Western

UNIVERSITY · CANADA

Chapter 11 – Mass-Storage Structure

Spring 2023

Overview

- Overview
- HDDs and NVMs
- HDD and NVM Scheduling
- Error Detection and Correction
- Storage Device Management
- Swap-Space Management
- Storage Attachment
- RAID Structure

Overview

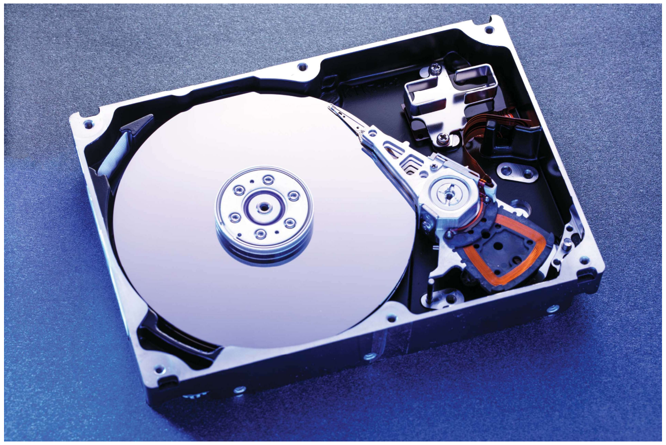
- Modern computers must
  - Store large amounts of data
  - Store this data beyond the lifetime of any processes (even if powered off)
  - Multiple processes must be able to use the data concurrently

Overview

- Secondary storage for modern computers uses either
  - **hard disk drives (HDD)** or
  - **nonvolatile memory (NVM)**
    - Most common version of NVM is **solid state drives (SSD)** and **USB drives**
- Some systems can use part of the volatile memory as if it were secondary storage. This is known as a **RAM drive**
  - This is typically used to hold temporary data, data in transit between disk and memory, or swap space
- Magnetic tape is used for inexpensive, long-lasting backup and archive

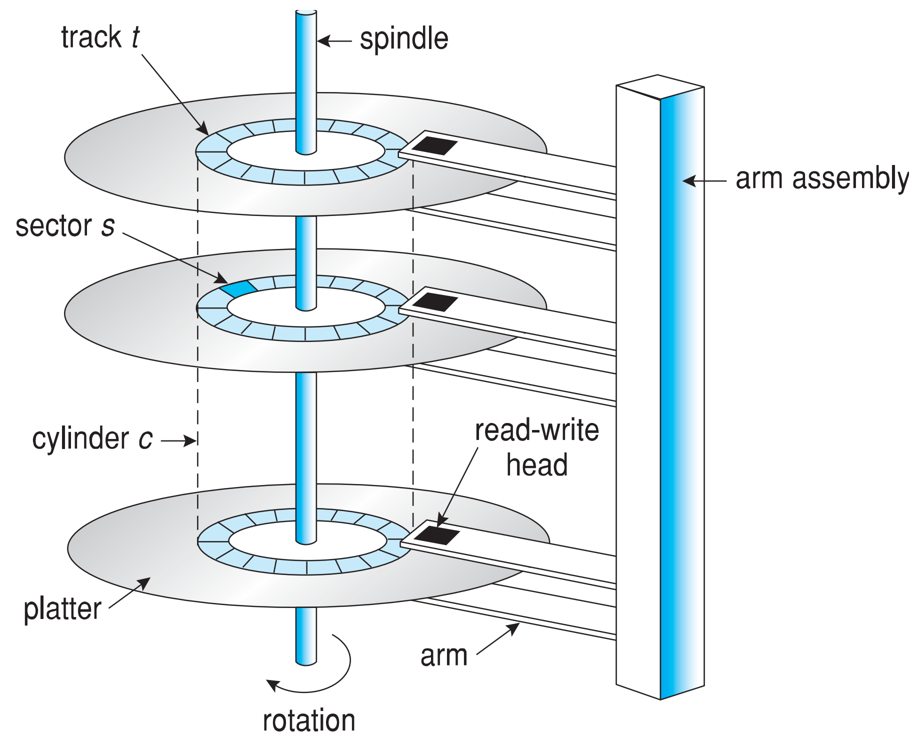
HDD and NVM

- HDDs spin platters of magnetically-coated material under moving read-write heads
  - The read-write head is separated from the platter using a very thin cushion of air or gas (e.g. helium)
  - Each platter is divided into **cylinders** divided into **tracks** divided into **sectors**
- Drives rotate at 60 to 250 times per second (rotations per minute – RPM)
  - 5,400 or 7,200 or 10,000 or 15,000 RPMs are common
- CDs, DVDs, and Blu-Ray discs are removable



