Distributed Video Systems

Chapter 3

Storage Technologies

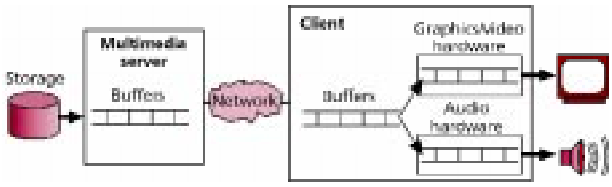
Jack Yiu-bun Lee
Department of Information Engineering
The Chinese University of Hong Kong

Contents

3.1 Introduction

Jack Y.B. Lee

- System Model



- Challenges
  - Real-time storage and retrieval:
    - Continuous media data must be presented using the same timing sequence with which they were captured.
    - Any deviation from this timing sequence can lead to artifacts such as jerkiness in video motion, pops in audio, or possibly complete unintelligibility.

3.1 Introduction

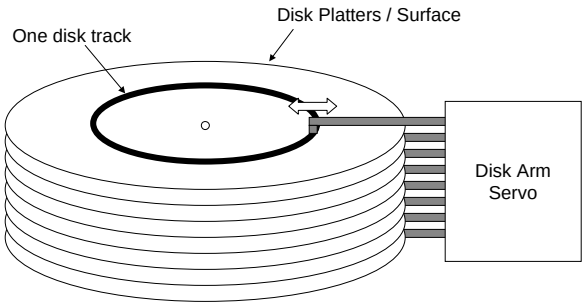
Jack Y.B. Lee

- Challenges
  - Real-time storage and retrieval:
    - Media components may also need *synchronization*. For example, a video stream must synchronize an audio stream in a movie.
  - High data transfer rate and large storage space:
    - Digital video and audio playback demands a high data transfer rate, so that storage space is rapidly filled. (E.g. MPEG1 ~ 1.5Mbps, MPEG2 ~ 4Mbps)
    - The server must efficiently store, retrieve, and manipulate data in large quantities at high speeds.

3.2 Magnetic Disks

Jack Y.B. Lee

- Disk Model



  - The disk platters spin at speed from 3600rpm to 10000rpm;
  - Disk heads in all platters move together.
  - A disk track is further divided into disk sectors.

3.2 Magnetic Disks

Jack Y.B. Lee

- Disk Model
  - Fixed Delays
    - Processing delay at disk controller;
    - Delay at data bus (e.g. SCSI) between disk and controller;
    - Head-switching time;
  - Variable Delays
    - Rotational Latency
      - Depends on position and spindle speed
    - Seek time
      - Depends on number of tracks to seek
    - Transfer Time
      - Depends on how much data to transfer to host

3.2 Magnetic Disks

- Disk Model
  - Disk-Seek Time Function:

$$T_{seek}(n) = \alpha + \beta\sqrt{n}$$

Number of tracks to seek
Seek-time constant (sec)
Fixed overhead (sec)

  - Total Disk-Read Time Function:

$$T_{read}(n) = \alpha + \beta\sqrt{n} + T_{latency} + \frac{Q}{R_{disk}}$$

Size of data to read (Bytes)
Disk transfer rate (Bytes/sec)
Rotational latency (sec)

3.2 Magnetic Disks

- Typical Disk Parameters
  - Seagate 4GB ST12400N (SCSI-2)

Disk Parameter	*Value*
Spindle speed	*5411 rpm*
Max latency (r)	*11ms*
Number of tracks	*2621*
Raw transfer rate	*3.35MB/s*
Single-track seek	*1ms*
Max full-stroke seek	*19ms*

