Computer Architecture and Organization

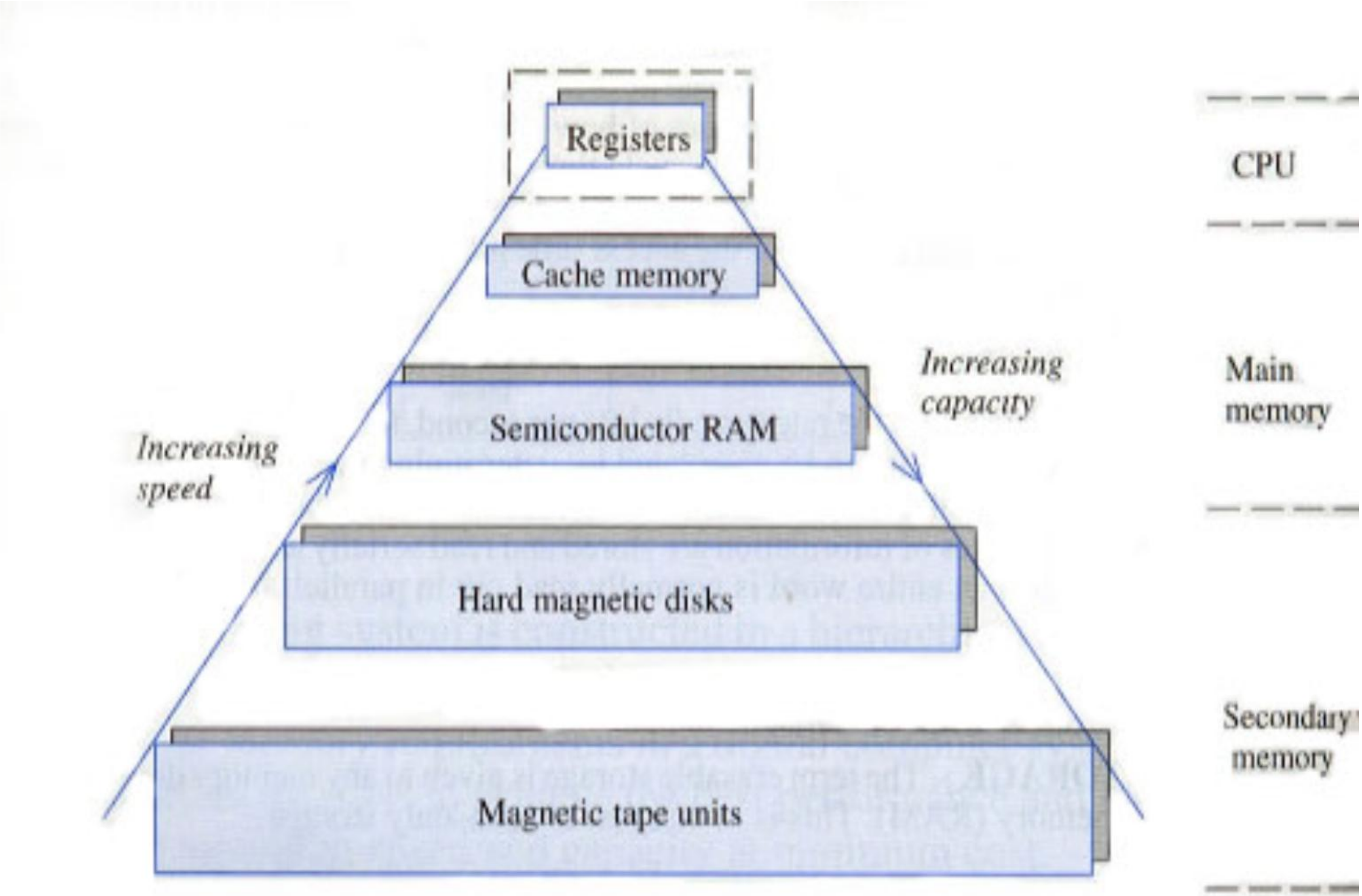
Memory System Design

Lecture 4

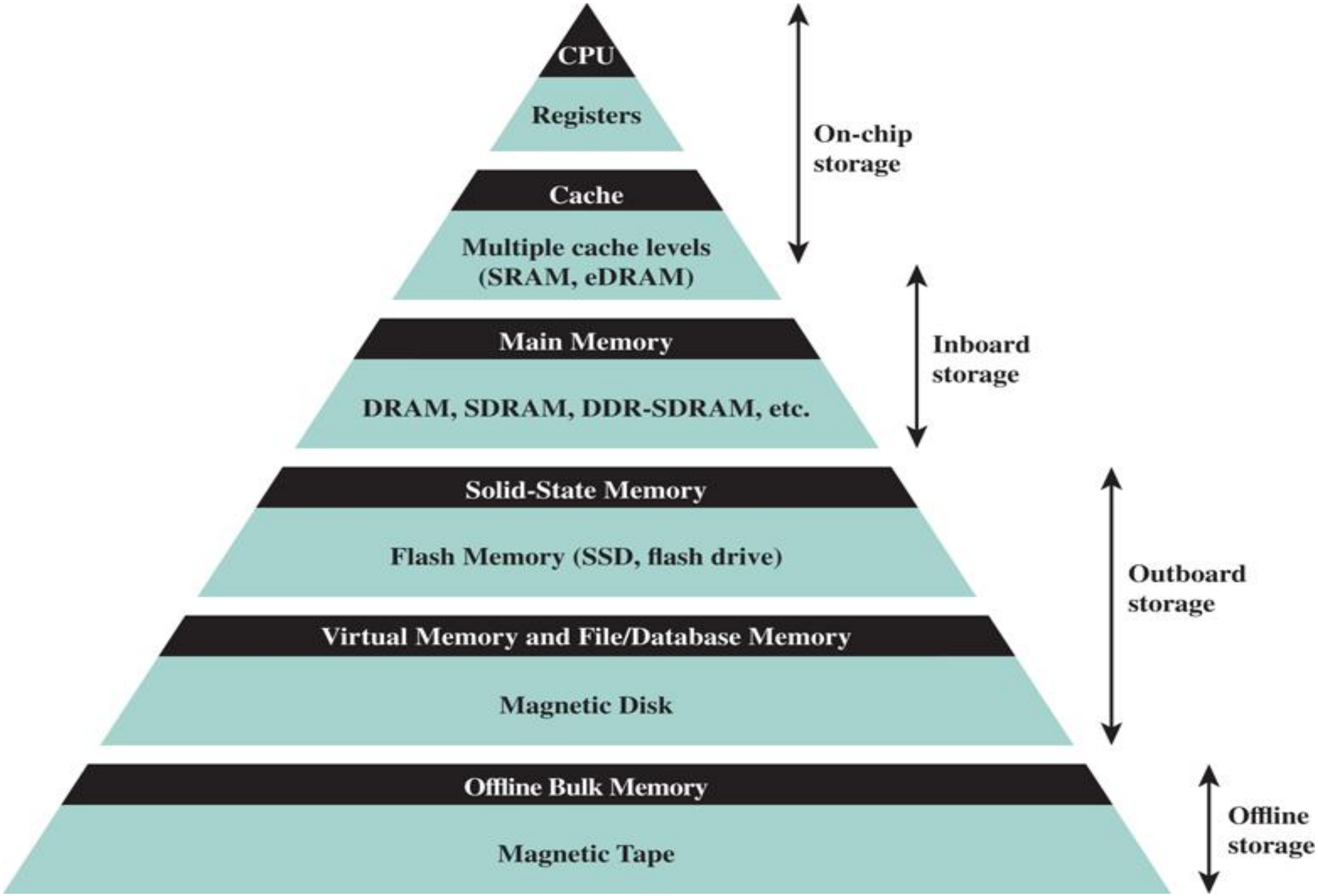
Characteristics of Memory Systems

- ❑ Computer memory is organized into a hierarchy.
- ❑ At the highest level (closest to the processor) are the processor registers.
- ❑ One or more levels of cache - L1, L2,.. .
- ❑ Main memory - All of these are considered internal to the computer system.
- ❑ The hierarchy continues with external memory - a fixed hard disk, and one or more levels below that consisting of removable media such as optical disks and tape.

