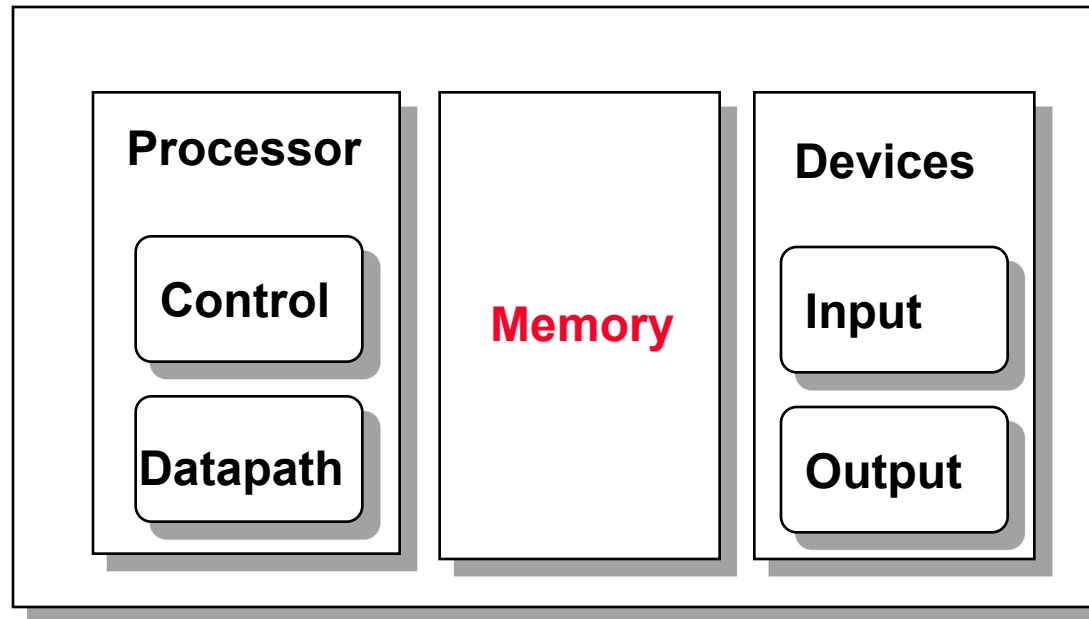
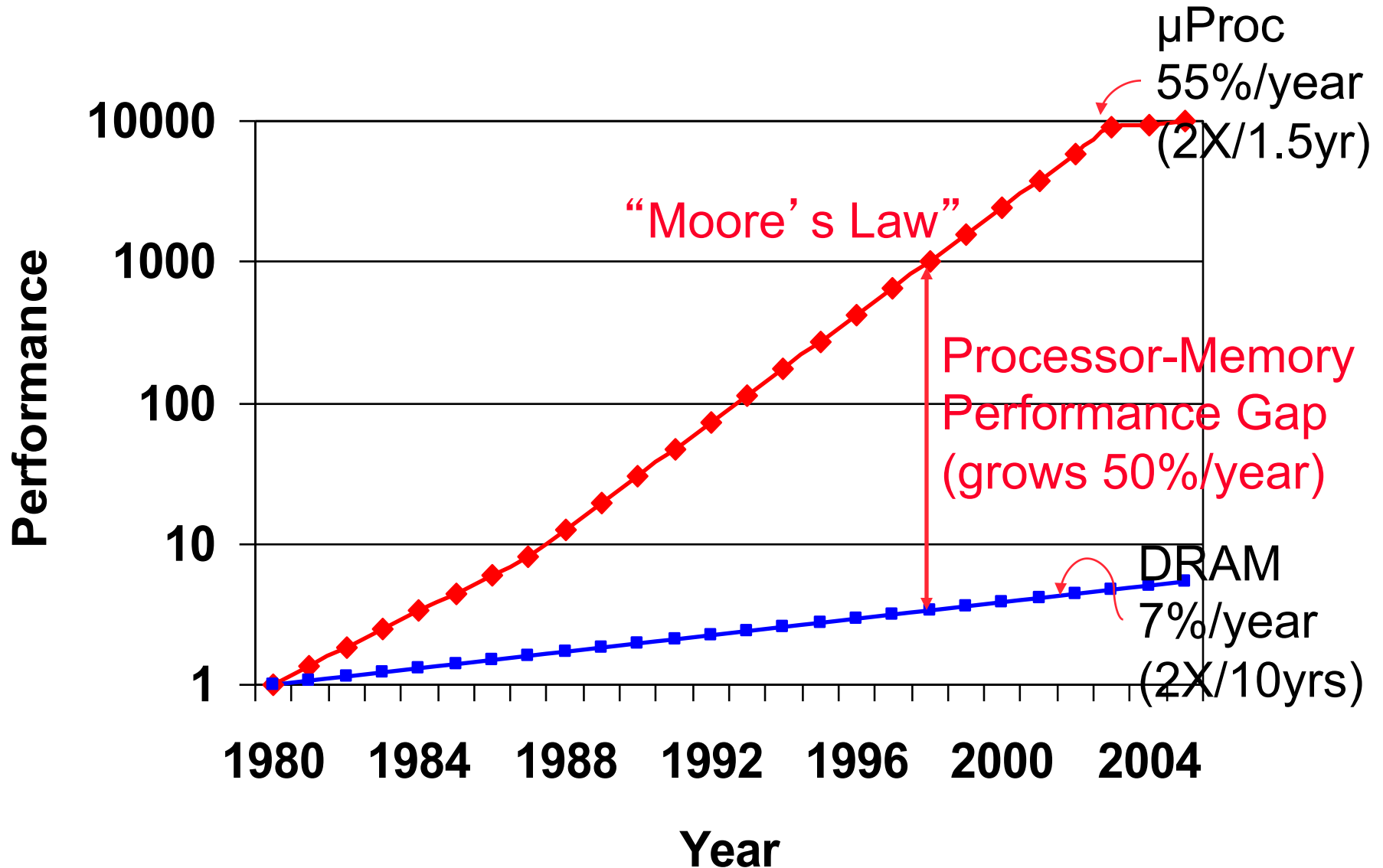

Memory Hierarchy Review

[Adapted from Mary Jane Irwin for
Computer Organization and Design,
Patterson & Hennessy, © 2005, UCB]

