Memory Organization

Characteristic of memory system

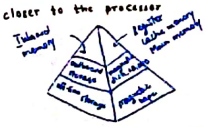
1. Location
   - CPU
   - Internal
   - External
2. Capacity
3. Unit of Transfer
   - Internal
   - External
4. Access Method
   - Sequential access
   - Direct access
   - Random access
   - Associative access
5. Performance
   - Access time
   - Memory Cycle Time
6. Physical Type
   - Semiconductor
   - Magnetic
   - Optical
7. Physical Characteristics
8. Organization

Bytes and Bits

- The byte is a unit of digital information that mostly consists of eight bits
- Infact, a byte was a number of bits used to encode a single character of text in a computer and for this reason it has become the basic addressable element in many computer architectures.
- The size of the byte has been hardware dependent and no definition exists

Memory Hierarchy

- It explains that the nearer the memory to the processor, faster is its access
- Costlier the memory becomes as it goes closer to the processor



SRAM

- A type of semiconductor memory that uses bi-stable latching circuitry (flip flop) to store each bit
- Stands for Static Random Access Memory
- Very Fast
- Does not require refresh cycles to retain data
- Requires refreshing, it has more complex circuitry and timing requirements
- Used for CPU cache
- Require minimum time to access data
- Complex Structure → has flip flops
- Has a lower density
- Expensive

DRAM

- A type of random access semiconductor memory that stores each bit of data in a separate tiny capacitor within an integrated circuit
- Stands for Dynamic Random Access Memory
- Not as fast as SRAM
- Requires periodical refresh cycles to retain data
- Not as complex as SRAM
- Used for the computers main memory
- Requires more time to access data
- Simple structures → has a transistor and a capacitor
- Has a higher density
- Less Expensive

Types of Memory

→ SPROM (Erasable Programmable Read Only Memory)

→ RAM (Random Access Memory)

ROM

- ROM is cheaper compared to RAM.
- Used for implementation of the secondary or the virtual memory
- Applications: Hard disks, External storage like CD/DVD and floppy Disks, etc
- Advantages:
  - Non volatile in nature
  - Cannot be accidentally changed
  - Cheaper than RAM's
  - Easy to test
  - More reliable than RAM
  - Static and do not require refreshing
  - Contents are always known / can be verified

Types of ROM

- MROM (Masked ROM)
- PROM (Programmable Read Only Memory)
- EPROM (Erasable and Programmable Read Only Memory)
- EEPROM (Electrically Erasable and Programmable Read Only Memory)

Magnetic Memory

- Magnetic Disks are cheap
- Used as External Storage and Hard Disks
- When used as Hard Disk they are called as Winchester Disk

Optical Memory

- Memory devices like Compact Disks (CD) & Digital Versatile Disk / Digital Video Disk (DVD) use the optical method to read the data written on them
- Devices for optical memory
  → CD ROM
  → DVD

Allocation Policies

- Best Fit
  - The smallest variable fragment is searched and the required data is stored in that fragment
  - The smallest fragment searched for should be ≥ size of data to be stored
- Worst Fit
  - The largest available block is used to store data
- Next Fit
  - Immediate next empty block of size ≥ size of data to be stored in searched sequentially and required data is stored there

Principles of Locality of reference

- Locality of reference is the term used to explain the characteristics of programs that run in relatively small loops in consecutive memory locations.
- The locality of reference principle comprises of: → Temporal Locality
  → Spatial Locality

Temporal Locality

- Since the programs have loops, the same instructions are required frequently i.e. the program tend to use the most recently used information again & again
- If for a long time an information in cache is not used then it is less likely to be used again
- This is known as the principle of temporal locality

Spatial Locality

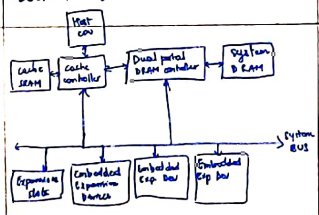
- Programs and the data accessed by the processor mostly reside in consecutive memory locations.
- This means that processor is likely to need code or data that are close to locations already accessed
- This is known as principle of Spatial Locality

