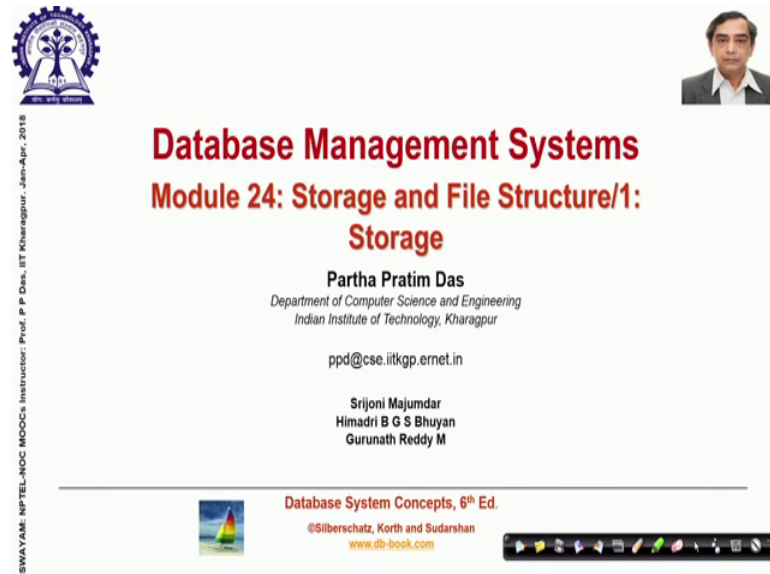


Database Management System
Prof. Partha Pratim Das
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 24
Storage and File Structure: Storage

(Refer Slide Time: 00:16)



Database Management Systems
Module 24: Storage and File Structure/1:
Storage

Partha Pratim Das
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

ppd@cse.iitkgp.ernet.in

Srijoni Majumdar
Himadri B G S Bhuyan
Gurunath Reddy M

Database System Concepts, 6th Ed.
©Silberschatz, Korth and Sudarshan
www.db-book.com

SWAYAM NPTEL-NOC Instructor: Prof. P P Das, IIT Kharagpur, Jan-Apr, 2018

Welcome to module 24 of database management systems in this module and the next we will take a look at the storage and file structure of database systems. So, we will start with the storage.

(Refer Slide Time: 00:30)

PPD

Module Objectives

- To take a look at various Physical Storage Media for high volume, fast, reliable and inexpensive options for data storage for databases
- To understand the structure and basic functionality of Magnetic Disks
- To understand RAID – array of redundant disks in parallel to enhance speed and reliability
- To understand the options of Tertiary Storage for high volume, inexpensive backup options

SWAYAM: NPTEL-NOC MOOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 8th Edition

24.3

©Silberschatz, Korth and Sudarshan

So, specifically we want to look at various physical storage medium because. So, far we have been talking only about the logical layer of the database design and now we want to actually look at the in physical terms how the data will be stored what could be the physical storage medium for high volume fast reliable inexpensive options for databases. We would like to understand the structure and basic functionality of magnetic disks, because they are they are most widely used we will try to take the glimpse about RAID which is a kind of a good option in terms of reliable databases and also look at options for the tertiary storage.

(Refer Slide Time: 01:12)

PPD

Module Outline

- Overview of Physical Storage Media
- Magnetic Disks
- RAID
- Tertiary Storage

SWAYAM: NPTEL-NOC MOOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

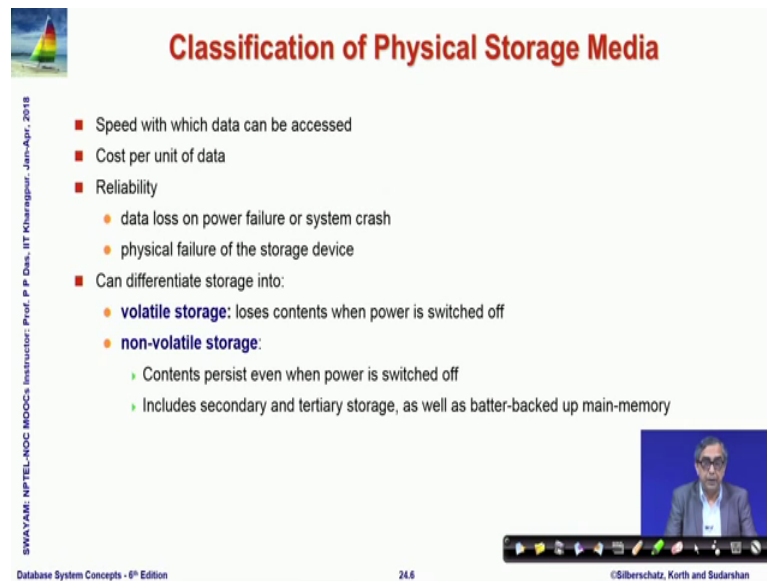
Database System Concepts - 8th Edition

24.4

©Silberschatz, Korth and Sudarshan

So, these are the topics that we will quickly cover in this.

(Refer Slide Time: 01:17)



The slide, titled "Classification of Physical Storage Media", lists the following factors and storage types:

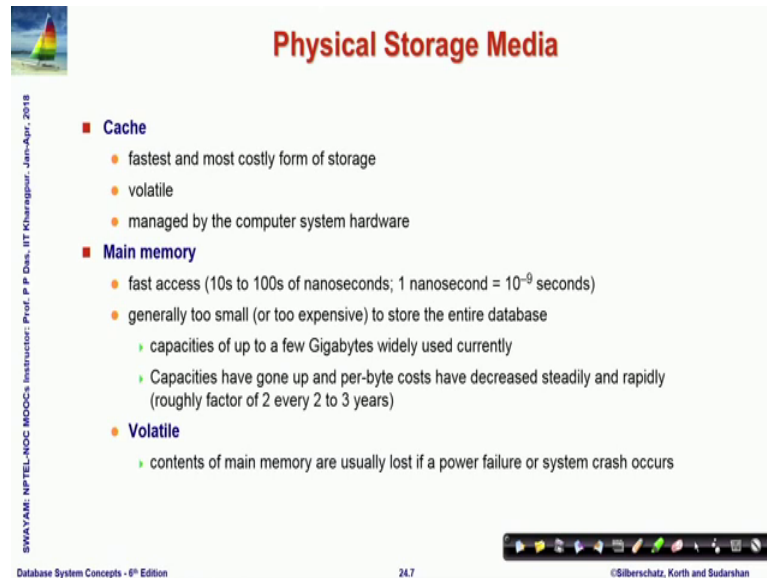
- Speed with which data can be accessed
- Cost per unit of data
- Reliability
 - data loss on power failure or system crash
 - physical failure of the storage device
- Can differentiate storage into:
 - **volatile storage:** loses contents when power is switched off
 - **non-volatile storage:**
 - Contents persist even when power is switched off
 - Includes secondary and tertiary storage, as well as batter-backed up main-memory

SWAYAM: NPTEL-NOC IMOOCs Instructor: Prof. P. Das, IIT Kharagpur, Jan-Apr, 2018
Database System Concepts - 8th Edition 24.6 ©Silberschatz, Korth and Sudarshan

So, first let us take an overview of the physical storage medium. I am sure all of you have known all or parts of this. So, this is, but this is more for completeness to look at from the perspective of a database application. So, some of the classification of storage media done on different factors. The factors include speed, which is the first thing, how fast the data can be accessed, the cost per unit of data, you can say rupees per bit or rupees per byte or rupees per kilobyte, something like that.

So, which is a cost per unit of data, the reliability, that is, if we will the data get lost if power fails or if the system crashes and or if this physical failure of the storage device and so on. So, what is the reliability on that; and broadly as you all know we can differentiate storage into volatile storage which loses contents. When the power is switched off and the non-volatile storage which are secondary and tertiary storage where the data will continue to stay even when power is off even you have some parts of the memory which may be battery backup which also will be non-volatile.

