Coa exam notes unit wise:

Unit 2

Memory Hierarchy:

- Memory unit is an essential component in any digital computer as it is needed for storing programs and data.
- A computer is equipped with a hierarchy of memory subsystems, some internal to the system (directly accessible by the processor), and some external (accessible by the processor via an I/O module)
- **External memory** consists of peripheral storage devices, such as disk and tape, that are accessible to the CPU via I/O controllers. External memory can also be referred as *secondary memory* or *auxiliary memory*.
- **Internal memory** is equated with main memory. But there are other forms of internal memory like CPU requires its own local memory in the form of registers. Internal memory is also called as **main memory/Primary memory**.

Main memory can be classified as

- **Volatile:**

  -- RAM (Random Access Memory)

  -- RAM is working memory. Data can be read or written in RAM with the help of address location and when the data is no longer needed we can use the storage location for writing again.

  -- Contents of volatile memory are vanished when power supply is switched off.

- **Non Volatile :**

  -- ROM (Read Only Memory)

  -- It is useful to have instructions that are used often , permanently stored inside the computer. Programs and data on the ROM are not lost if the computer is powered down

There are three key characteristics of memory:

- COST
- CAPACITY
- ACCESS TIME

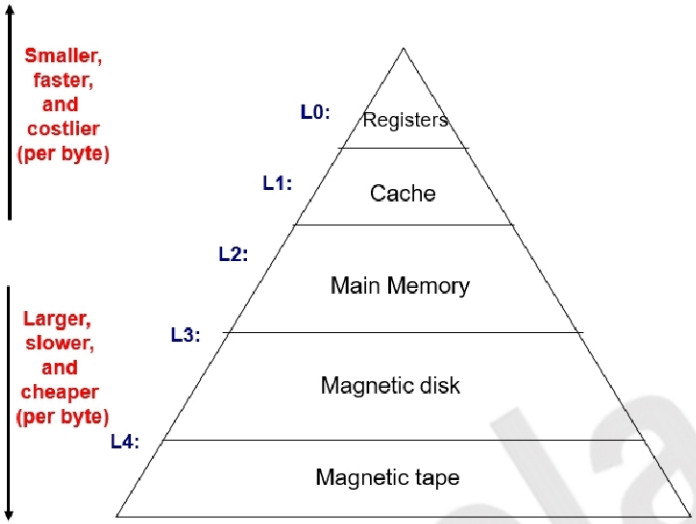
The relationship between them is as follows:

- ✓ Greater capacity, smaller cost per bit
- ✓ Greater capacity, greater access time
- ✓ Smaller access time, greater cost per bit

The overall goal of using a memory hierarchy is to obtain the highest possible average access speed while minimizing total cost of entire memory system.

Educlash Just Another Way To Learn

Memory hierarchy



In both the above figures as we go down following occurs :

a. Decreasing cost/bit
b. Increasing capacity
c. Increasing access time
d. Decreasing frequency of access of memory by the CPU

Random-Access Memory (RAM)

- **Key features**
  - RAM is packaged as a chip.
  - Basic storage unit is a cell (one bit per cell).
  - Multiple RAM chips form a memory.
- **Static RAM (SRAM)**

- Each cell stores bit with inverter, transistor circuit.
- Retains value indefinitely, as long as it is kept powered.
- Relatively insensitive to disturbances such as electrical noise.
- Faster and more expensive than DRAM.
- Access time is about 10 ns

- **Dynamic RAM (DRAM)**
  - Each cell stores bit with a capacitor and transistor.
  - Value must be refreshed every 10-100 ms.
  - Sensitive to disturbances.
  - Slower and cheaper than SRAM.
  - Access time is about 50 ns

Advanced DRAM Organization-explaination in book on page191

- Enhanced DRAM
  - Simplest of new DRAM architectures
  - Developed by Ramtron
  - Integrates a small SRAM cache onto a DRAM chip
  - Refresh operation can be conducted in parallel with cache read operation
  - It has separate read path and write path so it enables a subsequent read access to cache in parallel with the completion of write operation.

eduClash Just Another Way To Learn

- Cache DRAM
    - Developed by Mitsubishi
    - Similar to EDRAM, includes a larger SRAM cache than the EDRAM
    - SRAM on CDRAM can be used as either true cache or as a buffer to support the serial access to a block of data.
    - Buffer stores most recently accessed data.

