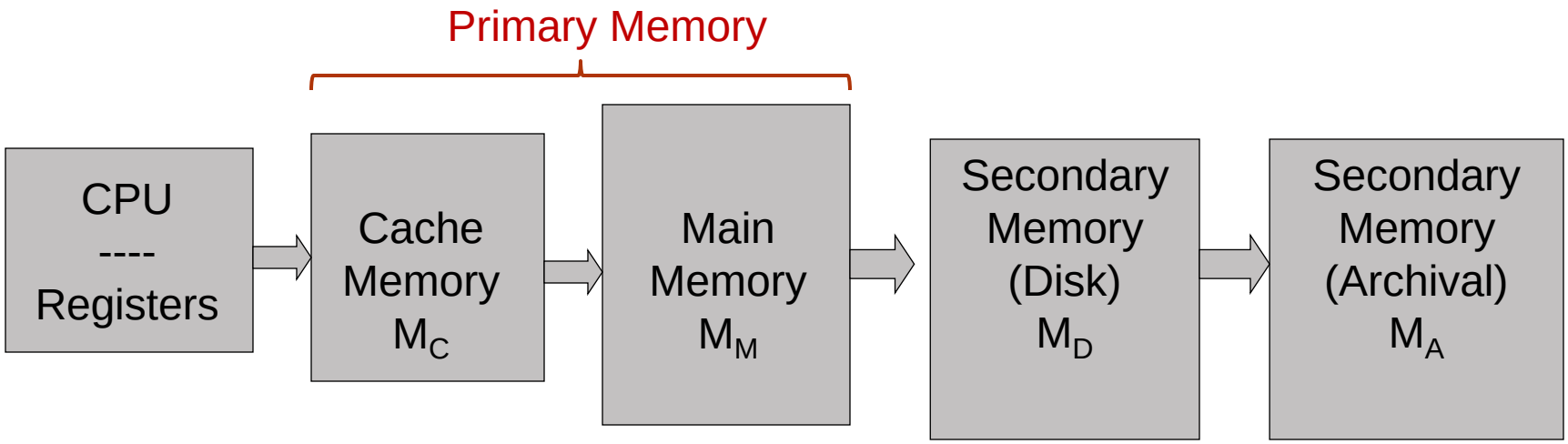
Memory Systems

Patterson & Hennessey

Chapter 5

Memory Hierarchy



Memory Content: $M_C \subseteq M_M \subseteq M_D \subseteq M_A$

Memory Parameters:

- Access Time: increase with distance from CPU
- Cost/Bit: decrease with distance from CPU
- Capacity: increase with distance from CPU

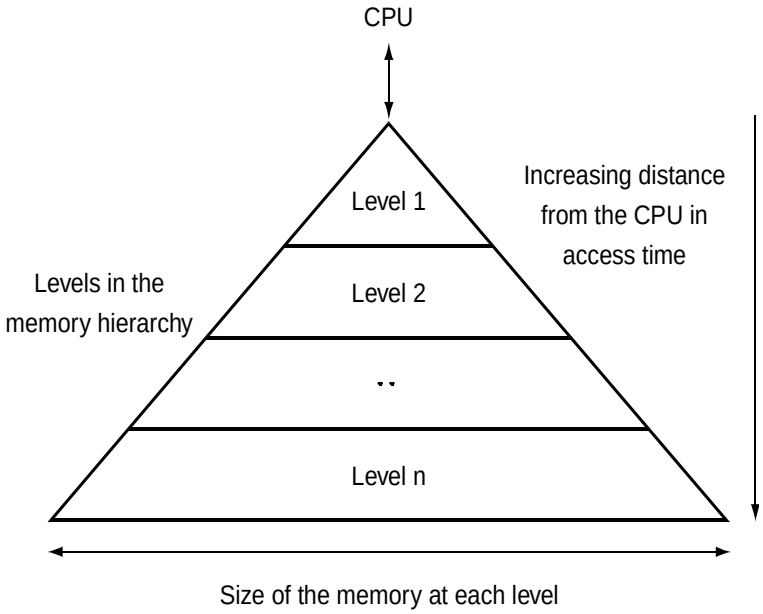
Other Factors:

- Energy consumption
- Reliability
- Size/density

Memory Technology

- Static RAM (SRAM)
  - 0.5ns – 2.5ns, $2000 – $5000 per GB
- Dynamic RAM (DRAM)
  - 50ns – 70ns, $20 – $75 per GB
- Magnetic disk
  - 5ms – 20ms, $0.20 – $2 per GB
- Ideal memory
  - Access time of SRAM
  - Capacity and cost/GB of disk

Memory Hierarchy



Locality of Reference

- Programs access a small proportion of their address space at any time
- If an item is referenced:
  - temporal locality: it will tend to be referenced again soon
  - spatial locality: nearby items will tend to be referenced soon