(Refer Slide Time: 02:29)



Physical Storage Media

- **Cache**
 - fastest and most costly form of storage
 - volatile
 - managed by the computer system hardware
- **Main memory**
 - fast access (10s to 100s of nanoseconds; 1 nanosecond = 10^{-9} seconds)
 - generally too small (or too expensive) to store the entire database
 - capacities of up to a few Gigabytes widely used currently
 - Capacities have gone up and per-byte costs have decreased steadily and rapidly (roughly factor of 2 every 2 to 3 years)
 - **Volatile**
 - contents of main memory are usually lost if a power failure or system crash occurs

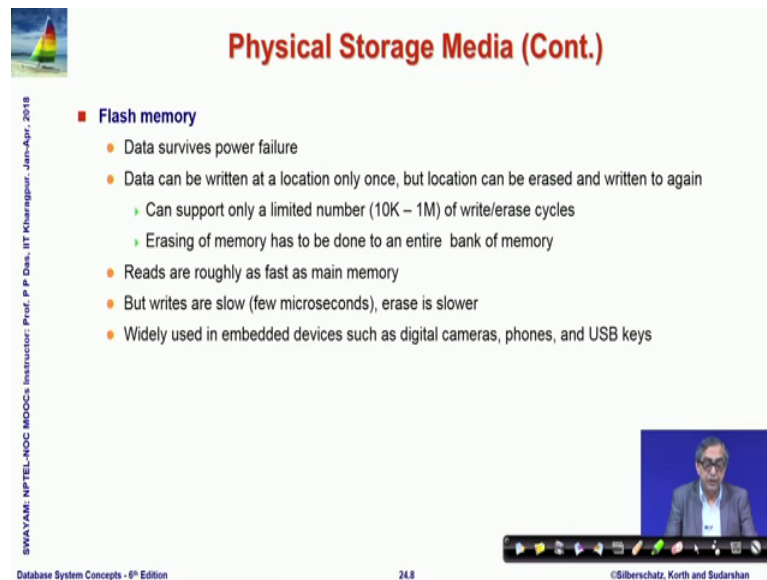
SWAYAM: NPTEL-NOC MDOCS Instructor: Prof. P. P. Das, IIT Khargpur, Jan-April, 2018

Database System Concepts - 8th Edition 24.7 ©Silberschatz, Korth and Sudarshan

So, in terms of the physical storage certainly the absolute starting point of the storage is registered in the CPU; we are not talking about that because they are primarily meant for temporary computations. So, in terms of data the first possible level is the cache which is the fastest and most costly form of storage it is volatile in nature and is managed by the computer system hardware.

So, cache typically is a fast semiconductor memory that exist between your main memory and the disk system and it is very fast to work with then comes the main memory which is which has fast access, but compare to cache it may be may be much bigger, but overall it is too small to store an entire database, but I mean every regularly the size of this main memory is increasing. So, the capacity of couple of gigabytes are common these days, but still it is small compare to the requirement of the databases and main memory typically is volatile. So, if the power goes up the system crashes all the data is lost.

(Refer Slide Time: 03:48)



The slide is titled "Physical Storage Media (Cont.)" and features a small image of a sailboat in the top left corner. The main content is a bulleted list under the heading "Flash memory". The list includes: "Data survives power failure", "Data can be written at a location only once, but location can be erased and written to again" (with sub-points: "Can support only a limited number (10K – 1M) of write/erase cycles" and "Erasing of memory has to be done to an entire bank of memory"), "Reads are roughly as fast as main memory", "But writes are slow (few microseconds), erase is slower", and "Widely used in embedded devices such as digital cameras, phones, and USB keys". The slide also contains a small video inset of a man speaking, a navigation bar at the bottom, and footer text: "Database System Concepts - 8th Edition", "24.8", and "©Silberschatz, Korth and Sudarshan".

Physical Storage Media (Cont.)

■ **Flash memory**

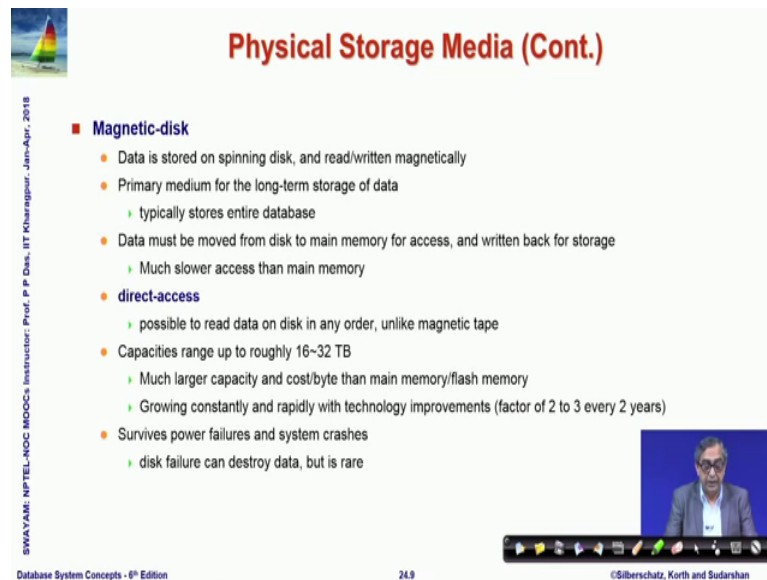
- Data survives power failure
- Data can be written at a location only once, but location can be erased and written to again
 - Can support only a limited number (10K – 1M) of write/erase cycles
 - Erasing of memory has to be done to an entire bank of memory
- Reads are roughly as fast as main memory
- But writes are slow (few microseconds), erase is slower
- Widely used in embedded devices such as digital cameras, phones, and USB keys

Database System Concepts - 8th Edition 24.8 ©Silberschatz, Korth and Sudarshan

We have flash memory where the data can survive across power failure; it can be they had data can be written at a location only once, but you can erase and write it again.

So, it is not like in the main memory where you can read write read write like that here you can write and then if you want to write again then you will have to erase and write it. So, the read is very fast in case of flash memory which is almost as fast as a main memory, but writes a slope particularly when you have erase and write it will be a slow process all the kinds of USB keys pen drives digital phone memory that we are often using are actually flash memory.

(Refer Slide Time: 04:35)



Physical Storage Media (Cont.)

- **Magnetic-disk**
 - Data is stored on spinning disk, and read/written magnetically
 - Primary medium for the long-term storage of data
 - typically stores entire database
 - Data must be moved from disk to main memory for access, and written back for storage
 - Much slower access than main memory
 - **direct-access**
 - possible to read data on disk in any order, unlike magnetic tape
 - Capacities range up to roughly 16-32 TB
 - Much larger capacity and cost/byte than main memory/flash memory
 - Growing constantly and rapidly with technology improvements (factor of 2 to 3 every 2 years)
 - Survives power failures and system crashes
 - disk failure can destroy data, but is rare

SWAYAM: NPTEL-NOC MOOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 8th Edition

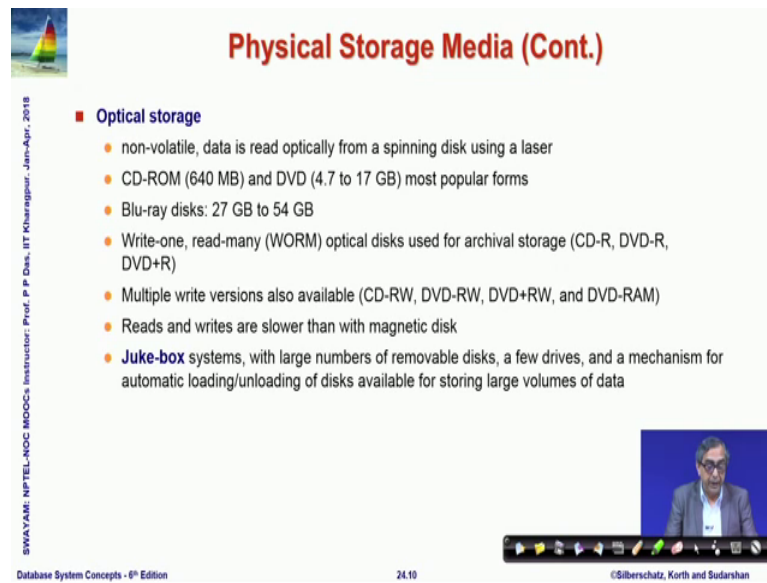
24.9

©Silberschatz, Korth and Sudarshan

Then you have the magnetic disk where the data is stored on spinning disk and it is typically written and read magnetically. So, this is the primary medium for long term storage of large volume of data. So, data needs to be moved from disk to the main memory and written back for permanent storage. So, it is ways slower compare to the main memory and, but it is has a kind of direct access which means that it is possible to read data on this disk in any arbitrary order in compare to magnetic disk.

Which is the serial device here it is a, it is kind of a I can do things in parallel at random in any order capacity is go up to tens of terabytes easily and it can survive for failure and system crashes, because it will the magnetic recording will still be there if the disk itself fails then it will the data will get distract, but such a situation is usually rear.

(Refer Slide Time: 05:40)



Physical Storage Media (Cont.)

