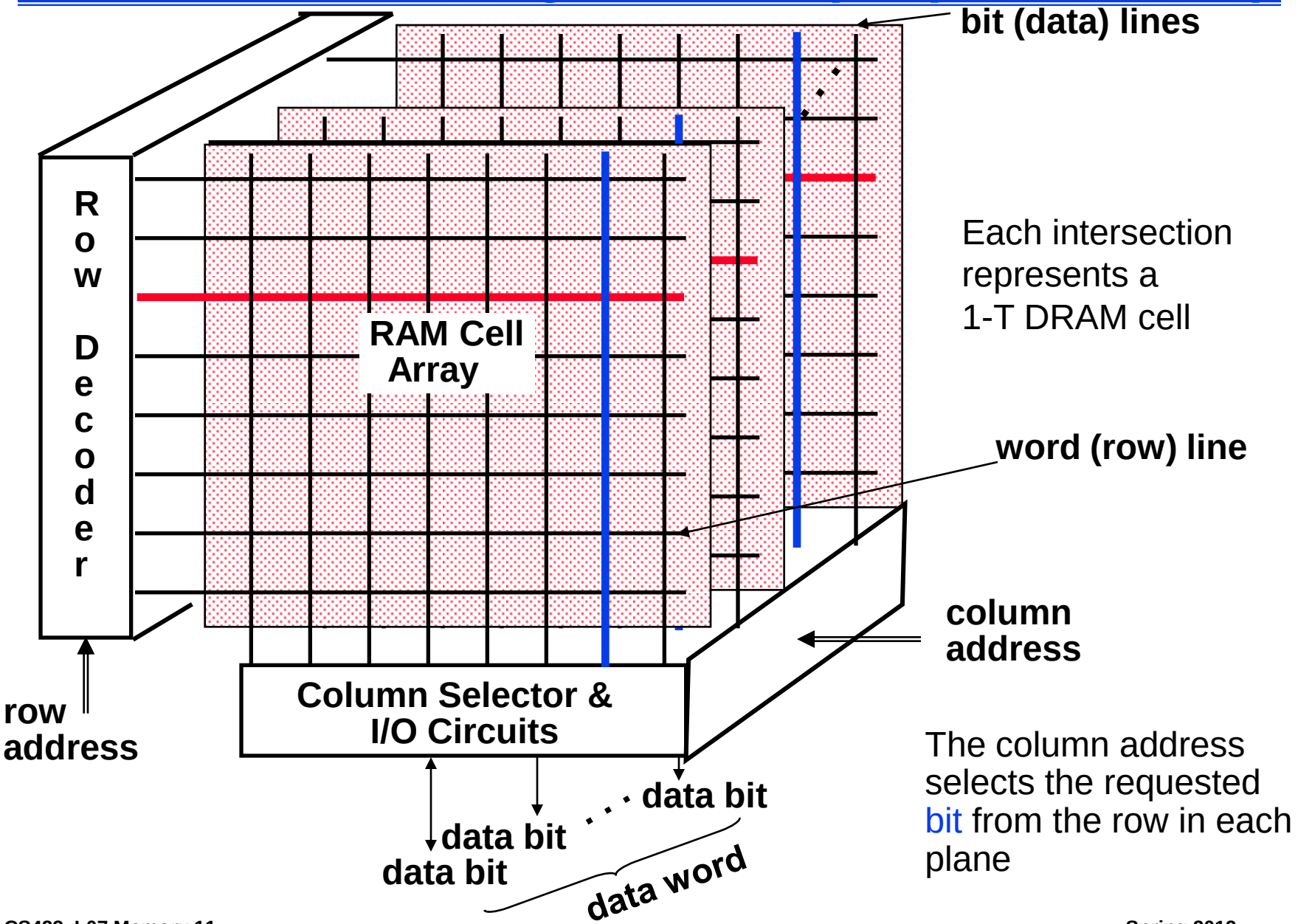
- Main Memory uses *DRAM* for size (density)
  - High density (1 transistor cells), low power, cheap, slow
  - Dynamic: needs to be "refreshed" regularly (~ every 8 ms)
    - 1% to 2% of the active cycles of the DRAM
  - Addresses divided into 2 halves (row and column)
    - *RAS* or *Row Access Strobe* triggering row decoder
    - *CAS* or *Column Access Strobe* triggering column selector

Classical RAM Organization (~Square)



One memory row holds a block of data, so the column address selects the requested bit or word from that block

