COMPUTER ORGANIZATION (IS F242)

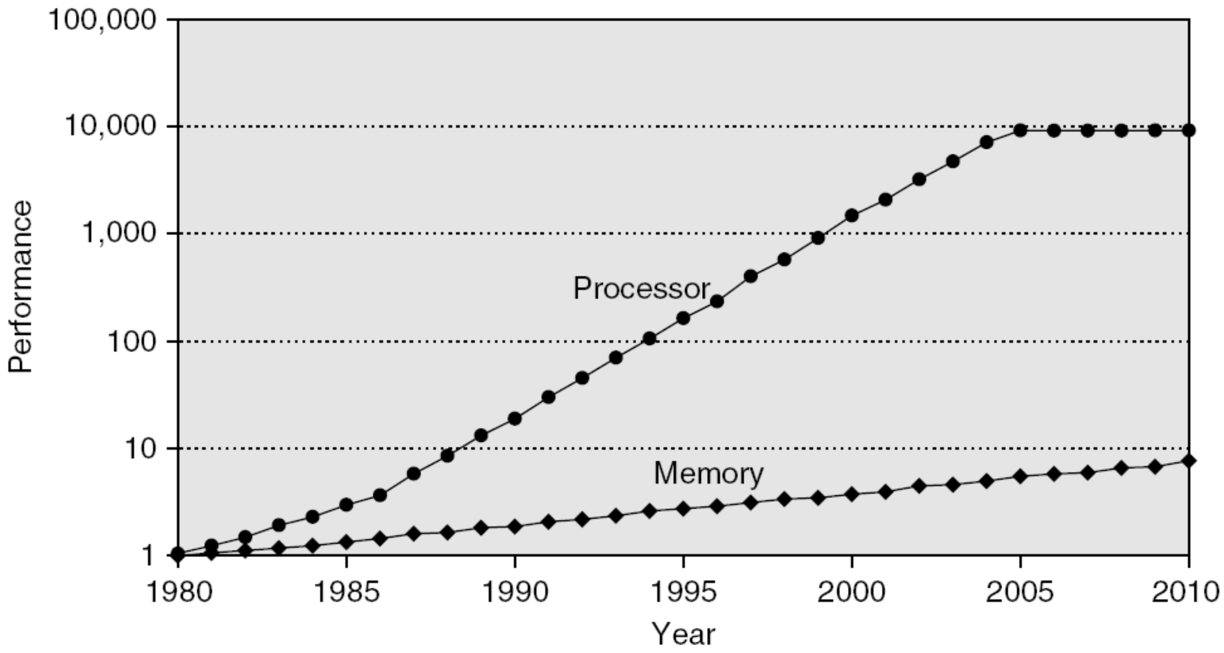
LECT 35: CACHE MEMORY

Memory

- Characteristics
  - Properties
  - Location
  - Capacity
  - Unit of transfer
  - Access method
  - Performance
  - Organisation
  - Semiconductor Memories

Need for organisation

- CPU performance improves by 60% per year
- Memory performance improves by 10% per year
- Gap between CPU performance and memory performance in terms of access time increases