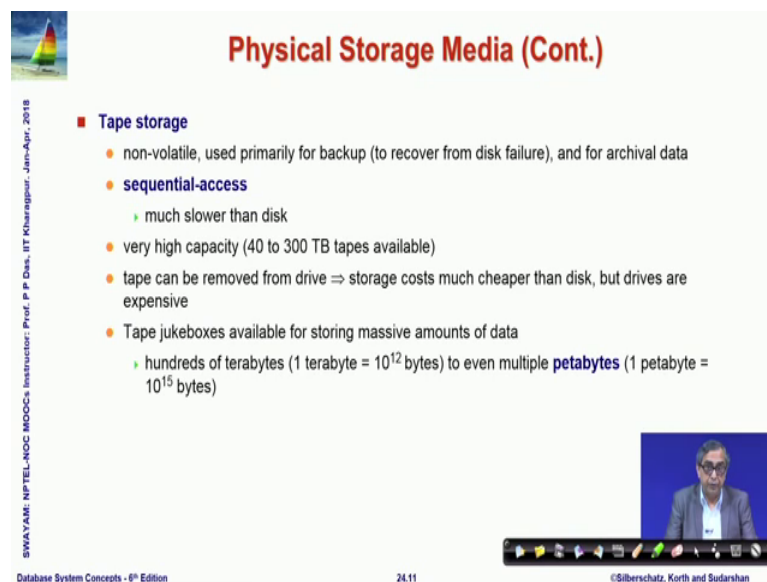
- **Optical storage**
 - non-volatile, data is read optically from a spinning disk using a laser
 - CD-ROM (640 MB) and DVD (4.7 to 17 GB) most popular forms
 - Blu-ray disks: 27 GB to 54 GB
 - Write-one, read-many (WORM) optical disks used for archival storage (CD-R, DVD-R, DVD+R)
 - Multiple write versions also available (CD-RW, DVD-RW, DVD+RW, and DVD-RAM)
 - Reads and writes are slower than with magnetic disk
 - **Juke-box** systems, with large numbers of removable disks, a few drives, and a mechanism for automatic loading/unloading of disks available for storing large volumes of data

SWAYAM: NPTEL-NOC MDOCS Instructor: Prof. P. P. Das, IIT Khargapur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.10 ©Silberschatz, Korth and Sudarshan

We have different optical storage devices CD-ROM, DVD and so, on the juke box systems and which are also non volatile and data is written optically here, that is by using a laser light on the spinning disk use a typically optical storages are removable media.

(Refer Slide Time: 06:03)



Physical Storage Media (Cont.)

- **Tape storage**
 - non-volatile, used primarily for backup (to recover from disk failure), and for archival data
 - **sequential-access**
 - ▶ much slower than disk
 - very high capacity (40 to 300 TB tapes available)
 - tape can be removed from drive \Rightarrow storage costs much cheaper than disk, but drives are expensive
 - Tape jukeboxes available for storing massive amounts of data
 - ▶ hundreds of terabytes (1 terabyte = 10^{12} bytes) to even multiple **petabytes** (1 petabyte = 10^{15} bytes)

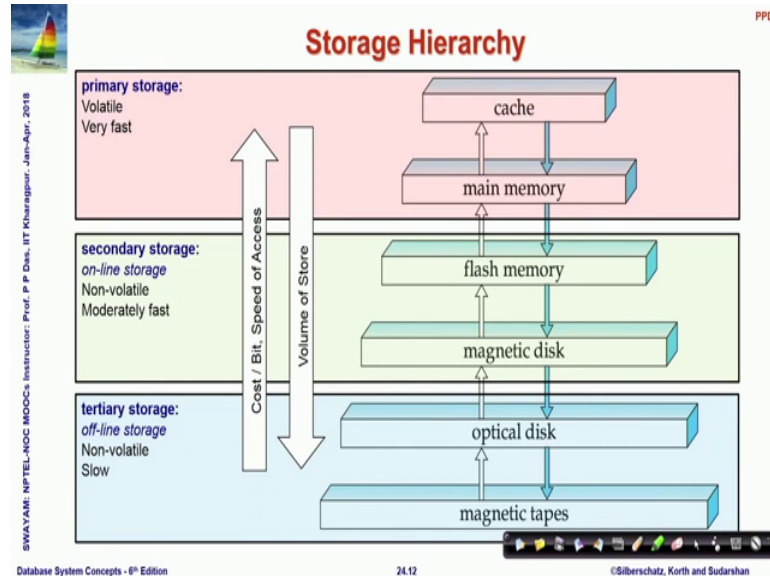
SWAYAM: NPTEL-NOC MDOCS Instructor: Prof. P. P. Das, IIT Khargapur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.11 ©Silberschatz, Korth and Sudarshan

Then you have the tape storage which usually is a largest volume of storage, but it is as a name suggests it is a tape it is a linear device. So, access can only be sequential. So, if you want to read the 6th record you have to skip of a record 1 to 5, but it can be a very high capacity usually it is slope.

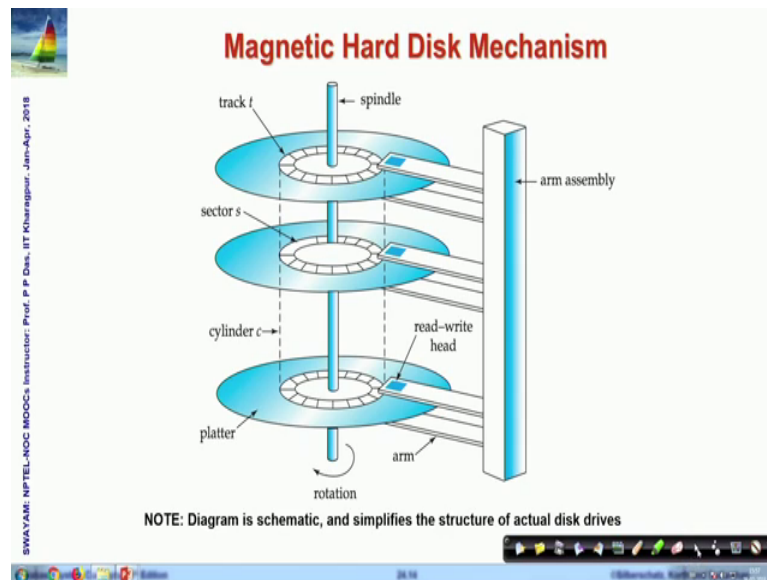
But very large in volume and tape juke boxes can support hundreds of terabytes or even multiple of petabyte. So, that is for offline storage.

(Refer Slide Time: 06:35)



This is a big medium. So, this is the basic storage hierarchy. So, we broadly classify them in to three groups primary storage which is volatile and very fast cache and main memory is within that or secondary storage which is an online store which is non volatile and moderately fast and tertiary storage is called the offline store which is non volatile and slow. So, flash memory and magnetic disk are secondary storage they are non volatile and moderately fast and they are online. So, they exist with the system whereas, optical disk and magnetic disk can be removed and taken elsewhere.

(Refer Slide Time: 07:12)



So, let us quickly take a look into the magnetic disk this is how a typical magnetic disk looks like; these are the different cylinders these are the different disks that we have and these are all different read write head.

So, as you can see all of them can work along a path backward forward like this and they can come and parallelly all of them parallelly can read from the different disks and this disks keep on spinning to help you look at the data anywhere on the disk. So, that is a typical structure of a magnetic.

(Refer Slide Time: 07:50)

Magnetic Disks

- Read-write head
 - Positioned very close to the platter surface (almost touching it)
 - Reads or writes magnetically encoded information
- Surface of platter divided into circular tracks
 - Over 50K-100K tracks per platter on typical hard disks
- Each track is divided into sectors
 - A sector is the smallest unit of data that can be read or written.
 - Sector size typically 512 bytes
 - Typical sectors per track: 500 to 1000 (on inner tracks) to 1000 to 2000 (on outer tracks)
- To read/write a sector
 - disk arm swings to position head on right track
 - platter spins continually; data is read/written as sector passes under head
- Head-disk assemblies
 - multiple disk platters on a single spindle (1 to 5 usually)
 - one head per platter, mounted on a common arm.
- Cylinder i consists of i^{th} track of all the platters

Legend:
A: Track
B: Geometrical sector
C: Track sector
D: Cluster / Block

Labels in diagrams: Track/Cylinder, Sector, Heads (8 Heads, 4 Platters), spindle, arm assembly, platter, cylinder, track, read-write head.

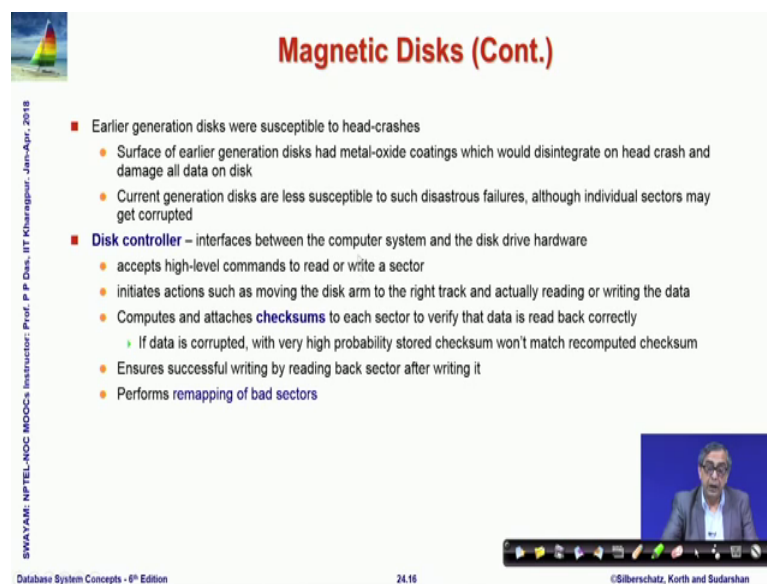
Database System Concepts - 6th Edition 24.15 ©Silberschatz, Korth and Sudarshan

Disk if you look specifically into. So, this is this is the structure we saw and if you specifically look into one particular disk then the disk is radically divided into different sectors portions. So, these are these are all separate portions and you can at a time the head can read one such sector.