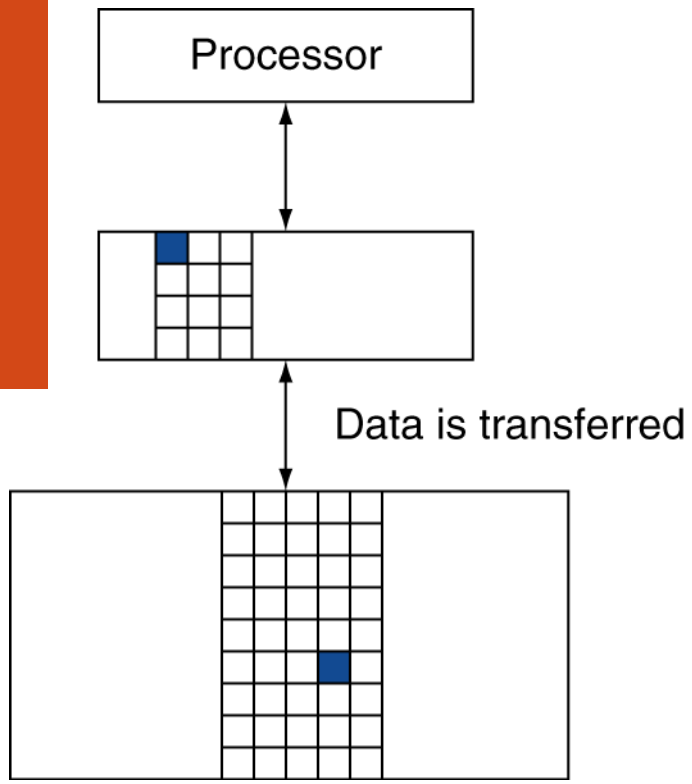
Why does code have locality?

- Our initial focus: two levels (upper, lower)
  - block: minimum unit of data transferred
  - hit: data requested is in the upper level
  - miss: data requested is not in the upper level

Taking Advantage of Locality

- Memory hierarchy
- Store everything on disk
- Copy recently accessed (and nearby) items from disk to smaller DRAM memory
  - Main memory
- Copy more recently accessed (and nearby) items from DRAM to smaller/faster SRAM memory
  - Cache memory attached to CPU

Memory Hierarchy Levels



- Block (aka line): unit of copying
  - May be multiple words
- If accessed data is present in upper level
  - Hit: access satisfied by upper level
    - Hit ratio: hits/accesses
- If accessed data is absent
  - Miss: block copied from lower level
    - Time taken: miss penalty
    - Miss ratio: misses/accesses
      = 1 – hit ratio
  - Then accessed data supplied from upper level

Memory Performance

- Access Time (latency) – from initiation of memory read to valid data returned to CPU
  - Cache: T_A typically = CPU cycle time
  - Main: T_A typically = multiple CPU cycles
  - Disk: T_A typically several orders of magnitude greater than CPU cycle time (software controlled)

- Bandwidth (throughput) = number of bytes transferred per second

  $$BW = (bytes/transfer) \times (transfers/second)$$

  Ex. Synchronous Dynamic RAM "burst transfers"

Memory Access Modes

- Random Access – locations accessed in any order with access time independent of location
  - SRAM
  - ROM/EPROM/EEPROM/Flash technologies
  - DRAM (some special modes available)
  - Cache

- Serial Access – locations must be accessed in a predetermined sequence (ex. tape)
  - Disk – position read element over a track, and then serial access within the track
  - CDROM – data recorded in a "spiral"