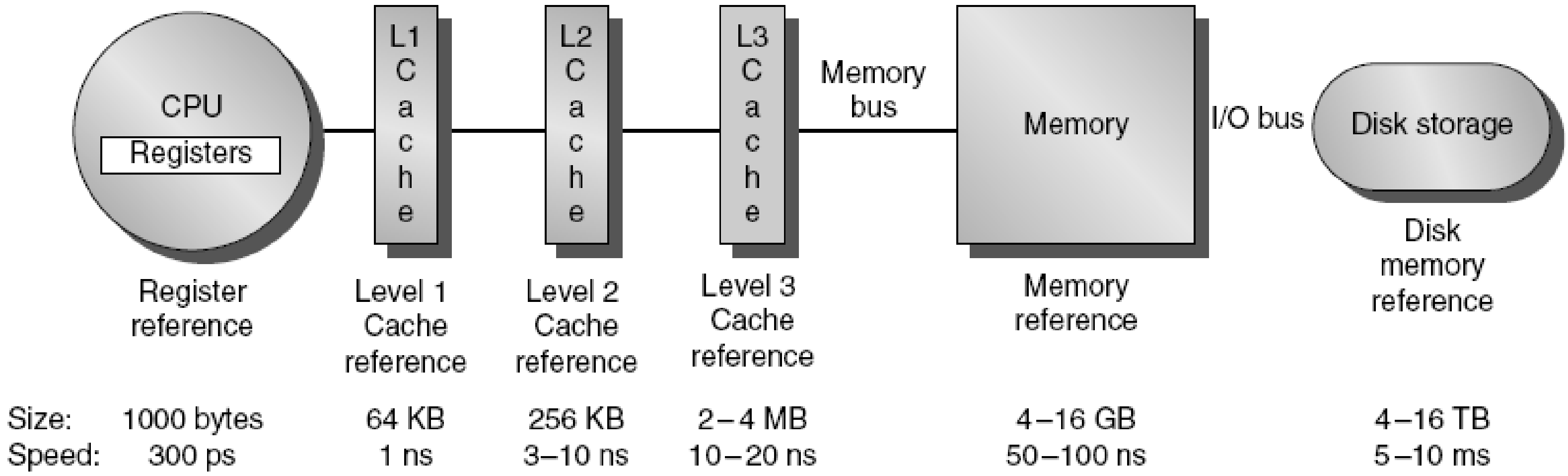


Memory Hierarchy Design

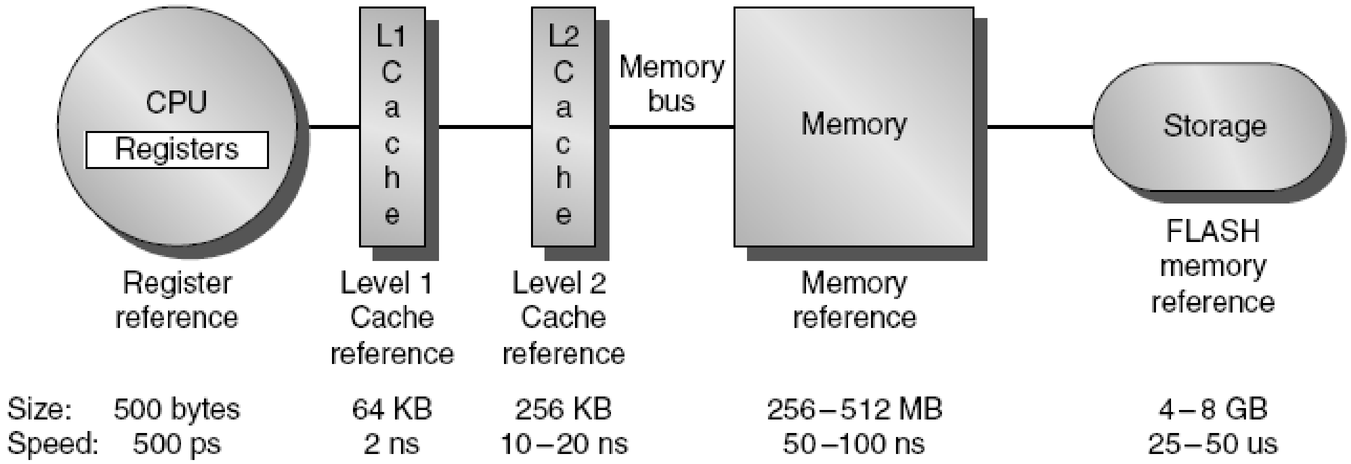
- Memory hierarchy design becomes more crucial with recent multi-core processors:
  - Aggregate peak bandwidth grows with # cores:
    - Intel Core i7 can generate two references per core per clock
    - Four cores and 3.2 GHz clock
      - 25.6 billion 64-bit data references/second +
      - 12.8 billion 128-bit instruction references
      - = 409.6 GB/s!
    - DRAM bandwidth is only 6% of this (25 GB/s)
    - Requires:
      - Multi-port, pipelined caches
      - Two levels of cache per core
      - Shared third-level cache on chip

Need for organisation

- Solution
  - Improve locality of reference somehow
  - Use faster memory (extra cost) closer to CPU to match the CPU speed.
  - Organize the memory in hierarchical fashion to improve the performance
  - Address other bottlenecks such as bus width etc ..