So, these are called the geometric sectors and the whole of the ring that you can see the cylindrical ring that you can see is called a track. So, this is a track sector the orange one is a track sector and often we take multiple sectors from a particular track and combine them into one unit this is called the block of data and we will often talk about the block of data.

So, here at the typical numbers of how these sizes of this different units turn out to be.

(Refer Slide Time: 08:44)



Magnetic Disks (Cont.)

- Earlier generation disks were susceptible to head-crashes
 - Surface of earlier generation disks had metal-oxide coatings which would disintegrate on head crash and damage all data on disk
 - Current generation disks are less susceptible to such disastrous failures, although individual sectors may get corrupted
- **Disk controller** – interfaces between the computer system and the disk drive hardware
 - accepts high-level commands to read or write a sector
 - initiates actions such as moving the disk arm to the right track and actually reading or writing the data
 - Computes and attaches **checksums** to each sector to verify that data is read back correctly
 - If data is corrupted, with very high probability stored checksum won't match recomputed checksum
 - Ensures successful writing by reading back sector after writing it
 - Performs remapping of bad sectors

SWAYAM: NPTEL-NOC IODC's Instructor: Prof. P. Das, IIT Kharagpur, Jan-Apr., 2018

Database System Concepts - 9th Edition 24.16 ©Silberschatz, Korth and Sudarshan

So, the early generation where of disk were susceptible to head crashes, but now a days it is moved it quite stable there are disk controllers which regularly check and manage the; read write into in terms of the sectors naturally the. So, many heads being in parallel the data can be written on to multiple sectors at the same time and so, for doing that.

(Refer Slide Time: 09:12)

Disk Subsystem

system bus

disk controller

disks

- Multiple disks connected to a computer system through a controller
 - Controllers functionality (checksum, bad sector remapping) often carried out by individual disks; reduces load on controller
- Disk interface standards families
 - ATA (AT adaptor) range of standards
 - SATA (Serial ATA)
 - SCSI (Small Computer System Interconnect) range of standards
 - SAS (Serial Attached SCSI)
 - Several variants of each standard (different speeds and capabilities)

SWAYAM: NPTEL-NOC MOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 6th Edition

24.17

©Silberschatz, Korth and Sudarshan

We the we will see what is the mechanism for that; and we also have disk subsystems where if you need a large a volume of data to be managed which is larger than a single disk. Then you can connect multiple disk and through a disk controller you can actually read write data on to them and there are different such disk interface standards that you may have heard of.

(Refer Slide Time: 09:38)

Disk Subsystem

- Disks usually connected directly to computer system
- In **Storage Area Networks (SAN)**, a large number of disks are connected by a high-speed network to a number of servers
- In **Network Attached Storage (NAS)** networked storage provides a file system interface using networked file system protocol, instead of providing a disk system interface

SWAYAM: NPTEL-NOC MOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 6th Edition

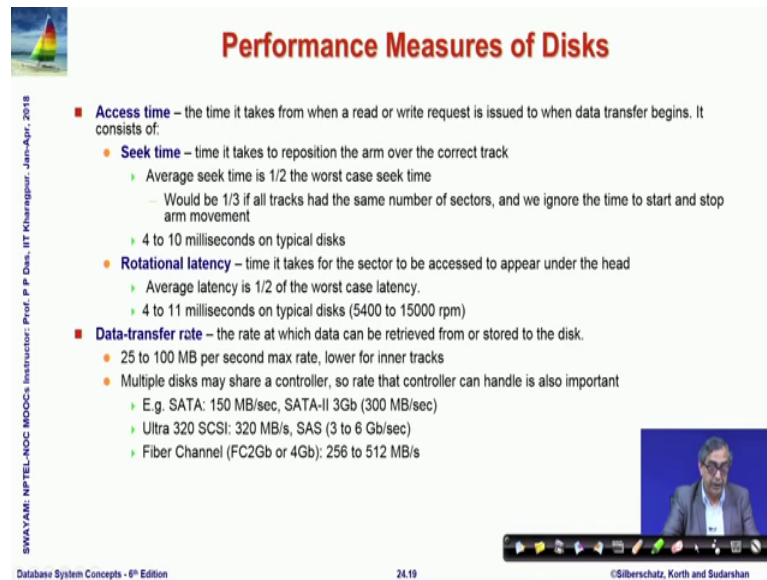
24.18

©Silberschatz, Korth and Sudarshan

We will just mentioned that these are the typical storages the san and are the typical storages.

So, if you come across these terms we will not go into details of that that takes us into a different course of hardware discussion, but these are the very common storages for database disk systems.

(Refer Slide Time: 09:56)



Performance Measures of Disks

- **Access time** – the time it takes from when a read or write request is issued to when data transfer begins. It consists of:
 - **Seek time** – time it takes to reposition the arm over the correct track
 - ▶ Average seek time is 1/2 the worst case seek time
 - Would be 1/3 if all tracks had the same number of sectors, and we ignore the time to start and stop arm movement
 - ▶ 4 to 10 milliseconds on typical disks
 - **Rotational latency** – time it takes for the sector to be accessed to appear under the head
 - ▶ Average latency is 1/2 of the worst case latency.
 - ▶ 4 to 11 milliseconds on typical disks (5400 to 15000 rpm)
- **Data-transfer rate** – the rate at which data can be retrieved from or stored to the disk.
 - 25 to 100 MB per second max rate, lower for inner tracks
 - Multiple disks may share a controller, so rate that controller can handle is also important
 - ▶ E.g. SATA: 150 MB/sec, SATA-II 3Gb (300 MB/sec)
 - ▶ Ultra 320 SCSI: 320 MB/s, SAS (3 to 6 Gb/sec)
 - ▶ Fiber Channel (FC2Gb or 4Gb): 256 to 512 MB/s

SWAYAM: NPTEL-NOC MOOC's Instructor: Prof. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.19 ©Silberschatz, Korth and Sudarshan

Now basic question is for doing this read write on the disk what kind of performance measures we should look at. So, one main measure is access time if I want to access a record from the disk, then how much time shall I need. So, this is based on two major components one is a seek time now naturally as we have seen that the disk platter as a whole range of read heads which are positioned.

So, to be able to get the data the platter has to the head has to move forward or backward to the right track on which the data is there. So, this time it takes to go from current position to the correct track where, I find the data is called the seek time. Now, even when it is come to the; correct position in terms of the correct track it may not actually be on the correct sector.

Because, it is at the whole track is a is kind of a circle. So, it may be at one part of the circle and the actual data may be in a sector which is in a different part of the circle. So, the disk will have to rotate. So, that the correct sector comes under the head which is already positioned through the seek. So, the time to seek plus the rotational latency the time it takes for the sake sector to be access is gives the total access time and then comes the other measure which is the data transfer rate as to you using this then what is

the rate how many megabytes and so, on can be transferred at from this. So, it that depends on a besides the access time you will have to see what is the rate at which the disk can be copied from the magnetic medium to the semiconductor medium and transferred.

(Refer Slide Time: 11:52)

Performance Measures (Cont.)