Review: Major Components of a Computer

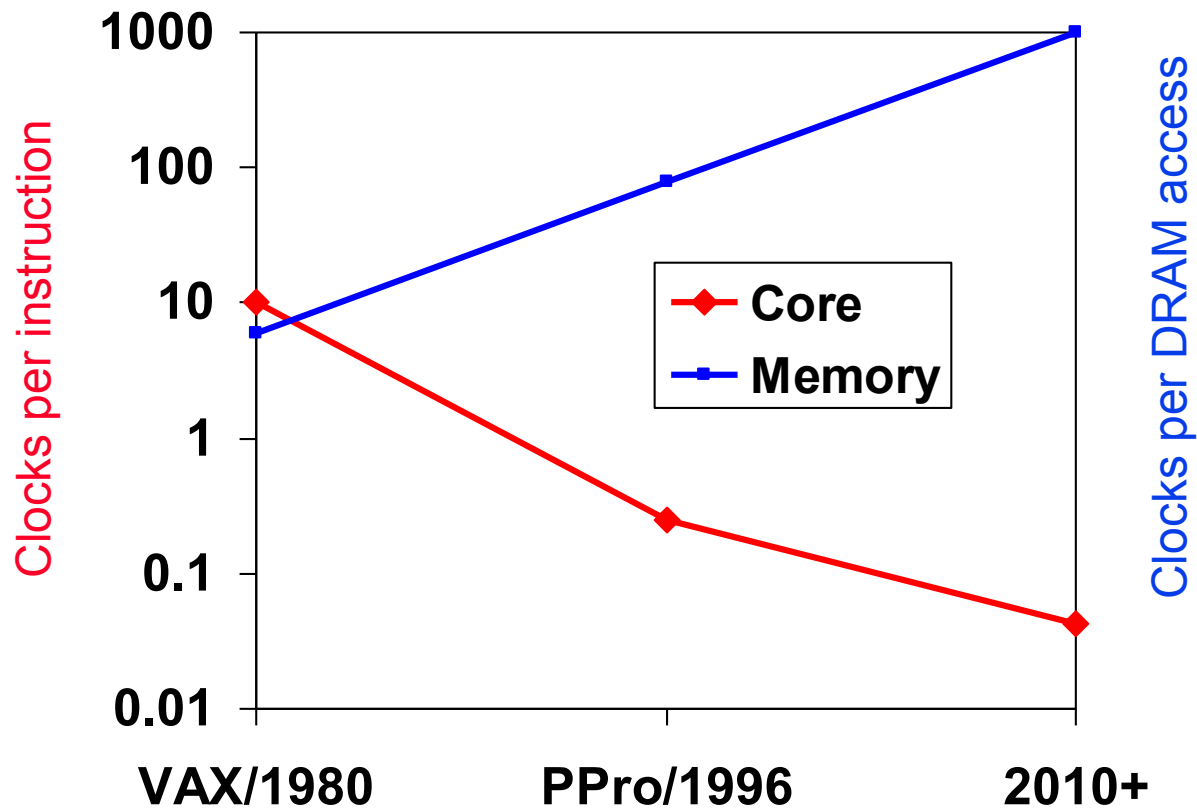


Processor-Memory Performance Gap



The “Memory Wall”