Cache Performance

% Hit Rate = $\frac{\text{Cache Hits}}{\text{Total Memory Accesses}} \times 100\%$

Cache Architecture

- Look Through Cache Design



Cache Memory

- It is a very high speed semiconductor memory which can speed up the CPU
- It acts as a buffer between the buffer and the main memory
- Used to hold most frequently used part or data used by CPU
- The parts of data and programs are transferred from the disk to cache memory by the OS from where the CPU can access them

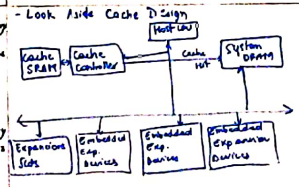
Advantage:

- Cache memory is faster than main memory
- It consumes less access time as compared to main memory
- It stores the program that can be executed within a short period of time
- It stores data for temporary use

Disadvantage:

- Cache memory has limited capacity
- It is very expensive

- Look Aside Cache Design



Cache Coherency

Snooping

- Shared memory multiprocessor environment
- Main memory is passive
- Cache distributes state transition to other caches and memory
- All caches listen to snoop messages and act on them
- Most machines used cache coherence protocols with different trade off

But, performance also depends on physical implementation
- Bus Design
- Cache Design
- Integration with memory

Write Policy

→ Write through cache designs
→ Buffered or posted write through designs
→ Write back cache designs

Bus master/Cache interaction for Cache Coherency

- When another device in system uses the buses, it must become bus master
- Since bus master can write to and read from memory, cache consistency problems may happen under 3 circumstances
  → Writes-to-memory (With write through cache)
  → Reads from memory (With write back cache)
  → Write to main memory (With write back cache)

Bus snooping / Snarfing

- It is the method used by cache controller to monitor memory accessed performed by other bus master
- There are 2 possibilities that may create the need to snoop the bus
  ① Memory writes by another bus master
  → Invalidate the line
  → Snarf the data
  ② Memory read by another bus master

- Replacement algo is required to replace a line from the cache memory with the new line
- Types of Replacement policies
  → Least Recently used (LRU)
  → First In First Out (FIFO)
  → Least Frequently used
  → Random

Cost and performance measurement of 2 level memory hierarchy

- Parameters considers for Performance Analysis
  → Average Cost (C_s) = $\frac{C_1 S_1 + C_2 S_2}{S_1 + S_2}$
  → Hit ratio (H) = $\frac{N_1}{N_1 + N_2}$
  → Average Access Time (t_a) = $H\,t_{A1} + (1-H)\,t_{A2}$
  → Efficiency (%) = $\frac{t_{A1}}{t_A}$

Cache Mapping Techniques

→ Direct Mapping Technique

Advantages
- Simple implementation
- Inexpensive

Disadvantages
- Fixed location for given block, hence if a program accessed 2 blocks that map to the same line repeatedly, cache misses are very high

→ Fully Associative Mapping

Advantages
- If a program accesses 2 blocks repeatedly, cache misses will not occur

Disadvantages
- Complex design for many parallel comparisons of tag
- Expensive due to implementation of parallel comparator

→ Set Associative Mapping

Associative Memory
- A memory unit accessed by content is called an associative memory
- It is also called as parallel search memory or multi-access memory
- Stored data can be identified for access by content of data rather than address

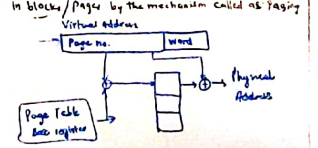
Interleaved Memory
- Main memory divided into 2 or more sections
- The CPU can access alternate sections immediately without waiting for memory to catch up (through wait states)
- Interleaved memory is one technique for compensating for the relative slow speed of dynamic RAM (DRAM)

Virtual Memory

- The term virtual memory refers to something which appears to be present but actually it is not.
- The virtual memory technique allows user to use more memory for a program than the real memory of computer
- So, virtual memory is the concept that gives the illusion to the user that they will have main memory equal to the capacity of secondary storage media

Paging

- The data required by the application is brought from the external slow memory to main memory in blocks / Pages by the mechanism called as Paging



Demand Paging

Segmentation

- It refers to logical division of the main memory so as to give modular structure mechanism and multitasking