

# From Physical Memory To Virtual Memory: Understanding the Memory Hierarchy

Alistair Rendell

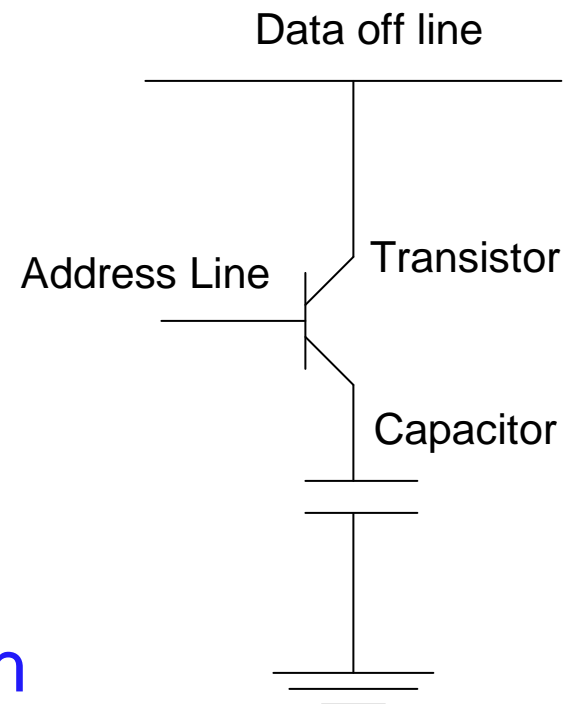
Computer Systems: A Programmers Perspective  
Randal E Bryant and David O'Hallaron  
Structured Computer Organization  
Andrew Tanenbaum

# Memory Chips: SRAM

- Remember the SR latch? (Tanenbaum 3.3)
  - Clocked SR Latch
  - Clocked D Latch
- Clocked D latch requires 11 transistors
  - More sophisticated (but less obvious) requires just 6
- Basis for Static RAM or SRAM
- State persists providing power is on

# Memory Chips: DRAM

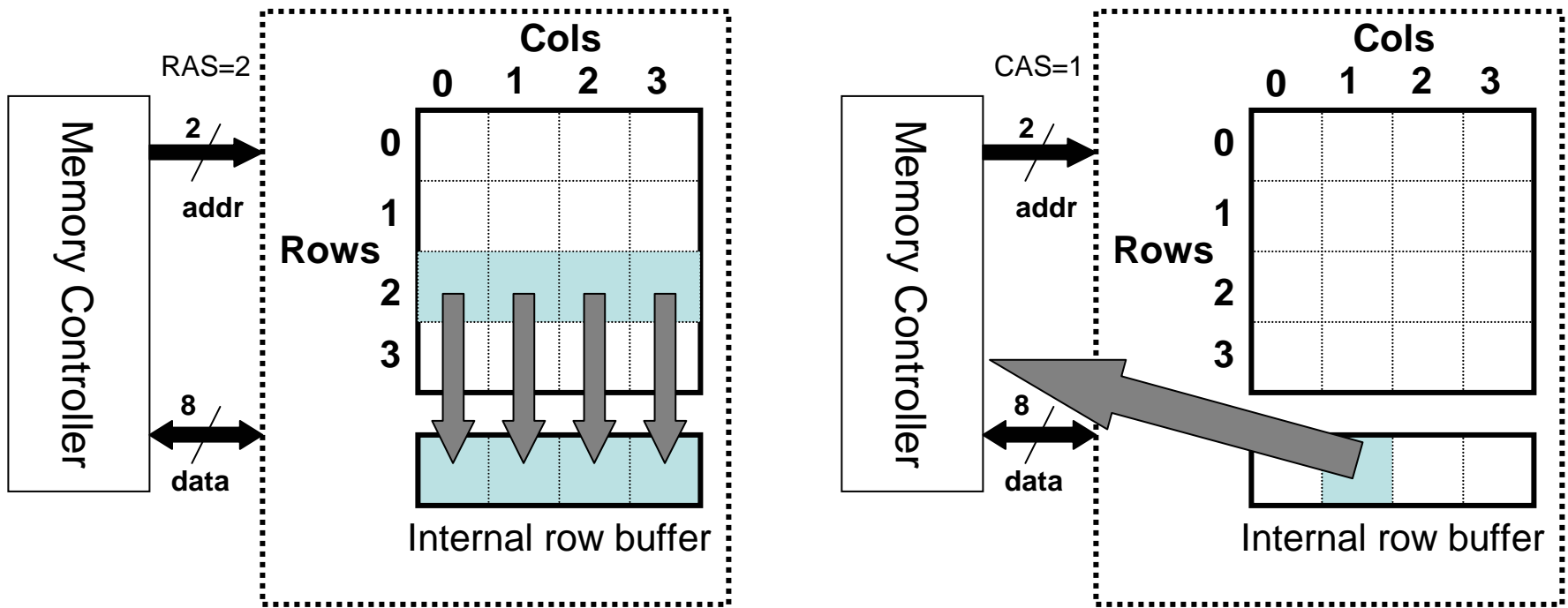
- Alternative to SRAM
  - 1 capacitor and 1 transistor
- Capacitor stores ( $10^{-15}\text{F}$ )
  - Just 40,000 electrons!
- Resistance  $10^{12}$  Ohms
  - Time constant of  $10^{-3}\text{sec}$
  - Hence DRAM requires refresh
- Typical DRAM refresh 10-100ms



