Database Management Systems

ICT1212

Introduction to Disk Storage and File Structures

Department of ICT
Faculty of Technology
University of Ruhuna

Lecture 10

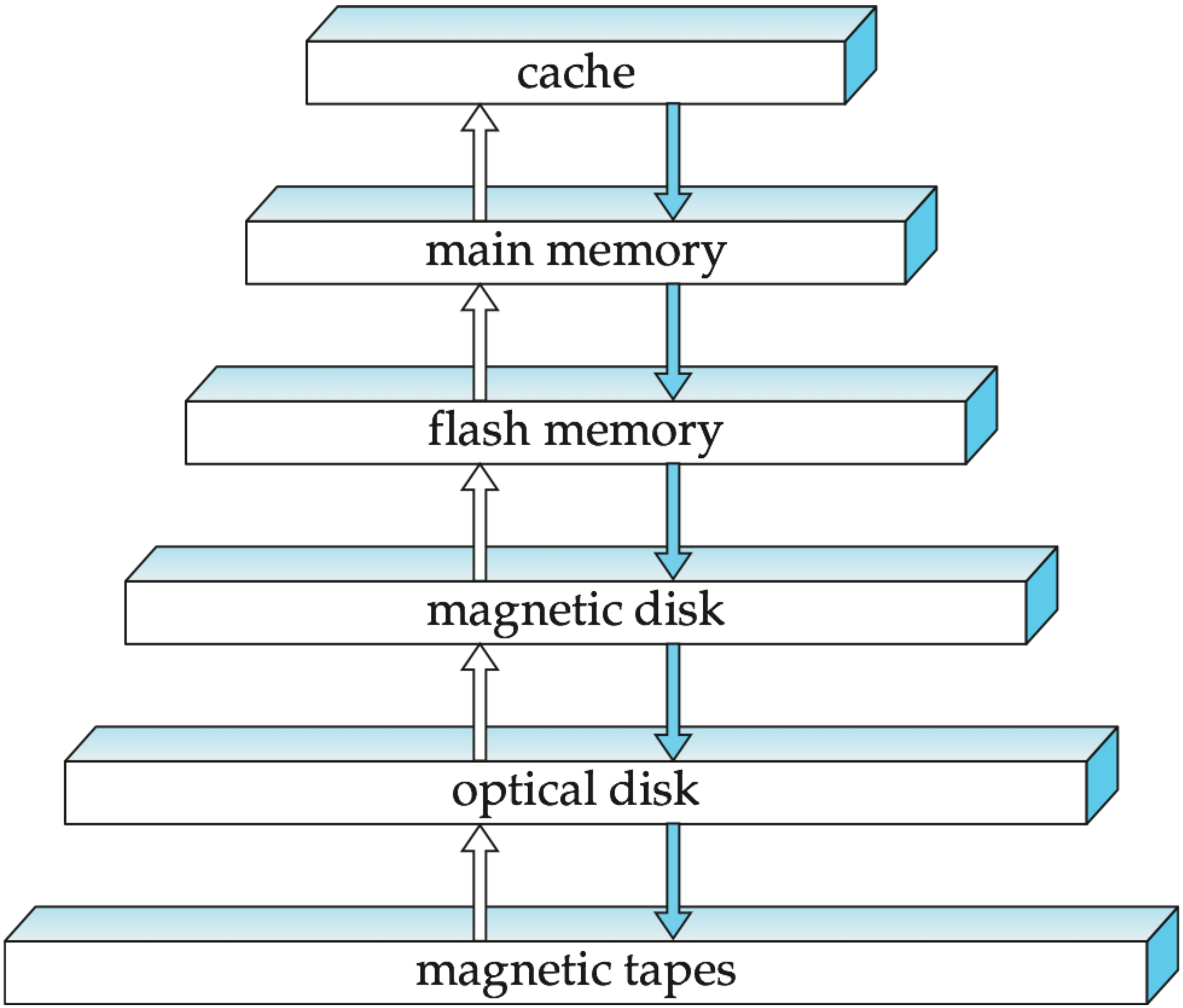
Chapter Outline

- Overview of Physical Storage Media
- Magnetic Disks
- RAID Technology
- Tertiary Storage
- Storage Access
- File Organization
- Organization of Records in Files
- Data-Dictionary Storage

Classification of Physical Storage Media

- 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.