- ❑ Logic vs DRAM speed gap continues to grow



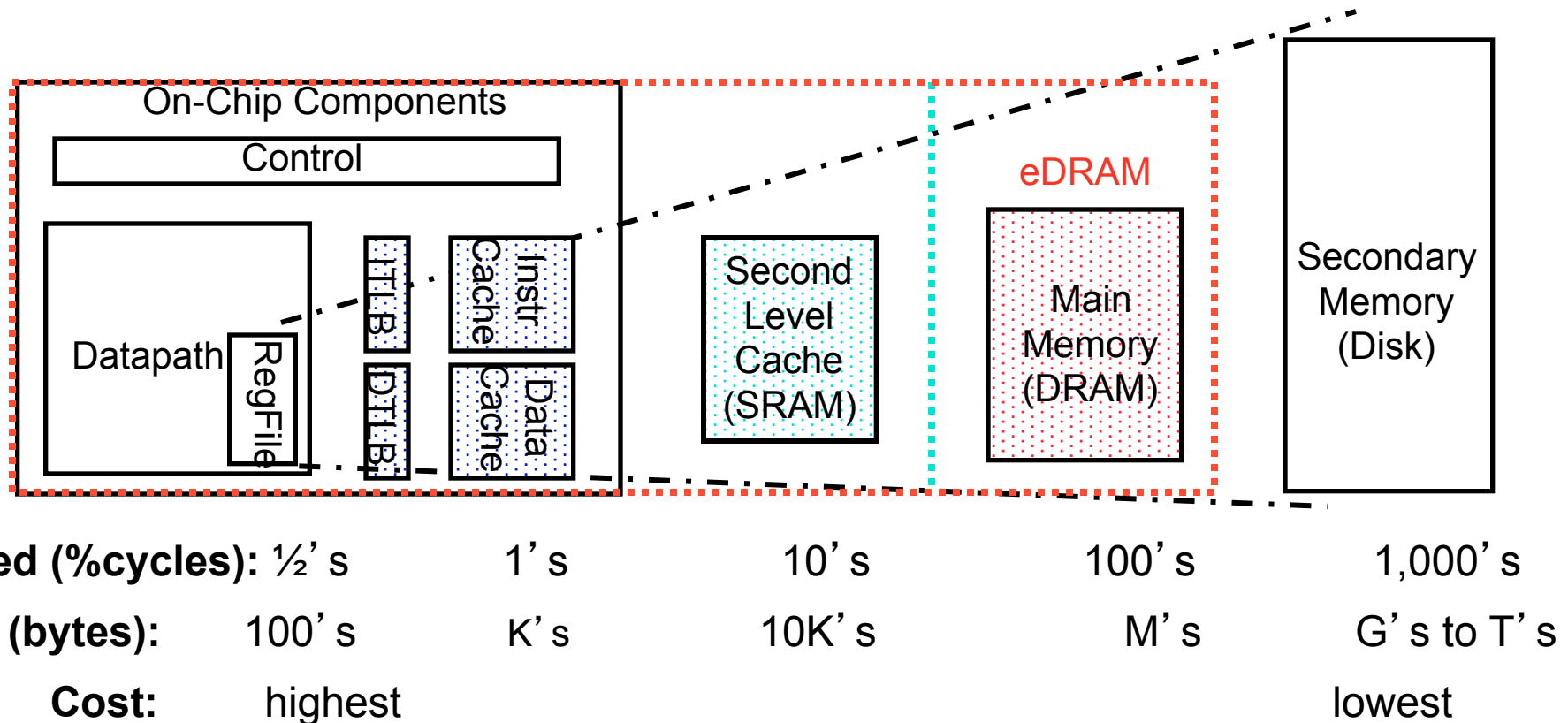
The Memory Hierarchy Goal

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

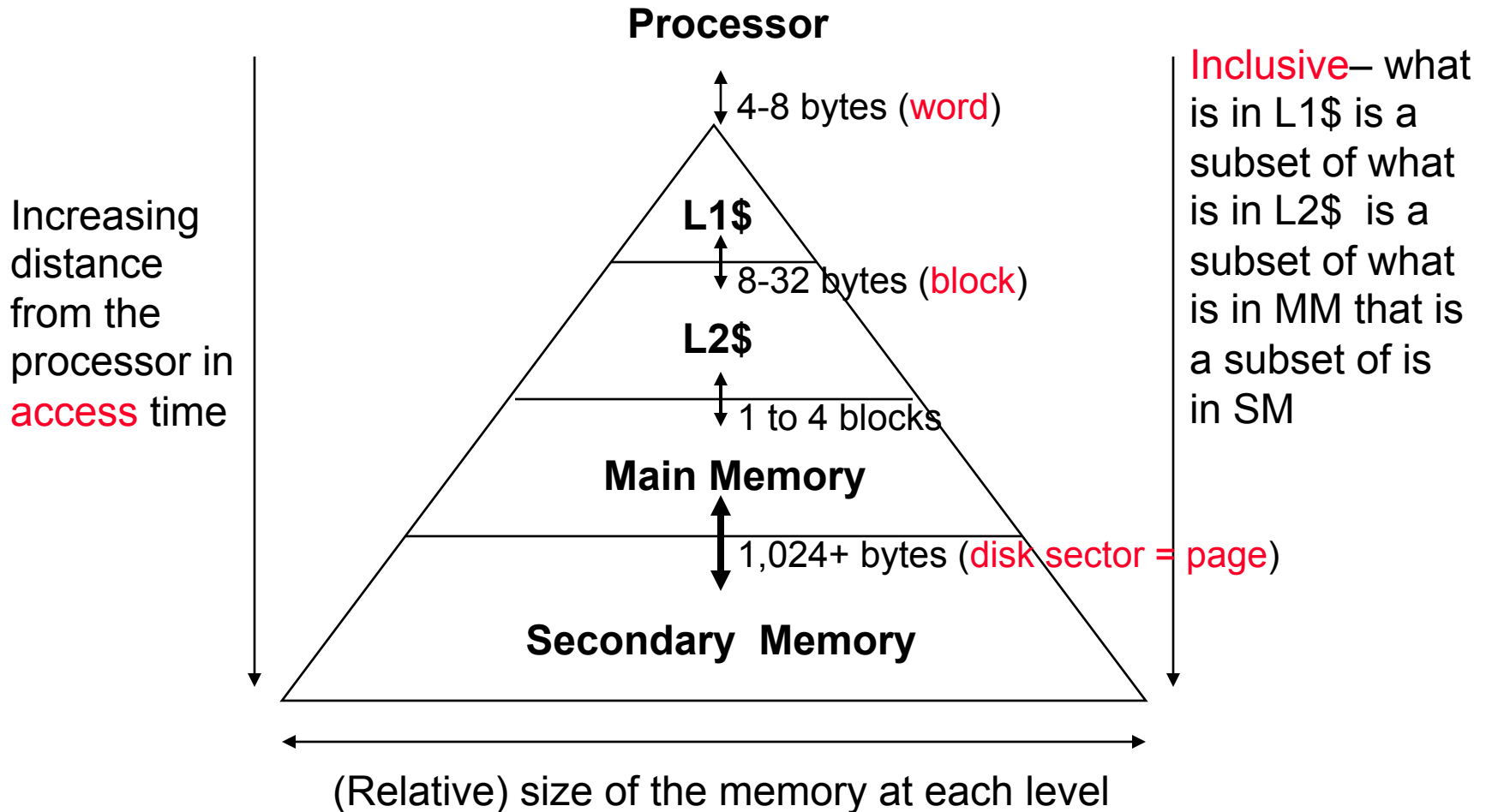
- ❑ How do we create a memory that gives the illusion of being large, cheap and fast (most of the time)?
 - With hierarchy
 - With parallelism

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



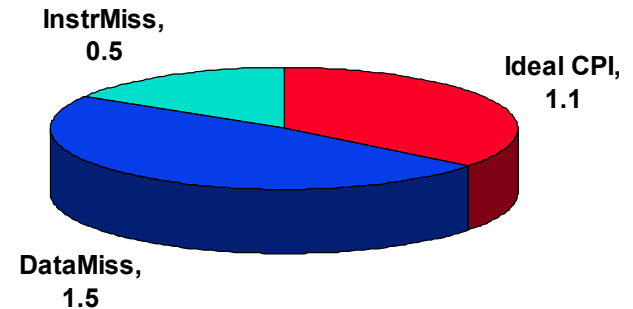
Characteristics of the Memory Hierarchy



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



- ❑
$$\begin{aligned}\text{CPI} &= \text{ideal CPI} + \text{average stalls per instruction} \\ &= 1.1(\text{cycle}) + (0.30 (\text{datamemops/instr}) \\ &\quad \times 0.10 (\text{miss/datamemop}) \times 50 (\text{cycle/miss})) \\ &= 1.1 \text{ cycle} + 1.5 \text{ cycle} = 2.6\end{aligned}$$

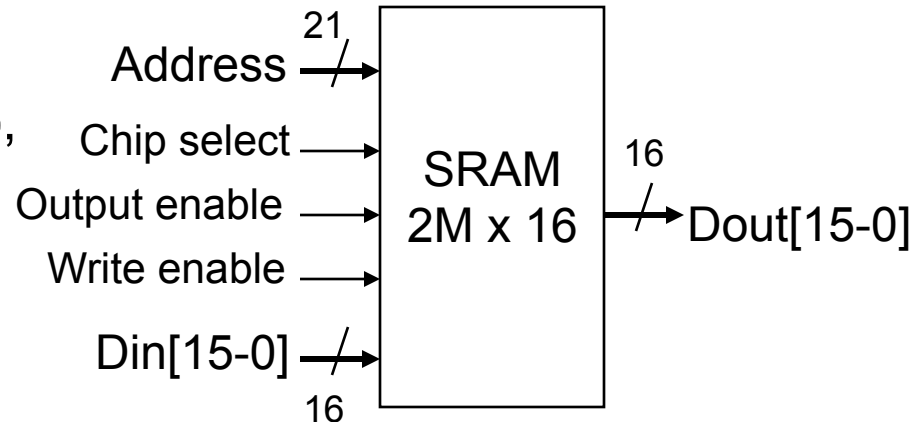
so 58% of the time the processor is stalled waiting for
memory!