# Comparison

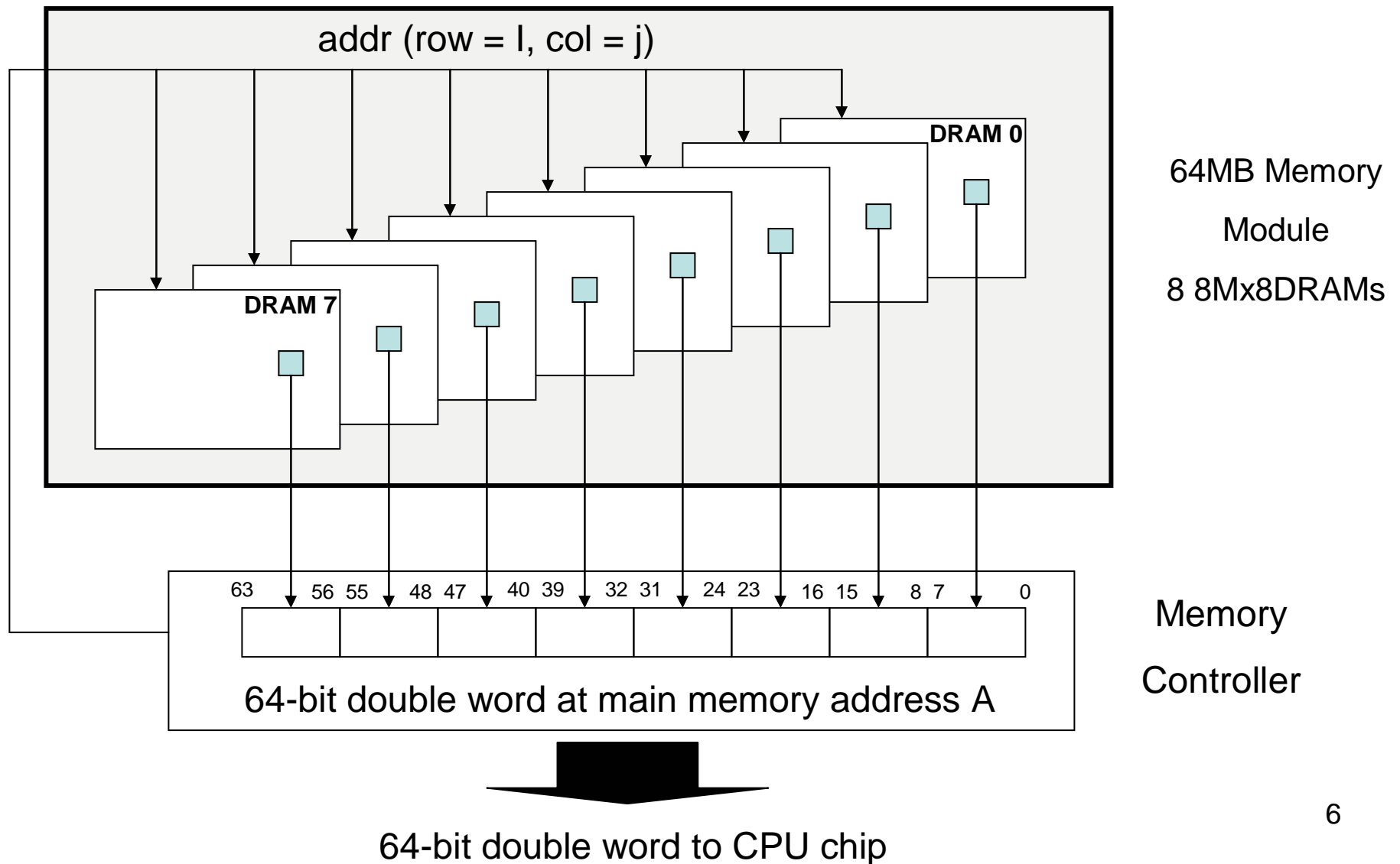
	<b>Transistors per bit</b>	<b>Relative Access Time</b>	<b>Persistent</b>	<b>Sensitive</b>	<b>Relative Cost</b>	<b>Use</b>
SRAM	6	1X	Yes	No	100x	?
DRAM	1	10X	No	Yes	1x	?

