CS343 - Operating Systems

Module-5C

Storage Arrays & RAID Implementation

Dr. John Jose

Assistant Professor

Department of Computer Science & Engineering

Indian Institute of Technology Guwahati, Assam.

http://www.iitg.ac.in/johnjose/

Disk Storage Systems Management

- Disk Organization & Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- Storage Arrays
- RAID Structure
- Stable-Storage Implementation

Objectives

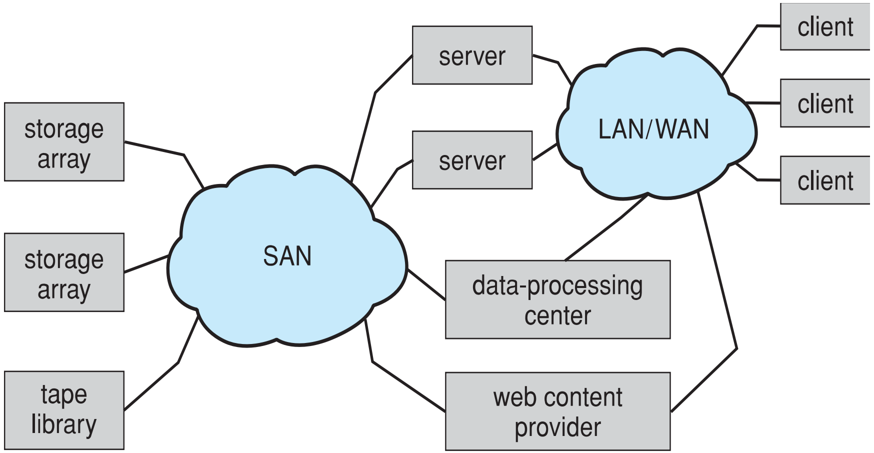
- To describe the physical structure of secondary storage devices and its effects on the uses of the devices
- To explain the performance characteristics of mass-storage devices
- To evaluate disk scheduling algorithms
- To discuss operating-system services provided for mass storage, including RAID

Storage Array

- Attach multiple disks - arrays of disks
- Storage array has controller that provides features to attached hosts
  - A few to thousands of disks
  - Ports to connect hosts to array
  - Memory, controlling software
  - Support RAID, hot spares, hot swap
  - Shared storage → more efficiency

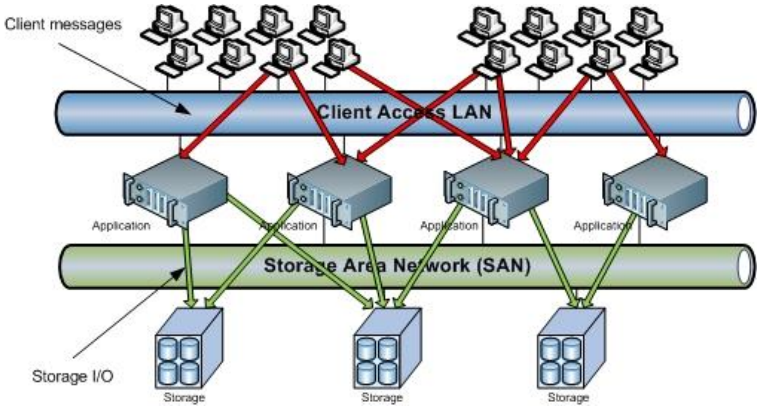
Storage Area Network

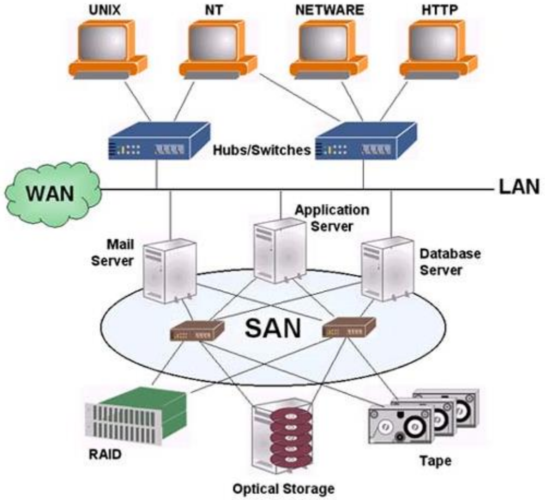
- Common in large storage environments
- Multiple hosts attached to multiple storage arrays - flexible



Storage Area Network

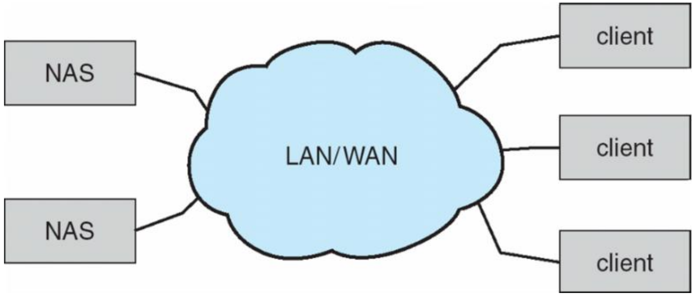
- SAN is one or more storage arrays
  - Connected to one or more Fibre Channel switches
- Hosts also attach to the switches
- Easy to add or remove storage, add new host and allocate it storage