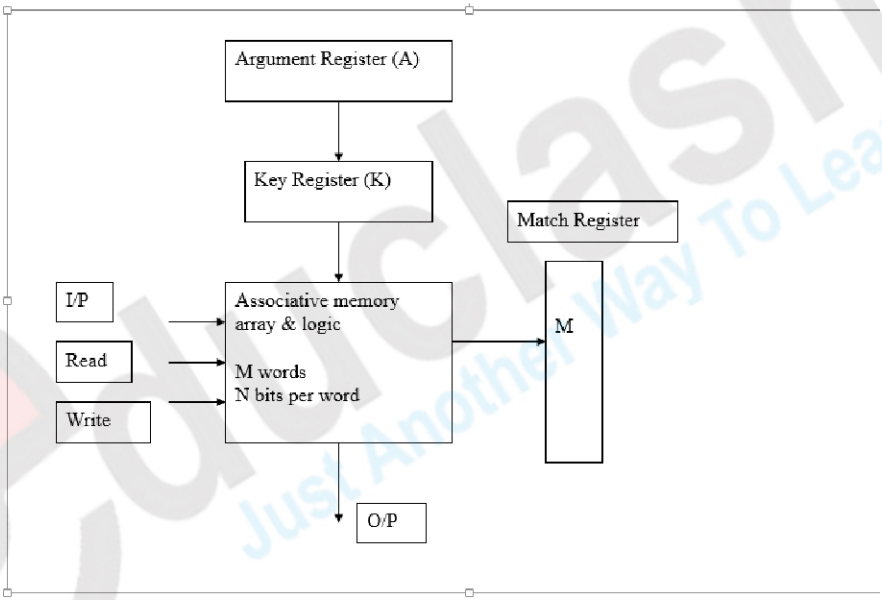
- Synchronous DRAM
    - Jointly developed by no of companies
    - It exchanges data with the processor synchronized to an external clock signal
    - It runs at speed equivalent to that of processor / memory bus.
    - Data moves in and out from DRAM under the control of system clock.
    - It has dual bank internal architecture
    - SDRAM includes important key features like Mode register and associated control logic. It provides a mechanism to customize SDRAM according to system needs.
    - It performs best when transferring large blocks of data serially, e.g. in word processing, multimedia etc.

- Rambus DRAM
    - Developed by Rambus
    - RDRAM chips are vertical packages, with all pins on one side.
    - fastest current memory technologies used by PCs. Normally SRAM can deliver data at a maximum speed of about 100 MHz, RDRAM transfers data at up to 800 MHz
    - RDRAM is used with Pentium III Xeon processors and more recently it is being used with Pentium 4 processors

Associative Memory

- Also called as **Content-addressable memory** (CAM), **associative storage**, or **associative array**
- Content-addressed or associative memory refers to a memory organization in which the memory is accessed by its content (as opposed to an explicit address).
- It is a special type of computer memory used in certain very high speed searching applications.
- In standard computer memory (random access memory or RAM) the user supplies a memory address and the RAM returns the data word stored at that address.
- In CAM the user supplies a data word and then CAM searches its entire memory to see if that data word is stored anywhere in it. If the data word is found, the CAM returns a list of one or more storage addresses where the word was found.
- CAM is designed to search its entire memory in a single operation.

- In CAM, each cell must have storage capability as well as logic circuits for matching its content with an external argument.
- Associative memories are expensive compared to RAMs because of the added logic associated with each cell.
- It is much faster than RAM in virtually all search applications.
- Associative memories are used in applications where the search time is very critical and short.