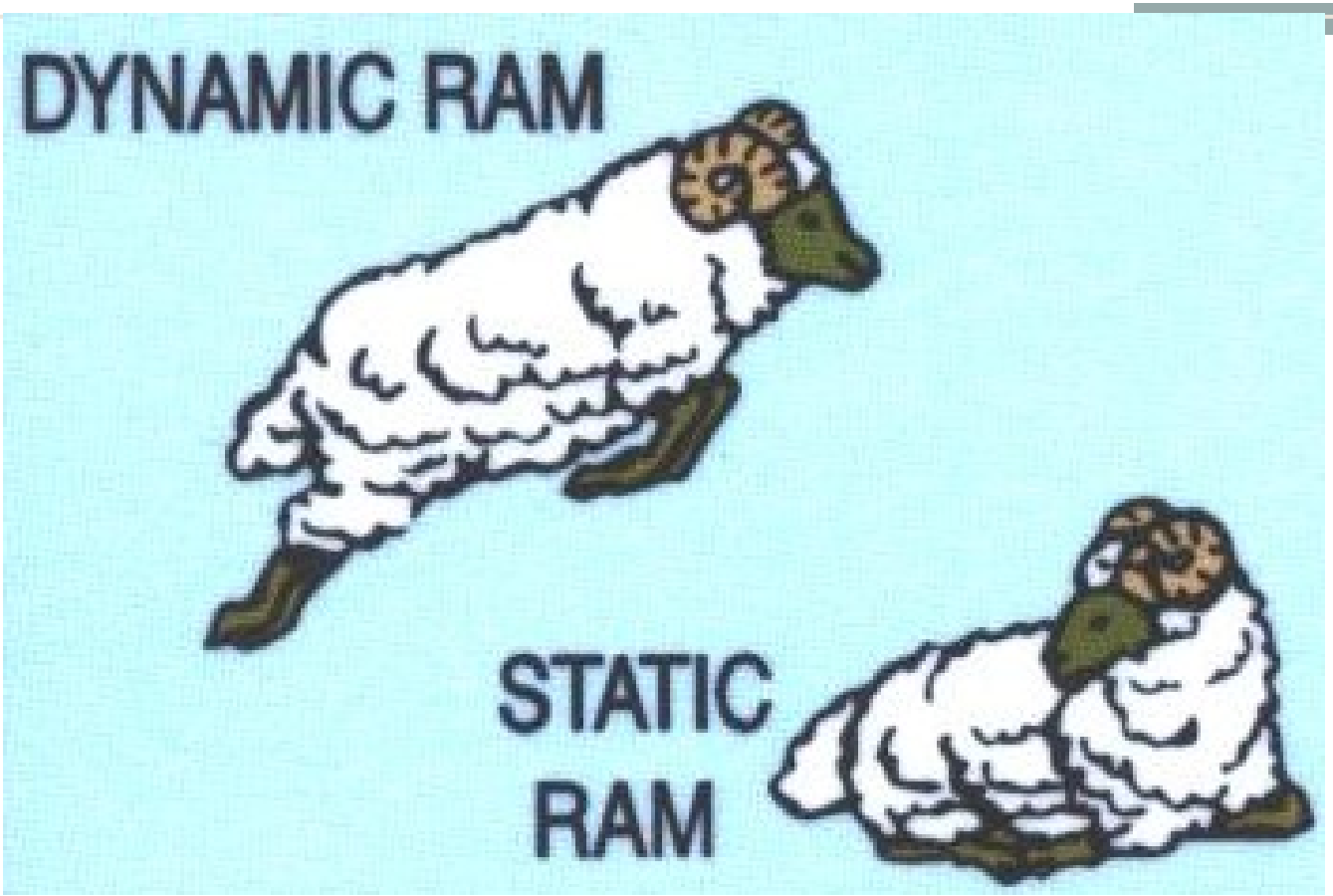
Memory Write Strategies

- Writing a new value of an item in a hierarchical memory which has copies in multiple memory levels

- "Write-through" strategy
  - Update all levels of memory on each write
  - All levels "consistent" or "coherent" at all times

- "Write-back" strategy
  - Update closest (fastest) copy during write operation
  - Copy the value to other levels at "convenient time"
  - Memory state temporarily inconsistent

Types of Computer Memories

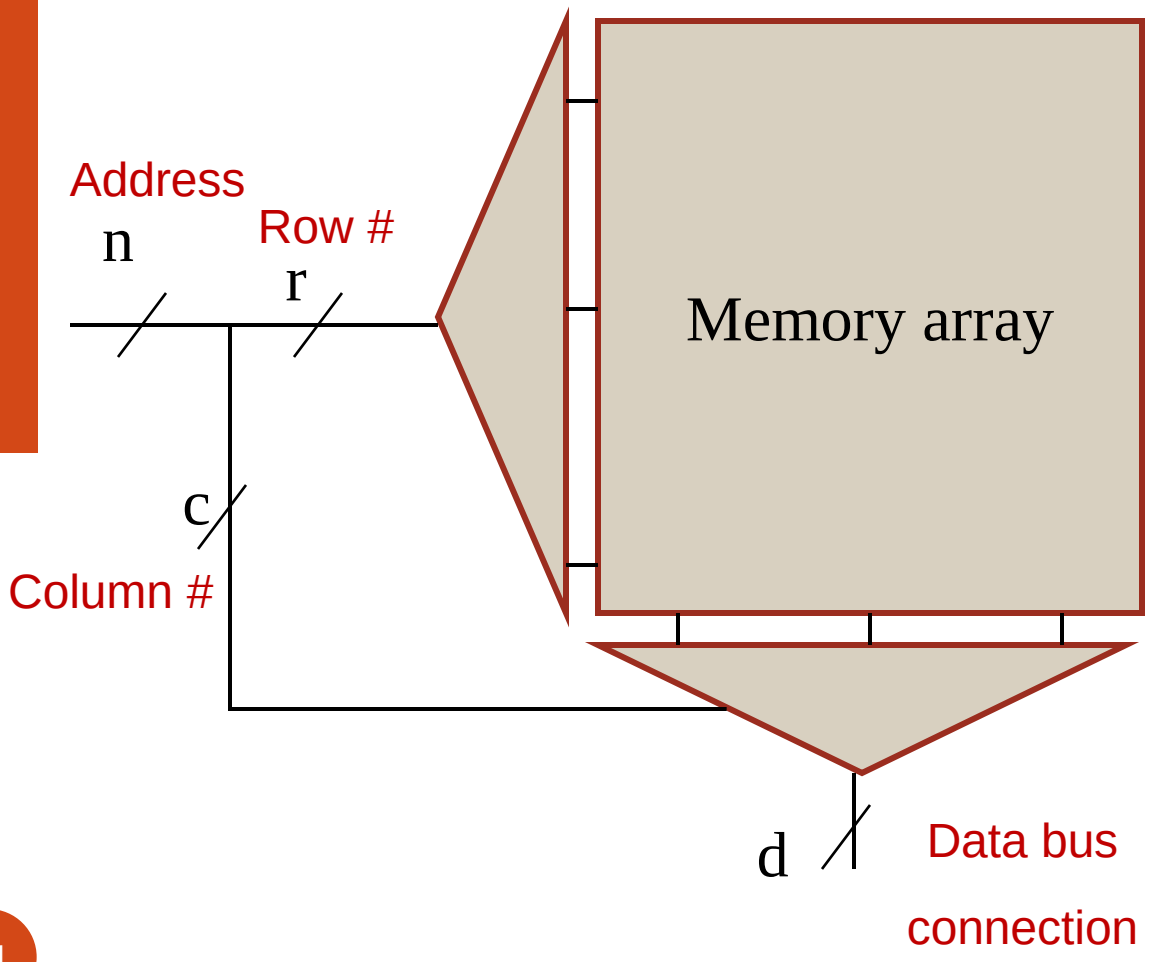


From the cover of:

A. S. Tanenbaum, *Structured Computer Organization, Fifth Edition*, Upper Saddle River, New Jersey: Pearson Prentice Hall, 2006.