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

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)



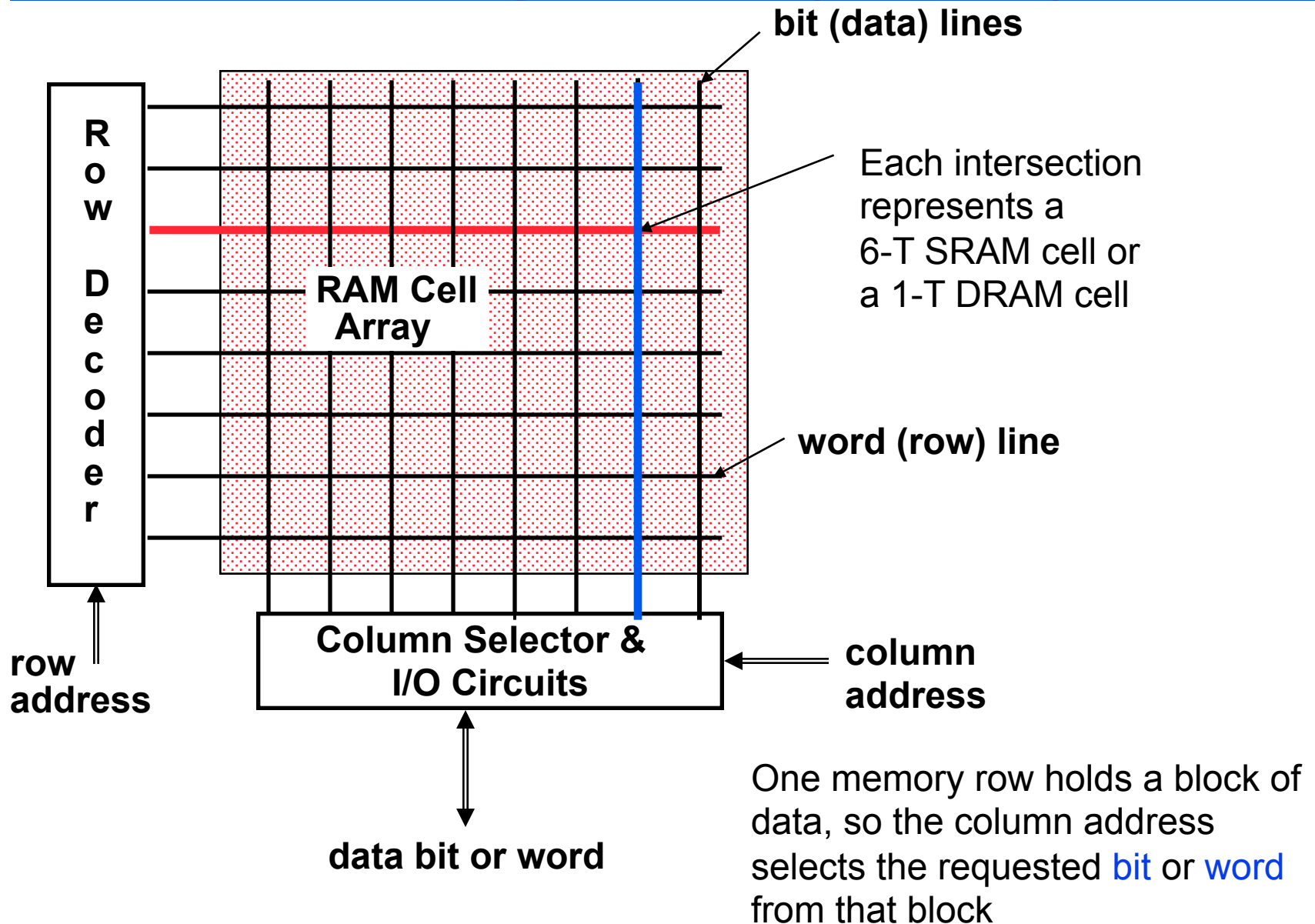
- ❑ Main Memory uses *DRAM* for size (density)