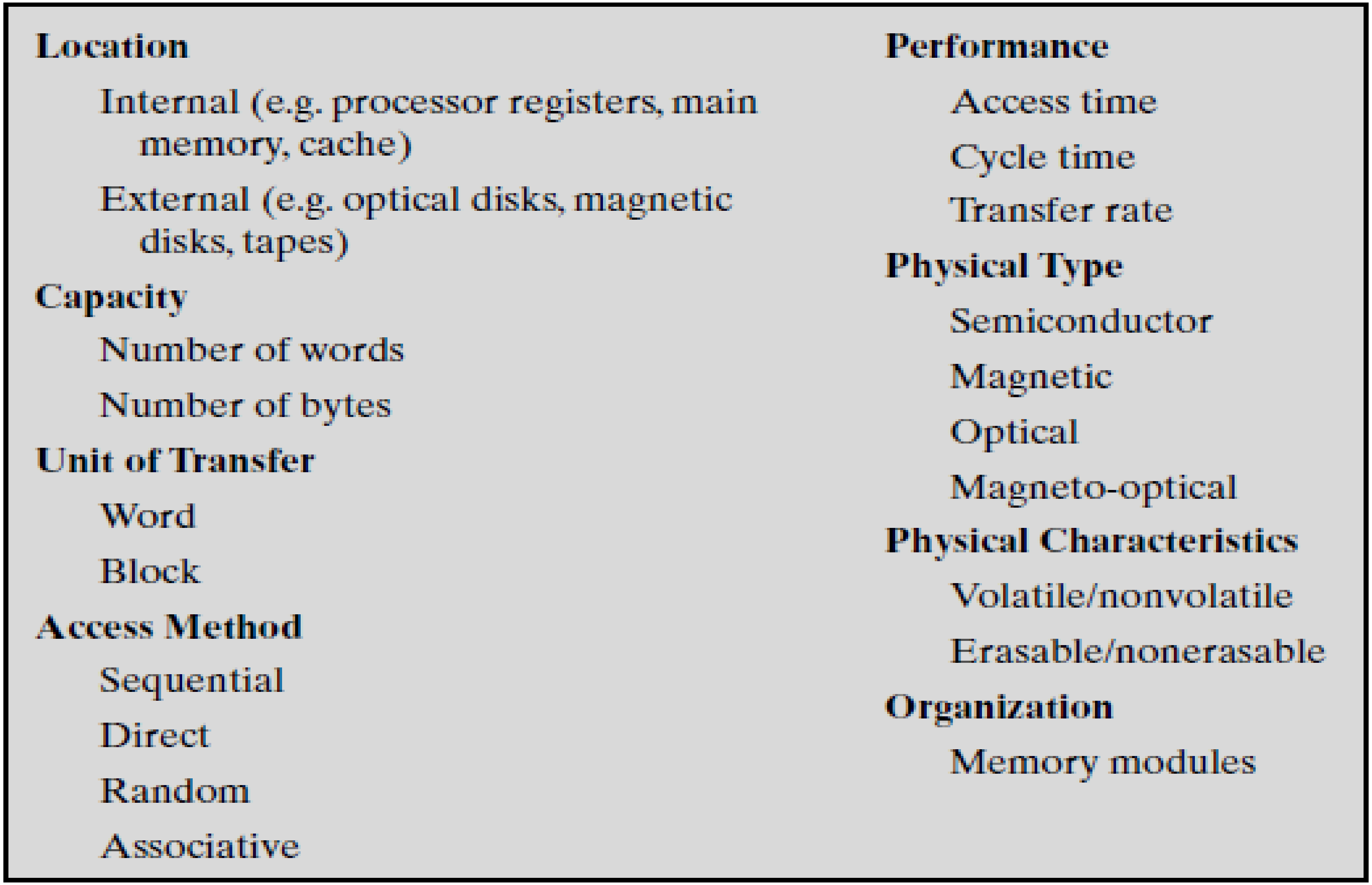
Memory Hierarchy



Memory Hierarchy



Characteristics of Memory Systems



Location

- Internal (e.g. processor registers, main memory, cache)
- External (e.g. optical disks, magnetic disks, tapes)

Capacity

- Number of words
- Number of bytes

Unit of Transfer

- Word
- Block

Access Method

- Sequential
- Direct
- Random
- Associative

Performance

- Access time
- Cycle time
- Transfer rate

Physical Type

- Semiconductor
- Magnetic
- Optical
- Magneto-optical

Physical Characteristics

- Volatile/nonvolatile
- Erasable/nonerasable

Organization

- Memory modules

Characteristics of Memory Systems

❑ **location -** refers to whether memory is internal and external to the computer.

❑ **Capacity-** For internal memory, this is typically expressed in terms of bytes (1 byte = 8 bits) or words. word lengths are(8, 16, and 32 bits). External memory capacity is typically expressed in terms of bytes.

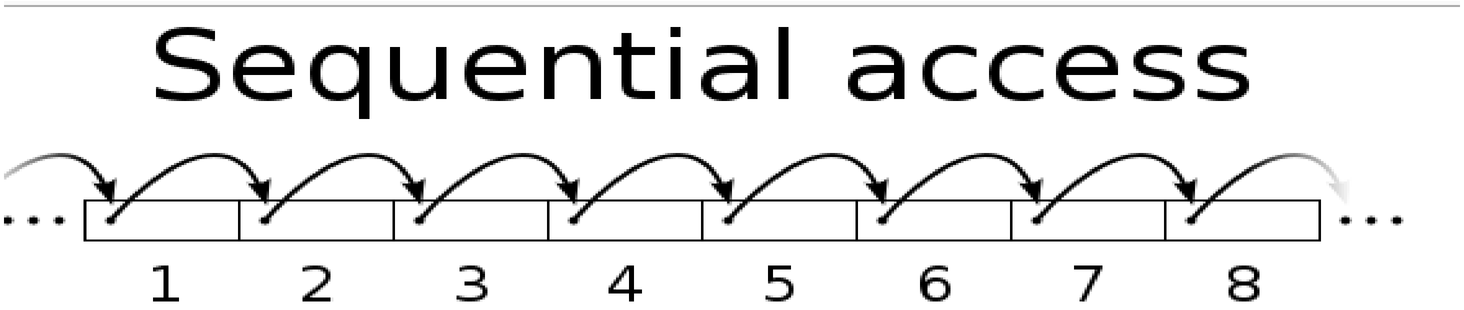
❑ **unit of transfer**- For internal memory, the unit of

transfer is equal to the number of electrical lines into and out of the memory module.

Characteristics of Memory Systems

❑ **Method of accessing units-** These include the following:

- Sequential Access: Memory is organized into units of data, called **records**. Data access is very slow, because the data will be sorted serially one by one. (Magnetic Tape)



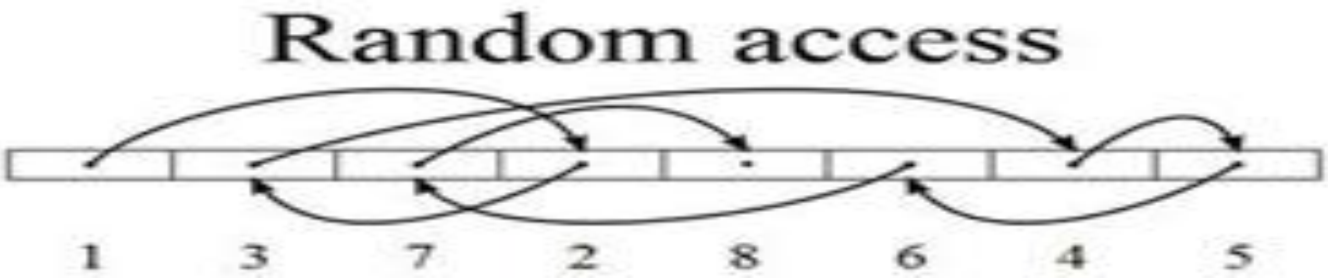
Characteristics of Memory Systems

❑ Method of accessing units :

- **Direct access** : obtain data from a storage device by going directly to where it is physically located on the device rather than by having to sequentially look for the data at one physical location after another. (Hard Disk, Floppy Disk).