Memory device organization

Memory "organization" = 2^n x d
(from system designer's perspective)

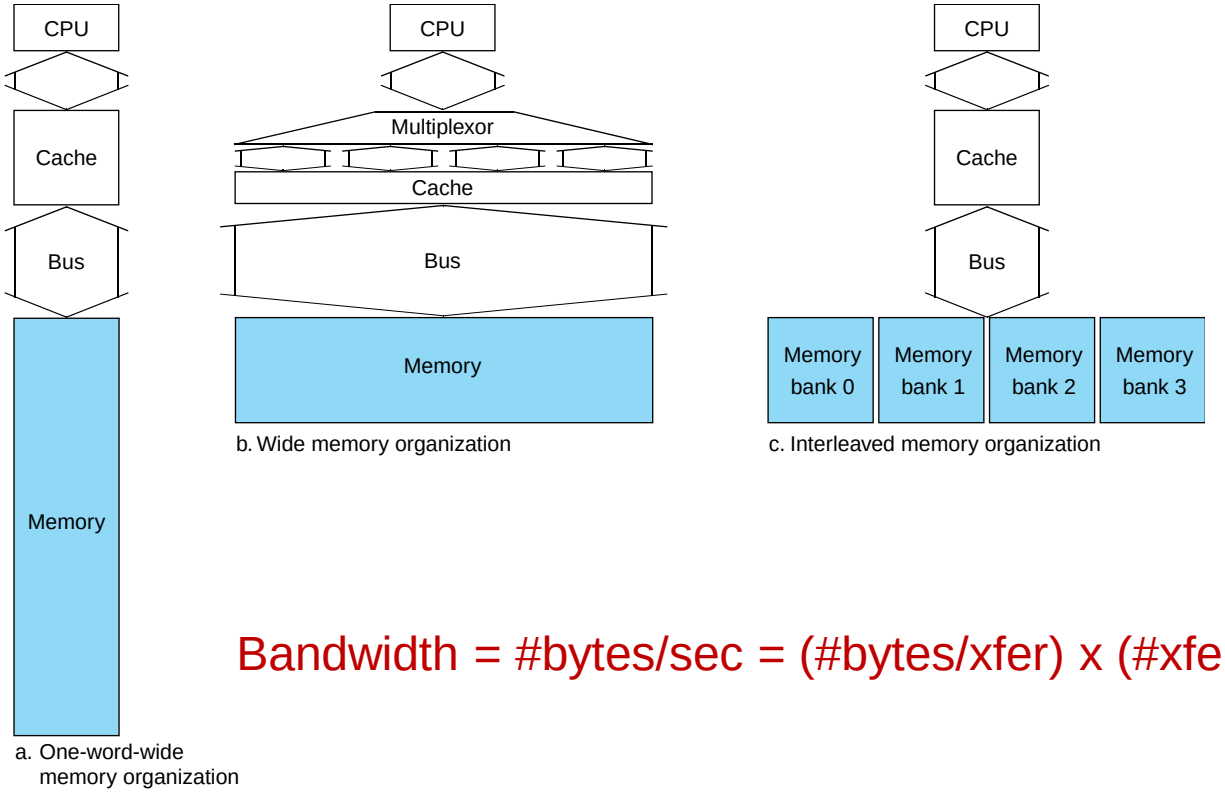


- Size.
  - Address width.

    $n = r + c$
- Aspect ratio.
  - Data width d.

