# I/O & RAID

COMP4611
Tutorial 10
19-23, Nov



## I/O Performance

Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead

- **Seek Time** depends on the number of tracks and mechanical characteristics of arm
- **Rotation Time** depends on how fast the disk rotates and how far sector is from head
- **Transfer Time** depends on data rate (bandwidth) of disk and size of request



## I/O Performance – Example 1

Compare the time to read and write a 64KB block to Flash memory and magnetic disk.

For Flash, assume it takes 65ns to read 1 byte, 1.5us to write 1 byte, and 5ms to erase 4KB.

For disk, average seek time = 12ms, rotation speed = 3600rpm and data transfer rate = 2.6-4.2MB/s.

Assume the measured seek time is one-third of the calculated average, the controller overhead is 0.1ms, and the data are stored in the outer-most track (the disk rotates in one direction).



## Example 1 - Analysis

**File to transfer: 64 KB**

- Magnetic Disk
  - average seek time = 12ms
  - rotation speed = 3600rpm
  - data transfer rate = 2.6-4.2MB/s
  - controller overhead = 0.1ms
- Flash
  - 65ns to read 1 byte
  - 1.5us to write 1 byte
  - 5ms to erase 4KB
- Some Key points
  - Data are stored in the outer-most track
  - We want to use the average rotational delay in order to find the time to read to or write from the disk



## Example 1 - Solution

- Average disk access is equal to measured seek time + average rotational delay + transfer time + controller overhead. The average time to read or write 64KB for the disk is

  ! !

  12 / 3ms + 0.5 / 3600RPM + 64KB / (4.2MB/s) + 0.1ms = 19.3ms
- Flash read time = 64KB / (1B / 65ns) = 4.3ms
- Flash write time = erase time + write time
  = 64KB / (4KB / 5ms) + 64KB / (1B / 1.5us) = 178.3ms
- Thus, Flash memory is about 4.5 times faster than disk for reading 64KB, and disk is about 9 times faster than Flash memory for writing 64KB.



## Impact of I/O on System Performance

Suppose we have a benchmark that executes in 100 seconds of elapsed time, where 90 seconds is CPU time and the rest is I/O time. If the CPU time improves by 50% per year for the next five years but I/O time does not improve, how much faster will our program run at the end of the five years?

**Answer:** Elapsed Time = CPU time + I/O time

| After n years | CPU time | I/O time | Elapsed time | % I/O time |
|---|---|---|---|---|
| 0 | 90 Seconds | 10 Seconds | 100 Seconds | 10% |
| 1 | $\frac{90}{1.5}=60$ Seconds | 10 Seconds | 70 Seconds | 14% |
| 2 | $\frac{60}{1.5}=40$ Seconds | 10 Seconds | 50 Seconds | 20% |
| 3 | $\frac{40}{1.5}=27$ Seconds | 10 Seconds | 37 Seconds | 27% |
| 4 | $\frac{27}{1.5}=18$ Seconds | 10 Seconds | 28 Seconds | 36% |
| 5 | $\frac{18}{1.5}=12$ Seconds | 10 Seconds | 22 Seconds | 45% |

*Over five years:*

CPU improvement = 90/12 = 7.5 **BUT** System improvement = 100/22 = 4.5

## I/O System Performance

- I/O system performance depends on many aspects of the system (limited by weakest link in the chain):
  - CPU
  - Memory system (internal and external caches, main memory)
  - Underlying interconnection (buses)
  - I/O controller
  - I/O device
  - Speed of I/O software (operating system)
  - Efficiency of the software's use of the I/O devices

7

## I/O Performance Metrics

- Throughput: I/O bandwidth
  - The number of bytes received by the server in unit time
  - In order to get the highest possible throughput
    - The server should never be idle
    - The bus should never be empty
- Response time: latency
  - Begins when a byte is transmitted by the server
  - Ends when it is received by another server
  - In order to minimize the response time
    - The bus should be empty
    - The server will be idle

8

## I/O Performance – Example 2

- A disk workload consisting of 64KB reads and writes where the user program executes 200,000 instructions per disk I/O operation and
  - a processor that sustains 3 billion instr/s and averages 100,000 OS instructions to handle a disk I/O operation (the device I/O operation's latency is close to 0)