(a) Memory hierarchy for server



(b) Memory hierarchy for a personal mobile device

Levels of the Memory Hierarchy

Capacity
Access Time
Cost/bit

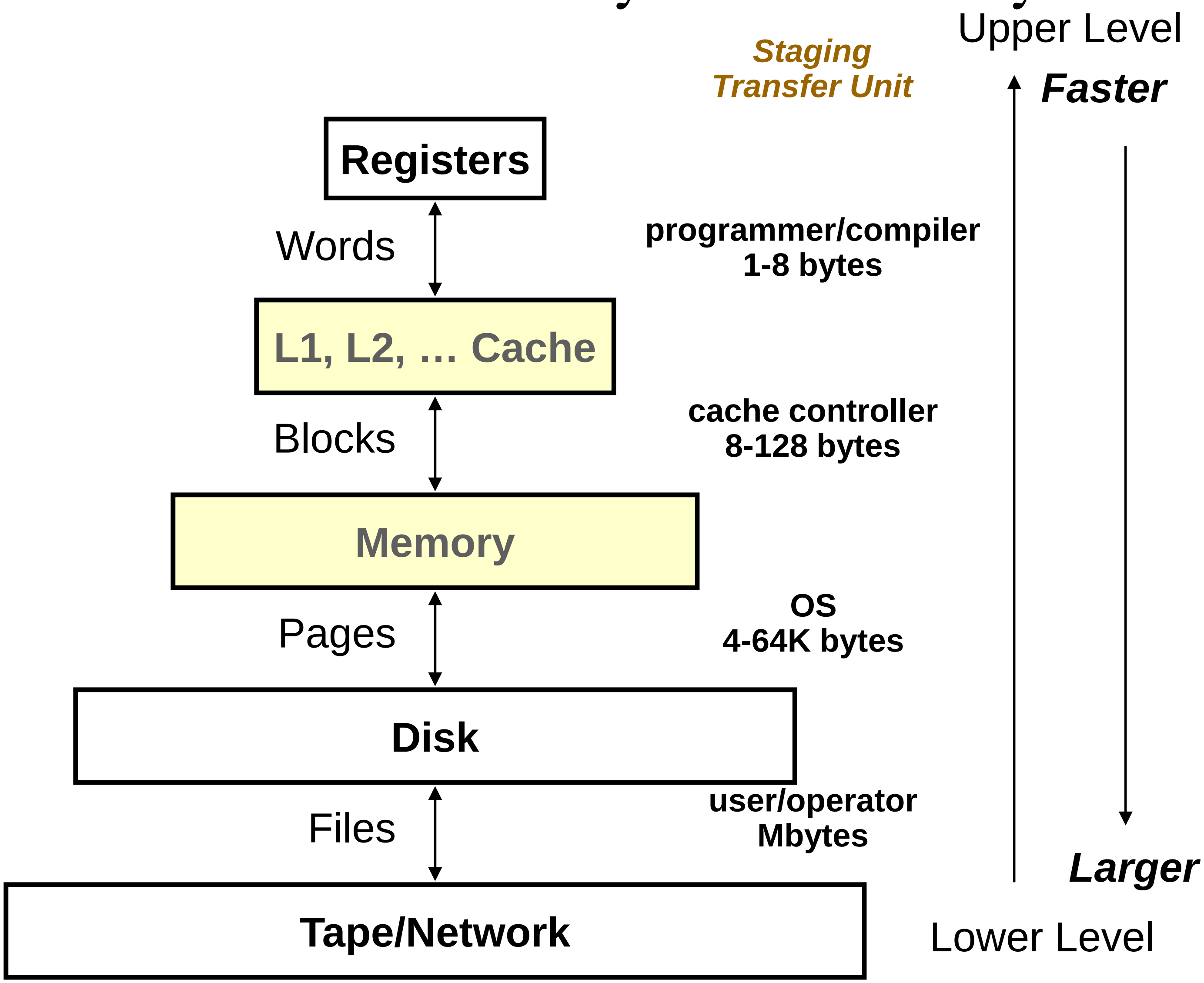
CPU Registers
1000 Bytes
300 ps
~$.01

Cache
16K-4M Bytes
1 ns
~$.0001

Main Memory
4G-16G Bytes
50 - 100ns
~$.0000001

Disk
1 – 16 T Bytes
5 – 10 ms
10^{-5}- 10^{-7} cents

Tape/Network
“infinite”
secs.
10^{-8} cents



Staging
Transfer Unit

Upper Level

Faster

Registers

Words

programmer/compiler
1-8 bytes

L1, L2, … Cache

Blocks

cache controller
8-128 bytes

Memory

Pages

OS
4-64K bytes

Disk

Files

user/operator
Mbytes

Tape/Network

Larger

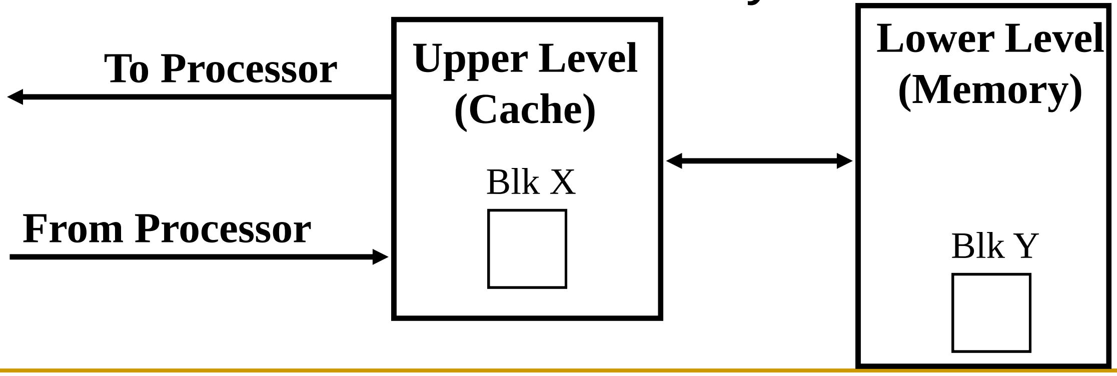
Lower Level

Comparison Chart

Level	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Name	Register	L1 Cache	L2 Cache	L3 Cache	Main Memory	Disk
Size	Depends	16KB	256KB	4MB	16GB	16TB
Implementation Technology	Custom memory with multiple ports CMOS	ON chip SRAM CMOS	ON chip SRAM CMOS	Off chip SRAM CMOS	CMOS DRAM	Magnetic
Access Time	300ps	1ns	10ns	20ns	100ns	10ms
BW (MB/s)	100,000	10,000	8,000	5,000	3,000	150
Managed by	Compiler	Hardware	Hardware	Hardware	OS	OS/Operator
Backed by	L1 Cache	L2 Cache	L3 Cache	Main Memory	Magnetic Disk	Magnetic Tape

Memory Hierarchy: Principles of Operation

- At any given time, data is copied between only 2 adjacent levels
  - Upper Level (Cache): the one closer to the processor
    - Smaller, faster, and uses more expensive technology
  - Lower Level (Memory): the one further away from the processor
    - Bigger, slower, and uses less expensive technology
- Block
  - The smallest unit of information that can either be present or not present in the two-level hierarchy



Memory Hierarchy: Terminology

- Hit: data appears in some block in the upper level
  - Hit Rate = fraction of memory access found in upper level
  - Hit Time = time to access the upper level
    - memory access time + Time to determine hit/miss
- Miss: data needs to be retrieved from a block in the lower level (e.g.: Block Y in previous slide)
  - Miss Rate = 1 - (Hit Rate)
  - Miss Penalty: includes time to fetch a new block from lower level
    - Time to replace a block in the upper level from lower level + Time to deliver the block the processor
- Hit Time: significantly less than Miss Penalty