- 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

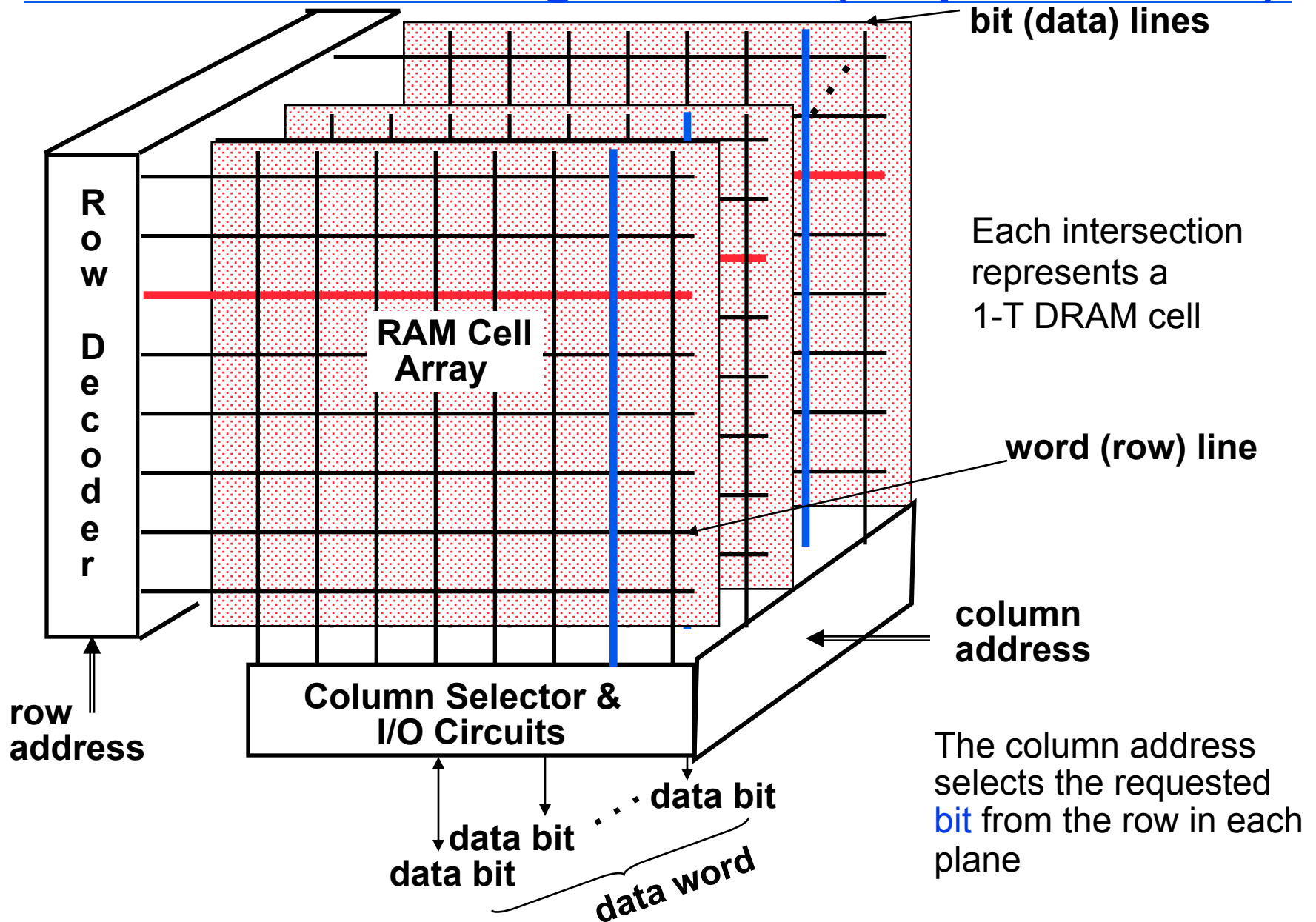
Memory Performance Metrics

- ❑ **Latency**: Time to access one word
 - *Access time*: time between the request and when the data is available (or written)
 - *Cycle time*: time between requests
 - Usually cycle time > access time
 - Typical read access times for SRAMs in 2004 are 2 to 4 ns for the fastest parts to 8 to 20ns for the typical largest parts
- ❑ **Bandwidth**: How much data from the memory can be supplied to the processor per unit time
 - width of the data channel * the rate at which it can be used

Classical RAM Organization (~Square)



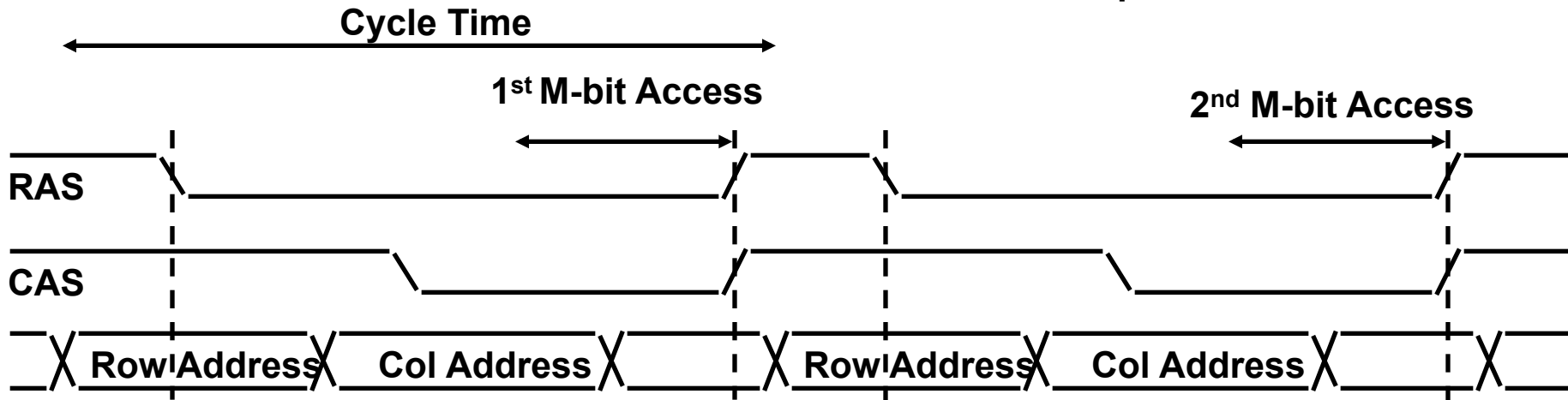
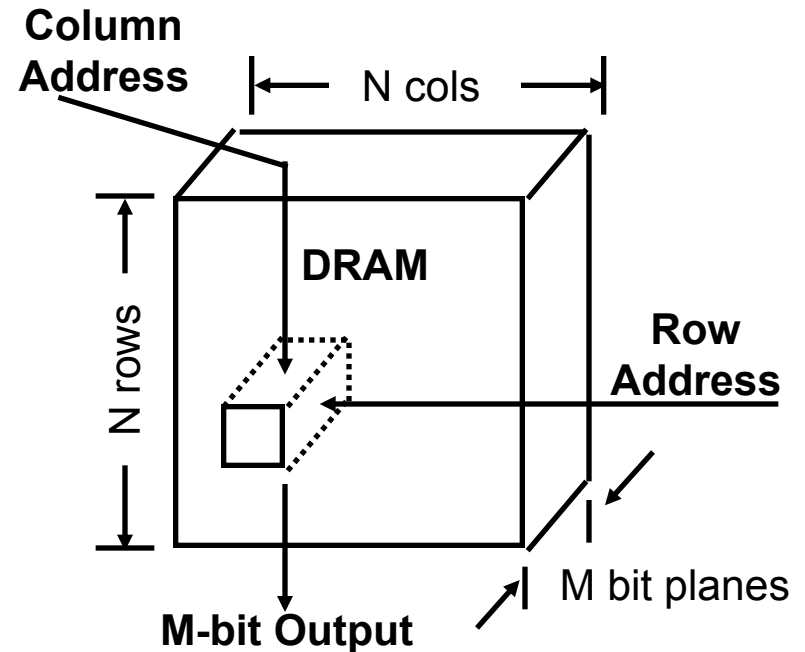
Classical DRAM Organization (~Square Planes)