  The maximum disk I/O rate (# I/O's/s) of the processor is

  - a memory-I/O bus that sustains a transfer rate of 1000 MB/s

  Each disk I/O reads/writes 64 KB so the maximum I/O rate of the bus is

9

## Example 2 – Solution

- A disk workload consisting of 64KB reads and writes where the user program executes 200,000 instructions per disk I/O operation and
  - a processor that sustains 3 billion instr/s and averages 100,000 OS instructions to handle a disk I/O operation (the device I/O operation's latency is close to 0)

  The maximum disk I/O rate (# I/O's/s) of the processor is

  $$\frac{\text{Instr execution rate}}{\text{Instr per I/O}} = \frac{3 \times 10^9}{(200 + 100) \times 10^3} = 10{,}000 \text{ I/O's/s}$$

  - a memory-I/O bus that sustains a transfer rate of 1000 MB/s

  Each disk I/O reads/writes 64 KB so the maximum I/O rate of the bus is

  $$\frac{\text{Bus bandwidth}}{\text{Bytes per I/O}} = \frac{1000 \times 10^6}{64 \times 10^3} = 15{,}625 \text{ I/O's/s}$$

10

## Example 2 (con't)

- A disk workload consisting of 64KB reads and writes where the user program executes 200,000 instructions per disk I/O operation and
  - SCSI disk I/O controllers with a DMA transfer rate of 320 MB/s that can accommodate up to 7 disks per controller
  - disk drives with a read/write bandwidth of 75 MB/s and an average seek plus rotational latency of 6 ms

What is the maximum sustainable I/O rate and what is the number of disks and SCSI controllers required to achieve that rate?

11

## Example 2 – Solution

So the processor is the bottleneck, not the bus

Disk I/O read/write time = seek + rotational time + transfer time = 6ms + 64KB/(75MB/s) = 6.9ms

Thus each disk can complete 1000ms/6.9ms or 146 I/O's per second. To saturate the processor requires 10,000 I/O's per second or 10,000/146 = 69 disks

To calculate the number of SCSI disk controllers, we need to know the average transfer rate per disk to ensure we can put the maximum of 7 disks per SCSI controller and that a disk controller won't saturate the memory-I/O bus during a DMA transfer

Disk transfer rate = (transfer size)/(transfer time) = 64KB/6.9ms = 9.56 MB/s

Thus 7 disks won't saturate either the SCSI controller (with a maximum transfer rate of 320 MB/s) or the memory-I/O bus (1000 MB/s). This means we will need 69/7 or 10 SCSI controllers.

12

## Reliability and Availability

- Reliability: Is anything broken?
  - Reliability can be improved by:
    - Enhancing environmental conditions
    - Building more reliable components
    - Building with fewer components
      - Improve ability may come at the cost of lower reliability
- Availability: Is the system still available to the user?
  - Availability can be improved by adding hardware:
    - Example: adding error-correcting code on memory

13

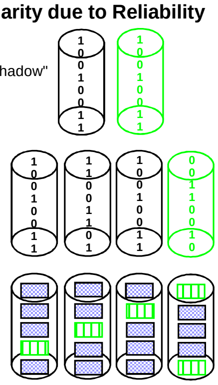
## Motivation for RAID

- As a first solution to increase disk performance we could use *Disk Arrays*
- Reliability of N disks = Reliability of 1 Disk ÷ N
  - 1,200,000 Hours ÷ 100 disks = 12,000 hours
  - 1 year = 365 * 24 = 8700 hours
- Disk system MTTF: Drops from 140 years to about 1.5 years!
- Problem: No redundancy between the disks – failed data cannot be retrieved

14

## RAID Techniques: Goal was Performance, Popularity due to Reliability

- ***Disk Mirroring, Shadowing (RAID 1)***
  - Each disk is fully duplicated onto its "shadow"
  - Logical write = two physical writes
  - 100% capacity overhead
- ***Parity Data Bandwidth Array (RAID 3)***
  - Parity computed horizontally
  - Logically a single high data bw disk
- ***High I/O Rate Parity Array (RAID 5)***
  - Interleaved parity blocks
  - Independent reads and writes
  - Logical write = 2 reads + 2 writes



15

## Example 3

Suppose we have a RAID 5 system with 5 disks. A disk has failed and is being replaced. Assume the remaining disks are error-free. Reconstruct the data for the new disk.

