High Performance Computing
Course Notes 2009-2010

High Performance Storage
Storage devices

→ Primary storage:
  - register (1 CPU cycle, a few ns)
  - Cache (10-200 cycles, 0.02-0.5us)
  - Main memory
    - Local main memory (0.2-4us)
    - NUMA (2-10 x latency of local memory)

→ Secondary storage:
  - Magnetic disk (2-20ms)
  - Solid state disk (0.05-0.5ms)
  - Cache in storage controller (0.05-0.5ms)

→ Tertiary storage
  - Removable media: tapes, floppies, CDs (ms-minutes), usb (0.05-0.5ms)
  - Tape library (few seconds – few minutes)
Hard disk vs. solid state drive

a) 2.5-inch hard disk

b) solid state drive
Tape library

- Consisting of
  - Tape
  - Tape drive
  - tape slots
  - Robot
  - Barcode reader
  - terabytes to petabytes data
Disks

- 1-12 platters per disk
- 2 heads per platter
- 2k – 40k tracks per platter
- 50 – 200kB per track
- 512 bytes per sector
Disk failure and metrics

- **Mean Time Between Failures (MTBF):** the average time between failures of a disk

\[
MTBF = \frac{\sum (\text{downtime} - \text{uptime})}{\text{number of failures}}
\]

- **Annual failure rate (AFR):** number of failures per year

\[
AFR = \frac{\text{running-time-per-year}}{MTBF}
\]

\[
AFR_{\text{disks}} = N_{\text{disks}} \times AFR_{\text{disk}}
\]
Solutions for disk failures

- Full redundancy
  - Replication (mirroring)

- Partial Redundancy
  - Parity information
Parity

- Parity calculation is performed using “XOR”.
  - XOR operator is "true" if and only if one of its two operands is true
  - Property of XOR:
    - If \( D_p = D_1 \oplus \ldots \oplus D_k \oplus \ldots \oplus D_n \),
      then \( D_k = D_p \oplus D_1 \oplus \ldots \oplus D_{k-1} \oplus D_{k+1} \oplus \ldots \oplus D_n \)

- Therefore, if any data is lost, we can recover the data from parity and the remaining data

- Advantages: only one of the "N+1" drives contains redundancy information

- Disadvantages: parity information has to be computed every time the data is updated
RAID

RAID: Redundant Arrays of Independent Disks

- **Goals**: increased data reliability, storage capacity and increased I/O performance

Main concepts in RAID

- Mirroring
- Parity
- Stripping: a file is divided into different parts, and stored in different disks

Advantages:

- High capability
- High performance: data stripe
- Graceful degrading
- One disk fails, only that disk needs to be replaced
RAID

Disadvantage: failures

$\text{AFR}_{\text{disks}} = N_{\text{disks}} \times \text{AFR}_{\text{disk}}$

Solution

Redundancy:

- 1) replication/mirroring: need more space
- 2) parity: recover from single disk failure; need more operations to maintain parity info and recover
Disk arrays taxonomy

→ RAID levels

- 0: stripping without redundancy
- 1: full copy mirroring
- 3: separate disk for parity
- 4: similar as RAID 3
- 5: rotated distributed parity
- 6: double parity

They are just classifications rather than a ordered list
RAID levels

- **RAID0**
  - Stripped without redundancy
  - Data can be read off in parallel
  - Any disk failure destroys the entire array

- **RAID1**
  - Mirrored
  - Array continues to operate so long as at least one drive is functioning
- **RAID3**
  - Striped set with dedicated parity
  - Single parity disk is a bottleneck for writing
  - Byte-level striping (typically under 1k)

- **RAID4**
  - Identical to RAID 3 but does block-level striping instead of byte-level striping
  - The block can be of any size
**RAID5**

- Striped set with distributed parity
- The array is not destroyed by a single drive failure
- Upon drive failure, any subsequent reads can be calculated from the distributed parity
- The array will have data loss in the event of a second drive failure
RAID6

- Striped set with dual parity.
- Provides fault tolerance from two drive failures
Disk arrays - mid-range architectures

- Mid-range array (e.g. HP FC60)
  - Sometimes separate controllers and disk boxes
  - Up to 1-2 TB disk, 0.5 GB cache memory
  - Can saturate a 100 MB/s FibreChannel link; $O(10,000 \text{ IOs/s})$
Disk arrays – high-end architectures

- High-end array (e.g. HP XP256, EMC Symmetrix)
  - Separate controllers and disk boxes
  - Few TB disk, few GB cache memory
  - Can saturate few 100 MB/s FibreChannel links; $O(50,000 \text{ IOs/s})$
Network Attached Storage (NAS)

→ Follows a client/server design

→ A NAS head acts as the interface between the NAS and network clients

→ The NAS appears on the network as a single "node" that is the IP address of the head device

→ Clients access a NAS over an Ethernet connection

→ The NAS devices require no monitor, keyboard or mouse and run an embedded os

→ NAS uses file-based application protocols such as NFS (Network File System)
Storage Area Networks (SANs)

- An architecture to attach remote storage devices to computer servers in such a way that the storage devices appear as locally attached to the OS.
- The data can be accessed in blocks.
- Use FibreChannel protocol to access data.
NAS vs. SAN

### NAS
- **Application software**
- **Network**
- **File System**
- **Storage**

### SAN
- **Application software**
- **Network**
- **File System**
- **FC/GbE**
- **Storage**