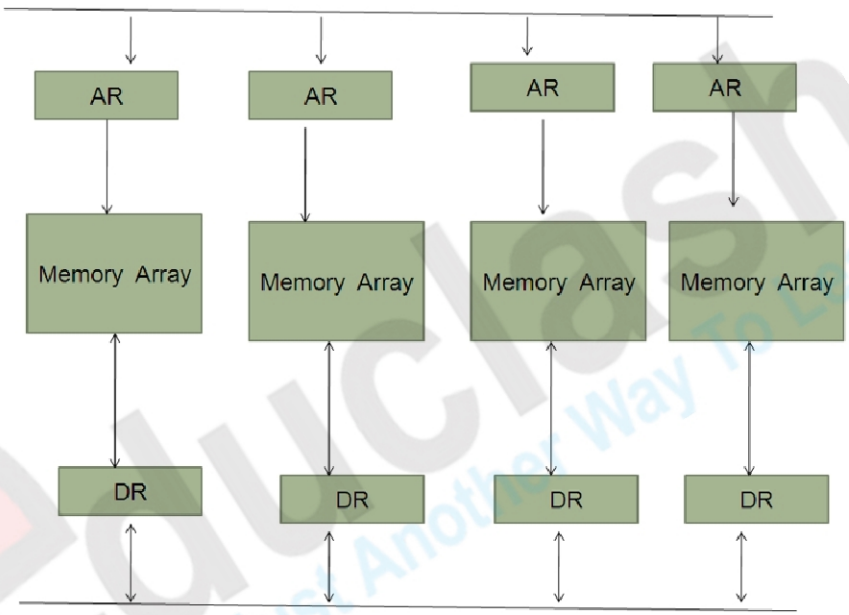
Block Diagram of Associative Memory

Interleaved Memory:

Main memory composed of a collection of DRAM memory chips. A number of chips grouped together to form a memory bank.

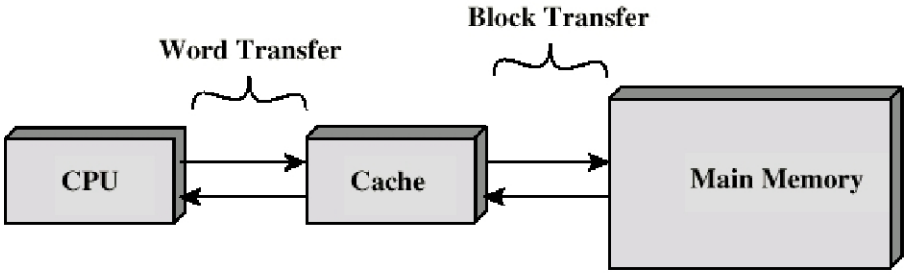
It is possible to organize the memory banks in a way known as interleaved memory.

Each bank is independently able to service a memory read and write request , so that a system with K banks can service K request simultaneously , increasing memory read or write rates by a factor of K. If consecutive words of memory are stored in different banks , then the transfer of block of memory is speeded up.



Cache

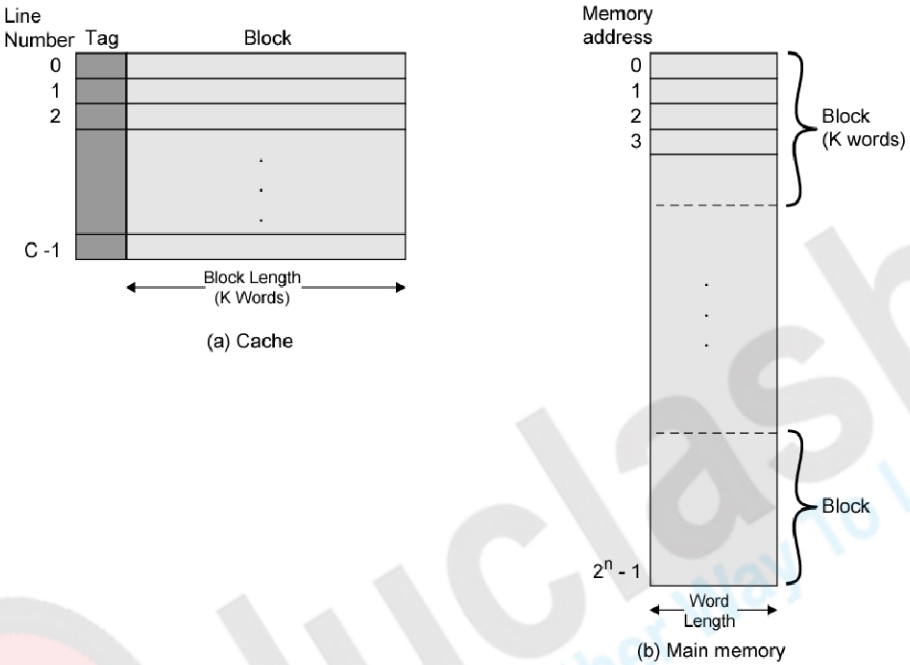
- Small amount of fast memory
- Sits between normal main memory and CPU
- May be located on CPU chip or module
- It is relatively expensive



Why Cache is needed?