- **Random access:** any location can be selected at random and directly addressed and accessed.

(Main mer



Characteristics of Memory Systems

❑ **Method of accessing units :**

- Associative : This is a random access type of memory that enables one to make a comparison of desired bit locations within a word for a specified match, and to do this for all words simultaneously.

❑ **Performance-**Three performance parameters are used:

1- Access time (latency)

2- Memory cycle time

3- Transfer rate

Characteristics of Memory Systems

❑**Performance :**

- **Access time** (latency): this is the time it takes to perform a read or write operation.

- **Memory cycle time** : consists of the access time plus any additional time required before a second access can commence.

- **Transfer rate :** This is the rate at which data can be transferred into or out of a memory unit. For random-access memory, it is equal to 1/(cycle time).

Characteristics of Memory Systems

- ❑ **Physical types of memory -** semiconductor memory, magnetic surface memory, used for disk and tape, and optical and magneto-optical.
- ❑ **physical characteristics :**

- volatile memory- that loose the stored information when power is turned off. **(semiconductor memories- RAM) .**

- nonvolatile memory-that retain the stored information even when the power is turned off . (Magnetic memories and some semiconductor memories ROM) .

- ❑ **Organization** - is meant the physical arrangement of bits to form words

The Memory Hierarchy

- The design constraints on a computer’s memory can be summed up by three questions: How much? How fast? How expensive?
- As one goes down the hierarchy, the following occur:

a. Decreasing cost per bit

b. Increasing capacity

c. Increasing access time

d. Decreasing frequency of access of the memory by the processor

CACHE MEMORY PRINCIPLES

❑ Cache memory :

✓ is a very high speed semiconductor memory which can speed up CPU.

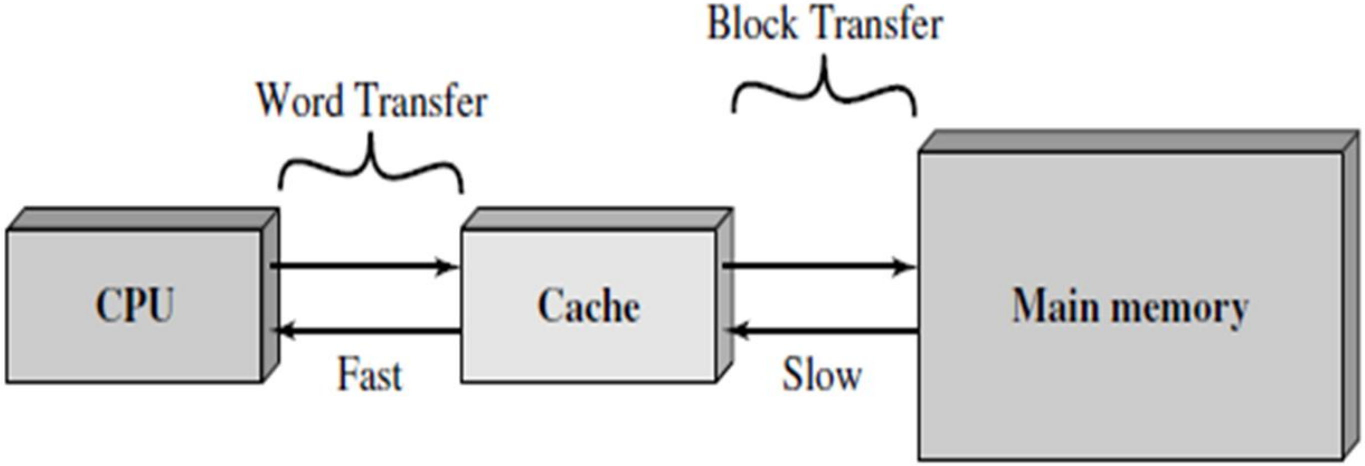
✓ It acts as a buffer between the CPU and main memory.

✓ It is used to hold those parts of data and program which are most frequently used by CPU.

✓ The parts of data and programs are transferred from disk to cache memory by operating system, from where CPU can access them.

CACHE MEMORY PRINCIPLES

- ✓ Cache memory is designed to combine the memory access time of expensive, high- speed memory combined with the large memory size of less expensive, lower- speed memory.