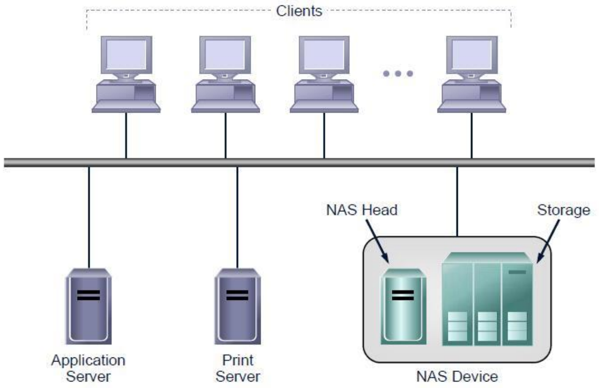


Network-Attached Storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
  - Remotely attaching to file systems
- Implemented via remote procedure calls (RPCs) between host and storage over typically standard computer network protocols.



NAS is shared storage on a network infrastructure



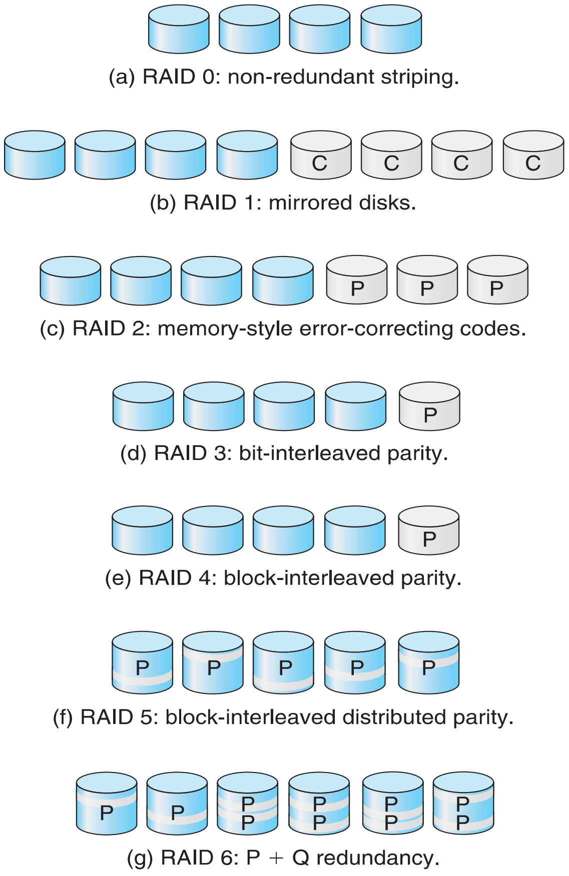
RAID Structure

- RAID – redundant array of inexpensive disks
- Multiple disk drives provides reliability via **redundancy**
- Use of multiple disks working cooperatively
- Increases the **mean time to failure**
- Frequently combined with **NVRAM** to improve write performance
- Disk **striping** (**RAID 0**) uses a group of disks as one storage unit
- RAID is arranged into six different levels

RAID Structure

- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data
  - **Mirroring** or **shadowing** (**RAID 1**) keeps duplicate of each disk
  - Striped mirrors (**RAID 1+0**) or mirrored stripes (**RAID 0+1**) provides high performance and high reliability
  - **Block interleaved parity** (**RAID 4, 5, 6**) uses much less redundancy
- RAID within a storage array can still fail if the array fails, so automatic **replication** of the data between arrays is common
- Frequently, a small number of **hot-spare** disks are left unallocated, automatically replacing a failed disk and having data rebuilt onto them

RAID Levels



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

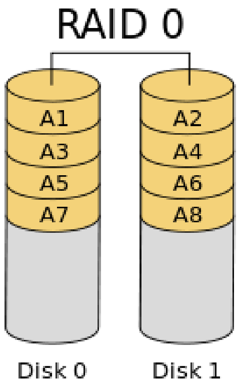
(c) RAID 2: memory-style error-correcting codes.

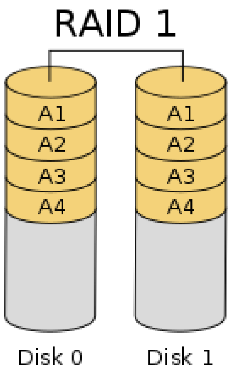
(d) RAID 3: bit-interleaved parity.