Hardware Issues

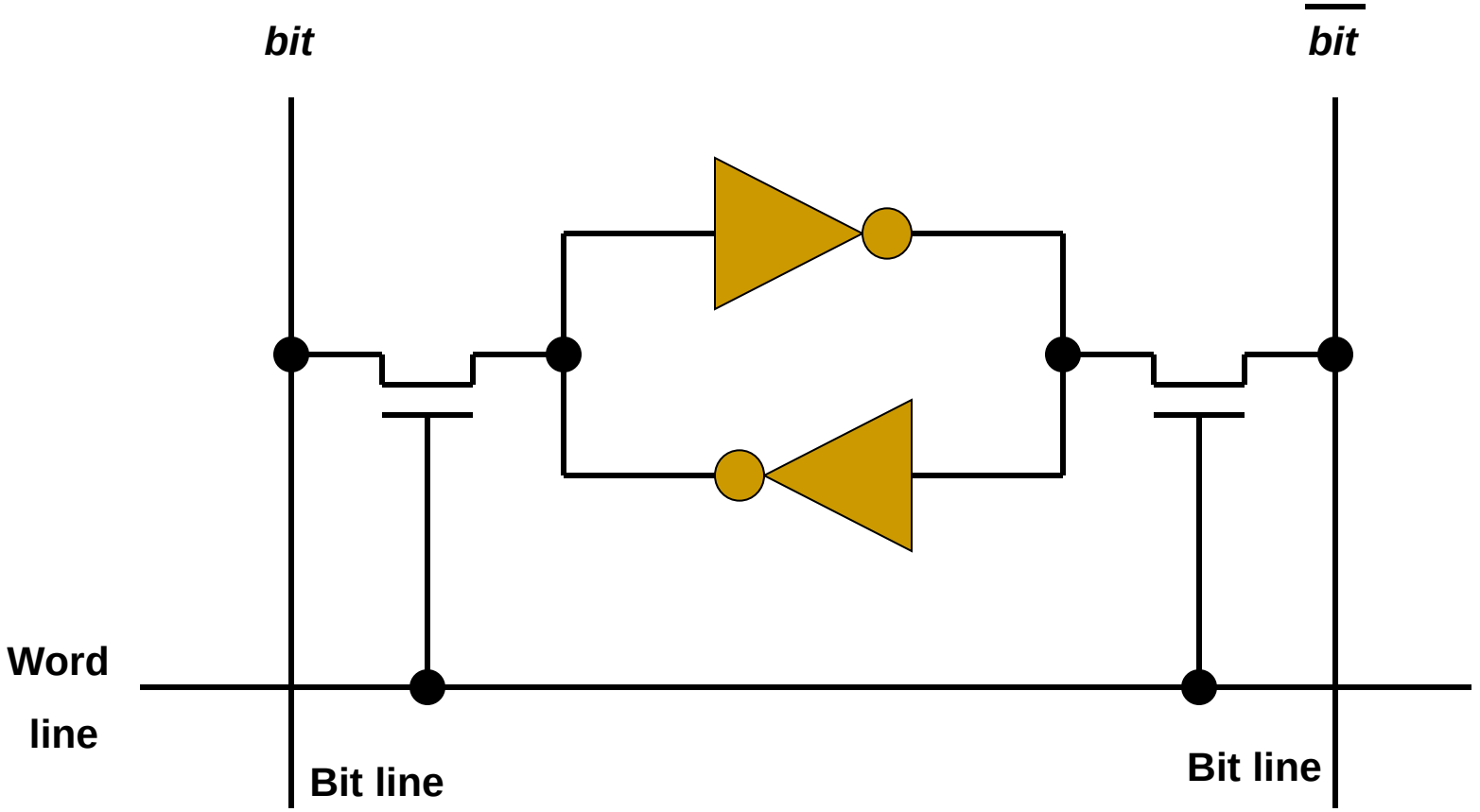
- Make reading multiple words easier by using banks of memory



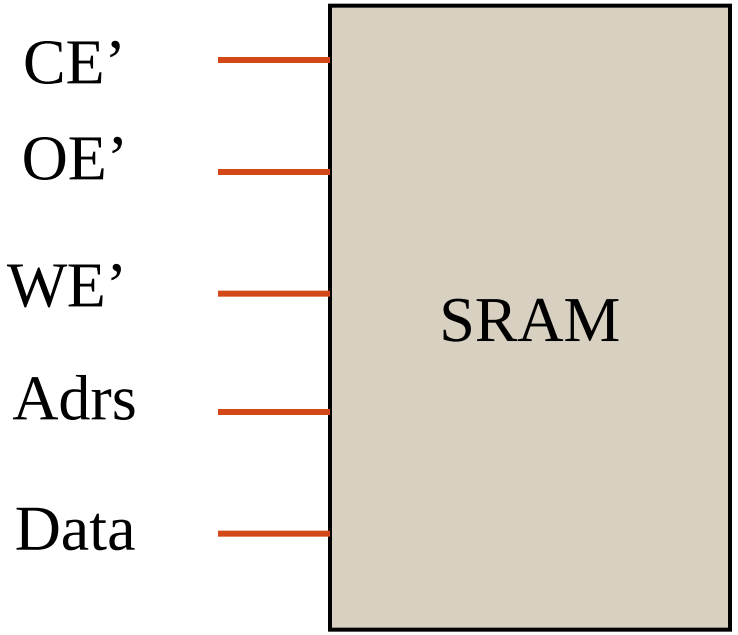
Bandwidth = #bytes/sec = (#bytes/xfer) x (#xfers/sec)