Classical DRAM Organization (~Square Planes)

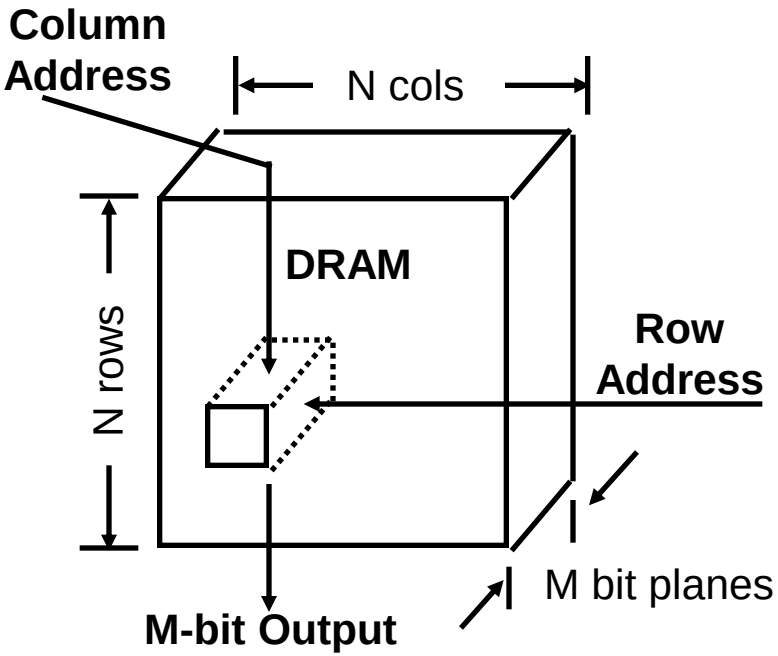


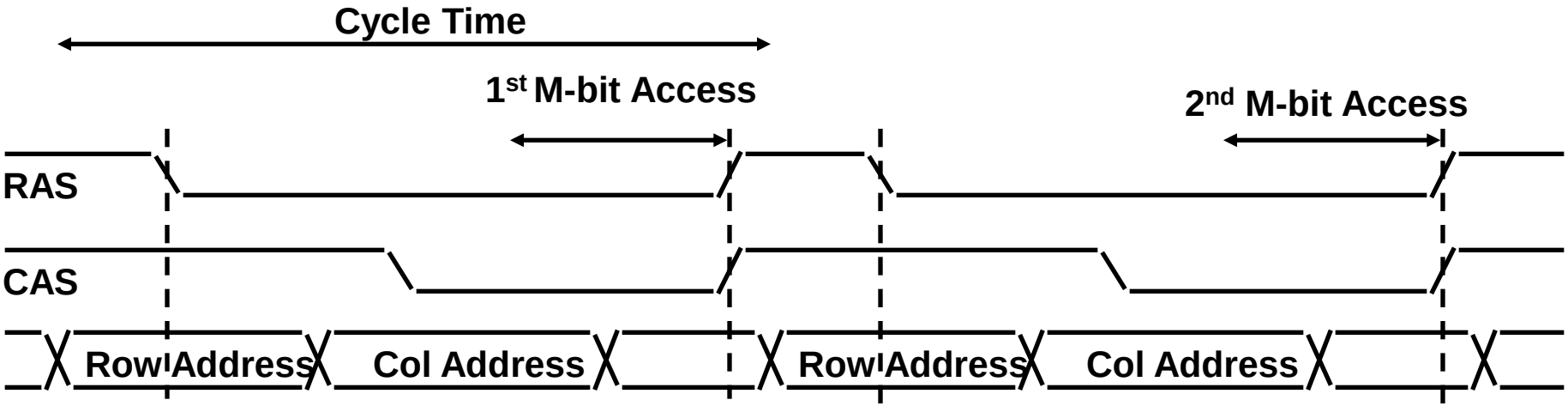
Each intersection represents a 1-T DRAM cell