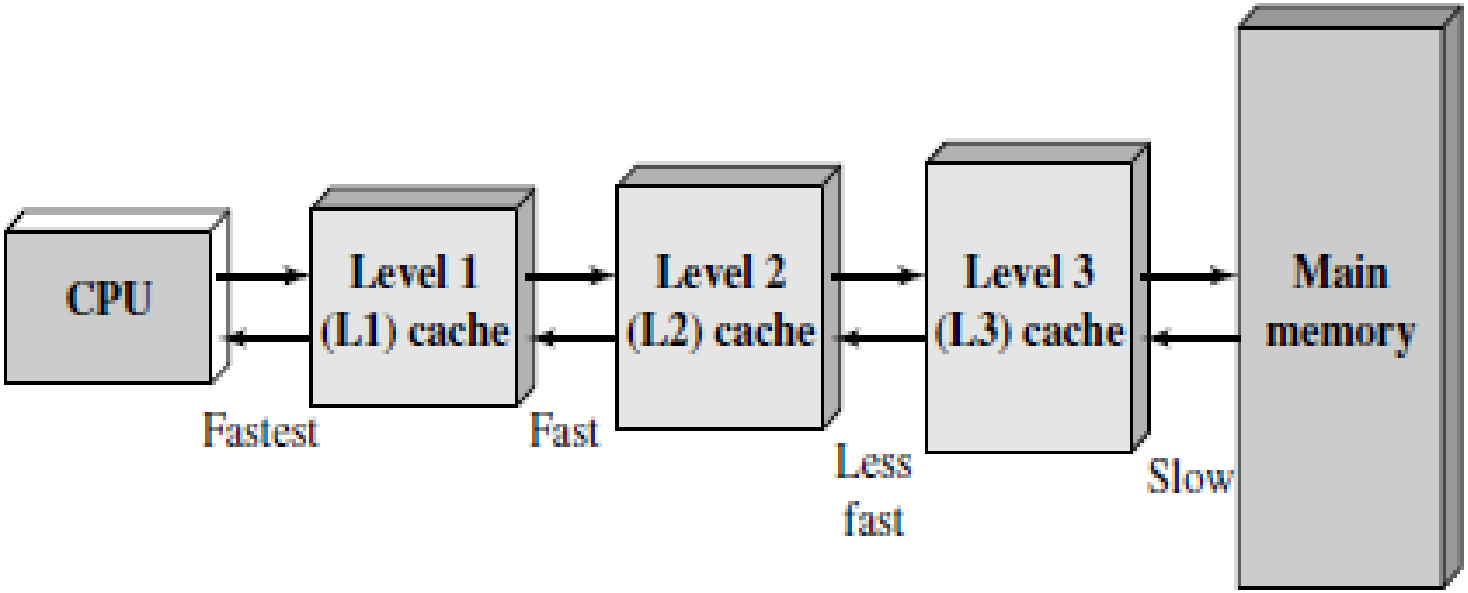


CACHE MEMORY PRINCIPLES

- The cache contains a copy of portions of main memory.
- When the processor attempts to read a word of memory, a check is made to determine if the word is in the cache. If so, the word is delivered to the processor .
- If not, a block of main memory, consisting of some fixed number of words, is read into the cache and then the word is delivered to the processor.
- Because of the phenomenon of locality of reference, when a block of data is fetched into the cache to satisfy a single memory reference, it is likely that there will be future references to that same memory location or to other words in the block.

CACHE MEMORY PRINCIPLES

✓ **Three-level cache organization** :The L2 cache is slower and typically larger than the L1 cache, and the L3 cache is slower and typically larger than the L2 cache.