| | | | | |
|---|---|---|---|---|
| 0111 | 1100 | | 0110 | 1101 |
| 1010 | 1111 | | 0011 | 0001 |
| 0100 | 1101 | | 1011 | 0011 |
| 0101 | 1111 | | 1001 | 1010 |
| 0000 | 1001 | | 1110 | 1111 |

16

## Example 3 - Solution

| | | | | |
|---|---|---|---|---|
| 0111 | 1100 | 0000 | 0110 | 1101 |
| 1010 | 1111 | 0111 | 0011 | 0001 |
| 0100 | 1101 | 0001 | 1011 | 0011 |
| 0101 | 1111 | 1001 | 1001 | 1010 |
| 0000 | 1001 | 1000 | 1110 | 1111 |

17

## Example 4

Suppose we want to build a RAID 0 system with 4000GB storage capacity. There are two options available:

SysA: 100 x 40GB and $400 per disk
SysB: 50 x 80GB and $1000 per disk

Assume the MTTF for every disk is 1,000,000 hours.

What is the cost and MTTF of the each option?

18

## Example 4 - Solution

Cost of SysA = 100 x $400 = $40000
Cost of SysB = 50 x $1000 = $50000

MTTF of SysA = 1000000 / 100 = 1000hrs
MTTF of SysB = 1000000 / 50 = 2000hrs

SysA has a lower cost while SysB has a better MTTF value



## Well Known Storage

- Hard drive: random access, magnetic, various density and speed
- Tape: sequential access, huge storage capacity, cheap
- Helical scan tapes: diagonal storage of bits to allow high speed tape rotation
  - Used also for VCR and camcorders
- Optical disk: high density, mostly read-only
- Flash memory: much faster than hard drive, expensive and limited in capacity



## Storage Example: Internet Archive

- Goal of making a historical record of the Internet
  - Internet Archive began in 1996
  - Wayback Machine interface performs time travel to see what a web page looked like in the past
- Contains over a petabyte ($10^{15}$ bytes)
  - Growing by 20 terabytes ($10^{12}$ bytes) of new data per month
- Besides storing historical record, same hardware crawls Web to get new snapshots



## Internet Archive Cluster

- 1U storage node PetaBox GB2000 from Capricorn Technologies
- Has 4 500-GB Parallel ATA (PATA) drives, 512 MB of DDR266 DRAM, G-bit Ethernet, and 1 GHz C3 processor from VIA (80x86)
- Node dissipates ≈ 80 watts
- 40GB2000s in a standard VME rack, ⇒ 80 TB raw storage capacity
- 40 nodes connected with 48-port Ethernet switch
- Rack dissipates about 3 KW
- 1 Petabyte = 12 racks



## Estimated Cost

- VIA processor, 512 MB of DDR266 DRAM, ATA disk controller, power supply, fans, and enclosure = $500
- 7200 RPM 500-GB PATA drive = $375 (in 2006)
- 48-port 10/100/1000 Ethernet switch and all cables for a rack = $3000
- Total cost $84,500 for a 80-TB rack
- 160 Disks are ≈ 60% of total



## Estimated Performance

- 7200 RPM drive
  - Average seek time = 8.5 ms
  - Transfer bandwidth 50 MB/second
  - PATA link can handle 133 MB/second
  - ATA controller overhead is 0.1 ms per I/O
- VIA processor is 1000 MIPS
  - OS needs 50K CPU instructions for a disk I/O
  - Network stack uses 100K instructions per data block
- Average I/O size
  - 16 KB for archive fetches
  - 50 KB when crawling Web
- Disks are limit
  - ≈ 75 I/Os/s per disk, thus 300/s per node, 12000/s per rack
  - About 200-600 MB/sec bandwidth per rack
- Switches must achieve 1.6-3.8 Gbps over 40 Gbps links

## Estimated Reliability

- CPU/memory/enclosure MTTF is 1,000,000 hours (x 40)
- Disk MTTF 125,000 hours (x 160)
- PATA controller MTTF 500,000 hours (x 40)
- PATA cable MTTF 1,000,000 hours (x 40)
- Ethernet switch MTTF 500,000 hours (x 1)
- Power supply MTTF 200,000 hours (x 40)
- Fan MTTF 200,000 hours (x 40)
- MTTF for system works out to 531 hours (≈ 3 weeks)
- 70% of failures in time are disks
- 20% of failures in time are fans or power supplies

25