- Cache is a special very high speed memory used to increase the speed of processing by making current programs and data available to CPU at a rapid rate.
- CPU logic is usually faster than main memory access time.
- To compensate for the mismatch in operating speeds is to use, extremely fast & small cache b/w CPU & main memory.
- Access time of cache is close to processor logic clock cycle time.
- Cache is used to store segments of programs currently being executed in CPU & temporary data frequently needed to do calculations.

Cache/Main Memory Structure:



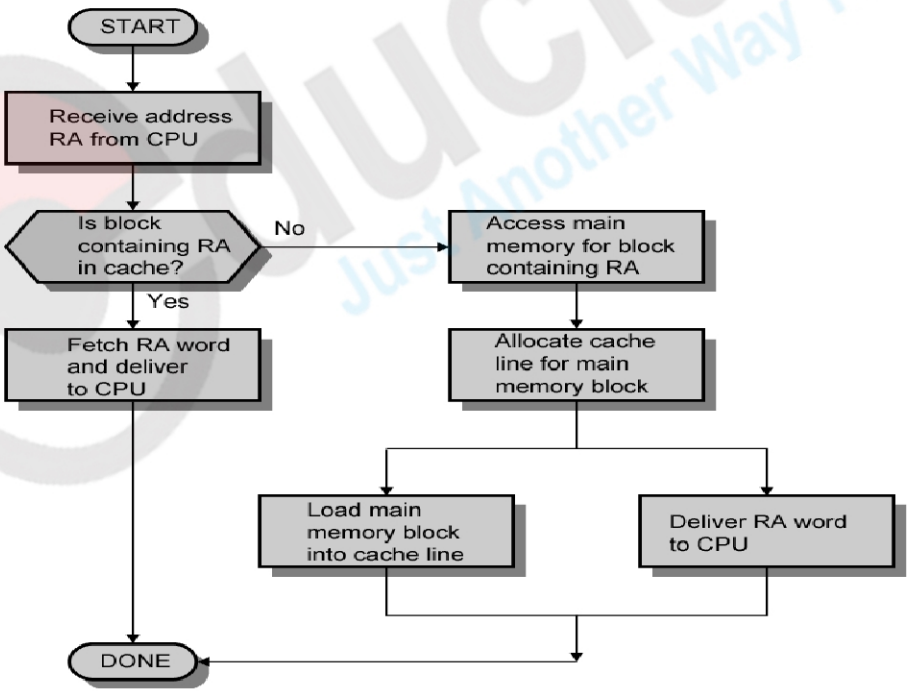
- Main memory consists of up to 2^n addressable words, each word has a unique n-bit address.
- For mapping purpose, this memory is considered to consist of a number of fixed length blocks of K words each.
- Total no of blocks M = 2^n / K
- Cache is divided into C no of slots, each slot can contain K words.
- No of slots in cache is considerably less than no of blocks in main memory (C<<M).
- An individual slot cannnot be dedicated to a particular block , so each slot includes a tag that identifies which block is

currently being stored. This is a portion of the main memory address.

Cache operation – overview

- CPU requests contents of memory location
- Check cache for this data
- If present, get from cache (fast)
- If not present, read required block from main memory to cache
- Then deliver from cache to CPU
- Cache includes tags to identify which block of main memory is in each cache slot

Cache Read Operation – Flowchart



Elements of Cache Design:

1. **Size**

- **Size of cache should be small enough so that overall average cost per bit is close to that of main memory alone and large enough so that overall average access time is close to that of the cache alone.**
- **Larger is cache, larger is the no of gates involved in addressing the cache, as a result performance decreases i.e. cache becomes slower.**
- **Cache size also depend on the available chip and board area.**