- It can get a lot more complicated…

Six-Transistor SRAM Cell

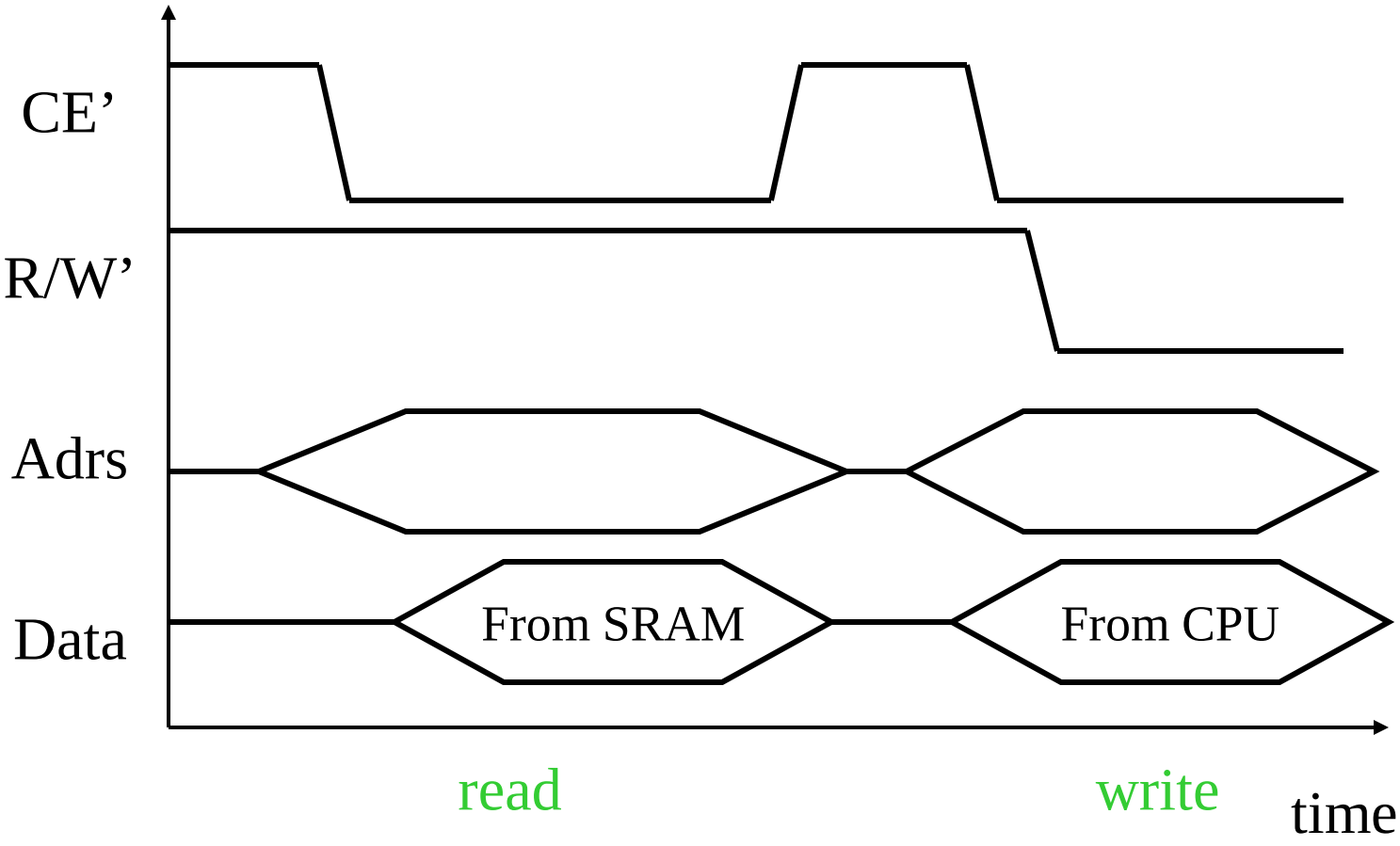


Typical generic SRAM

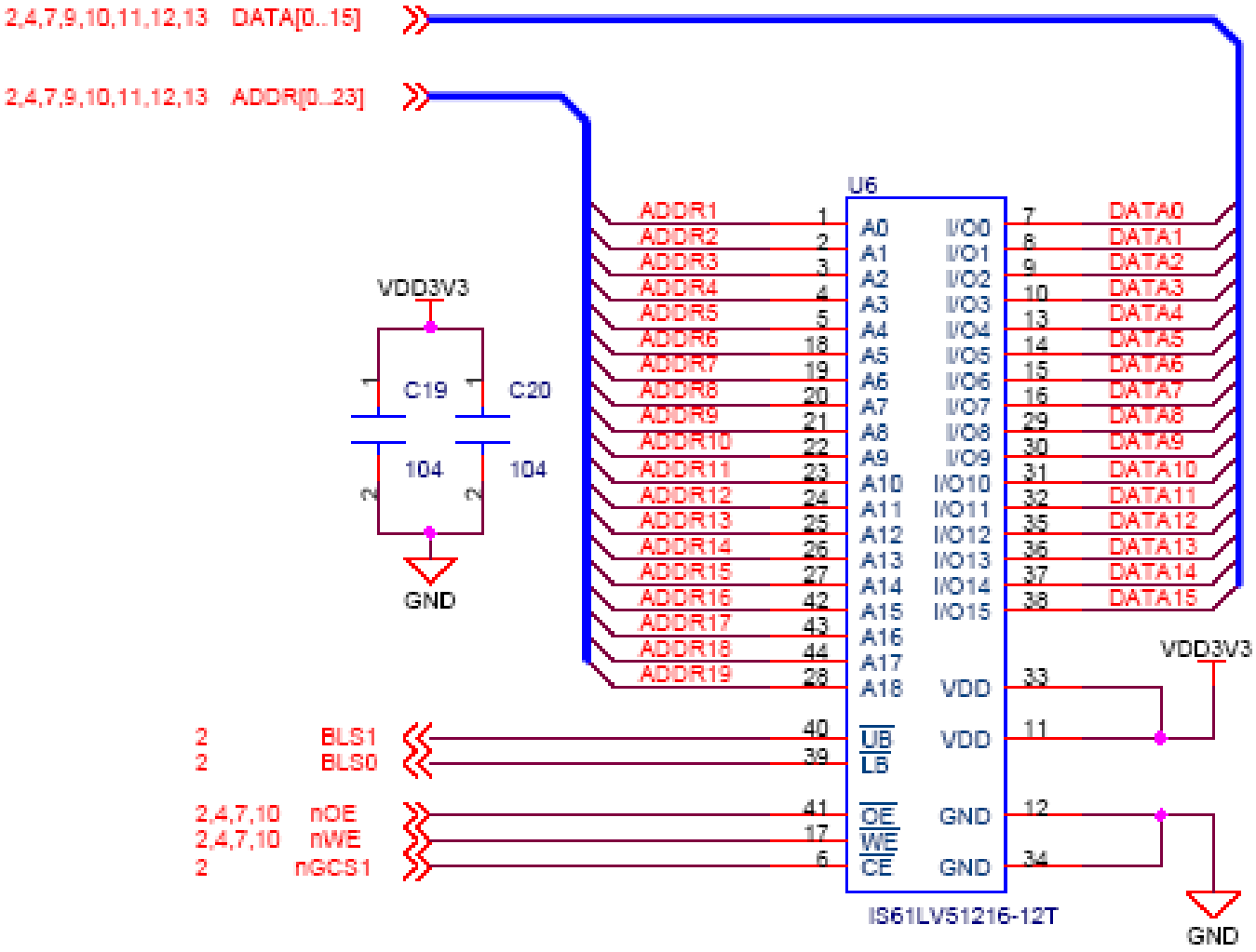


- Often have a single R/W’ signal instead of OE’ and WE’.
