SNS COLLEGE OF TECHNOLOGY

(Autonomous)

COIMBATORE-35

RAID ,disk attachment ,stable storage

RAID

RAID, or “Redundant Arrays of Independent Disks” is a technique which makes use of a combination of multiple disks instead of using a single disk for increased performance, data redundancy or both.

Key evaluation points for a RAID System

Reliability

Availability

Performance

Capacity

RAID

- **RAID** – **R**edundant **A**rray of **I**ndependent **D**isks
  - Multiple disk drives provides **reliability** via **redundancy**.
- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively.
- **Disk striping** uses a group of disks as one storage unit.
- RAID schemes improve performance and / or improve the reliability of the storage system by **storing redundant data.**
  - ***Mirroring*** or ***shadowing*** keeps duplicate copy of each disk.
  - ***Block interleaved parity*** requires much less extra space for redundancy.

Redundancy for Improved Reliability

Why do we care?

- **Systems becoming larger and larger** – now can have hundreds of disks attached, storing terabytes of data.
  - **Recovery time following a failure is much longer** for large systems than for smaller systems
  - Large systems are often **more mission critical** than small system
- Even with highly reliable hardware, as the number of instances of the hardware increases, the **probability of failure increases**
  - E.g., if mean time between failure for a disk drive is one failure every five years, if have 100 disks on the system, probability is that will have a disk failure every three weeks
  - This is not good for your bank (or hospital, or inventory control system, or e-commerce system, or . . .)
  - And many systems now have hundreds of drives attached
- **Use data redundancy to reduce impact** of hardware failures