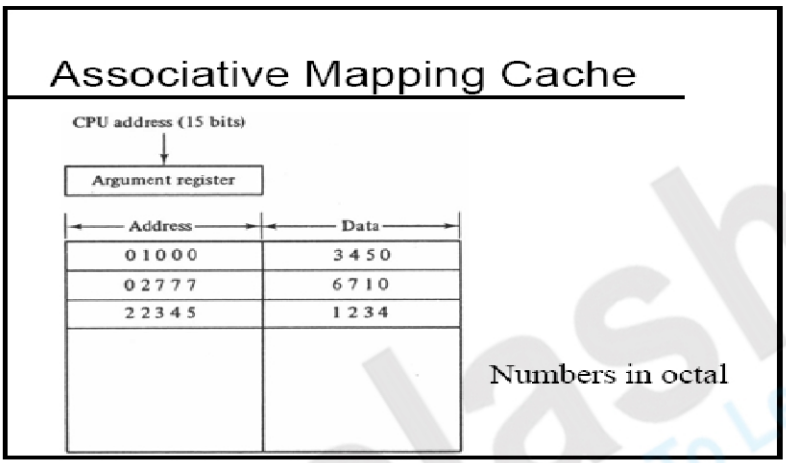
2. **Mapping Function**

- **Basic characteristics of cache is its fast access time, so time should not be wasted while searching for words in cache.**
- **There are fewer cache lines than main memory blocks, an algorithm is needed for mapping main memory blocks into cache lines. i.e. a means is needed to determine which main memory blocks occupies a cache line.**
- **The transformation of data from main memory to cache is referred to as a *mapping process*.**
- **Mapping function determines how the cache is organized.**
- **There are 3 types of mappings:**

    a) Associative mapping

- **- Any block location in Cache can store any block in memory**
- **-> Most flexible**
- **- Mapping Table is implemented in an associative memory**
- **-> Fast, very Expensive**

- **- Mapping Table**
- **Stores both address and the content of the memory word**

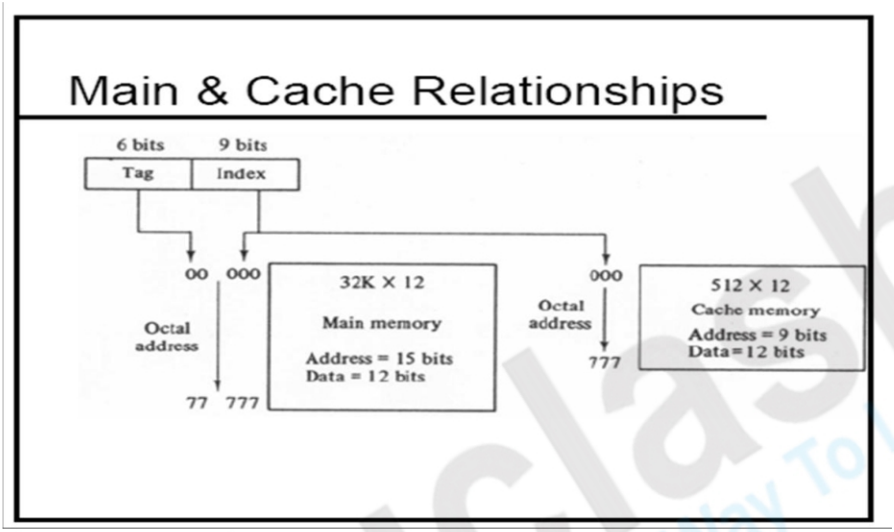


a) Direct mapping

- **- Each memory block has only one place to load in Cache**
- **- Mapping Table is made of RAM instead of CAM**
- **- n-bit memory address consists of 2 parts; k bits of Index field and**
- **n-k bits of Tag field**
- **- n-bit addresses are used to access main memory**
- **and k-bit Index is used to access the Cache**

Operation:

- CPU generates a memory request with (TAG;INDEX)



Addressing relationships b/w main and cache memories

- Access Cache using INDEX ; (tag; data)