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



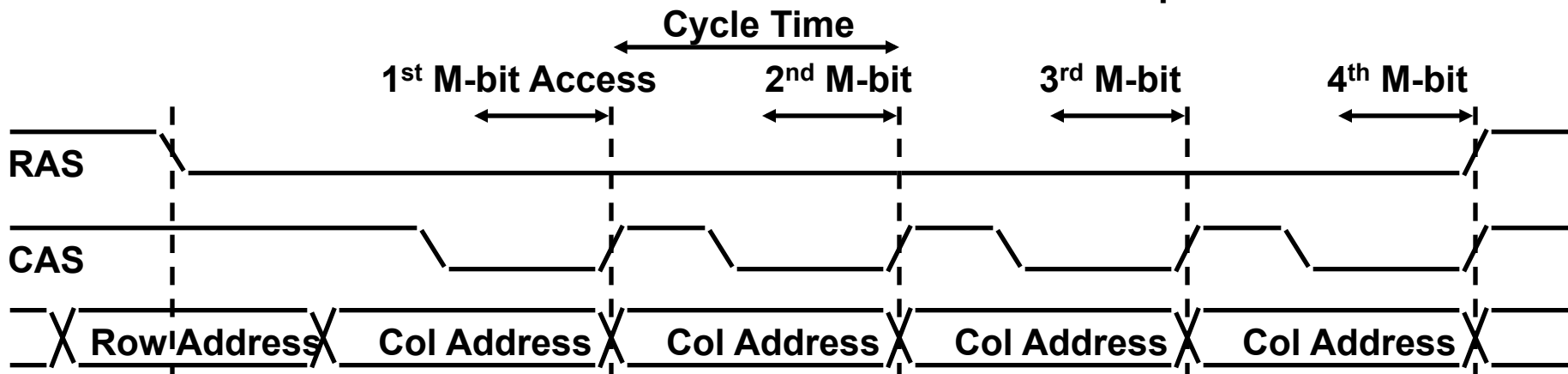
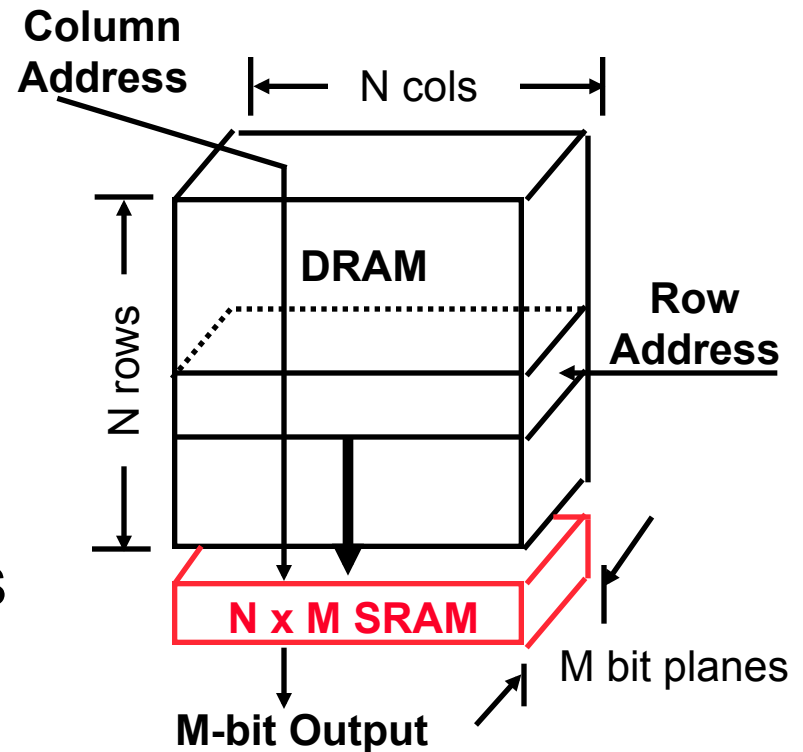
Page Mode DRAM Operation

❑ Page Mode DRAM

- $N \times M$ 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

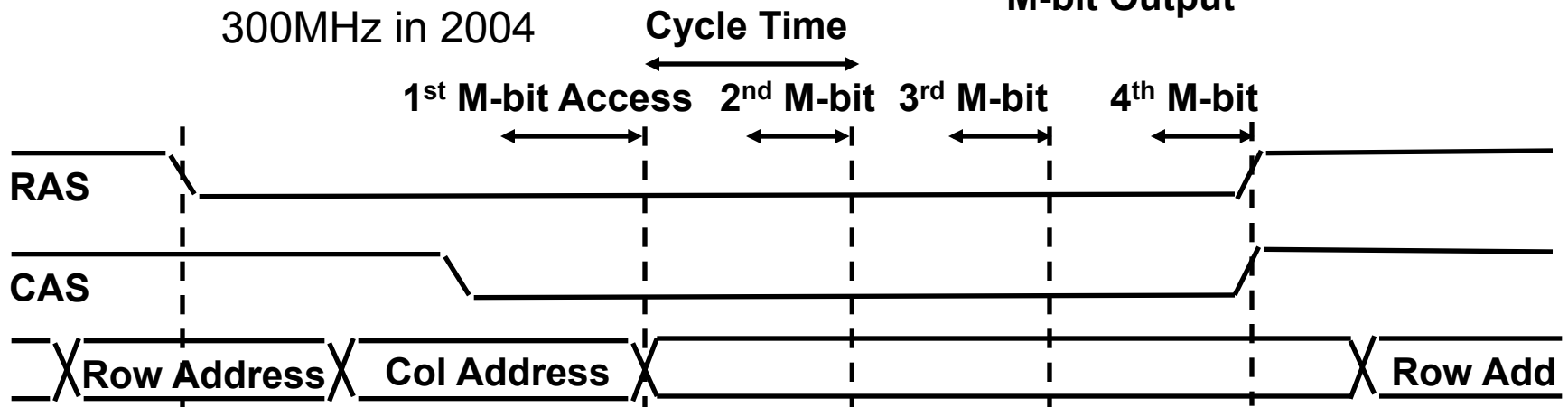
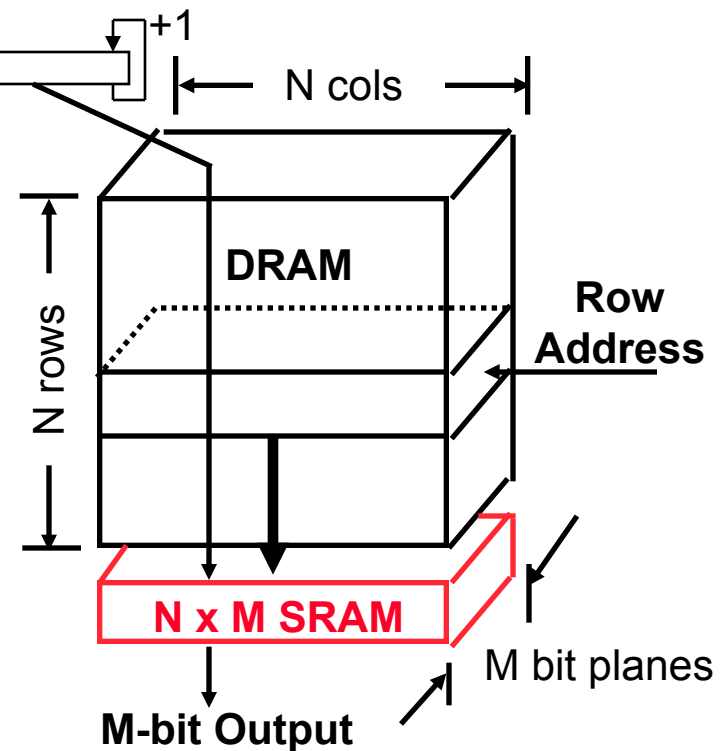