HDD and NVM

- Platters, Cylinders, Tracks, and Sectors
  - Sectors were typically 512 bytes each until about 2010
  - Sectors are now 4KB each
  - Each track used to hold the same number of sectors. Data on outer tracks were less dense with lots of inaccessible space
  - Variable number sectors allow more sectors on outer tracks. This increases data density and therefore capacity



HDD and NVM

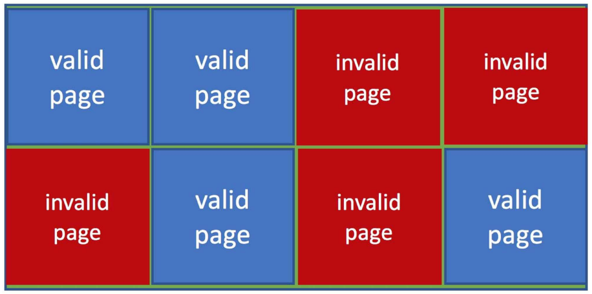
- Performance
  - **Transfer rate** is rate at which data flow between drive and computer
  - **Positioning time** (random-access time) affects the actual transfer rate
    - **Seek time** – is time to move disk arm to desired cylinder (this is usually the longest wait time)
    - **Rotation latency** – time for desired sector to rotate under the disk head
  - **Head crash** can occur if the disk head contacts the disk surface
    - Physical damage cannot be repaired and is usually unrecoverable unless using backups or RAID protection

HDD and NVM

- NVM
  - No moving parts and therefore no seek time or rotational latency
  - More reliable but may have a shorter life-span than HDDs
  - More expensive and less capacity than HDDs
  - Some busses may be slower than connecting directly to the system bus (e.g. USB bus)
  - Better suited for mobile devices than HDDs (laptops, phones, etc.)

HDD and NVM

- Data is read and written in "pages" (similar to the sector idea)
- NVM is the most efficient storage medium for reads
- Writing pages requires the existing data to be erased first
  - Erasing and writing new data slowly degrades the memory cell
  - Algorithms keep track of all pages to find free pages and evenly distribute writes



HDD and SSD scheduling

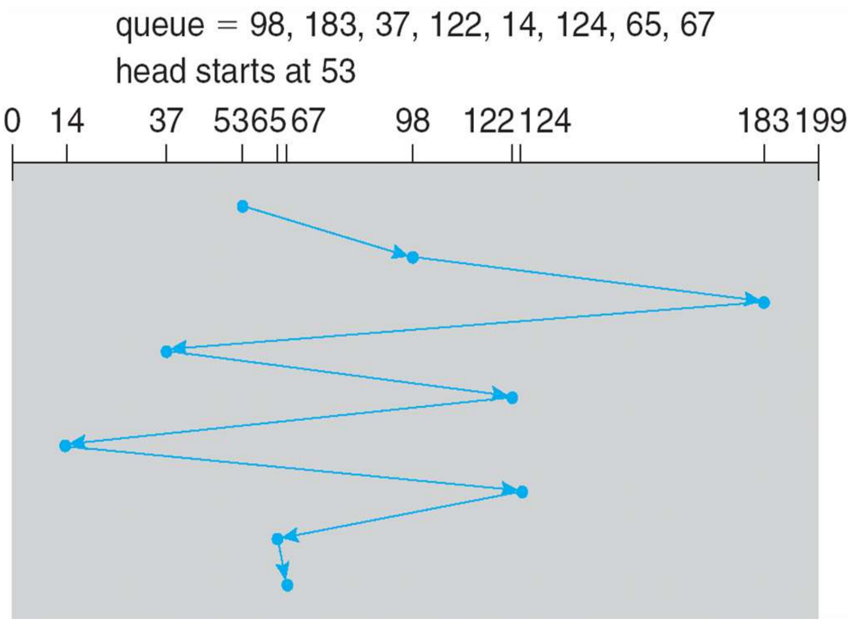
- Disk drives are addressed as one-dimensional arrays of logical blocks. Each block is mapped to a sector.
  - HDDs
    - Sector 0 is the first sector on the first track on the first outermost cylinder
    - Then mapping proceeds through each track in the same cylinder
    - Continue from the outermost cylinder to the innermost
  - NVM
    - Simply address through each chip, block, and page