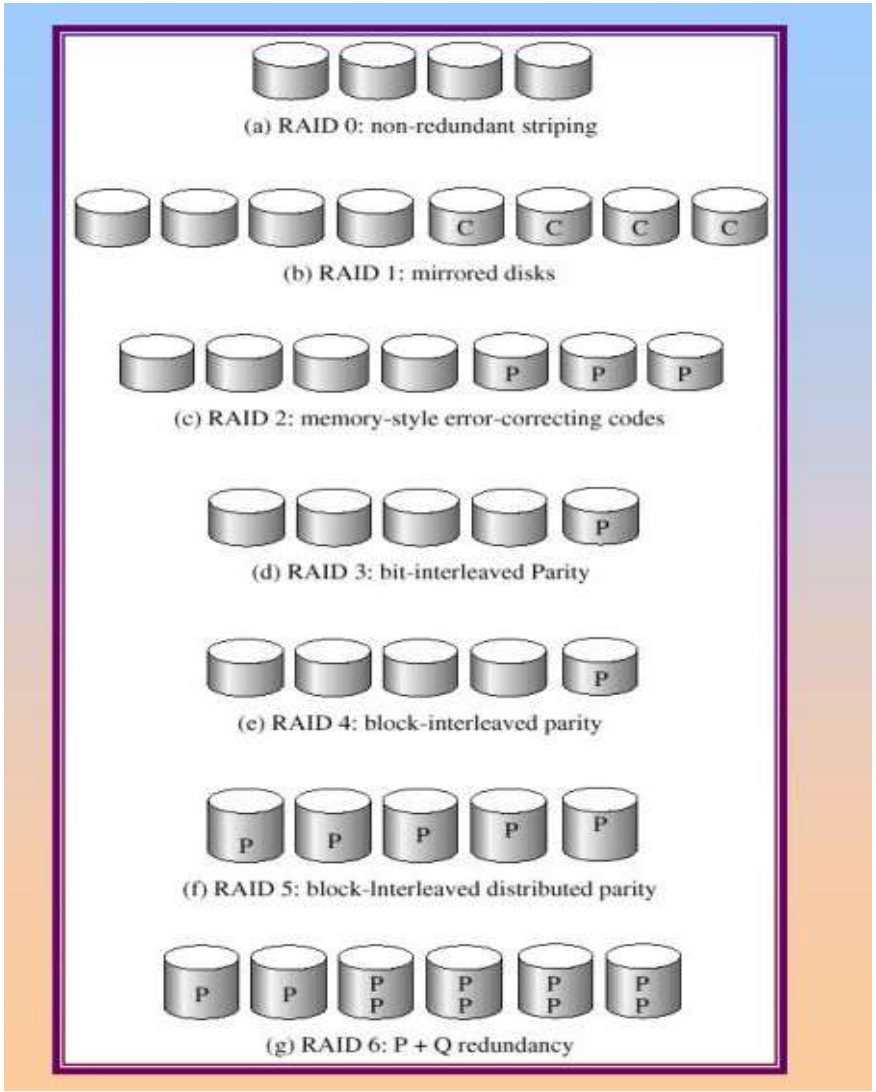
www.snsgroups.com

RAID Comes in Six Levels

- **Level 0** – non-redundant striping
  Data spread across all disks – no redundancy
- **Level 1** – mirroring
  Have duplicate copies of all disks
- Level 2 – memory-style ECC organization
  Store two or more extra bits to detect and correct errors
- Level 3 – bit interleaved parity organization
  Parity on one disk, no striping of data disks
- Level 4 – block interleaved parity organization
  Parity on one disk, data spread across n disks
- **Level 5** – block interleaved distributed parity
  Parity and data spread across all disks – quite common
- Level 6 – P+Q redundancy
  Stores extra parity to protect against multiple failures
- Level 0 + 1
  Spread data across all disks and mirror

Levels 0, 1, & 5 most common, 6 starting to be used
Also work on 5 + 1

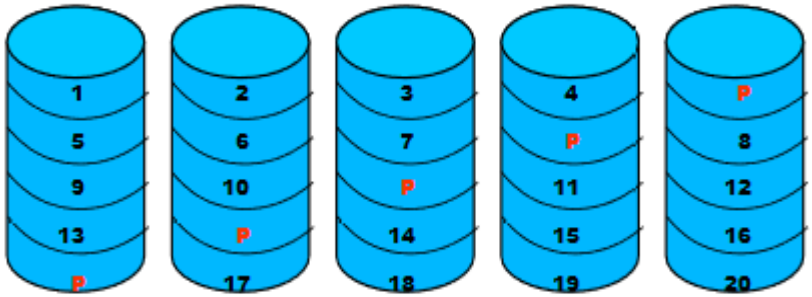
RAID Levels


(a) RAID 0: non-redundant striping
C
(b) RAID 1: mirrored disks
P
(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 Details

Data and parity are **striped and interleaved** across a set of disks

- **Requires one additional disk per parity set**
- **Parity is spread across disks** in the set to avoid "hot spots"



- Parity is the XOR of data in corresponding sectors on other disks of set

If a disk in set fails, **can recreate data** by XORing together the corresponding sectors of remaining disks in the set

- If two disks fail, lose all the data on the entire set

Reading data requires one I/O (just read the data)

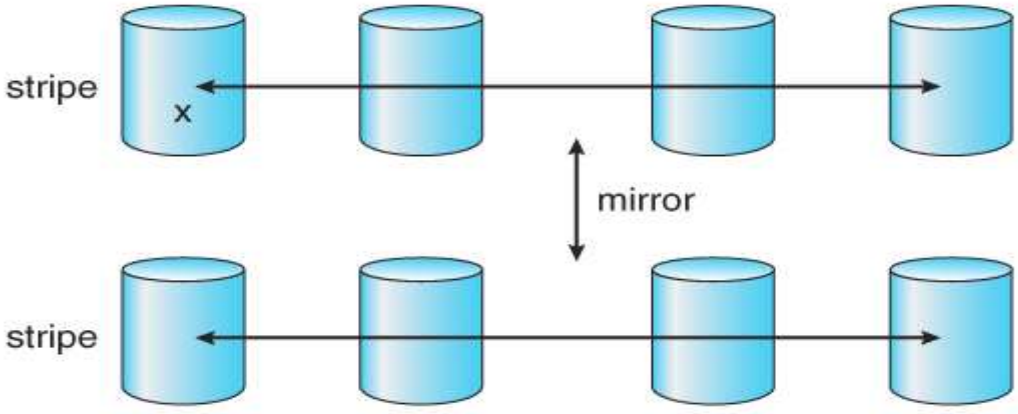
Writing data requires 4 I/Os plus system processing

- Read disk block that will be written, read corresponding parity block, XOR data block and parity block together, XOR result with data to be written to produce new parity value, write new data, write new parity