Synchronous DRAM (SDRAM) Operation

❑ 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



Other DRAM Architectures

- ❑ Double Data Rate SDRAMs – DDR-SDRAMs (and DDR-SDRAMs)
 - Double data rate because they transfer data on both the rising and falling edge of the clock
 - Are the most widely used form of SDRAMs

- ❑ DDR2-SDRAMs

DRAM Memory Latency & Bandwidth Milestones

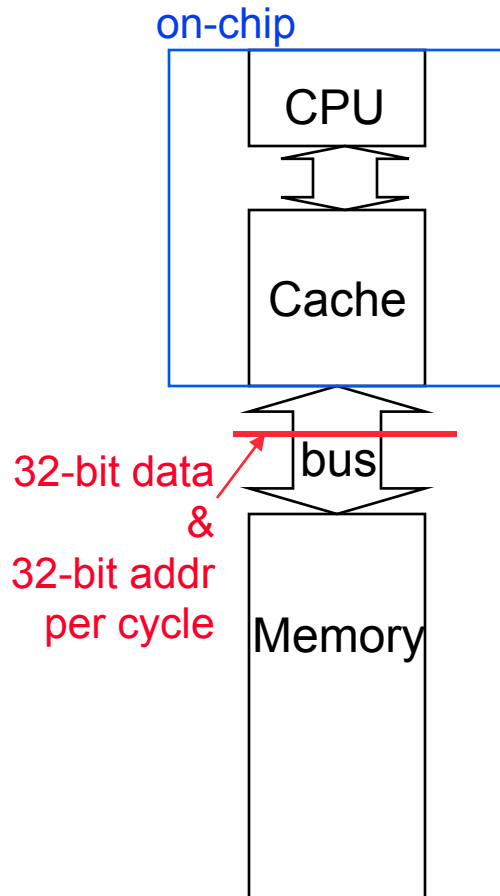
	DRAM	Page DRAM	FastPage DRAM	FastPage DRAM	Synch DRAM	DDR SDRAM
Module Width	16b	16b	32b	64b	64b	64b
Year	1980	1983	1986	1993	1997	2000
Mb/chip	0.06	0.25	1	16	64	256
Die size (mm ²)	35	45	70	130	170	204
Pins/chip	16	16	18	20	54	66
BWidth (MB/s)	13	40	160	267	640	1600
Latency (nsec)	225	170	125	75	62	52

Patterson, CACM Vol 47, #10, 2004

- ❑ In the time that the memory to processor **bandwidth doubles** the memory **latency** improves by a factor of only **1.2 to 1.4**
- ❑ To deliver such high bandwidth, the internal DRAM has to be organized as interleaved memory banks

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)

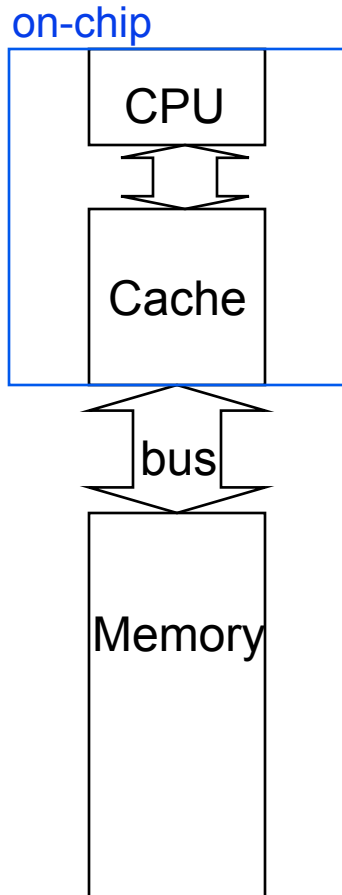
- ❑ Assume

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

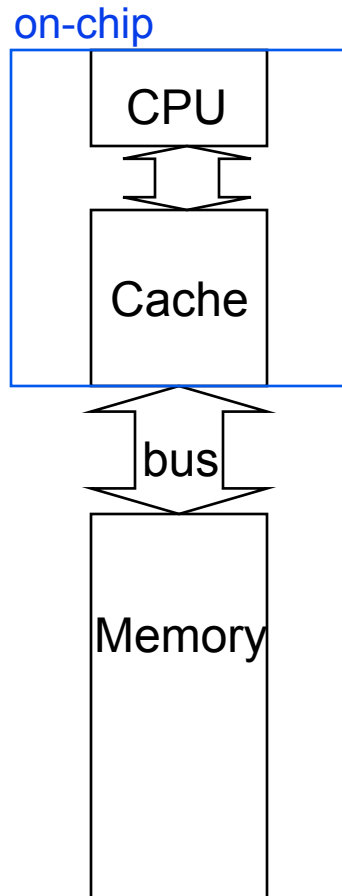
1	cycle to send address
25	cycles to read DRAM
1	cycle to return data
<hr/> 27	total clock cycles miss penalty