# Accessing Memory



- Select row (row access strobe=2)
- Select column (CAS=1)
- Why address bus 2 wide?
- Why not a single unique index per cell (0

# Dual Inline Memory Module (DIMM 168 pins)



# Memory Types

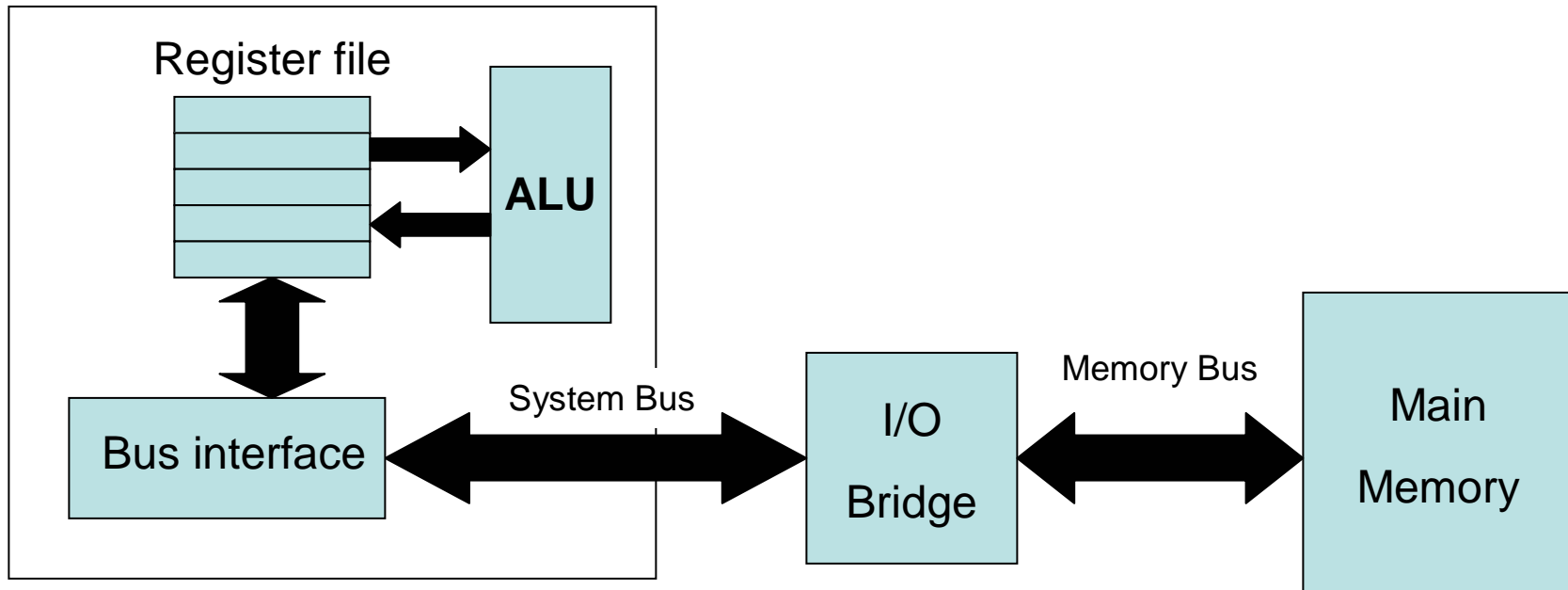
## Volatile Memory

- Fast page mode DRAM (FPM DRAM)