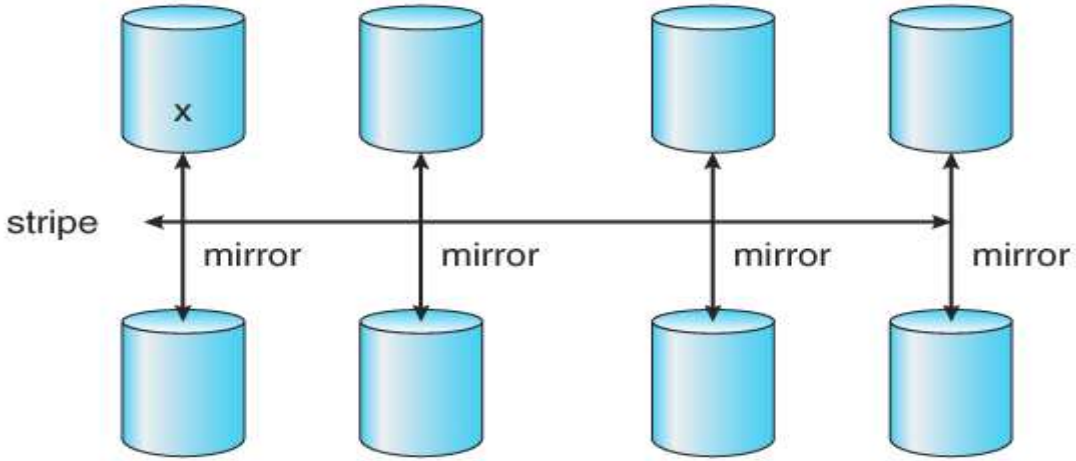
Selecting A Protection Level

- **RAID 0 (striping)** – popular for performance and ease of management
  no real protection from redundancy
- **RAID 1 (mirroring)** – provides best redundancy protection
  - But most expensive in terms of cost for extra disks
- **RAID 0 and RAID 1 combined** – best of both worlds
- Because the hardware required for **RAID 1 is expensive**, **RAID 5 is popular** as an alternative
  - However, **will not tolerate double failure in same RAID set** and often does not perform as well as RAID 1 (since a write requires two reads and two writes – data and parity – although caches help)
  - **Hot spares help avoid double failures**, but must have spare for each RAID set (and replace spare before a double failure occurs)
- **RAID 6 is an improvement**, since will tolerate a double failure without requiring 2x the disk space, but not widely available yet

RAID 0+1 and 1+0



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



RAID 1 + 0 with a single disk failure.

Additional Data Protection Options

Redundancy through **clustering**

Maintain a **redundant copy of data** on another system in cluster

- Use **remote journaling** or similar sort of replication
- Or **remote mirroring**

When primary copy of data fails, **switch to redundant copy** on different system in the cluster

Can also use clustering to **protect against system failure** (not just disk failure)

Disk attachment

Disk may be attached one of two ways

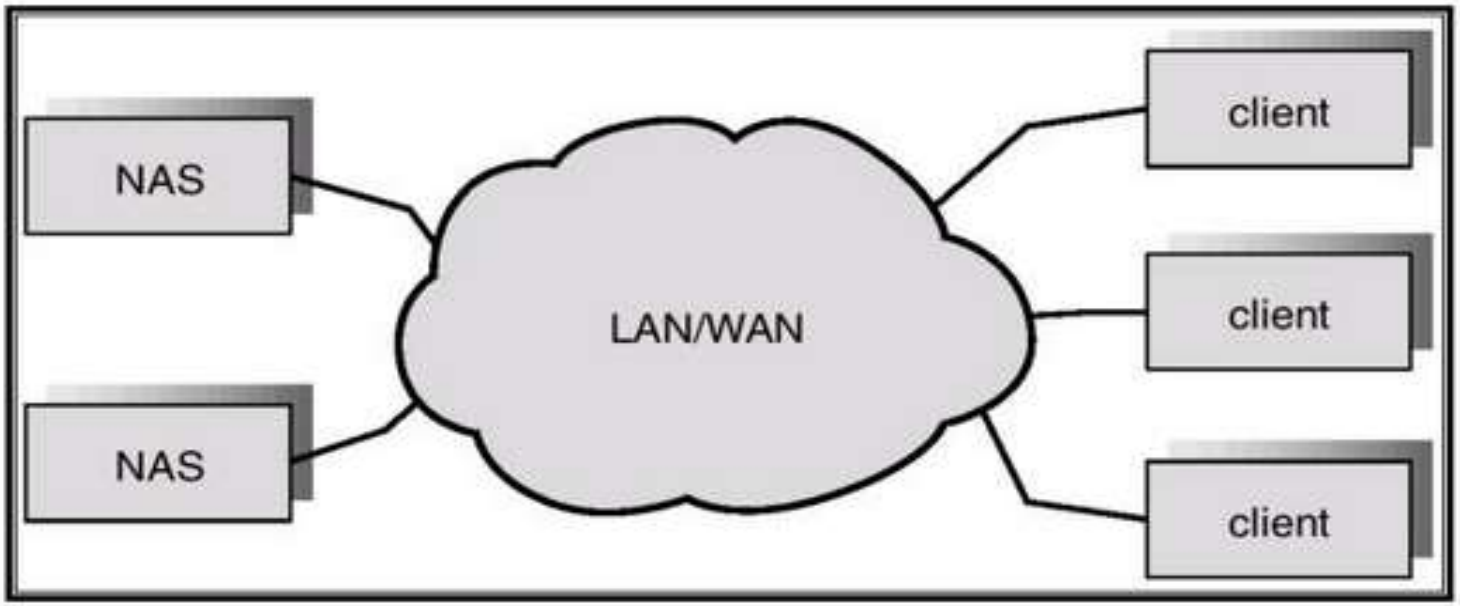
- Host attached via I/O Port
- Network attached via a network connection