The column address selects the requested bit from the row in each plane

Classical DRAM Operation

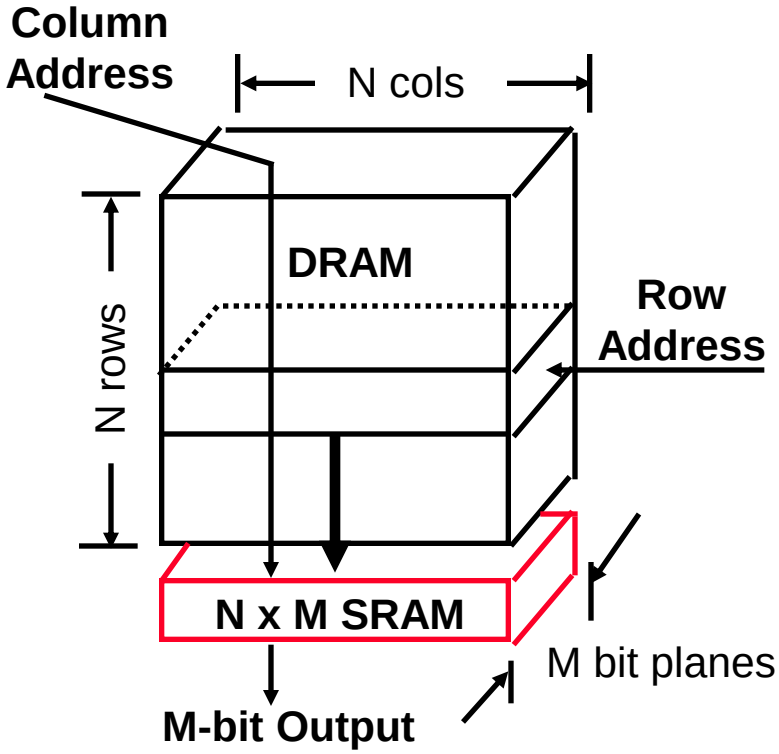
- DRAM Organization:
  - N rows x N column x M-bit
  - Read or Write M-bit at a time
  - Each M-bit access requires a RAS / CAS cycle