Compare TAG and tag

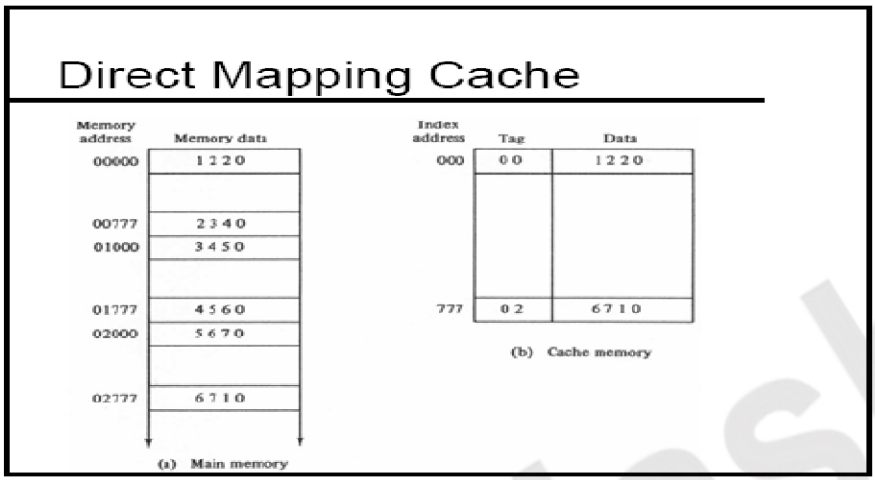
- If matches -> Hit

Provide Cache[INDEX](data) to CPU

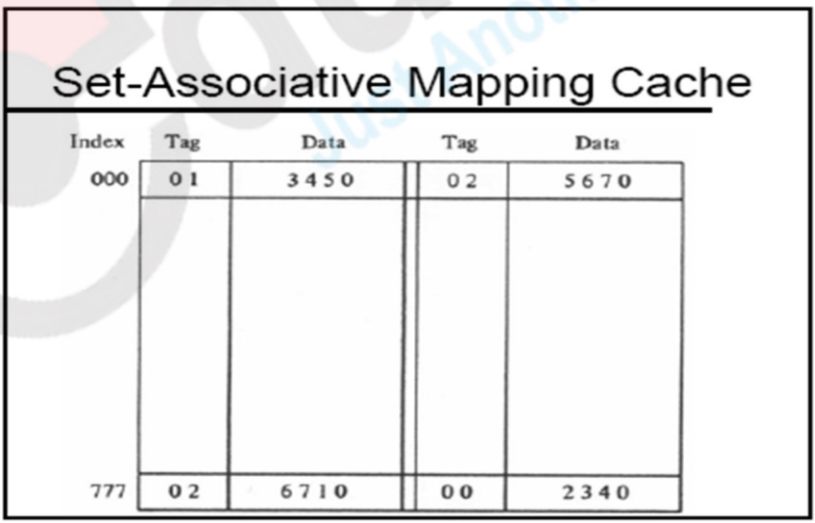
- If not match -> Miss

copy required data from main memory to cache

CPU <- Cache[INDEX](data)



a) Set-associative mapping



Two way set associative mapping

3.

Content is in book pg 147

Replacement Algorithm:pg 154

Write Policy :pg 154

Block Size:

Number of Caches: pg 156

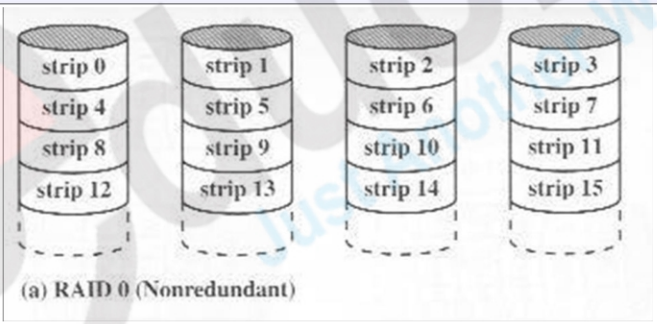
ExternaL Memory:

Magnatic disk: Pg 203

Optical memory: Pg 221

Raid: Pg 212 refer

RAID



RAID Level 0 requires a minimum of 2 drives to implement

Characteristics & advantages:

- RAID 0 implements a striped disk array, the data is broken down into blocks and each block is written to a separate disk drive

- I/O performance is greatly improved by spreading the I/O load across many channels and drives
- Best performance is achieved when data is striped across multiple controllers with only one drive per controller
- No parity calculation overhead is involved
- Very simple design
- Easy to implement

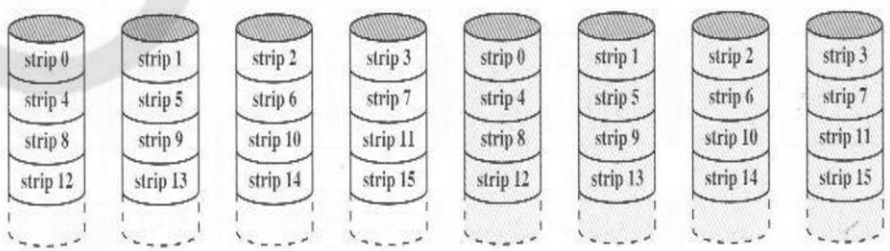
Disadvantages

- Not a "True" RAID because it is NOT fault-tolerant
  The failure of just one drive will result in all data in an array being lost
- Should never be used in mission critical environments