HDD and SSD scheduling

- The operating system should leverage disks efficiently. Knowing the underlying structure helps decrease access time and increase disk bandwidth
- There are many sources of disk I/O (OS, system processes, user processes)
    - Processes waiting for I/O move to a waiting queue
    - Disk requests are held in a queue and sent to the disk
        - Modern disks have their own queue instead of just the OS
- Each request contains a number of factors (e.g. read or write, disk address, memory address, number of sectors, etc.)
- Managing the queue of requests intelligently can make disk I/O more efficient

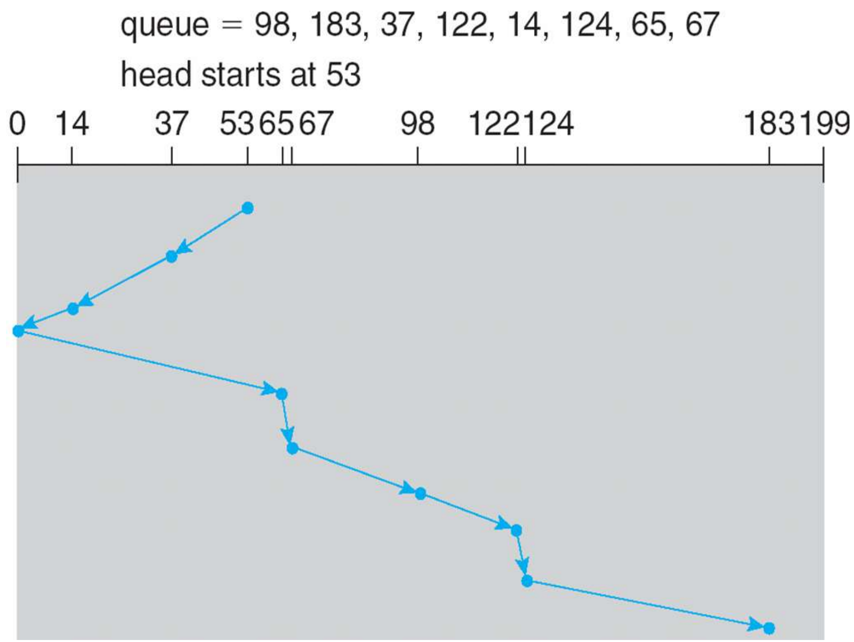
HDD and SSD scheduling

- First come first served scheduling



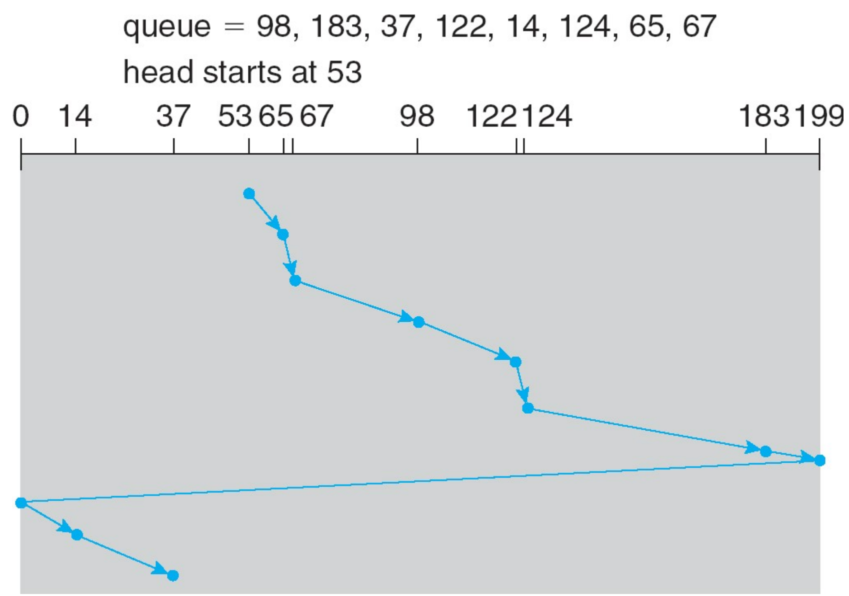
HDD and SSD scheduling

- SCAN scheduling (elevator scheduling)



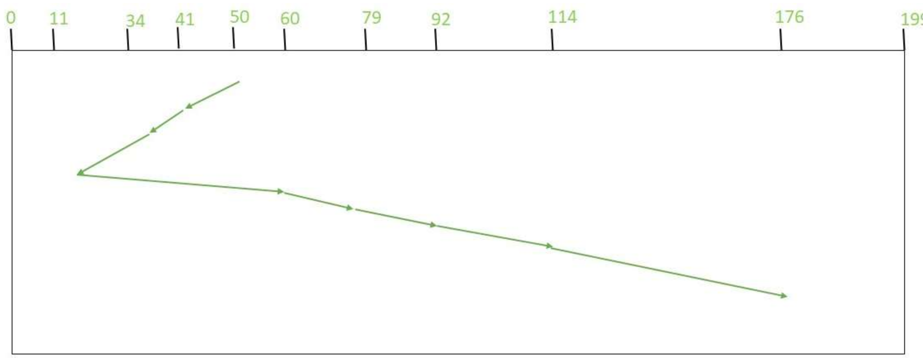
HDD and SSD scheduling

- C-SCAN scheduling



HDD and SSD scheduling

- Shortest seek time first – Select the address closest to the head's current position
  - Queue = 176, 79, 34, 60, 92, 11, 41, 114
  - Head starts at 50
  - Starvation can occur