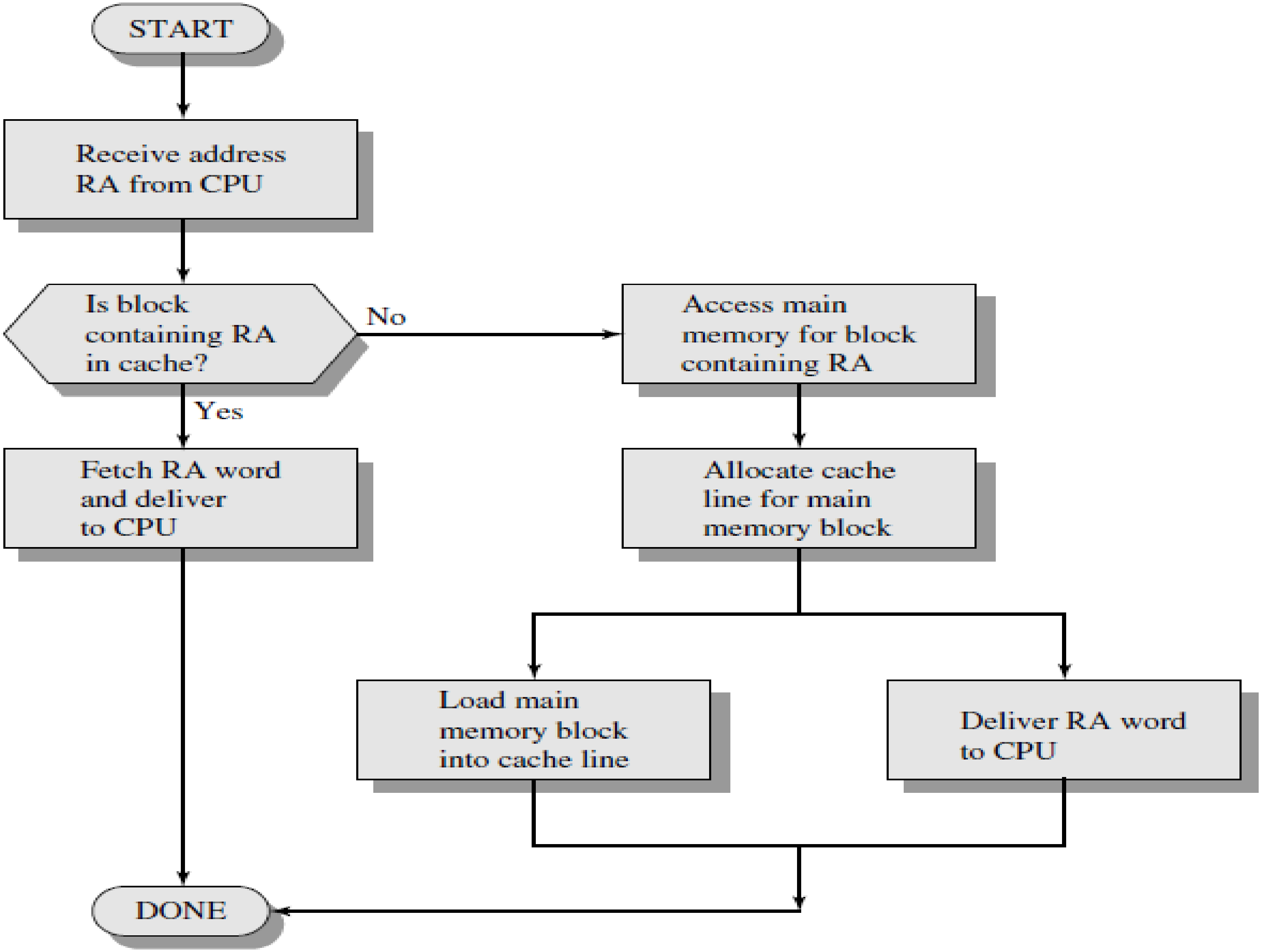


CACHE MEMORY PRINCIPLES

✓ **Cache Operation**

- When the processor finds data in the cache that it is looking for **Hit Cache**.
- When the processor looks for data in the cache, but the data is not available it is looking for **Miss Cache**.
- In the event of a miss, the cache controller unit must gather the data from the main memory, which can cost more time for the processor.

Cache Read Operation



START
Receive address RA from CPU
Is block containing RA in cache?
No
Yes
Access main memory for block containing RA