3.2 Magnetic Disks

Jack Y.B. Lee

- Typical Disk Parameters
  - SCSI Variants

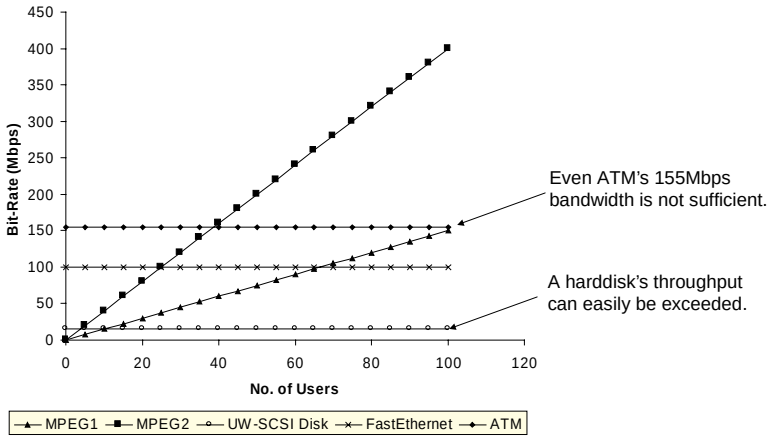
Types	Variants	Max. Speed	Number of Devices	Max. Cable Length
SCSI-1	-	5 MB/s	8	6m
SCSI-2	Fast SCSI	10 MB/s	8	1.5m~3m
	Fast Wide SCSI	20 MB/s	16	1.5m~3m
SCSI-3	Ultra SCSI	20 MB/s	8	1.5m
	Wide Ultra SCSI	40 MB/s	16	1.5m
	Ultra2 SCSI	40 MB/s	8	12m
	Wide Ultra2 SCSI	80 MB/s	16	12m
	Ultra3 SCSI	80 MB/s	8	12m
	Wide Ultra3 SCSI	160 MB/s	16	12m
Fibre Channel	FC-AL	100~200MB/s	126	30m~10km

  - Note that the "Max. Speed" is the top speed of the interface.
  - The actual achievable speed depends on the performance of the connected disks.

3.3 Video Retrieval

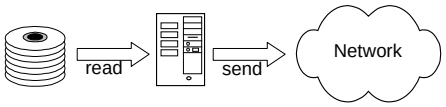
Jack Y.B. Lee

- The Bandwidth Landscape:



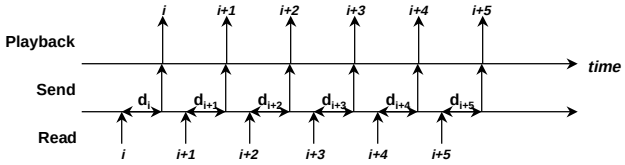
3.3 Video Retrieval

- Single-Stream Retrieval



  - Ideal Disk (Constant Service Time)

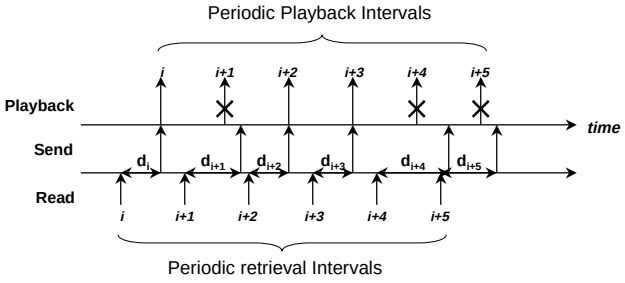
Constant delay: $d_i = d_j \ \forall i, j$



Assumes zero transmission time in network.

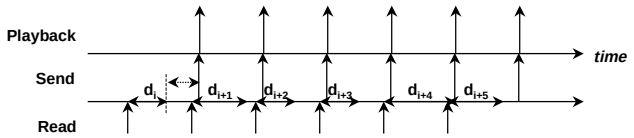
3.3 Video Retrieval

- Single-Stream Retrieval
  - In Practice (Variable Service Time)
    - Variable delay can cause playback glitches:

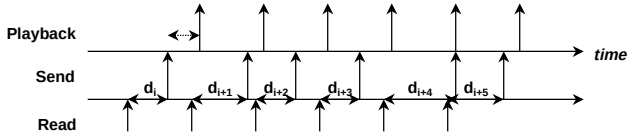


3.3 Video Retrieval

- Single-Stream Retrieval
  - In Practice (Variable Service Time)
    - Buffering At Server:



    - Buffering At Receiver:



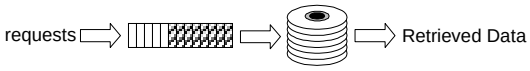
3.3 Video Retrieval

- Multi-Stream Retrieval
  - One Disk Per Stream
    - Simple but wasteful because disk bandwidth is usually much larger than video bit-rate.
    - E.g. >10Mbps for HD, but MPEG2 only ~4Mbps.
  - Multiple Streams Per Disk
    - A *disk scheduling algorithm* is required to ensure that the individual streams will not interfere with each other, and the delay constraint is met.
    - There are many disk scheduling algorithms, each with its own strengths and weaknesses.

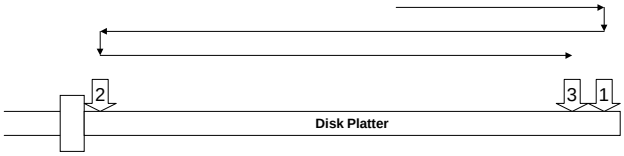
3.4 Disk Scheduling

Jack Y.B. Lee

- Conventional Disk Scheduling Algorithms
  - First-Come-First-Serve (FCFS)
    - Service requests in the order they arrive.



    - Simple but poor disk utilization.
      - Example:

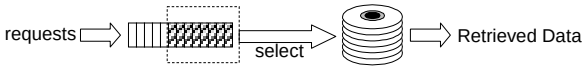


Very long seek time in this example.

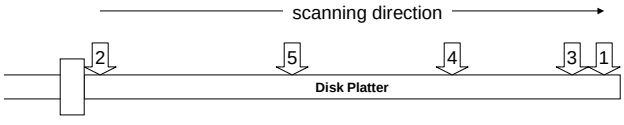
3.4 Disk Scheduling

Jack Y.B. Lee

- Conventional Disk Scheduling Algorithms
  - SCAN
    - Service requests along *scanning* direction.



    - Better disk utilization but potentially long *round time*.
      - Example:



Service Order: 2 5 4 3 1

Note request 1 has to wait longer even it arrives first!

3.4 Disk Scheduling

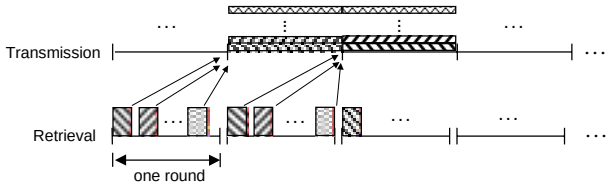
Jack Y.B. Lee

- Multimedia Disk Scheduling Algorithms
  - Earliest Deadline First (EDF)
    - This algorithm schedules the media block with the earliest deadline for retrieval.
    - Likely to yield *excessive* seek time and rotational latency, and *poor* server-resource utilization can be expected.
  - Scan-EDF
    - Same as EDF except using SCAN to schedule requests having the same deadline.

3.4 Disk Scheduling

Jack Y.B. Lee

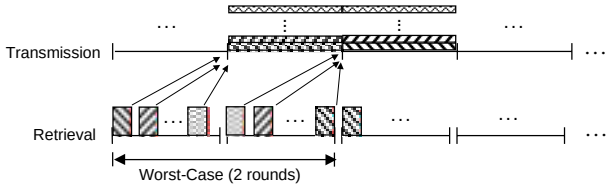
- Disk Scheduling Algorithms for VoD Servers
  - Characteristic of Continuous Media
    - Periodic retrieval of fixed-size data blocks;
    - The entire retrieval schedule is known beforehand.
  - Round-Based Disk Scheduling
    - Read one data block for each video stream in each round.
    - Retrievals in a round are serviced using SCAN.