Network-attached storage – implemented as a RAID array with software that implements RPC interface for NFS (UNIX machines) or CIFS (Windows machines)

Storage-area network – a private network (using storage protocols rather than network protocols) among servers and storage units, separate from LAN or WAN connecting clients and servers

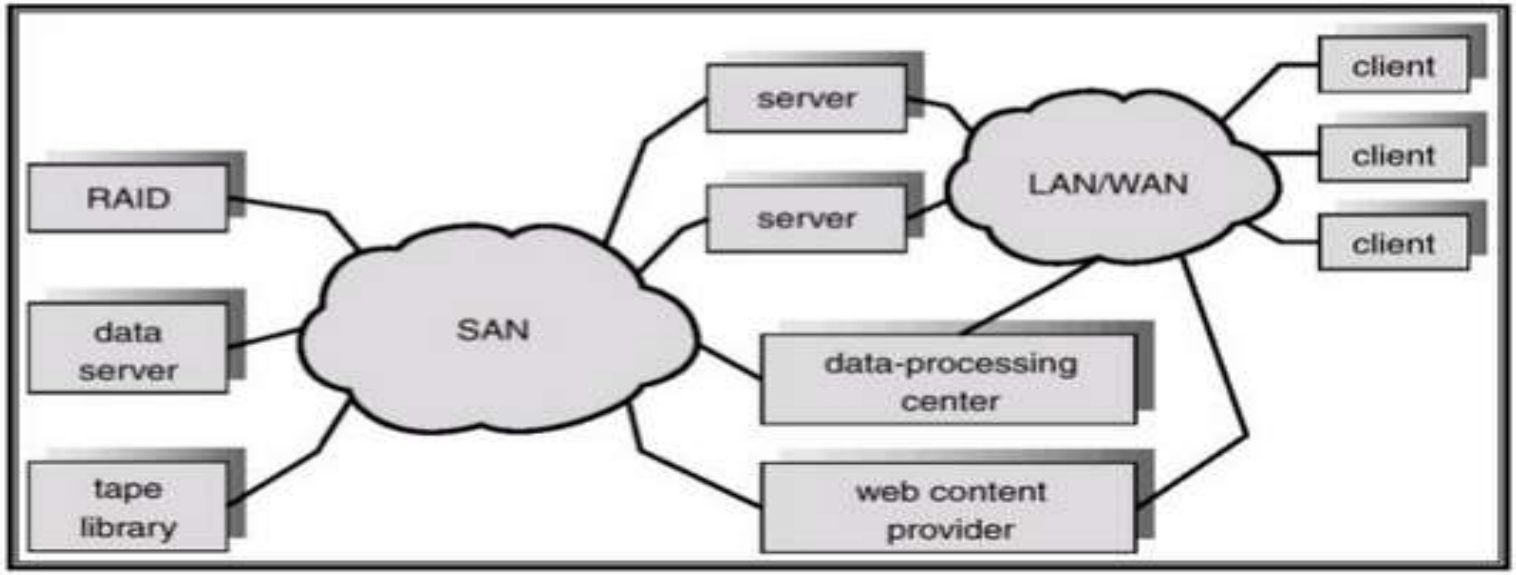
Disk attachment

Network-Attached Storage



Disk attachment

Storage-Area Network



Stable-Storage Implementation

- Information in ***Stable Storage*** is never lost
- Stable storage is **useful** or necessary for a number of scenarios
  - Write-ahead log scheme requires stable storage.
- **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 the stable data can be recovered after any failure during data transfer or recovery.
    - Successful completion
    - Partial failure
    - Total failure
    - Failures can occur anywhere in I/O hardware path

References

1. Silberschatz, Galvin, and Gagne, “Operating System Concepts”, Ninth Edition, Wiley India Pvt Ltd, 2009.
2 . Andrew S. Tanenbaum, “Modern Operating Systems”, Fourth Edition, Pearson Education, 2010.

Summarization