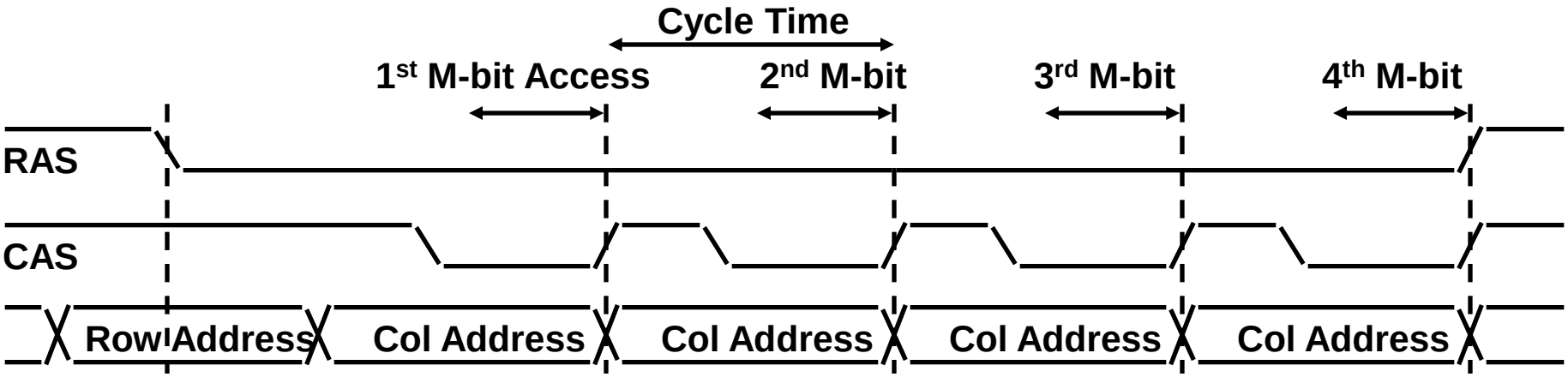




Page Mode DRAM Operation

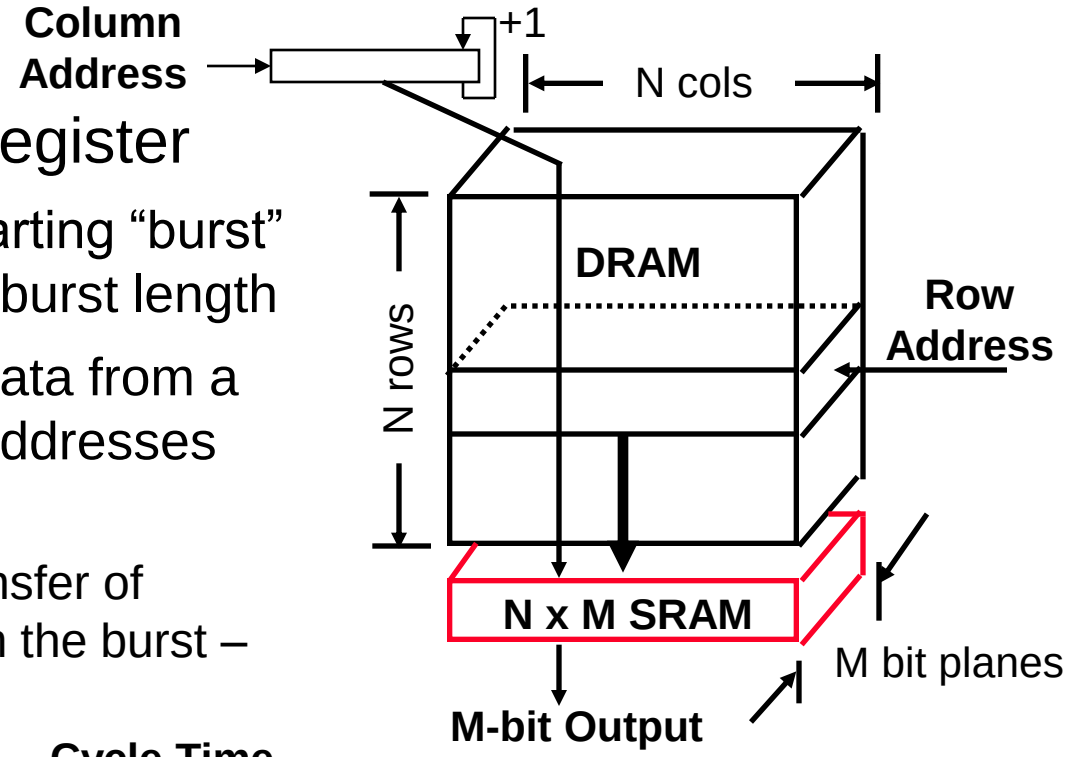
- Page Mode DRAM
  - N x M SRAM to save a row
- After a row is read into the SRAM "register"
  - Only CAS is needed to access other M-bit words on that row
  - RAS remains asserted while CAS is toggled