3.4 Disk Scheduling

Jack Y.B. Lee

- Disk Scheduling Algorithms for VoD Servers
  - Round-Based Disk Scheduling
    - To ensure the continuity of data flow for transmission, we need **two buffers per video stream**.
    - Limitations
      - All video streams must have the same data rate; or
      - The data rate must be an integer multiple of a base data rate.



3.5 Admission Control

Jack Y.B. Lee

- Admission Control
  - Motivation
    - A VoD system only have finite capacity. Hence a mechanism must be used to admit and reject users to avoid system overload.
  - Types of Admission Control Algorithms
    - Deterministic
      - Worst-case scenarios are used to guarantee the service of existing users.
    - Statistical
      - Statistical behaviors of the system are used to provide *probabilistic* guarantee. E.g. meeting deadline 99% of the time.
    - Observational
      - Current system status like utilizations are used to evaluate the admission of new users.

3.5 Admission Control

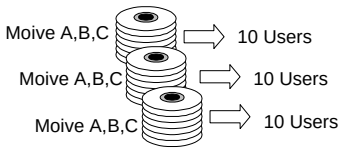
Jack Y.B. Lee

- Dealing with Missed Deadlines
  - Why?
    - Deadlines could be missed if the admission control algorithm is statistical or some other unexpected events occur.
  - What to do?
    - Ignore It
      - Causes service degradations such as jerky video, decoding error, scrambled video, audio clicks, etc.
      - Depends on how much and what kind of data is missed.
    - Error Concealment
      - Repeating data (previous frame, audio packet, etc.)
      - Skipping video frame
      - Lower the resolution (temporary)

3.6 I/O Bandwidth

Jack Y.B. Lee

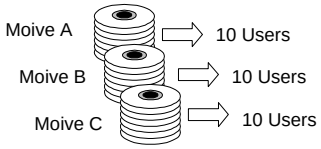
- Increasing Disk Throughput
  - Background
    - A single disk's through can serve a very limited number of concurrent users.
    - For example, a SCSI harddisk can serve around 10 MPEG1 video streams and 3~4 MPEG2 video streams.
  - Replication
    - Use multiple disks, each carry a separate copy of a movie.
    - Expensive since movie is large in size.



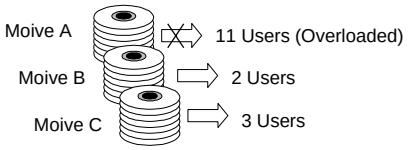
3.6 I/O Bandwidth

Jack Y.B. Lee

- Increasing Disk Throughput
  - Partition
    - Use multiple disks, each carry different movie titles.



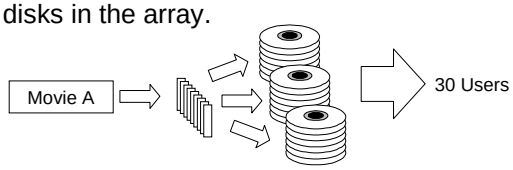
    - Same total storage but poor load-balancing.



3.6 I/O Bandwidth

Jack Y.B. Lee

- Increasing Disk Throughput
  - Disk Striping (Disk Array)
    - Divides a video stream into units and distributes over all disks in the array.



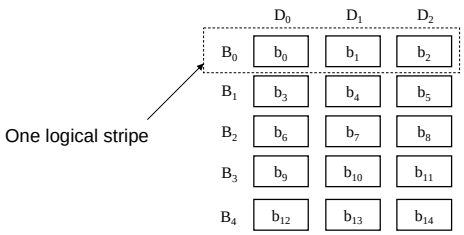
Data Organization:

	D_0	D_1	D_2
B_0	b_0	b_1	b_2
B_1	b_3	b_4	b_5
B_2	b_6	b_7	b_8
B_3	b_9	b_{10}	b_{11}
B_4	b_{12}	b_{13}	b_{14}

3.6 I/O Bandwidth

Jack Y.B. Lee

- Increasing Disk Throughput
  - Disk Striping (Disk Array)
    - One logical *stripe* is retrieved per stream per round.