HDD and SSD scheduling

- Other scheduling algorithms
  - Random
  - Last in first out
  - Priority
    - Treat reads and writes with different priority
  - Etc.
- Scheduling selection depends on the anticipated workload
- Most operating systems use a combination of algorithms

HDD and SSD scheduling

- NVM scheduling
  - There are no disk heads or rotational latencies
  - Random is the best approach
  - Optimizing for reads is ideal because
    - Reads are much faster than writes
    - Read times are uniform, write times are variable

Error Detection and Correction

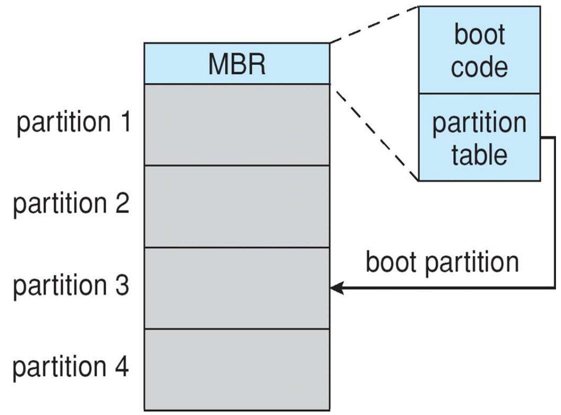
- Detection and correction is fundamental to memory, network, storage and others
- Bits can change spontaneously while stored on disk
- Reserve one bit in a byte as a **parity bit**
  - Set to 1 if the number of bits set to 1 is even
  - Set to 0 if the number of bits set to 1 is odd
- If an error occurs, the parity does not match so the user can be informed
- Error-correction code (ECC) can be used to detect and sometimes correct multiple bytes
  - A soft error can be automatically corrected. Too many errors is a hard error