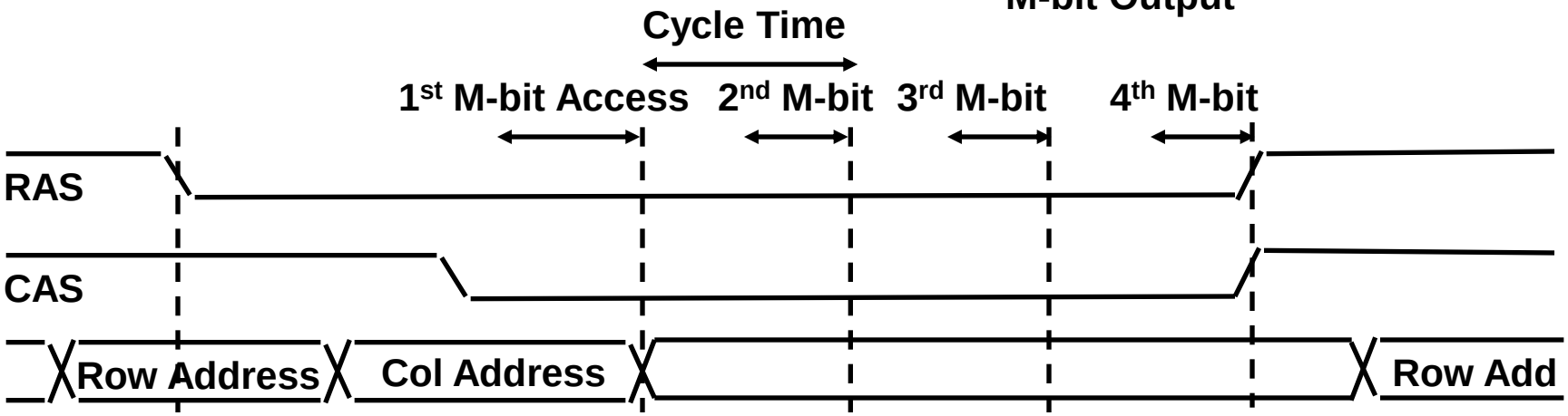




Synchronous DRAM (SDRAM) Operation

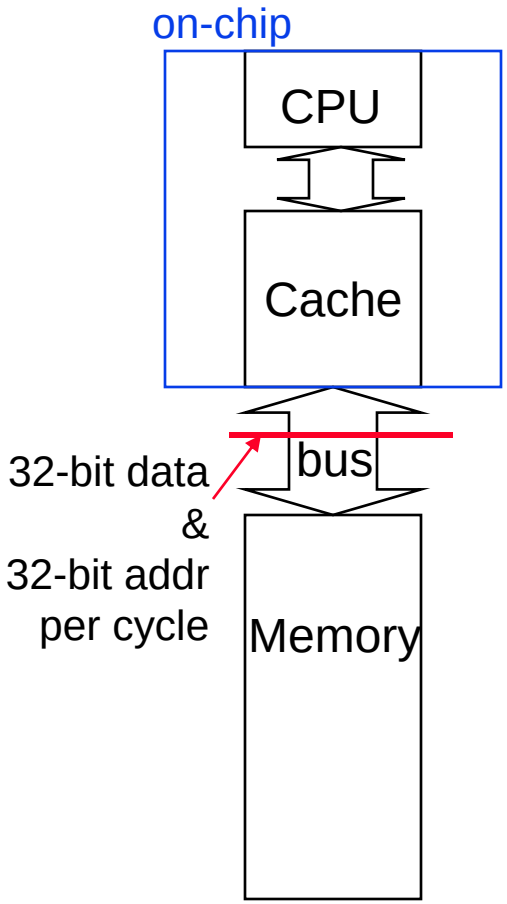
- After a row is read into the SRAM register
  - Inputs CAS as the starting "burst" address along with a burst length
  - Transfers a burst of data from a series of sequential addresses within that row
    - A clock controls transfer of successive words in the burst – 300MHz in 2004





Memory Systems that Support Caches

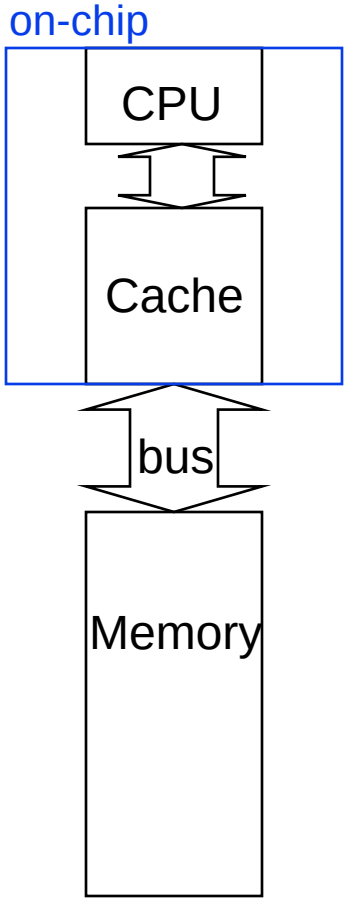
- The off-chip interconnect and memory architecture can affect overall system performance in dramatic ways



One word wide organization
(one word wide bus and
one word wide memory)

- Assume
  1. 1 clock cycle to send the address
  2. 25 clock cycles for DRAM cycle time, 8 clock cycles access time
  3. 1 clock cycle to return a word of data
- Memory-Bus to Cache bandwidth
  - number of bytes accessed from memory and transferred to cache/CPU per clock cycle

One Word Wide Memory Organization



- If the block size is one word, then for a memory access due to a cache miss, the pipeline will have to stall the number of cycles required to return one data word from memory

  cycle to send address

  cycles to read DRAM

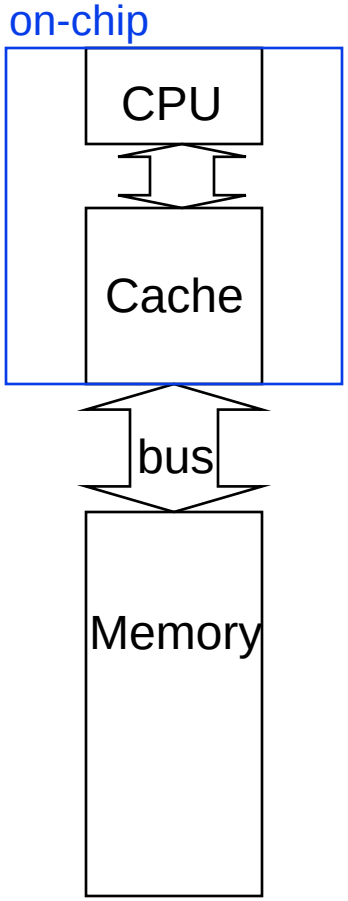
  cycle to return data

  —— total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

  bytes per clock

One Word Wide Memory Organization, con't



- What if the block size is four words?