- Multi-byte Data bus devices usually have byte-select signals.

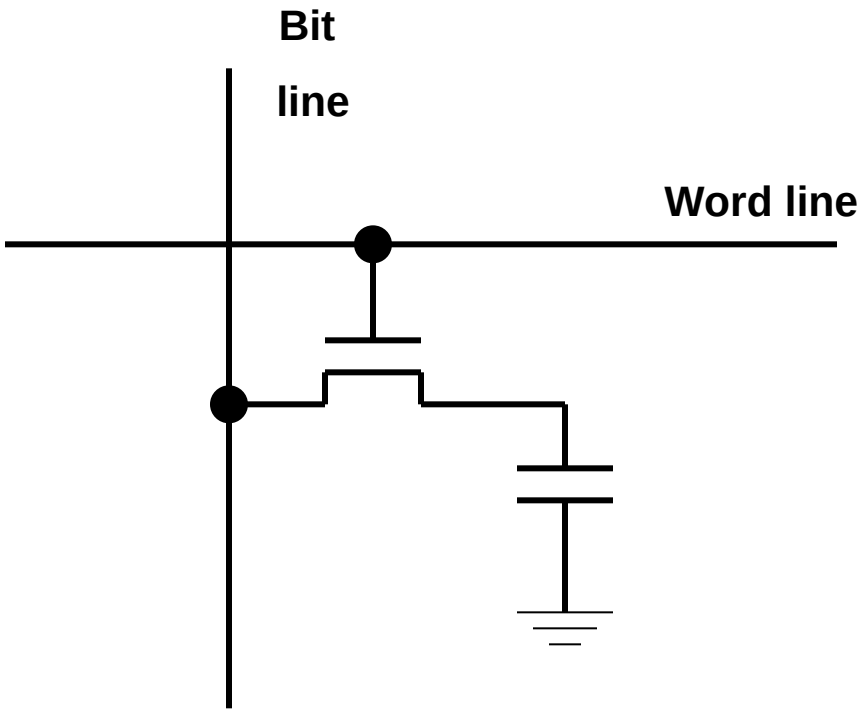
Generic SRAM timing



512K x 16 SRAM (on uCdragon board)



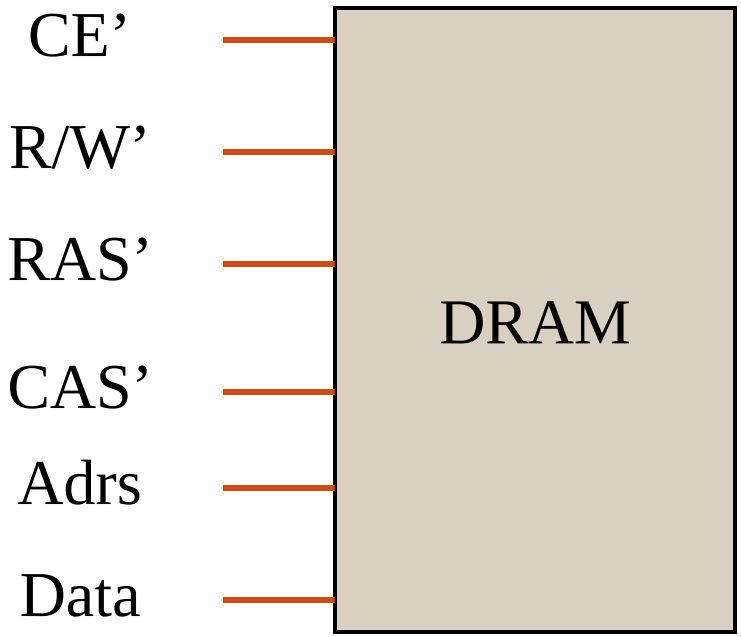
Dynamic RAM (DRAM) Cell



"Single-transistor DRAM cell"