- **Mean time to failure (MTTF)** – the average time the disk is expected to run continuously without any failure
 - Typically 3 to 5 years
 - Probability of failure of new disks is quite low, corresponding to a "theoretical MTTF" of 500,000 to 1,200,000 hours for a new disk
 - E.g., an MTTF of 1,200,000 hours for a new disk means that given 1000 relatively new disks, on an average one will fail every 1200 hours
 - MTTF decreases as disk ages

SWAYAM NPTEL-NOC MOC® Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.20 ©Silberschatz, Korth and Sudarshan

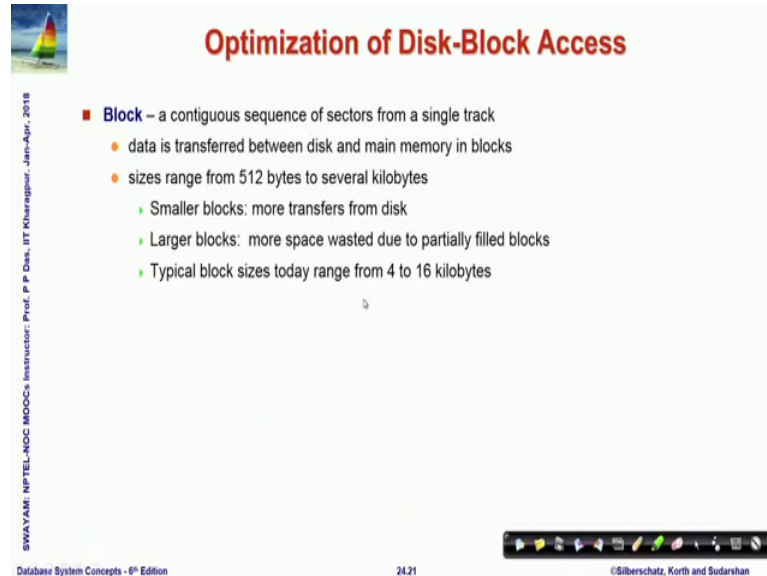
Ah the other performance measure that we are concerned with in the database system is known as MTTF which is mean time to failure.

Because, database systems as you know are dealing with persistent data. So, data has to exist and therefore, the disk on which we keep the data must be very very reliable. So, meantime to failure is conceptually that if you consider two failures of the database of the disk to consecutive failures then what is the time the time elapse between them the average of the time that has elapse between them now. So, if we say the, it is typically now mean time to failure for this magnetic disk that we use today at typically 2 to 5 years.

So, the probability of the failure of a new disk is very very low. So, it is kind of a theoretical MTTF we which will be said like this which; obviously, in terms of hours it it does not really make a make a physical science. So, what it in technical in in more practical terms, what it means that if you are given thousand relatively new disks then on average one of them will fail every 1200 hours.

So, this is what is a. So, MTTF certainly decrease it will it will decrease as a disk becomes old. So, it decreases with age. So, they will start failing sooner than it is to do.

(Refer Slide Time: 13:22)



Optimization of Disk-Block Access

- **Block** – a contiguous sequence of sectors from a single track
 - data is transferred between disk and main memory in blocks
 - sizes range from 512 bytes to several kilobytes
 - ▶ Smaller blocks: more transfers from disk
 - ▶ Larger blocks: more space wasted due to partially filled blocks
 - ▶ Typical block sizes today range from 4 to 16 kilobytes

Database System Concepts - 9th Edition 24.21 ©Silberschatz, Korth and Sudarshan

Now we have to basic objective is to transfer the data at a fast rate. So, what we try to optimize is a; what is called a block the contiguous sequence of sectors from a single track which I mentioned earlier. So, this is what we will be read at one go. So, once you access the data one block will be read block size can range from 512 bytes to several kilobytes. If the blocks are smaller than naturally will need more transfers from the disk if the blocks are larger then you might waste lot of space because your part of the block will not can be used with the data.

So, with all that consideration the typical blocks size that use today is 4 to 16 kilobytes. So, you will see that in all our subsequent discussions particularly with the access and the file organization and the indexing we will consider that a block is one unit.

Which can be fetched in one go and that determines the size of the basic information node where the information about the records will be maintained.

(Refer Slide Time: 14:36)

Optimization of Disk-Block Access (Cont.)

- **Disk-arm-scheduling** algorithms order pending accesses to tracks so that disk arm movement is minimized: Example: Queue 95, 180, 34, 119, 11, 123, 62, 64 with the Read-write head initially at the track 50 and the tail track being at 199

SWAYAM: NPTEL-NOC MOC's Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018
Source: <http://www.cs.iit.edu/~cs561/cs450/disk sched/disk sched.html>

Database System Concepts - 6th Edition 24.22 ©Silberschatz, Korth and Sudarshan

This is the more details in terms of how you move your disk head. So, I think will skip this is more advance material.

(Refer Slide Time: 14:46)

Optimization of Disk Block Access (Cont.)

- **File organization** – optimize block access time by organizing the blocks to correspond to how data will be accessed
 - E.g. Store related information on the same or nearby cylinders
 - Files may get **fragmented** over time
 - ▶ E.g. if data is inserted to/deleted from the file
 - ▶ Or free blocks on disk are scattered, and newly created file has its blocks scattered over the disk
 - ▶ Sequential access to a fragmented file results in increased disk arm movement
 - Some systems have utilities to **defragment** the file system, in order to speed up file access

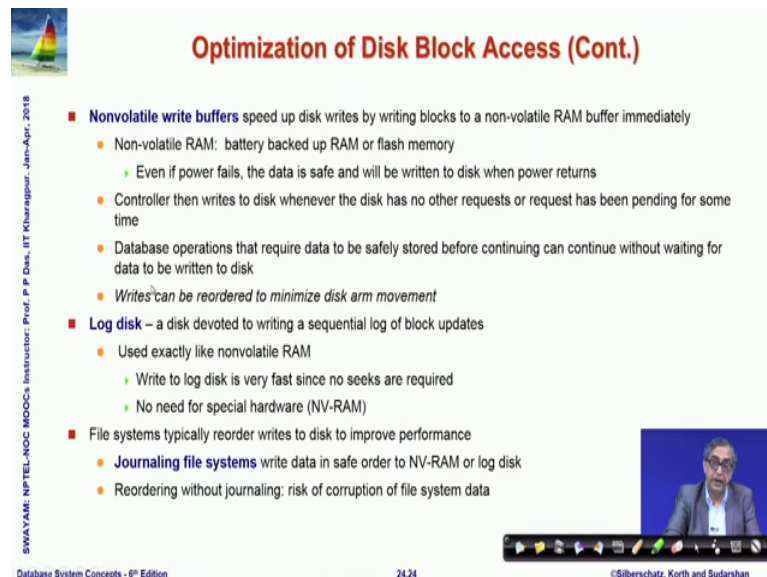
Database System Concepts - 6th Edition 24.23 ©Silberschatz, Korth and Sudarshan

So, to optimize the block access we need to organize the blocks corresponding to how the data will be access. So, certainly if the related data are kept in nearby blocks then naturally you are disk head and the rotation will have to be mean will get minimized. So, the basic idea is store the related information in nearby cylinders in the nearby cylindrical practice, but as you keep on using you may start with that, but as you keep on

using for various types of insert delete and overflows that keeps on happening. So, the data gets very fragmented which means that the data gets spread over a whole lot of widely separated cylinders and so, on.

So, the time to actually seek the data increases. So, we can correct that by doing what is called a defragmentation process. So, by defragmentation what you do is your data which is got distributed all over the disk you try to bring them together again to logically continuous physically contiguous blocks so, that their access time can improve. So, databases often will periodically defragment the file system.

(Refer Slide Time: 16:06)



Optimization of Disk Block Access (Cont.)

- **Nonvolatile write buffers** speed up disk writes by writing blocks to a non-volatile RAM buffer immediately
 - Non-volatile RAM: battery backed up RAM or flash memory
 - ▶ Even if power fails, the data is safe and will be written to disk when power returns
 - Controller then writes to disk whenever the disk has no other requests or request has been pending for some time
 - Database operations that require data to be safely stored before continuing can continue without waiting for data to be written to disk
 - Writes can be reordered to minimize disk arm movement
- **Log disk** – a disk devoted to writing a sequential log of block updates
 - Used exactly like nonvolatile RAM
 - ▶ Write to log disk is very fast since no seeks are required
 - ▶ No need for special hardware (NV-RAM)
 - File systems typically reorder writes to disk to improve performance
- **Journaling file systems** write data in safe order to NV-RAM or log disk
- Reordering without journaling: risk of corruption of file system data

SWAYAM: NPTEL-NOC IODCC, Instructor: Prof. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.24 ©Silberschatz, Korth and Sudarshan

The other way look at the optimize block access is by using buffers. So, idea of the buffer is suppose you want to write some data to the disk. So, naturally for writing the data also you will need a seek time you will need the rotational latency then we will have to data do the data transfer.

So, if you are trying to do some write you have you can take another option that you actually write that to a buffer a buffer which is in the memory in the it is a in the a semiconductor buffer where you can very quickly write and then when you have enough data in the buffer then you can take them in a single go to the disk.