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

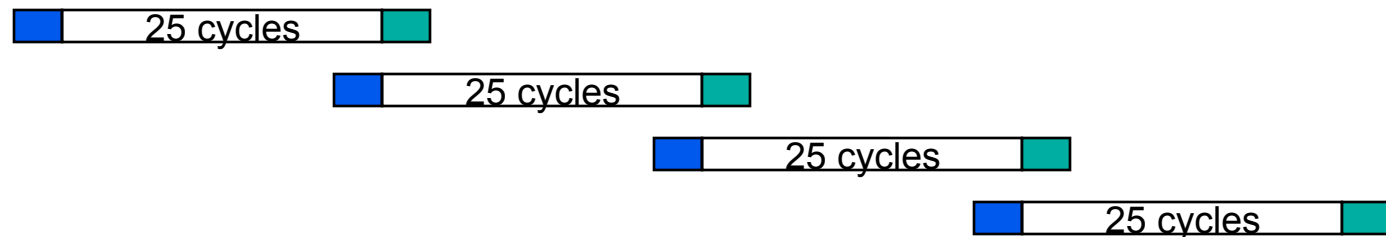
$$4/27 = 0.148 \text{ bytes per clock}$$

One Word Wide Memory Organization, con' t

❑ What if the block size is four words?



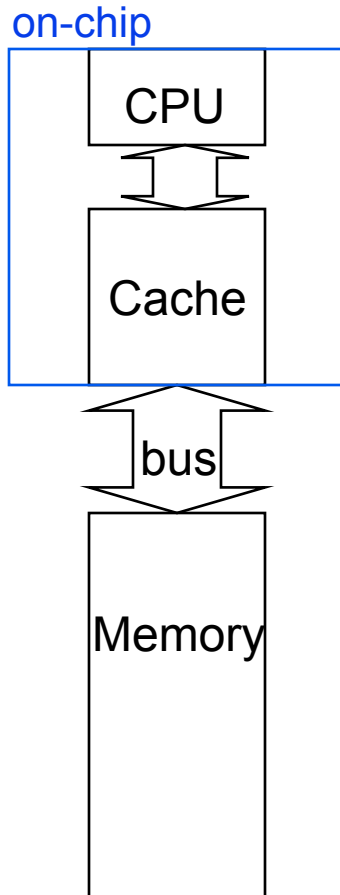
1 cycle to send 1st address
 $4 \times 25 = 100$ cycles to read DRAM
 1 cycles to return last data word
 $\underline{\quad 1 \quad}$
 102 total clock cycles miss penalty



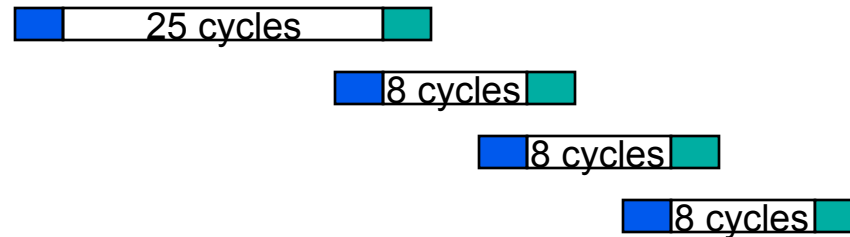
❑ Number of bytes transferred per clock cycle (bandwidth) for a single miss is
 $(4 \times 4)/102 = 0.157$ bytes per clock

One Word Wide Memory Organization, con' t

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



$$\begin{array}{rcl}
 & 1 & \text{cycle to send 1}^{\text{st}} \text{ address} \\
 25 + 3 \cdot 8 = & 49 & \text{cycles to read DRAM} \\
 & 1 & \text{cycles to return last data word} \\
 \hline
 & 51 & \text{total clock cycles miss penalty}
 \end{array}$$



- Number of bytes transferred per clock cycle (bandwidth) for a single miss is $(4 \times 4)/51 = 0.314$ bytes per clock

Interleaved Memory Organization

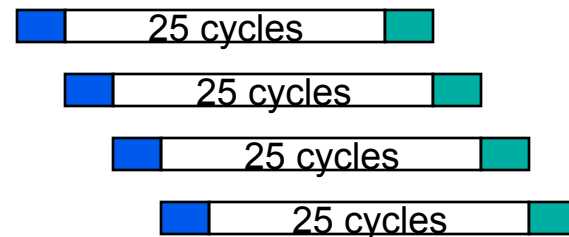
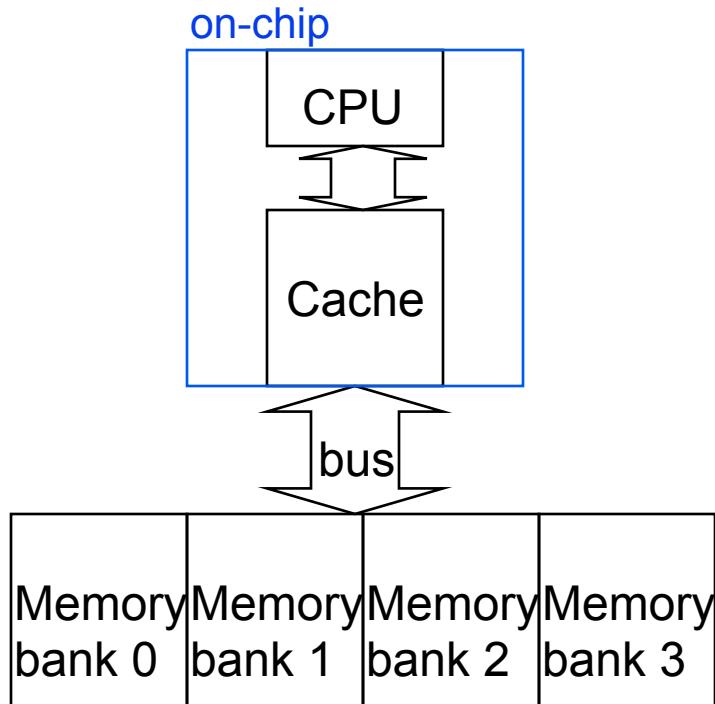
❑ For a block size of four words

1 cycle to send 1st address

25 + 3 = 28 cycles to read DRAM

1 cycles to return last data word

30 total clock cycles miss penalty



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

$$(4 \times 4)/30 = 0.533 \text{ bytes per clock}$$

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