Applications

- Video Production and Editing
- Image Editing
- Pre-Press Applications
- Any application requiring high bandwidth

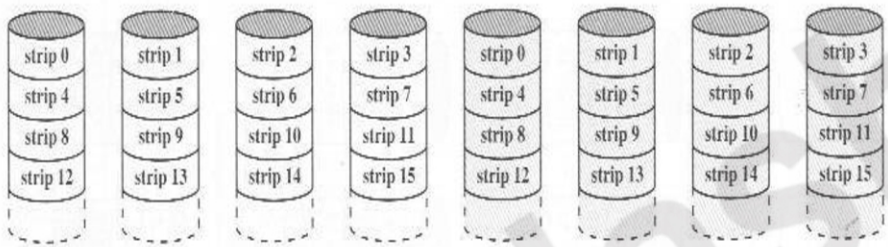
Raid 1:



(b) RAID 1 (Mirrored)

For Highest performance, the controller must be able to perform two concurrent separate Reads per mirrored pair or two duplicate Writes per mirrored pair.

RAID Level 1 requires a minimum of 2 drives to implement



(b) RAID 1 (Mirrored)

RAID Level 1 requires a minimum of 2 drives to implement

Characteristics & advantages:

- A read operation can be carried out from any one of the two disks that contains the requested data, whichever one requires minimum seek time + rotational delay.
- A write request requires both corresponding strips to be updated, but this can be done in parallel.
- Recovery from a failure is simple. When a drive fails, the data may still be accessed from second drive.

Disadvantages

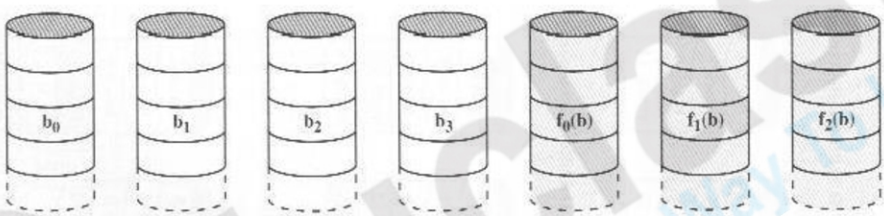
- Cost, as it requires twice the disk space
- In transaction oriented environment, performance of RAID 1 is approx. double that of RAID 0 in case of bulk of read requests.

But in case of write requests there may be no performance gain over RAID 0.

Applications

- Accounting
- Payroll
- Financial
- Any application requiring very high availability

Raid 2:



(c) RAID 2 (Redundancy through Hamming code)

Each bit of data word is written to a data disk drive (4 in this example: 0 to 3). Each data word has its Hamming Code ECC word recorded on the ECC disks. On Read, the ECC code verifies correct data or corrects single disk errors.

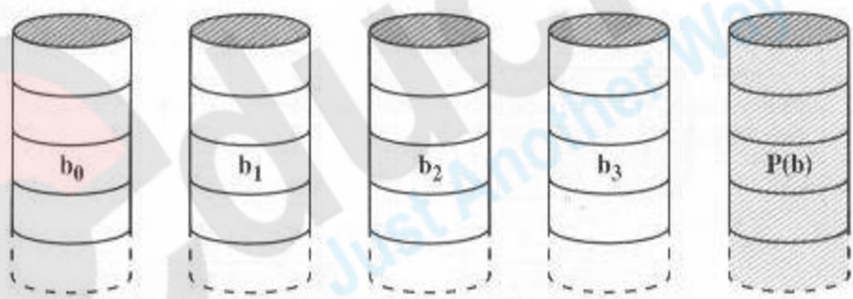
Characteristics & advantages:

- "On the fly" data error correction
- Extremely high data transfer rates possible
- The higher the data transfer rate required, the better the ratio of data disks to ECC disks
- Relatively simple controller design compared to RAID levels 3,4 & 5

Disadvantages

- Very high ratio of ECC disks to data disks with smaller word sizes – inefficient
- Entry level cost very high - requires very high transfer rate requirement to justify
- Transaction rate is equal to that of a single disk at best (with spindle synchronization)
- No commercial implementations exist / not commercially viable

Raid 3:



(d) RAID 3 (Bit-interleaved parity)

The data block is subdivided ("striped") and written on the data disks. Stripe parity is generated on Writes, recorded on the parity disk and checked on Reads.