Fetch RA word and deliver to CPU
Allocate cache line for main memory block
Load main memory block into cache line
Deliver RA word to CPU
DONE

Cache Read Operation

- ❑ The processor generates the read address **(RA)** of a word to be read. If the word is contained in the cache, it is delivered to the processor. Otherwise, the block containing that word is loaded into the cache, and the word is delivered to the processor.

❑ Typical Cache Organization

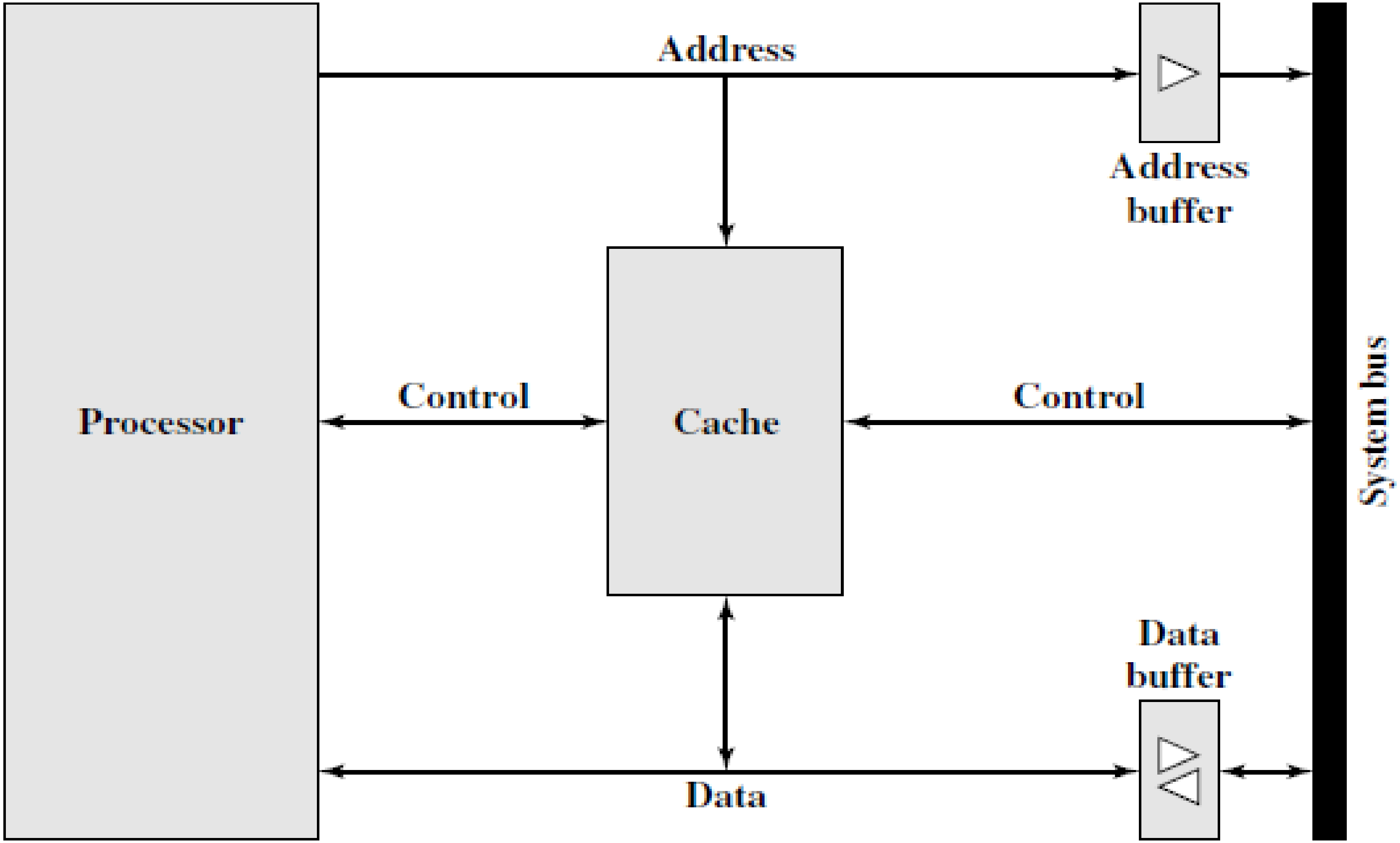
- The cache connects to the processor via data, control, and address lines .