Storage Device Management

- Low-level formatting (physical formatting) – Divide a disk into sectors
  - Each sector holds both data and metadata (eg. Size, ECC, etc.)
- The operating system will then record its own data structures on the disk
  - **Partition** the disk into one or more groups of cylinders, each treated as a logical disk (eg. C drive, D drive, /boot, /usr, /home, etc.)
  - Each partition is **logically formatted** to a file system type (eg. NTFS, ext4, zfs, etc.)
    - File-systems keep data and their metadata in similar sectors to reduce seek time
  - Partitions contain metadata including whether or not it is bootable

Storage Device Management

- A computer has a bootstrap program built into the firmware
- The last thing this program does is read the first block in the first partition on secondary storage to find the OS bootloader. This is the **master boot record**
- The boot loader passes control to a bootstrap program on the first sector on a bootable partition. This program loads the rest of the OS



Swap-Space Management

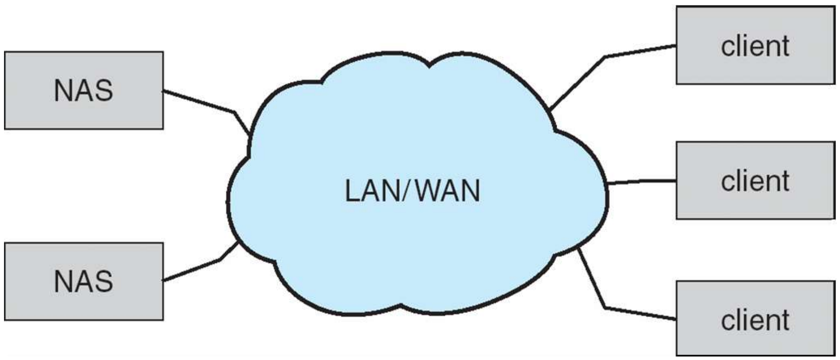
- We know that entire processes are no longer swapped out to disk
- However, individual pages may be swapped to disk
  - Swapping and paging are interchangeable terms
- Swapping may be sent to a (raw) partition or a file(s) on a partition
  - If a raw partition is used, algorithms that optimize for speed over storage efficiency might be used
  - Files can be added and removed as desired but are subject to the underlying file-system's storage structure

Storage Attachment

- Host-attached storage
  - Disks are attached to the computer over a set of wires known as the **I/O bus**
  - Data is transferred on either end of the I/O bus by controllers
    - **Host controller** is the controller on the motherboard
    - **Data controller** is built into the disk
  - Serial Advanced Technology Attachment (SATA) is the most common I/O bus type
  - Other common technologies include USB FireWire and Thunderbolt
  - High-end workstations and storage arrays may use optical based Fiber Channel

Storage Attachment

- Network-attached storage (NAS)
  - Storage made available across a network (Typically the Local Area Network)
  - Common protocols include NFS and CIFS (UDP/TCP layer)
  - iSCSI presents the SCSI protocol remotely (IP layer)

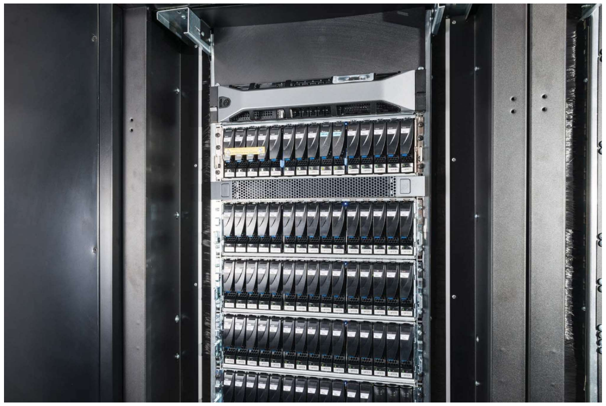


Storage Attachment

- Cloud storage
    - Similar to NAS except provided over the Internet or a WAN rather than LAN
    - Due to higher latency and packet loss across the WAN, commonly presented via APIs at the application layer
        - Common examples include Amazon S3, Dropbox, Microsoft OneDrive, Apple iCloud