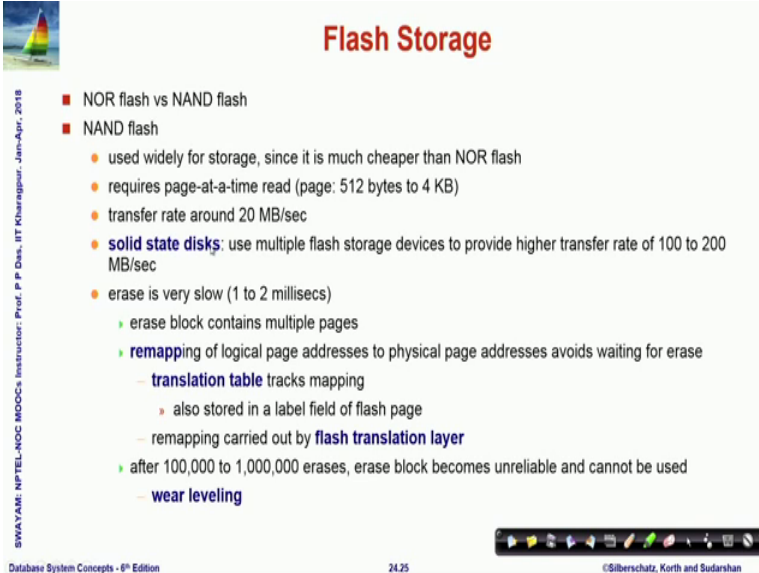
And write that or when the disk is not actually doing something some access you can use those cycles to write that now naturally if you write things onto the buffer then it is

possible that while your buffer has some data which has actually not been written to the disk if the power fails then you will lose that data. So, parts of the data which you think have been written actually have not gone to the disk. So, to take care of that often non volatile memory ram are used which are battery backed up or flash memory.

So, that even if power fails the right buffers will not be lost. So, we say that use non volatile buffers and other option that is used often is you maintain a log disk a log disk is nothing, but when you are writing to the disk you make a sequential log of the updates that you are doing. So, this kind of is a report. So, you are saying that I have written this data to this block written this data to this block.

So, if there is a failure, then if you can go through the log and you can actually retrieve the information of what was lost how far it actually worked correctly and you can used exactly like the non volatile ram and using the log you can find out. So, you are keeping a kind of a general link system you are keeping information of this is what I did this is what I did. So, the all this write information are maintained in terms of the log.

(Refer Slide Time: 18:30)



Flash Storage

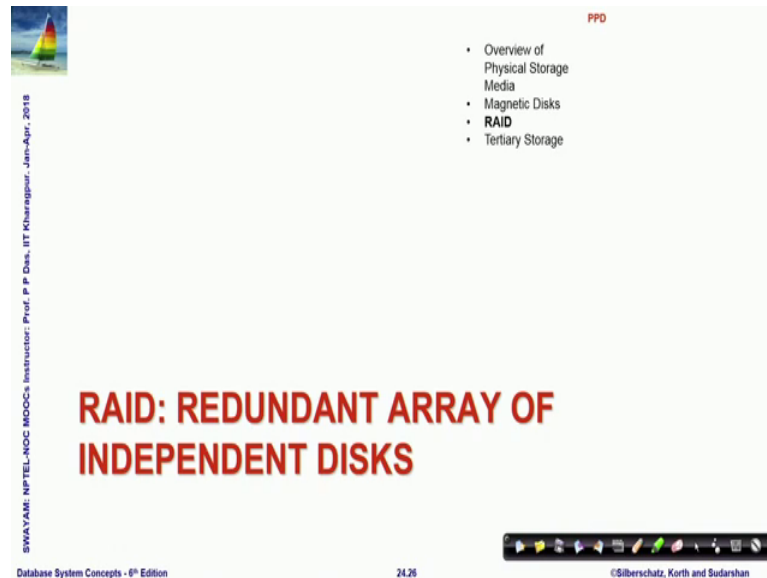
- NOR flash vs NAND flash
- NAND flash
 - used widely for storage, since it is much cheaper than NOR flash
 - requires page-at-a-time read (page: 512 bytes to 4 KB)
 - transfer rate around 20 MB/sec
 - **solid state disks**: use multiple flash storage devices to provide higher transfer rate of 100 to 200 MB/sec
 - erase is very slow (1 to 2 millisecs)
 - erase block contains multiple pages
 - **remapping** of logical page addresses to physical page addresses avoids waiting for erase
 - **translation table** tracks mapping
 - also stored in a label field of flash page
 - remapping carried out by **flash translation layer**
 - after 100,000 to 1,000,000 erases, erase block becomes unreliable and cannot be used
 - **wear leveling**

SWAYAM: NPTEL-NOC MOOCs Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 6th Edition 24.25 ©Silberschatz, Korth and Sudarshan

You can use flash storage nowadays the NAND based flash storage is are very common.

(Refer Slide Time: 18:40)



PPD

- Overview of Physical Storage Media
- Magnetic Disks
- **RAID**
- Tertiary Storage

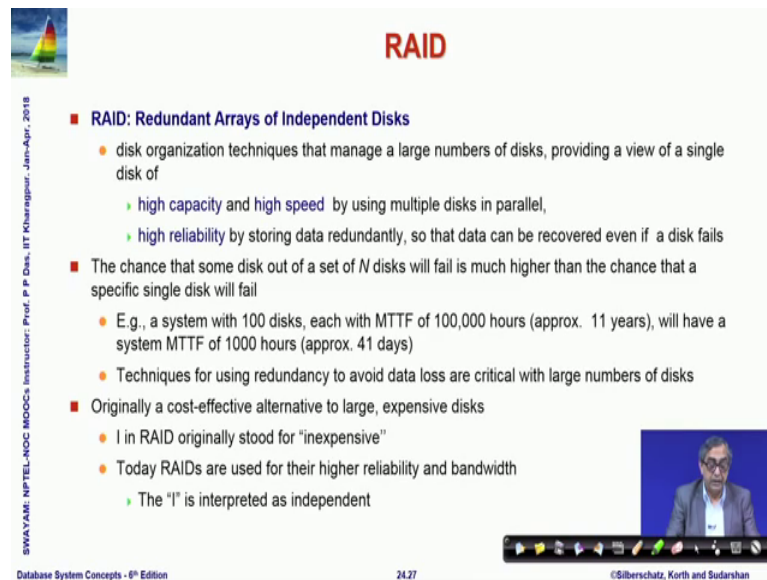
RAID: REDUNDANT ARRAY OF INDEPENDENT DISKS

SWAYAM: NPTEL-NOC MOOC's Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 6th Edition 24.26 ©Silberschatz, Korth and Sudarshan

So, here are some details on that let us move on to understanding. So, we just took a look into the basic storage hierarchy and the use of magnetic disks and what are the parameters that control the performance in terms of the read write in terms of the disk RAID is a the full form is redundant array of independent disks.

(Refer Slide Time: 19:05)



RAID

- **RAID: Redundant Arrays of Independent Disks**
 - disk organization techniques that manage a large numbers of disks, providing a view of a single disk of
 - high capacity and high speed by using multiple disks in parallel,
 - high reliability by storing data redundantly, so that data can be recovered even if a disk fails
- The chance that some disk out of a set of N disks will fail is much higher than the chance that a specific single disk will fail
 - E.g., a system with 100 disks, each with MTTF of 100,000 hours (approx. 11 years), will have a system MTTF of 1000 hours (approx. 41 days)
 - Techniques for using redundancy to avoid data loss are critical with large numbers of disks
- Originally a cost-effective alternative to large, expensive disks
 - I in RAID originally stood for "inexpensive"
 - Today RAID's are used for their higher reliability and bandwidth
 - The "I" is interpreted as independent

SWAYAM: NPTEL-NOC MOOC's Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018

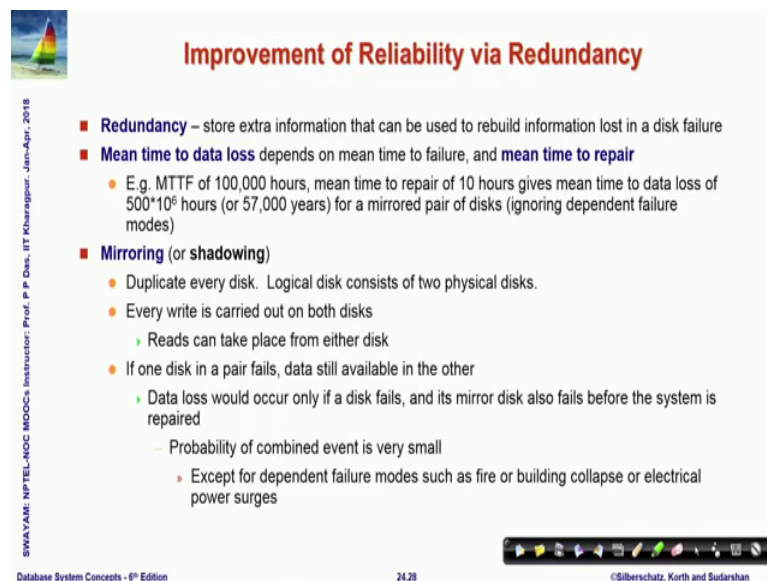
Database System Concepts - 6th Edition 24.27 ©Silberschatz, Korth and Sudarshan

So, the disk organization that manage a large number of disk, but the view that you get you do not get to see the multiple disk you get to see a single disk.