- Extended data out DRAM (EDO DRAM)
- Synchronous DRAM (SDRAM)
- Double Data-Rate Synchronous DRAM (DDR SDRAM)
- Rambus DRAM (RDRAM)
- Video RAM (VRAM)

## Nonvolatile Memory

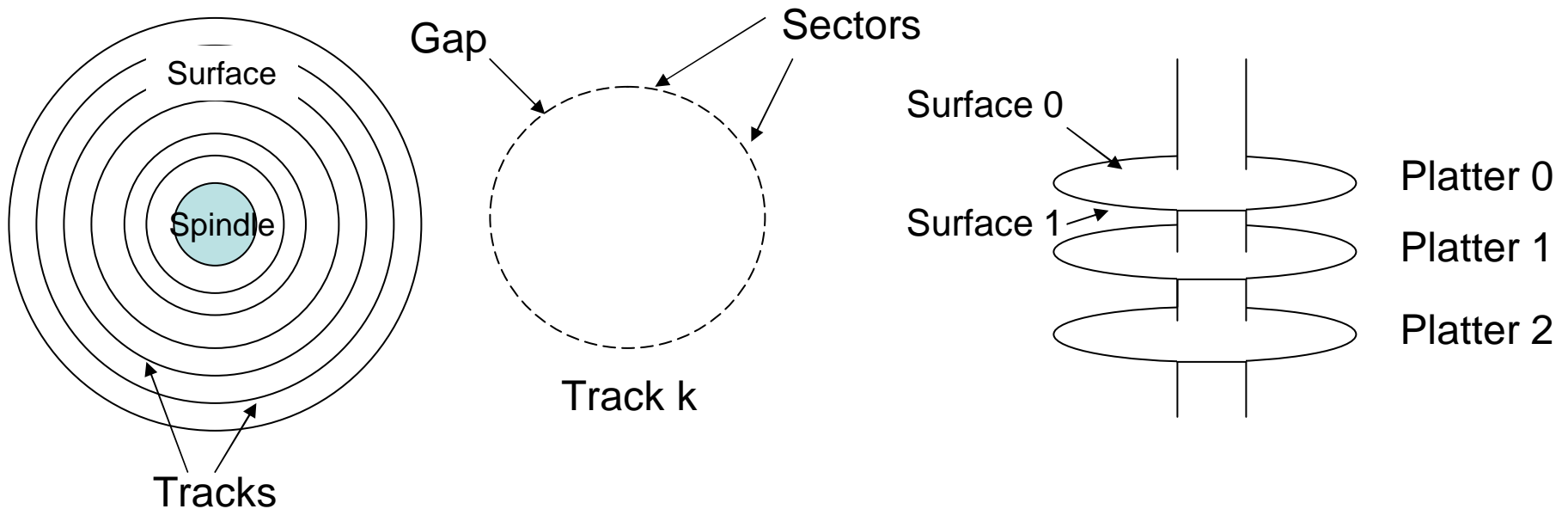
- Read Only Memory (ROM)
  - PROM, EPROM etc
- Programs stored in ROM are termed “firmware”