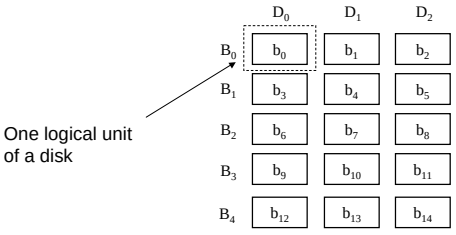


    - Hence the throughput is *N* times those of a single disk if there are *N* disks in the array.
    - The disks are *spindle synchronized*.

3.6 I/O Bandwidth

Jack Y.B. Lee

- Increasing Disk Throughput
  - Disk Interleaving
    - Same as disk striping except one logical unit is retrieved from one of the disk per stream per round.



    - Hence each disk can serve a different stream at the same time, or multiple streams are served concurrently.
    - The disks are not spindle synchronized and operates independently.

3.7 Storage Capacity

Jack Y.B. Lee

- Tertiary Storage and Storage Hierarchies
  - Motivation
    - While magnetic disks are suitable for use in VoD systems due to the high throughput and low latency, *they are still expensive*.
    - For applications like video library where large number of videos must be archived, storing all video in disks will become prohibitively expensive (and unnecessary).
  - Tertiary Storage

Feature	Magnetic Disk	Optical Disk	Low-end Tape	High-end Tape
Capacity	9GB	200GB	500GB	10TB
Mount time	None	20 secs	60 secs	90 secs
Transfer Rate	2MBps	300KBps	100KBps	1MBps
Cost	$5,000	$50,000	$50,000	$0.5M to $1M
Cost/GB	$555	$125	$100	$50

3.7 Storage Capacity

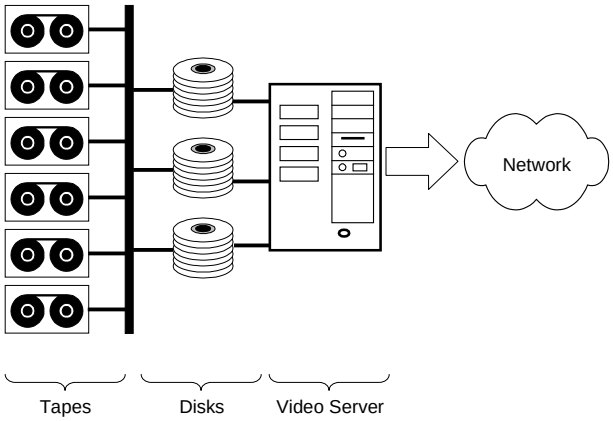
Jack Y.B. Lee

- Tertiary Storage and Storage Hierarchies
  - Tertiary Storage
    - Pros
      - Removable media like optical disks and tapes are less expensive in terms of cost per GB.
    - Cons
      - Lower data transfer rate;
      - Very long random access time.
  - Storage Hierarchy
    - Combines the cost-effectiveness of tertiary storage with the performance of magnetic disks.
    - Tertiary storage are used for permanent storage and the magnetic disks used as a cache for video delivery.

3.7 Storage Capacity

Jack Y.B. Lee

- Tertiary Storage and Storage Hierarchies
  - Storage Hierarchy



3.7 Storage Capacity

Jack Y.B. Lee

- Tertiary Storage and Storage Hierarchies
  - Storage Hierarchy
    - Scheme 1:
      - Store the beginning segments of videos in magnetic disk and the rest in tertiary storage;
      - Starts delivery from magnetic disk while downloading the rest of the video from the tertiary storage.
    - Scheme 2:
      - Downloads an entire video from tertiary storage to magnetic disks for delivery.
      - Manage the disk storage using most-recently-used policy.
      - Long startup time for uncached video but the caching should perform well since only a small number of video will be popular at any one time.