So, by this you can create very high capacity and very high speed and. So, because you have multiple disks so, you organize a data in such a way that when you are writing or you are reading actually you can parallelly read or write from multiple different disks. So, that not only increases capacity, but actually increases a throughput and you can get high reliability also by storing the data redundantly; that is keeping multiple copies and that is what RAID is often quite known for.

So, originally when RAID was originally designed actually the, I in red stood for inexpensive. So, it was kind of a disk array which was inexpensive to afford, but now it is not particularly the expenses is not the primary factor for which we do go for red, but we go for red for the high capacity high speed and high reliability. So, the I is now interpreted as independent array.

(Refer Slide Time: 20:21)



Improvement of Reliability via Redundancy

- **Redundancy** – store extra information that can be used to rebuild information lost in a disk failure
- **Mean time to data loss** depends on mean time to failure, and **mean time to repair**
 - E.g. MTTF of 100,000 hours, mean time to repair of 10 hours gives mean time to data loss of 500×10^6 hours (or 57,000 years) for a mirrored pair of disks (ignoring dependent failure modes)
- **Mirroring (or shadowing)**
 - Duplicate every disk. Logical disk consists of two physical disks.
 - Every write is carried out on both disks
 - Reads can take place from either disk
 - If one disk in a pair fails, data still available in the other
 - ▶ Data loss would occur only if a disk fails, and its mirror disk also fails before the system is repaired
 - Probability of combined event is very small
 - ▶ Except for dependent failure modes such as fire or building collapse or electrical power surges

SWAYAM: NPTEL-NOC MOCs- Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.28 ©Silberschatz, Korth and Sudarshan

So, we can improve. So, now, the main issue in red is to improve reliability. So, the main the key idea is have redundancy to improve the reliability that is stored the data in multiple copies and can rebuild from the copies once a disk has failed. So, we earlier looked at the MTTF mean time to failure. Now, we are look at another parameter which we say is a mean time to data loss. So, which depends on mean time to failure, because if you are if you are failed then you have lost the data, but when you have failed you have a possibility now, to recover the data to repair it; because you have redundant copies.

So, mean time to data loss it depends on the MTTF plus the mean time to repair how quickly can we use your redundant copies and get back the original data. So, mean time to data loss is the actual key factor which needs to be minimized and for this red does a mirroring or shadowing that is for every disk there is a duplicate there is a clone.

So, it the logically you have one disk, but physically you have actually two disk. So, every write that you do is carried out to both the disks and when you read the read happens on either of the disk usually it it switches between the disks. So, if one of the disk in the pair would fail, then the data can still be recovered from the other and the data loss would occur if a disk fail, but the mediate disk also has to fail before the system has been repair.

So, if one of the disk fail you get the data from the middle disk you continue to do that from the mirror disk you restore the other disk you may be replace and put a new one mirror it again and. So, on, but you can restore that. So, in between this time if the mirror disk also fails; then you have actually lost data, but the probability of that is very very small. So, mirroring gives a at the expense of naturally having lot more of redundant storage the mirroring can actually give you a much higher reliability.

(Refer Slide Time: 22:39)

The slide is titled "Improvement of Reliability via Redundancy" and contains the following content:

- **Bit-level striping** – split the bits of each byte across multiple disks
 - In an array of eight disks, write bit i of each byte to disk i
 - Each access can read data at eight times the rate of a single disk
 - But seek/access time worse than for a single disk
 - Bit level striping is not used much any more
- **Block-level striping** – with n disks, block i of a file goes to disk $(i \bmod n) + 1$
 - Requests for different blocks can run in parallel if the blocks reside on different disks
 - A request for a long sequence of blocks can utilize all disks in parallel

SWAYAM: NPTEL-NOC MOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 6th Edition 24.29 ©Silberschatz, Korth and Sudarshan

That we expect today another some of the other techniques which improve reliability is bit level and byte level block level striping techniques. So, what you in the basic bit level striping what you do is a say every byte has 8 bits. So, when you are writing a byte you

do not write all the bytes to the same disk you write them to multiple disks. So, you take an array of 8 disks. So, write bit I of each byte to disk I it is. So, did interesting concept you have fragmenting it in a very peculiar way.

So, you have 8 disks and every byte first byte first bit is written to one disk second byte is second bit is written to the second disk, third bit is written to the third disk and so, on. And when you access you can access from all these 8; now naturally which means that this will decrease your throughput to some extent, because you have to collect from all of that reconstruct. So, bits levels striping is not much in use any more instead you have block level striping where with end disk a block I of a file goes to the disk $i \bmod n + 1$.

So, circularly so, the first goes if you have say five disks in the first block goes to disk one second to disk two-fifth to disk 5 and the 6th again back to disk 1. So, the request to different blocks can run in parallel and reside on different disk and if they can easily utilize this parallelism to improve the throughput.

(Refer Slide Time: 24:30)

The slide is titled "Improvement of Reliability via Redundancy" and contains the following content:

- Bit-Interleaved Parity** – a single parity bit is enough for error correction, not just detection, since we know which disk has failed
 - When writing data, corresponding parity bits must also be computed and written to a parity bit disk
 - To recover data in a damaged disk, compute XOR of bits from other disks (including parity bit disk)
- Block-Interleaved Parity**: Uses block-level striping, and keeps a parity block on a separate disk for corresponding blocks from N other disks
 - When writing data block, corresponding block of parity bits must also be computed and written to parity disk
 - To find value of a damaged block, compute XOR of bits from corresponding blocks (including parity block) from other disks.

SWAYAM: NPTEL-NOC MOOCs Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-April, 2018

Database System Concepts - 6th Edition 24.30 ©Silberschatz, Korth and Sudarshan

At the same time improve the reliability now how do you improve the reliability. So, for improving the reliability you use the basic you know error correcting coding concept you just use a bit level that again two options one is you can do a bit inter lift parity which means a single parity bit is used which is good for error correction. Usually, we know that if we use a single parity bit we can know a single error we can detect a single error,

but here with a single bit you can correct the single error also because you know in case of a failure you know which particular disk has failed. So, then you can exalt with a data from the other disk and reconstruct the error bit.

So, the other is naturally block inter leaving of the parity which uses block level striping and keeps a parity block on a separate disk for corresponding blocks from n other disks and you can reconstruct in a very similar manner. So, by using block interleaved parity with block striping you can really have a higher throughput with a better reliability.

(Refer Slide Time: 25:44)

Choice of RAID Level

- Factors in choosing RAID level
 - Monetary cost
 - Performance: Number of I/O operations per second, and bandwidth during normal operation
 - Performance during failure
 - Performance during rebuild of failed disk
 - Including time taken to rebuild failed disk
- RAID 0 is used only when data safety is not important
 - E.g. data can be recovered quickly from other sources
- Level 2 and 4 never used since they are subsumed by 3 and 5
- Level 3 is not used anymore since bit-striping forces single block reads to access all disks, wasting disk arm movement, which block striping (level 5) avoids
- Level 6 is rarely used since levels 1 and 5 offer adequate safety for most applications

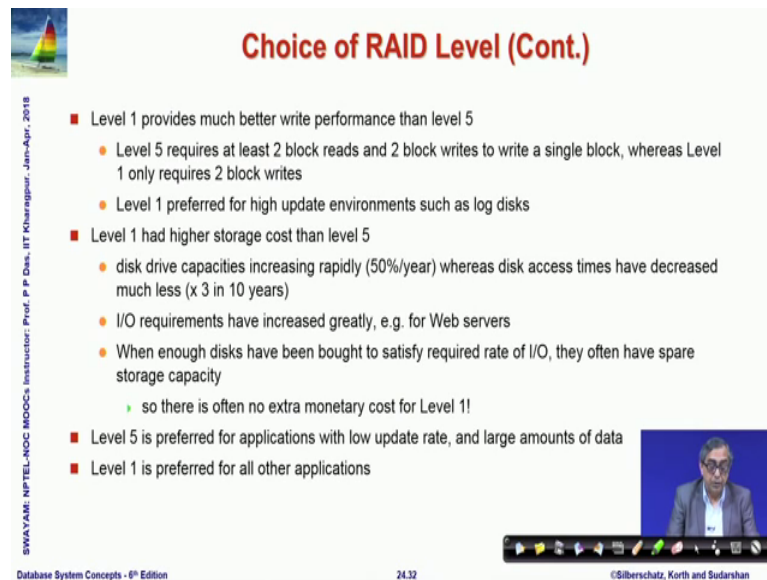
SWAYAM: NPTEL-NOC MOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 6th Edition 24.31 ©Silberschatz, Korth and Sudarshan

So, it is a win situation now naturally when you go for red you will find that there are different levels of read that are available today goes up to level 6 right. Now, and the factors that certainly has to be considered is I mean different RAID at different kind of cost the what is the performance what is the performance during failure; that is performance during failure is typically the MTTF and performance during rebuilt is a mean time to data loss so, including the rebuilding and so, on. So, based on these factors different rate levels can be looked at RAID 0 is used only when data safety is not important. So, that is not very common the RAID level 2 and 4 are never used.