Robert Dennard's 1967 invevention

Generic DRAM device



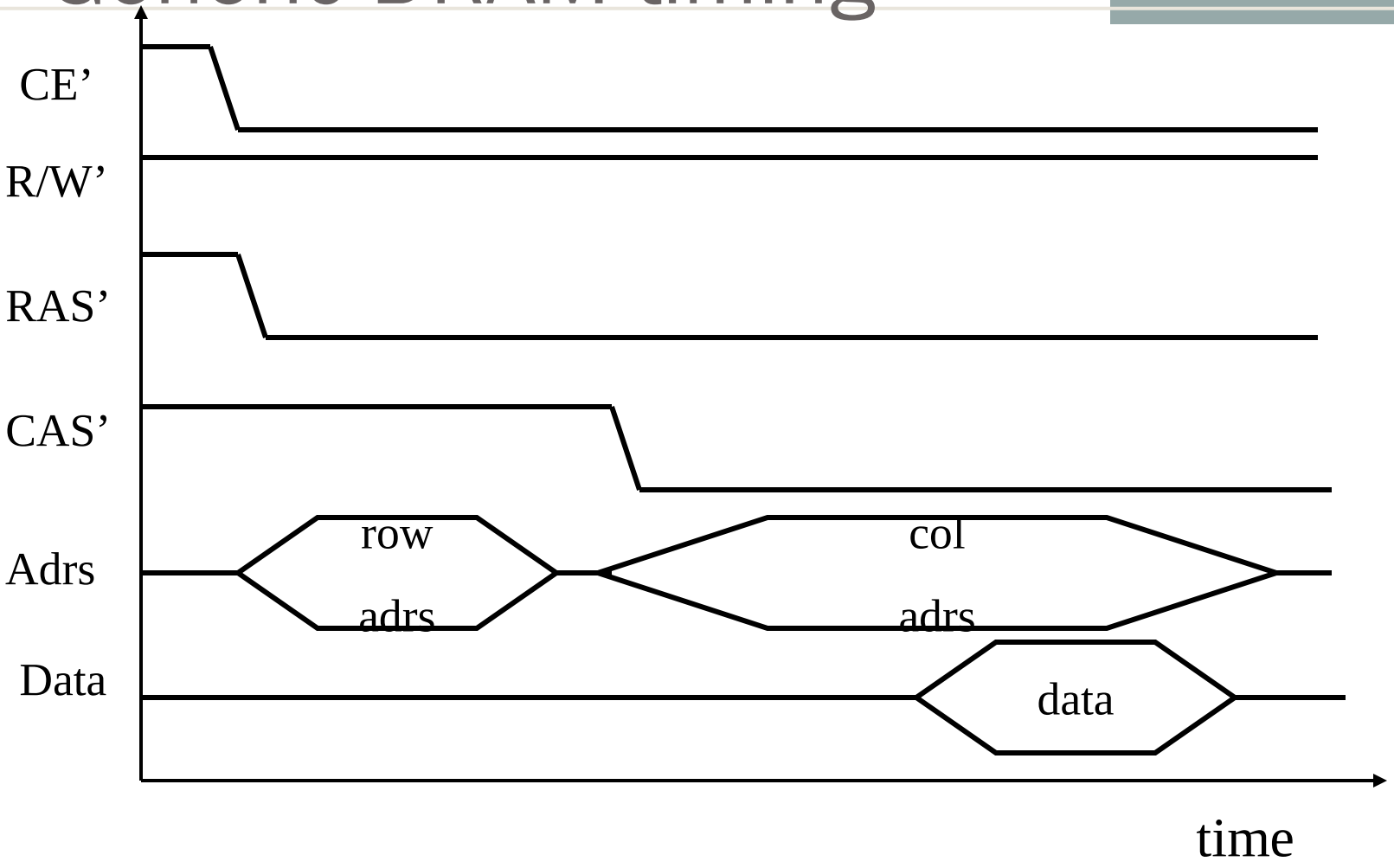
RAS: Row Address Strobe

CAS: Column Address Strobe

Adrs: Multiplexed Row/Column address

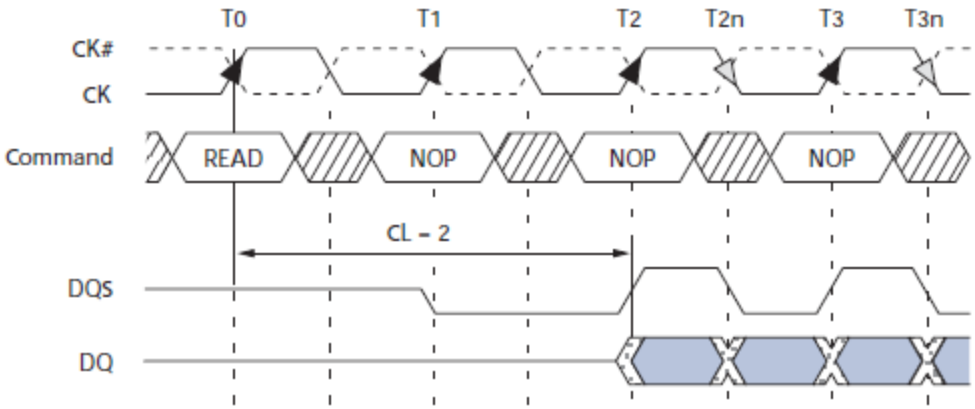
CE: Chip-enable (if present)

Generic DRAM timing



Micron 256Mbit DDR SDRAM

- **MT46V64M4 – 16 Meg x 4 x 4 banks**
- **MT46V32M8 – 8 Meg x 8 x 4 banks**
- **MT46V16M16 – 4 Meg x 16 x 4 banks**



Tcycle = 7.5ns

- Internal, pipelined double-data-rate (DDR)
- 4 internal banks for concurrent operation
- Programmable burst lengths: 2, 4, 8
- Auto refresh 64ms, 8192 cycles