RAID Level 3 requires a minimum of 3 drives to implement

Characteristics & advantages:

- Very high Read data transfer rate

- Very high Write data transfer rate
- Disk failure has an insignificant impact on throughput
- Low ratio of ECC (Parity) disks to data disks means high efficiency

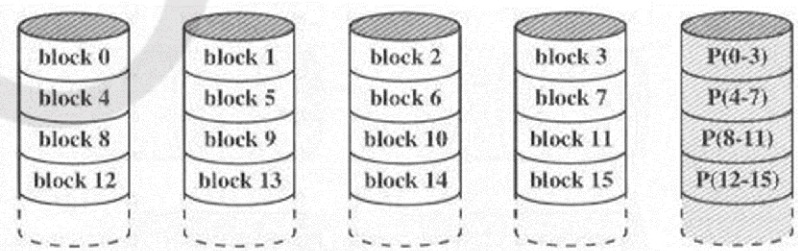
Disadvantages

- Transaction rate equal to that of a single disk drive at best (if spindles are synchronized)
- Controller design is fairly complex
- Very difficult and resource intensive to do as a "software" RAID

Applications

- Video Production and live streaming
- Image Editing
- Video Editing
- Prepress Applications
- Any application requiring high throughput

Raid 4:



(e) RAID 4 (Block-level parity)

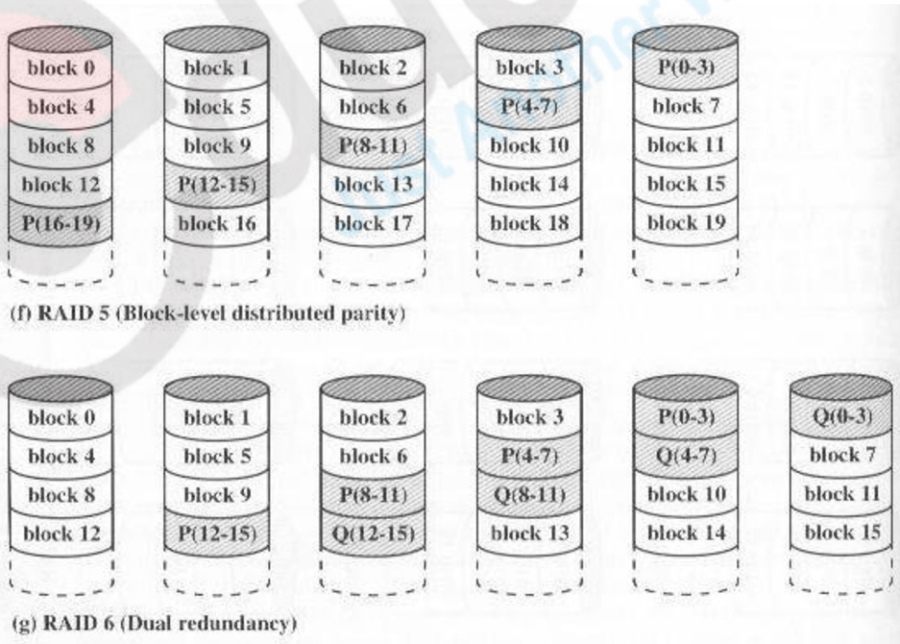
Characteristics & advantages

- Very high Read data transaction rate
- Low ratio of ECC (Parity) disks to data disks means high efficiency
- High aggregate Read transfer rate

Disadvantages:

- Quite complex controller design
- Worst Write transaction rate and Write aggregate transfer rate
- Difficult and inefficient data rebuild in the event of disk failure
- Block Read transfer rate equal to that of a single disk

Raid 5 and 6:



Characteristics & advantages

- Highest Read data transaction rate
- Medium Write data transaction rate
- Low ratio of ECC (Parity) disks to data disks means high efficiency
- Good aggregate transfer rate

Disadvantages

- Disk failure has a medium impact on throughput
- Most complex controller design
- Difficult to rebuild in the event of a disk failure (as compared to RAID level 1)
- Individual block data transfer rate same as single disk

Application

- File and Application servers
- Database servers
- Web, E-mail, and News servers
- Intranet servers
- Most versatile RAID level

Educlash Just Another Way To Learn