- The data and address lines also attach to data and address buffers, which attach to a system bus from which main memory is reached .

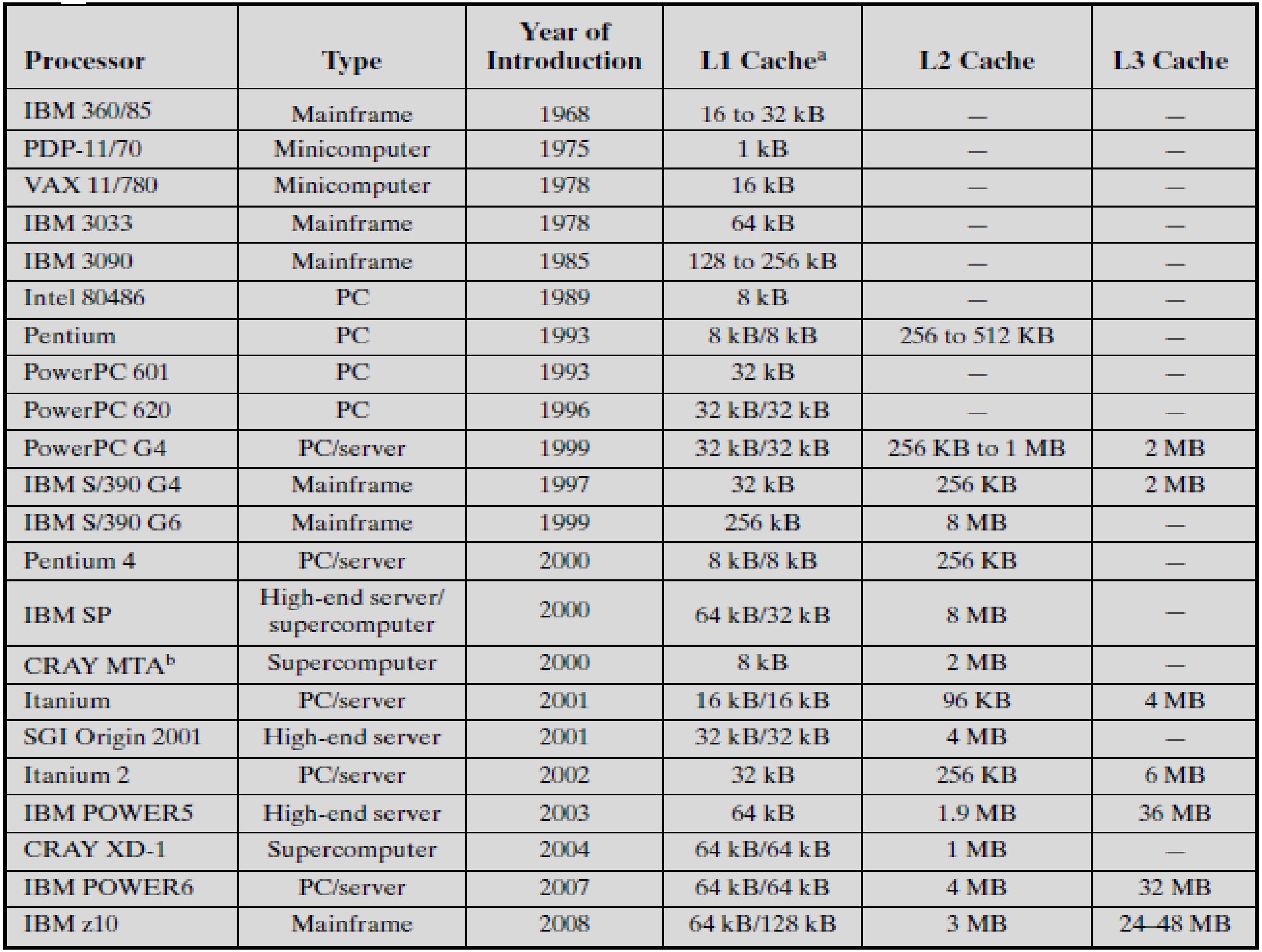
- When a cache hit occurs, the data and address buffers are disabled and communication is only between processor and cache with no system bus traffic .

- When a cache miss occurs, the desired address is loaded onto the system bus and the data are returned through the data buffer to both the cache and the processor.