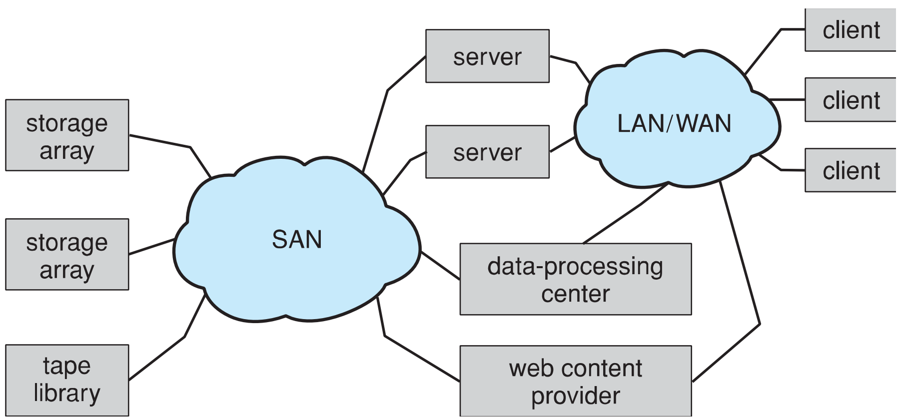
Storage Attachment

- Storage Area Network (SAN) and storage arrays
  - Storage arrays
    - custom built storage servers that manage multiple storage types, networking, permissions, etc.
    - Disks can be added and replaced transparently
    - Redundancy built in
    - Added features such as snapshots, clones, thin provisioning, deduplication, etc.



Storage Attachment

- Storage Area Network (SAN) and storage arrays
    - SANs are a private network connecting multiple hosts to multiple storage units
    - Typically used over short distances



RAID Structure

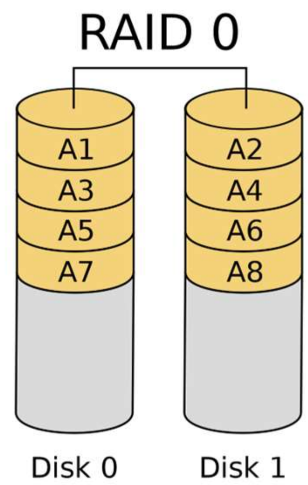
- All disks are prone to error and failure
- Losing data could be catastrophic
- Copies of the data should be made for redundancy
- **Redundant Array of Inexpensive Disks (RAID)** – Use multiple disks
    - To provide **reliability and redundancy**
    - Spreading data across multiple disks can also create parallelism which can improve **speed**
- (Disks these days are high speed and highly reliable and therefore no longer "inexpensive", so some use "independent" instead for the "I")

RAID Structure

- RAID logic is implemented at
  - Hardware level (a chip dedicated to managing RAID)
    - Expensive. Best suited for servers and super computers
  - Software level (ie. By the operating system)
    - Inexpensive. The boot process takes some care. Best suited for desktops
  - Firmware level (Low-level support in the BIOS with driver support for the OS)
    - Depends on OS and driver support.
    - Also known as "Fake RAID" or "Hybrid"

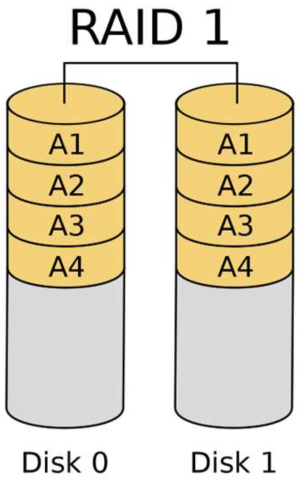
RAID Structure

- There are (generally) 6 different RAID types
  - RAID 0 – Spread data out across two disks instead of one. This is known as **striping**. Double the read and write speed but no redundancy.
  - Suitable for systems where speed is important, but reliability is not (E.g. Gaming)