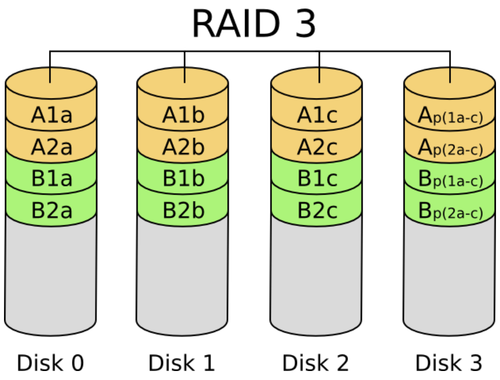
(e) RAID 4: block-interleaved parity.

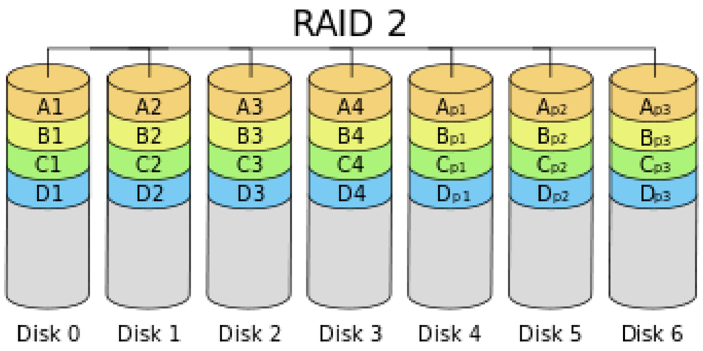
(f) RAID 5: block-interleaved distributed parity.

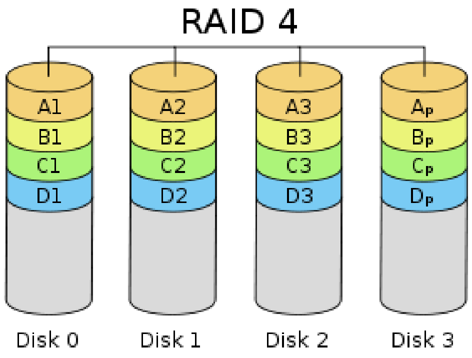
(g) RAID 6: P + Q redundancy.



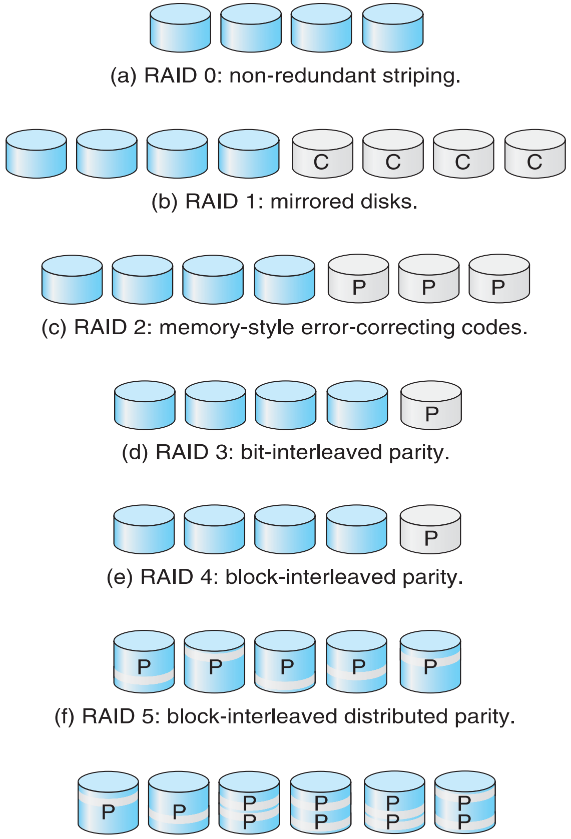








RAID Levels


(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

(c) RAID 2: memory-style error-correcting codes.

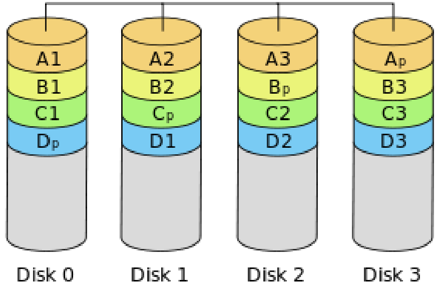
(d) RAID 3: bit-interleaved parity.

(e) RAID 4: block-interleaved parity.

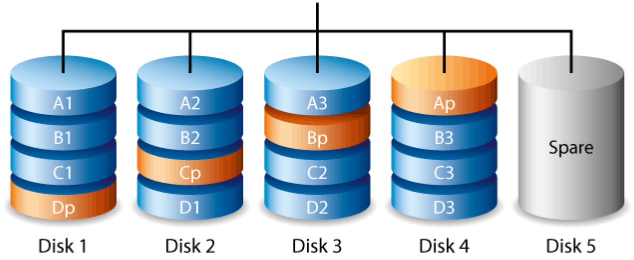
(f) RAID 5: block-interleaved distributed parity.