# Accessing Memory



- Processor/Memory communicate over shared bus (transactions)
- Read transaction
  - CPU places address on system bus, I/O bridge forwards to memory bus
  - Main memory reads address and places content on memory bus
  - CPU reads word from system bus and copies to register
- Write transaction
  - As above, memory waits to receive data

# Anatomy of a Disk

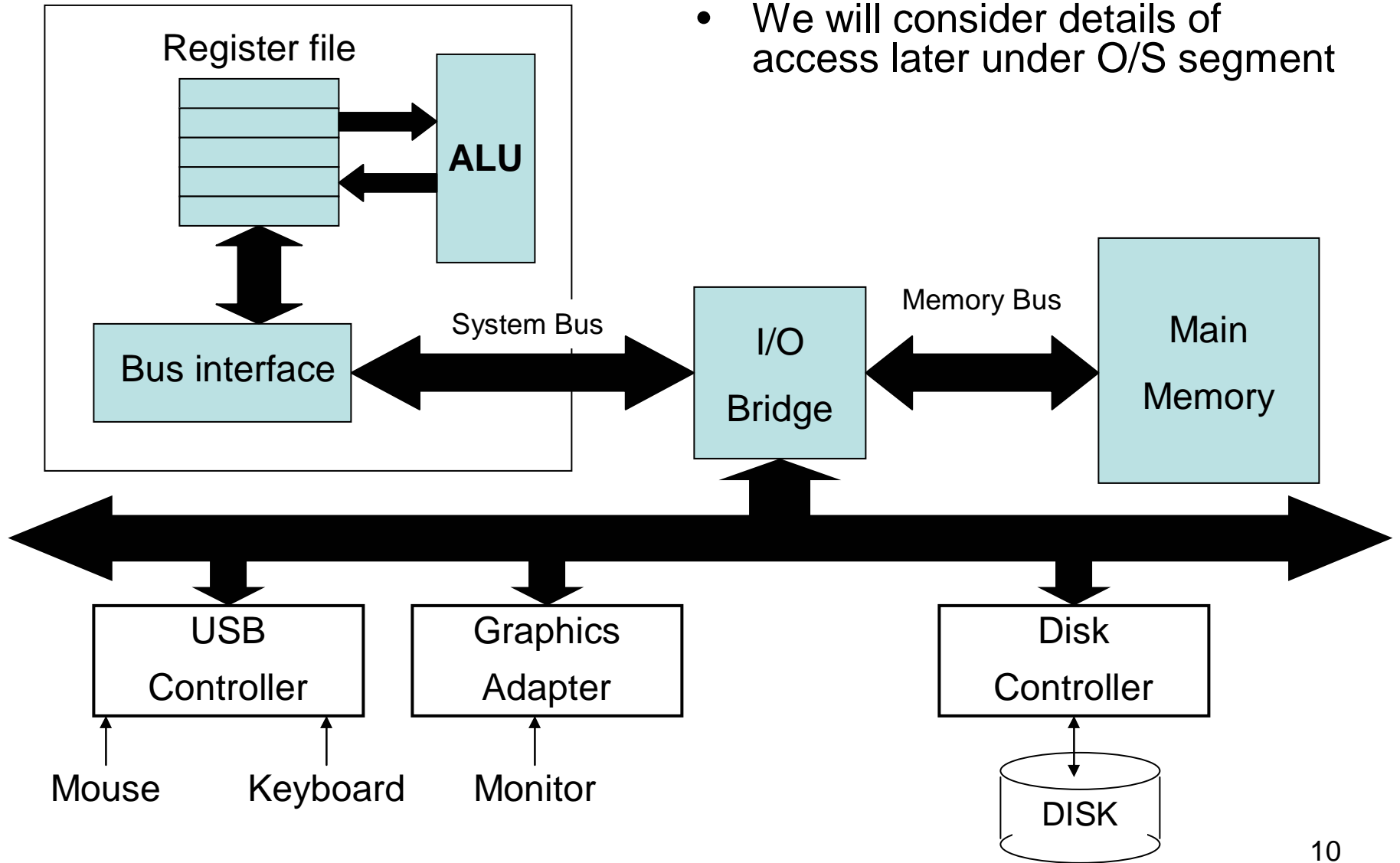


- Coated with magnetic material
- Typical rotation rates?

- Sectors typical 512bytes
- Cylinders:
  - all tracks equidistant from spindle

# Disk Configuration

- We will consider details of access later under O/S segment



# Disk Capacity

- Recording density (bits/in)
- Track density (tracks/in)
- Areal density (bits/in<sup>2</sup>)
- Early disks equal sectors per track
  - Multiple zone recording relaxed this
  - Floppy disks still use old format

$$Capacity = \frac{\#bytes}{sector} \times \frac{avg\#sector}{track} \times \frac{\#tracks}{surface} \times \frac{\#surfaces}{platter} \times \frac{\#platters}{disk}$$

# Disk Operation

- Read/write head connected to actuator arm
  - Arm  $10^{-7}$ m above surface and moves at 80km/h!
  - Hence disks come in airtight containers
- Total Access Time ( $T_{\text{access}}$ )
  - Seek time ( $T_{\text{seek}}$ )
  - Rotation latency ( $T_{\text{avg rot}}$ )
  - Transfer time ( $T_{\text{avg trans}}$ )

Parameter	Value
Rotation Rate	7200RPM
$T_{\text{avg seek}}$	9ms
Avg # sect/track	400

- $T_{\text{avg rot}} \approx 4\text{ms}$
- $T_{\text{avg trans}} \approx 0.02\text{ms}$

$$\Rightarrow T_{\text{access}} \approx 13.02\text{ms}$$

# Consider

- What dominates access time?
- How does access time compare with SRAM and DRAM?