Storage Hierarchy



Storage Hierarchy

- **Primary Storage:**
  - Fastest media but volatile
    - cache, main memory
- **Secondary Storage:**
  - Non-volatile, moderately fast access time
  - **on-line storage**
    - flash memory, magnetic disks
- **Tertiary Storage:**
  - Non-volatile, slow access time
  - **off-line storage**
    - magnetic tape, optical storage

Physical Storage Media

- **Cache**
  - fastest and most costly form of storage
  - Volatile
  - managed by the computer system hardware
- **Main memory**:
  - fast access
  - generally, too small (or too expensive) to store the entire database
  - Volatile

Physical Storage Media

- **Flash memory**
  - Non - Volatile
  - Data can be written at a location only once, but location can be erased and written to again
  - Reads are roughly as fast as main memory
  - But writes are slow
    - erase is slower
  - Widely used in embedded devices
    - digital cameras, phones etc

Physical Storage Media

- **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
  - Larger Capacities
    - larger capacity and cost/byte less than main memory/flash memory
  - Non-Volatile
    - Survives power failures and system crashes
    - disk failure can destroy data, but is rare

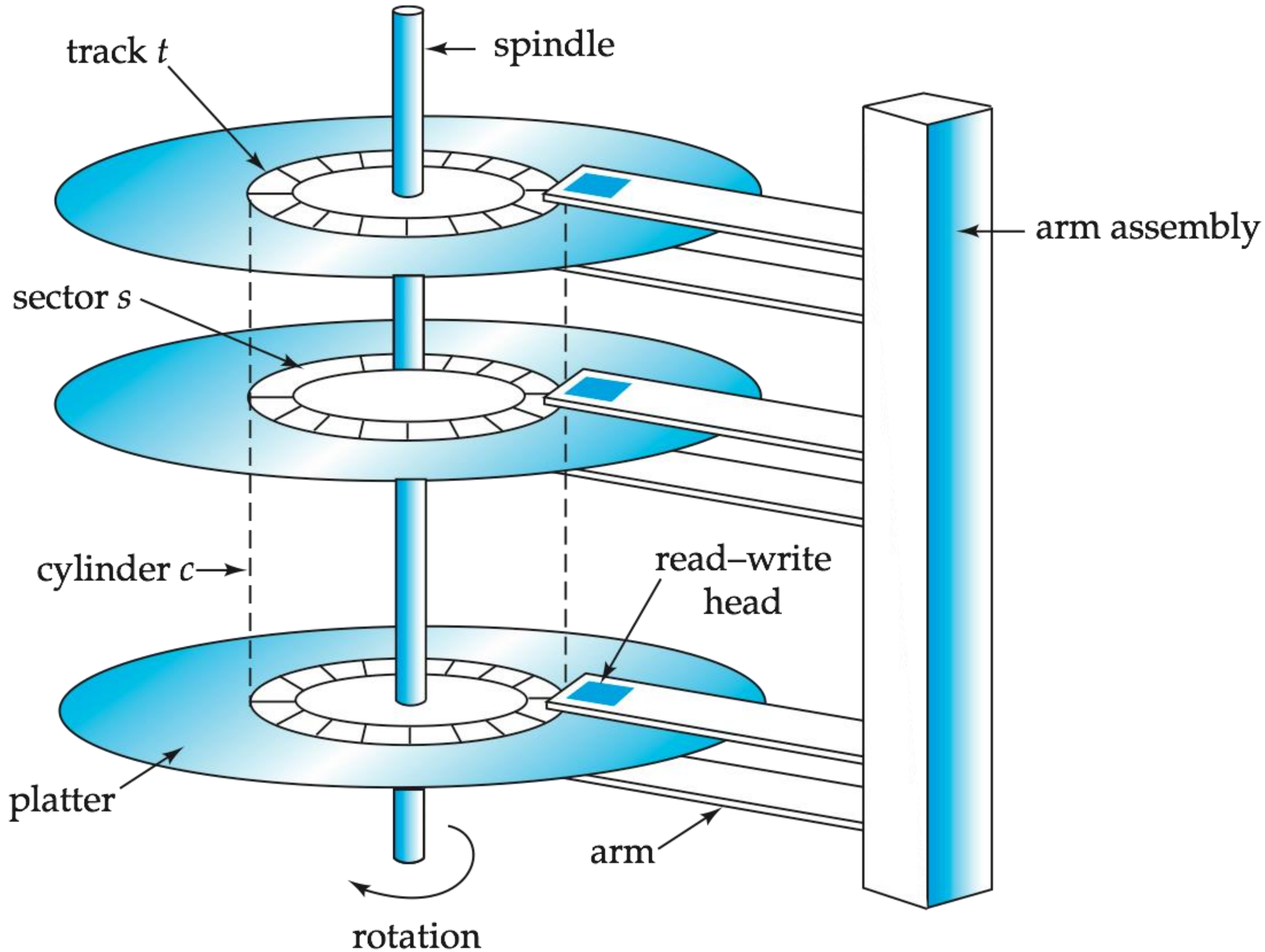
Physical Storage Media

- **Optical storage**
  - Non-Volatile
  - Data is read optically from a spinning disk using a laser
    - CD-ROM, DVD, Blu-ray disks
    - CD-R, DVD-R, DVD+R used for archival storage
    - CD-RW, DVD-RW etc
  - Reads and writes are slower than with magnetic disk

Physical Storage Media

- **Tape storage**
  - Non-Volatile
  - Used primarily for backup and archival data
  - sequential-access
    - much slower than disk
  - very high capacity
  - storage costs much cheaper than disk, but drives are expensive
  - Tape jukeboxes available for storing massive amounts of data

Magnetic Hard Disk Mechanism



simplified structure of a hard disk drive

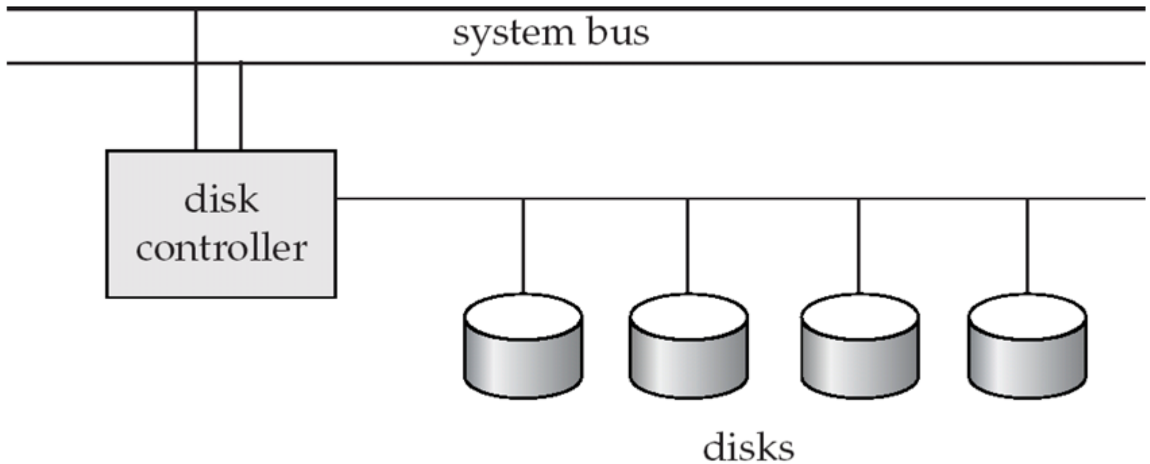
Magnetic Disks

- **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**
- Each track is divided into **sectors.**
  - A sector is the smallest unit of data that can be read or written.
- Head-disk assemblies
  - multiple disk platters on a single spindle
  - one head per platter, mounted on a common arm.
- **Cylinder** *i* consists of i^{th} track of all the platters