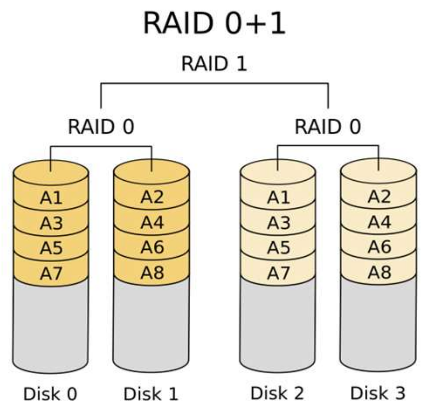
RAID Structure

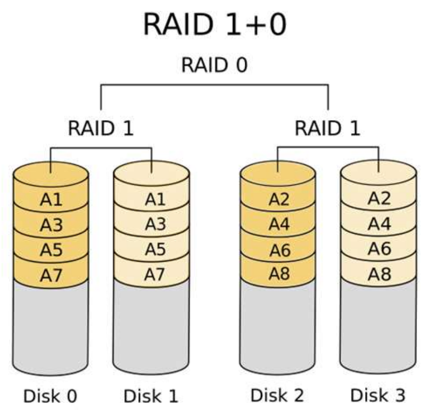
- There are (generally) 6 different RAID types
  - RAID 1 – Write all data to two disks (**mirroring**). One disk perfectly mirrors the other. Reads may be improved but write performance suffers. Disk errors or failures can be tolerated and corrected.
  - Suitable for systems where reliability is important (eg. Mission-critical)



RAID Structure

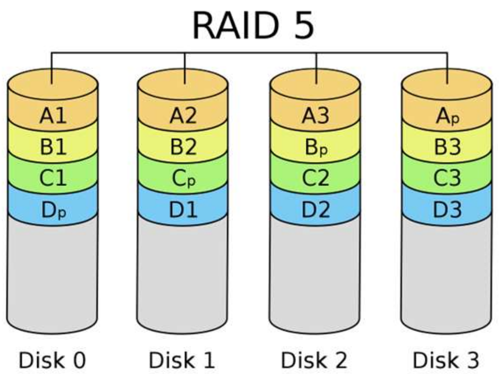
- There are (generally) 6 different RAID types
  - Raid 0+1 or RAID 01 – Mirror two striped disks to two other disks
  - Raid 1+0 or RAID 10 – Stripe two mirrored disks to two other disks
  - Brings the best of both striping and mirroring
  - Requires 4 disks at minimum





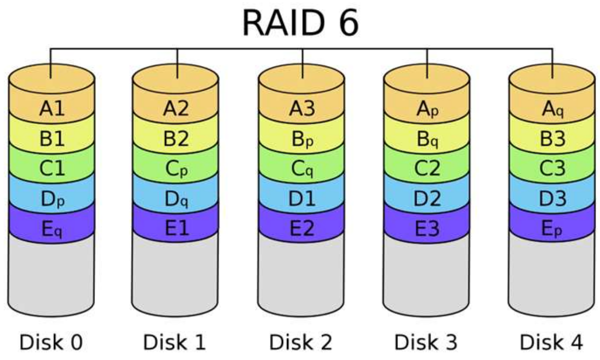
RAID Structure

- There are (generally) 6 different RAID types
  - RAID 5 – Stripe data across disks. Also stripe 1 parity block across disks.
  - The loss of one disk can be tolerated. The missing block can be rebuilt from the parity block
  - RAID 5 and RAID 10 are the most common implementations. They provide efficient striping and mirroring
    - In the event of disk failure in RAID 5, all disks are involved in a rebuild. In RAID 10, not all disks need to be involved.



RAID Structure

- There are (generally) 6 different RAID types
  - RAID 6 – Stripe data across disks. Also stripe 2 parity blocks across disks.
  - The loss of two disks can be tolerated. The missing blocks can be rebuilt from the parity blocks



RAID Structure

- There are (generally) 6 different RAID types
  - There are many other variants such as RAID 2, RAID 3, RAID 4, RAID 50, RAID 60, RAID 100, JBOD, etc.
- RAID cannot protect you from user error. Backups are still needed
- RAID disks need to be the same size (or purposely hold as much as the smallest disk). It is difficult to add or shrink space
- Multiple disks are more expensive
- **Hot spares** can be used to provide immediate rebuilds in the case of disk failure
  - Rebuilds take time and temporarily affect performance

Western

UNIVERSITY · CANADA