Because, they are they got subsumed in level 3 and 5. So, you can ignore that level three is also not used anymore, because it used bits striping and naturally we talked of that that is not that is words compare to the block striping which level 5 uses and level 6 is also really used. Since level 1 and 5 adequately support all applications.

(Refer Slide Time: 27:08)



Choice of RAID Level (Cont.)

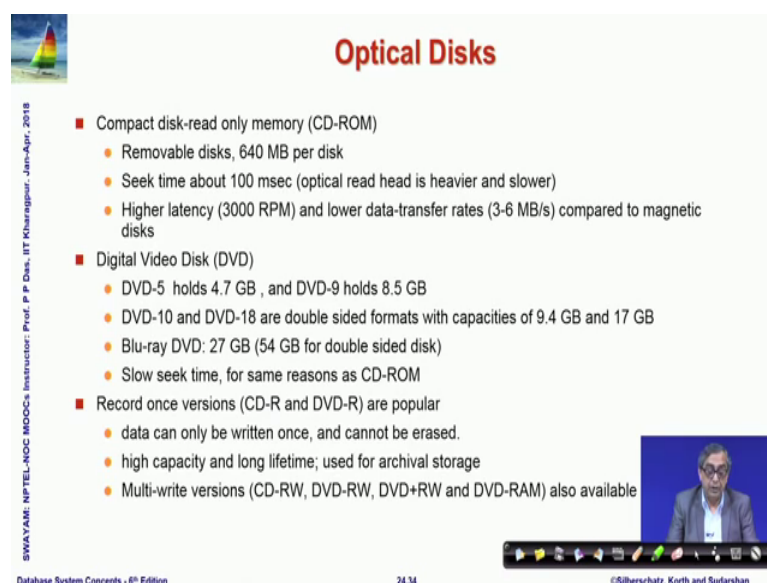
- Level 1 provides much better write performance than level 5
 - Level 5 requires at least 2 block reads and 2 block writes to write a single block, whereas Level 1 only requires 2 block writes
 - Level 1 preferred for high update environments such as log disks
- Level 1 had higher storage cost than level 5
 - disk drive capacities increasing rapidly (50%/year) whereas disk access times have decreased much less (x 3 in 10 years)
 - I/O requirements have increased greatly, e.g. for Web servers
 - When enough disks have been bought to satisfy required rate of I/O, they often have spare storage capacity
 - ▶ so there is often no extra monetary cost for Level 1!
- Level 5 is preferred for applications with low update rate, and large amounts of data
- Level 1 is preferred for all other applications

SWAYAM: NPTEL-NOC MOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 8th Edition 24.32 ©Silberschatz, Korth and Sudarshan

So, the conclusion simply is that you either use RAID level 1 or you use RAID level 5. So, RAID level 1 gives a better right performance, than RAID 5 and it is certainly level 1. So, therefore, level 1 is preferred for high update environments such as log disks and so, on whereas, level one has higher storage cost than 5 also and level 5 is preferred for applications that has low update rate and large volume of data. So, if you have very high update you go for level one rate. So, which will give you which will cost you more, but if you have a level 5, then you will be able to get a low update, but have large amount of data stored reliably with less amount of money invested into that.

(Refer Slide Time: 28:15)



Optical Disks

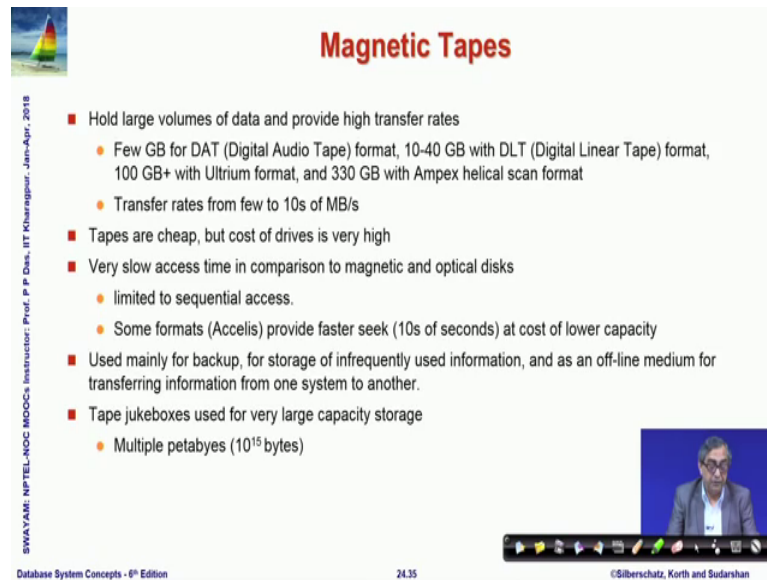
- Compact disk-read only memory (CD-ROM)
 - Removable disks, 640 MB per disk
 - Seek time about 100 msec (optical read head is heavier and slower)
 - Higher latency (3000 RPM) and lower data-transfer rates (3-6 MB/s) compared to magnetic disks
- Digital Video Disk (DVD)
 - DVD-5 holds 4.7 GB, and DVD-9 holds 8.5 GB
 - DVD-10 and DVD-18 are double sided formats with capacities of 9.4 GB and 17 GB
 - Blu-ray DVD: 27 GB (54 GB for double sided disk)
 - Slow seek time, for same reasons as CD-ROM
- Record once versions (CD-R and DVD-R) are popular
 - data can only be written once, and cannot be erased.
 - high capacity and long lifetime; used for archival storage
 - Multi-write versions (CD-RW, DVD-RW, DVD+RW and DVD-RAM) also available

SWAYAM: NPTEL-NOC MOCs Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 8th Edition 24.34 ©Silberschatz, Korth and Sudarshan

And so, these are the RAID levels then of course, finally there are different tertiary storages compact disk CD-ROM we all are familiar that DVD the record once versions where you just record and use that particularly for different kind of distribution these.

(Refer Slide Time: 28:34)



Magnetic Tapes

- Hold large volumes of data and provide high transfer rates
 - Few GB for DAT (Digital Audio Tape) format, 10-40 GB with DLT (Digital Linear Tape) format, 100 GB+ with Ultrium format, and 330 GB with Ampex helical scan format
 - Transfer rates from few to 10s of MB/s
- Tapes are cheap, but cost of drives is very high
- Very slow access time in comparison to magnetic and optical disks
 - limited to sequential access.
 - Some formats (Accelis) provide faster seek (10s of seconds) at cost of lower capacity
- Used mainly for backup, for storage of infrequently used information, and as an off-line medium for transferring information from one system to another.
- Tape jukeboxes used for very large capacity storage
 - Multiple petabytes (10^{15} bytes)

SWAYAM: NPTEL-NOC MOOCs Instructor: Prof. P. P. Das, IIT Kharagpur, Jan-Apr, 2018

Database System Concepts - 9th Edition 24.35 ©Silberschatz, Korth and Sudarshan

storages are used there are magnetic tapes which are very large volume and provide a high transfer rate and they are go currently they go into different couple of orders of terabytes even petabytes in size, but the tapes are not really very expensive. So, we can use them to hold really really large databases they are good for Backups and so, on, but the tape drives are quite expensive. So, that will have to keep in mind.

(Refer Slide Time: 29:08)

Module Summary

- Looked at the options of Physical Storage Media for high volume, fast, reliable and inexpensive options for data storage for databases
- Understood the structure and basic functionality of Magnetic Disks
- Understood RAID – array of redundant disks in parallel to enhance speed and reliability
- Understood the options of Tertiary Storage for high volume, inexpensive backup options

SWAYAM: NPTEL-NOC MDOCS Instructor: Prof. P. P. Das, IIT Khargpur, Jan-Apr, 2018

Database System Concepts - 8th Edition

24.36

©Silberschatz, Korth and Sudarshan

So, to summarize we have looked at different physical storage media. So, this was I mean besides all the details in the discussion the key take way point for us here is to understand that primarily.

We have a memory which is expensive and which is small in size very high speed. So, all operations that we need that need to be done we will finally, have to happen when the once the data is in memory and on the other side we have all the persistent data in a in some kind of a magnetic disk in certain structure and that is that can support large storage it is persistent it is reliable, but it is relatively slow to access and so, it needs to be used in a intelligent manner and one point that you have specifically noted that there is some unit for every disk system.

There is some unit called a block or disk block, which is a basic unit of data that will be transferred every time you access the disk. So, you are if your design is aligned with a size of the disk block which is couple of kilobytes then it will be easier to be able to design more optimal physical storage for your files and you can speed up the whole process of search and update that you want to do in the database.