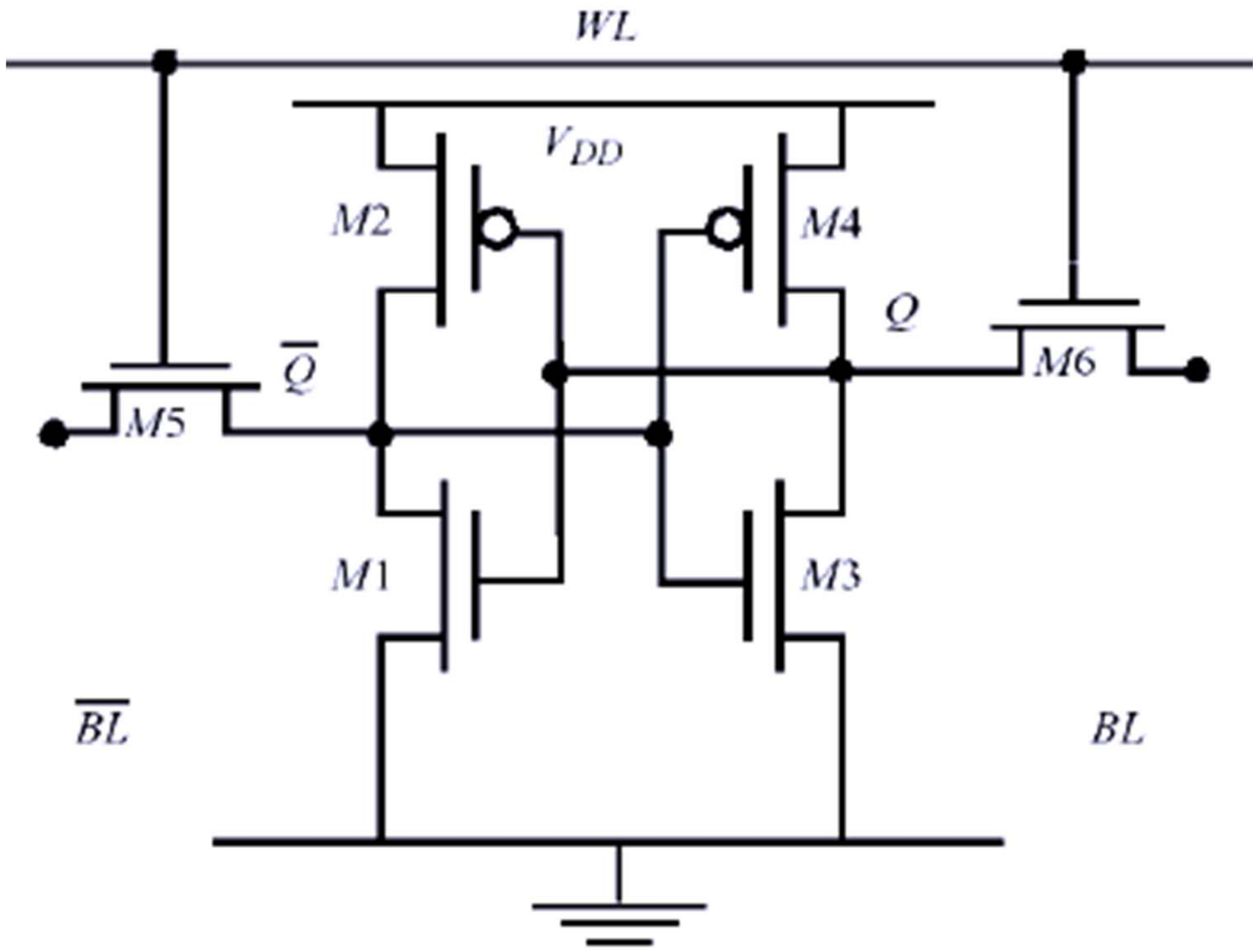
Semiconductor Memory

- RAM
  - Misnamed as all semiconductor memory is random access
  - Read/Write
  - Volatile
  - Temporary storage
  - Static or dynamic

Static RAM

- Capable of retaining state as long as power is applied
- No refreshing needed when powered
  - Retain value indefinitely as long as it is kept powered
  - No charges to leak
- Bits stored as on/off switches (bistable memory cell)
  - Relatively insensitive to disturbance
  - Faster response
  - Used for Cache Memory
- More complex construction
  - 6 transistor circuit
  - More expensive
  - Larger in size per bit

SRAM



WL
V_{DD}
M2
M4
$\overline{Q}$
Q
M5
M6
M1
M3
$\overline{BL}$
BL