❑ Typical Cache Organization

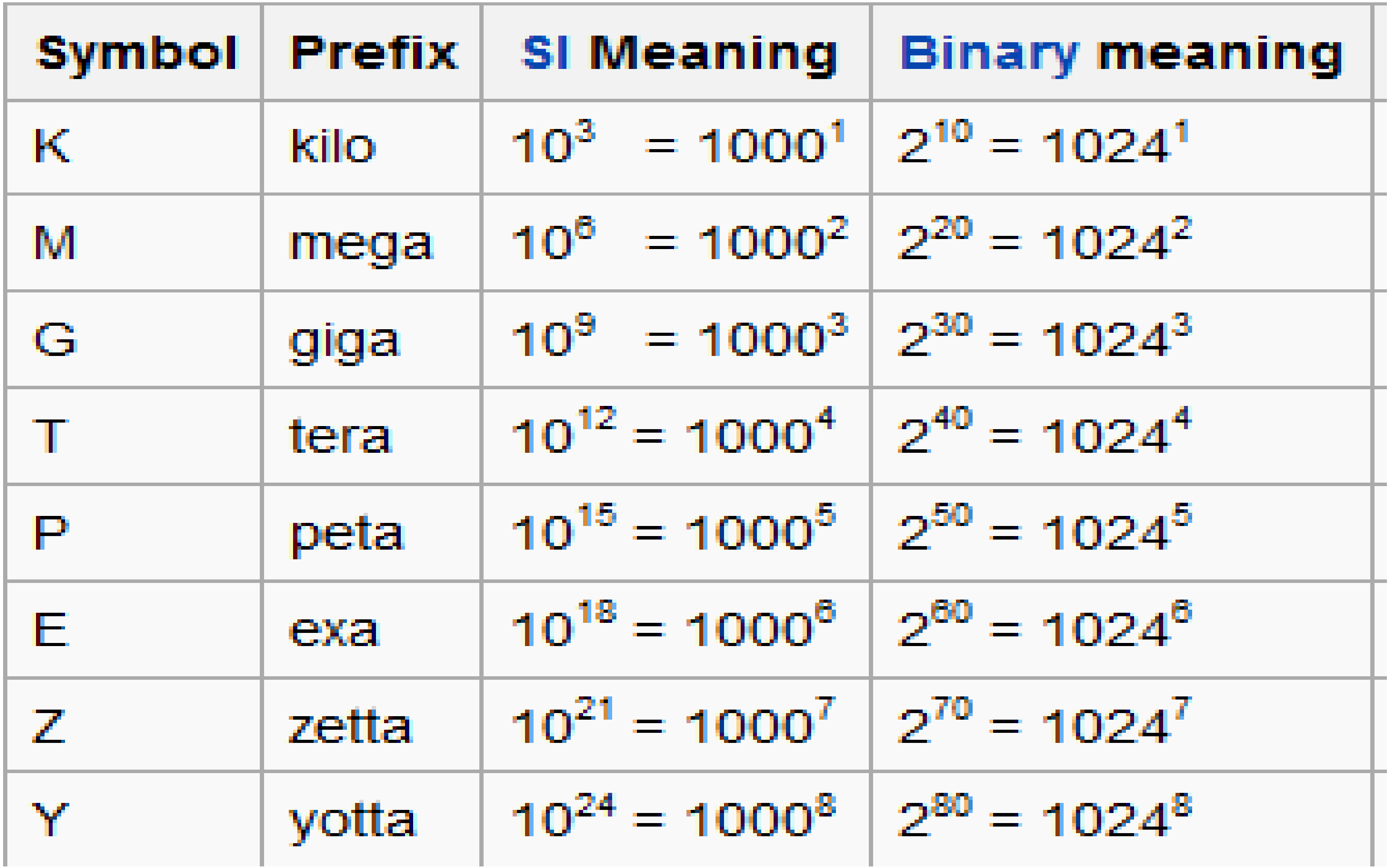


Cache Sizes of Some Processors



Processor	Type	Year of Introduction	L1 Cache[a]	L2 Cache	L3 Cache
IBM 360/85	Mainframe	1968	16 to 32 kB	—	—
PDP-11/70	Minicomputer	1975	1 kB	—	—
VAX 11/780	Minicomputer	1978	16 kB	—	—
IBM 3033	Mainframe	1978	64 kB	—	—
IBM 3090	Mainframe	1985	128 to 256 kB	—	—
Intel 80486	PC	1989	8 kB	—	—
Pentium	PC	1993	8 kB/8 kB	256 to 512 KB	—
PowerPC 601	PC	1993	32 kB	—	—
PowerPC 620	PC	1996	32 kB/32 kB	—	—
PowerPC G4	PC/server	1999	32 kB/32 kB	256 KB to 1 MB	2 MB
IBM S/390 G4	Mainframe	1997	32 kB	256 KB	2 MB
IBM S/390 G6	Mainframe	1999	256 kB	8 MB	—
Pentium 4	PC/server	2000	8 kB/8 kB	256 KB	—
IBM SP	High-end server/ supercomputer	2000	64 kB/32 kB	8 MB	—
CRAY MTA[b]	Supercomputer	2000	8 kB	2 MB	—
Itanium	PC/server	2001	16 kB/16 kB	96 KB	4 MB
SGI Origin 2001	High-end server	2001	32 kB/32 kB	4 MB	—
Itanium 2	PC/server	2002	32 kB	256 KB	6 MB
IBM POWER5	High-end server	2003	64 kB	1.9 MB	36 MB
CRAY XD-1	Supercomputer	2004	64 kB/64 kB	1 MB	—
IBM POWER6	PC/server	2007	64 kB/64 kB	4 MB	32 MB
IBM z10	Mainframe	2008	64 kB/128 kB	3 MB	24–48 MB

standard units of measurement used for data storage



Symbol	Prefix	SI Meaning	Binary meaning
K	kilo	$10^3 = 1000^1$	$2^{10} = 1024^1$
M	mega	$10^6 = 1000^2$	$2^{20} = 1024^2$
G	giga	$10^9 = 1000^3$	$2^{30} = 1024^3$
T	tera	$10^{12} = 1000^4$	$2^{40} = 1024^4$
P	peta	$10^{15} = 1000^5$	$2^{50} = 1024^5$
E	exa	$10^{18} = 1000^6$	$2^{60} = 1024^6$
Z	zetta	$10^{21} = 1000^7$	$2^{70} = 1024^7$
Y	yotta	$10^{24} = 1000^8$	$2^{80} = 1024^8$