  cycle to send 1st address

  cycles to read DRAM

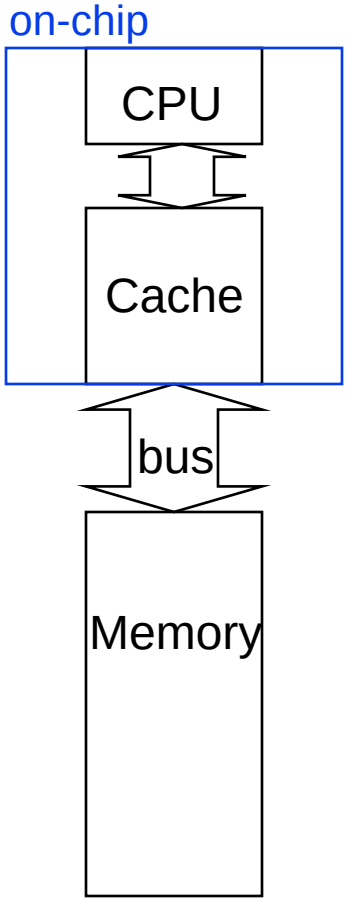
  cycles to return last data word

  total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

  bytes per clock

One Word Wide Memory Organization, con't



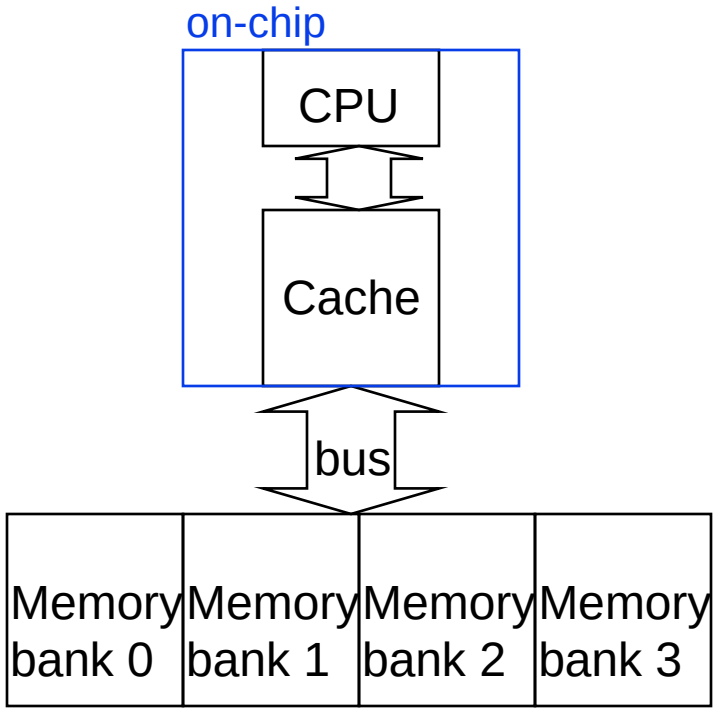
- What if the block size is four words and if a fast page mode DRAM is used?

	cycle to send 1st address
	cycles to read DRAM
	cycles to return last data word
—	total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

bytes per clock

Interleaved Memory Organization



- For a block size of four words

  cycle to send 1st address

  cycles to read DRAM

  cycles to return last data word

  total clock cycles miss penalty

- Number of bytes transferred per clock cycle (bandwidth) for a single miss is

  bytes per clock

DRAM Memory System Summary

- Its important to match the cache characteristics
  - caches access one block at a time (usually more than one word)
- with the DRAM characteristics
  - use DRAMs that support fast multiple word accesses, preferably ones that match the block size of the cache
- with the memory-bus characteristics
  - make sure the memory-bus can support the DRAM access rates and patterns
  - with the goal of increasing the Memory-Bus to Cache bandwidth

The Memory Hierarchy: Why Does it Work?

- Temporal Locality (Locality in Time):
  - ⇒ Keep most recently accessed data items closer to the processor
- Spatial Locality (Locality in Space):
  - ⇒ Move blocks consisting of contiguous words to the upper levels