(g) RAID 6: P + Q redundancy.

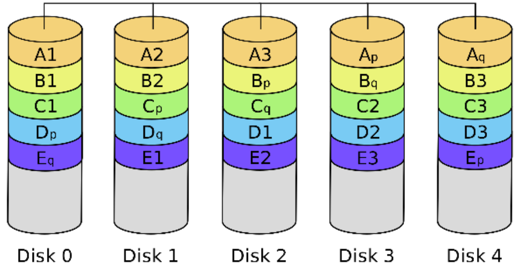
RAID 5



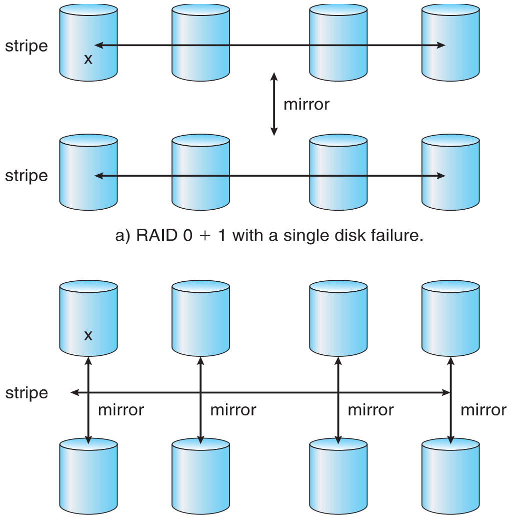
RAID 5+Spare



RAID 6

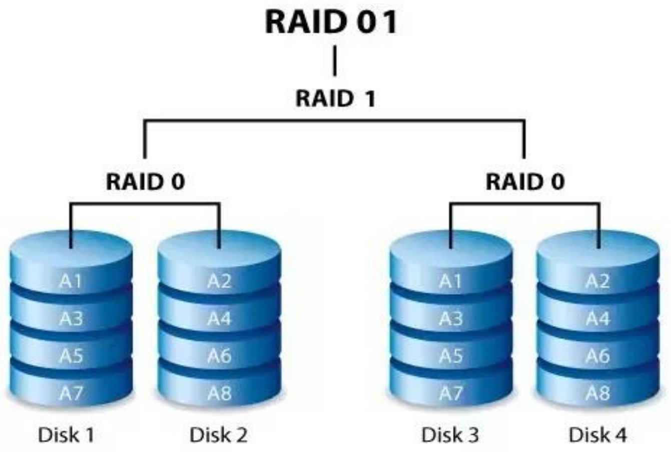


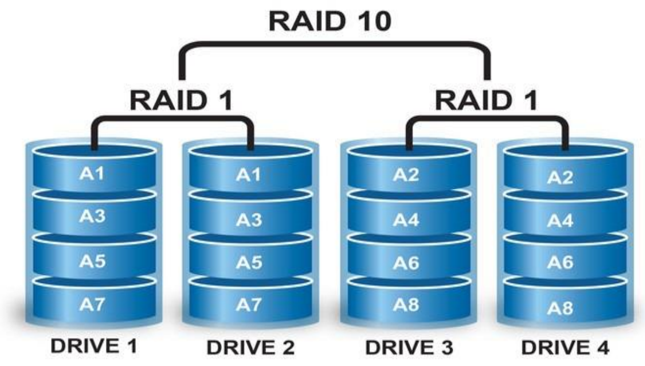
RAID Levels



a) RAID 0 + 1 with a single disk failure.

b) RAID 1 + 0 with a single disk failure.





Stable-Storage Implementation

- Write-ahead log scheme requires stable storage
- Stable storage means data is never lost (due to failure, etc)
- To implement stable storage:
    - Replicate information on more than one nonvolatile storage media with independent failure modes
    - Update information in a controlled manner to ensure that we can recover the stable data after any failure during data transfer or recovery

Stable-Storage Implementation

- Disk write has 1 of 3 outcomes
  - **Successful completion -** The data were written correctly on disk
  - **Partial failure -** A failure occurred in the midst of transfer, so only some of the sectors were written with the new data, and the sector being written during the failure may have been corrupted
  - **Total failure -** The failure occurred before the disk write started, so the previous data values on the disk remain intact
- If failure occurs during block write, recovery procedure restores block to consistent state
  - System maintains 2 physical blocks per logical block
    - Write to 1st physical, When successful, write to 2nd physical
    - Declare complete only after second write completes successfully

Thank you

johnjose@iitg.ac.in

http://www.iitg.ac.in/johnjose/

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी • Indian Institute of Technology Guwahati •