# Trends

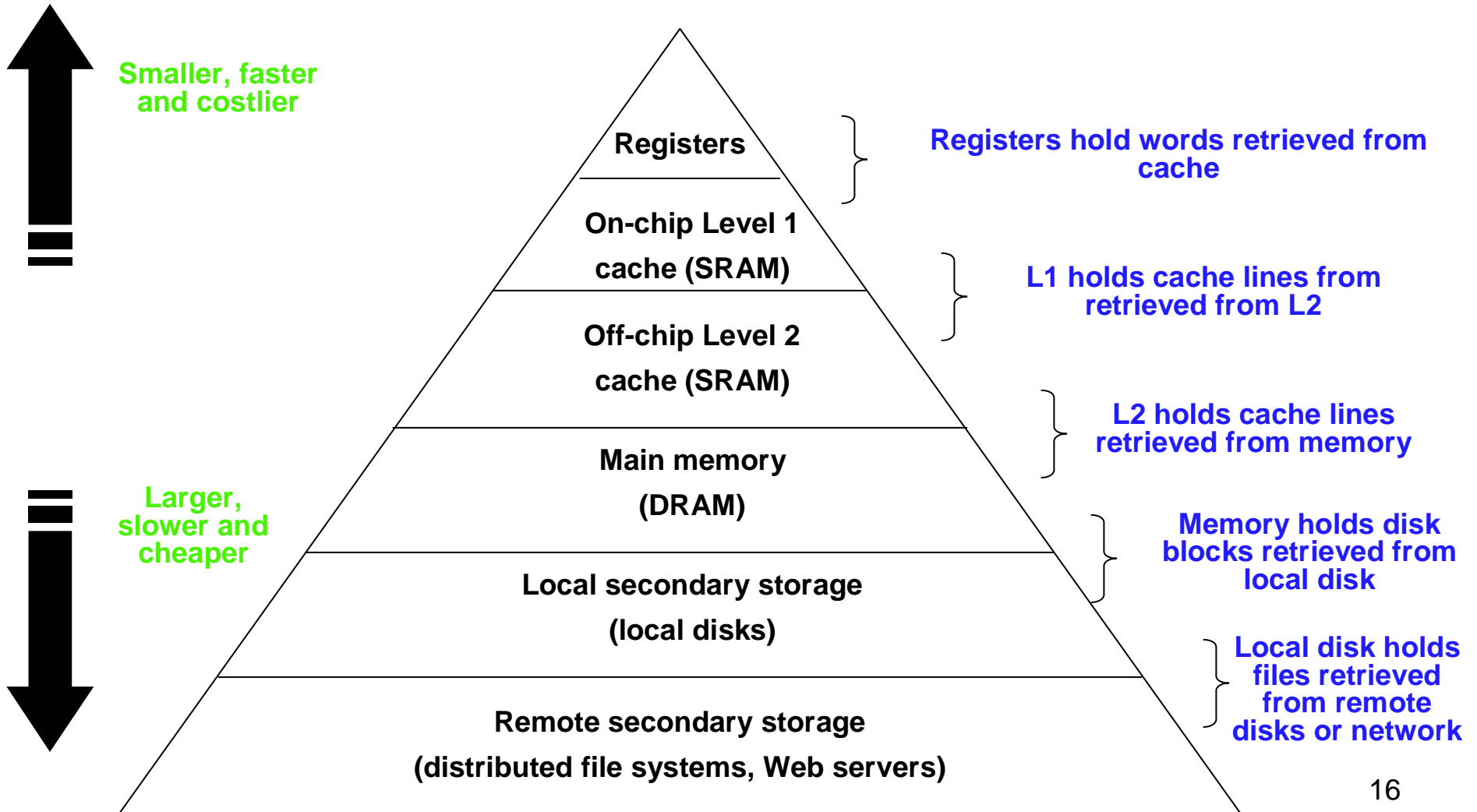
	1980	1985	1990	1995	2000	2000:1980
<b>SRAM</b>						
\$/MB	19200	2900	320	256	100	190
Access(ns)	300	150	35	15	3	100
<b>DRAM</b>						
\$/MB	8000	880	100	30	1	8000
Access(ns)	375	200	100	70	60	6
Typ Size (MB)	0.064	0.256	4	16	64	1000
<b>DISK</b>						
\$/MB	500	100	8	0.3	0.01	50000
Seek time (ms)	87	75	28	10	8	11
Typ Size (MB)	1	10	160	1000	20000	20000
<b>CPU</b>						
Intel	8080	80286	80386	Pentium	P-III	
Clock(MHz)	1	6	20	150	600	600
Cycle (ns)	1000	166	50	6	1.6	600

# The Take Home Message

- Different storage technologies have different prices and performance
- Price performance of different technologies changing at different rates
- DRAM and disk access is lagging CPU cycle times

How do we bridge the processor memory performance gap?

# The Memory Hierarchy



# Two Key Factors

- Spatial locality
- Temporal locality

# Caching

Type	What	Where	Latency (Cycles)	Managed
Registers	4-byte word	On-chip	0	Compiler
TLB	Address Translation	On-chip	0	Hardware/MMU
L1	32-byte block	On-chip	1	Hardware
L2	32-byte block	Off-chip	10	Hardware
Virtual Memory	4KB page	Main Memory	100	Hardware/OS
Buffer cache	Parts file	Main Memory	100	OS
Network buffer cache	Parts file	Local Disk	10,000,000	NFS client
Browser	Web pages	Local Disk	10,000,000	Web browser
Web	Web pages	Remote Disk	1,000,000,000	Web proxy server