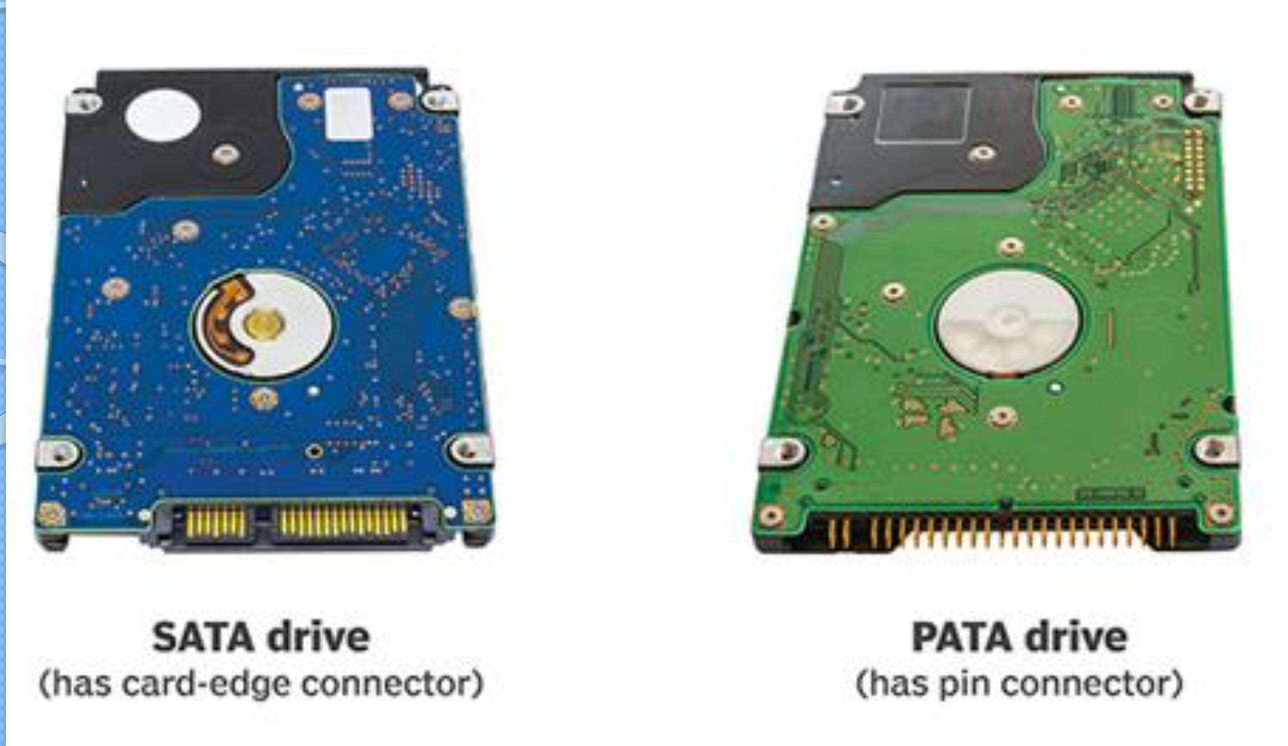
Magnetic Disks

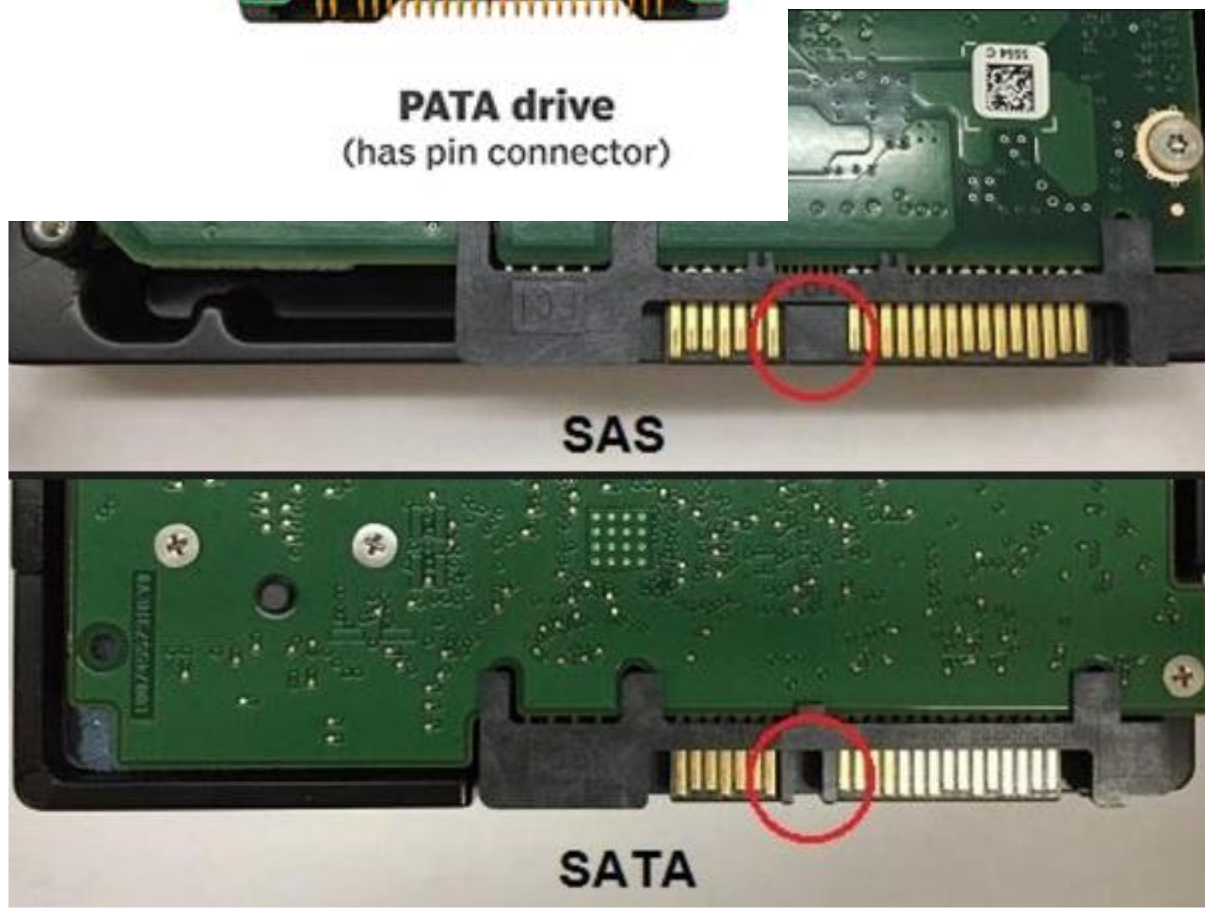
- Earlier generation disks were susceptible to head-crashes
- **Disk controller**
  - Interfaces between the computer system and the disk drive hardware.
  - Computes and attaches **checksums** to each sector to verify that data is read back correctly
  - Ensures successful writing by reading back sector after writing it
  - Performs remapping of bad sectors

Disk Subsystem



- Multiple disks connected to a computer system through a controller
- Disk interface standards families
  - ATA (AT adaptor) range of standards
    - PATA
  - SATA (Serial ATA)
  - SCSI (Small Computer System Interconnect) range of standards
  - SAS (Serial Attached SCSI)
  - Several variants of each standard


SATA drive
(has card-edge connector)
PATA drive
(has pin connector)


SAS
SATA

Disk Subsystem

- Disks usually connected directly to computer system
- **Directly Attached Storage (DAS)**
- **Network Attached Storage (NAS)**
- **Storage Area Networks (SAN)**

Performance Measures of Disks

- **Access time**
  - the time it takes from when a read or write request is issued to when data transfer begins
  - **Seek time**
    - time it takes to reposition the arm over the correct track.
  - **Rotational latency**
    - time it takes for the sector to be accessed to appear under the head.
- **Data-transfer rate**
  - the rate at which data can be retrieved from or stored to the disk.

Performance Measures

- **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
  - MTTF decreases as disk ages

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
- **Disk-arm-scheduling**
  - algorithms order pending accesses to tracks so that disk arm movement is minimized

Optimization of Disk Block Access

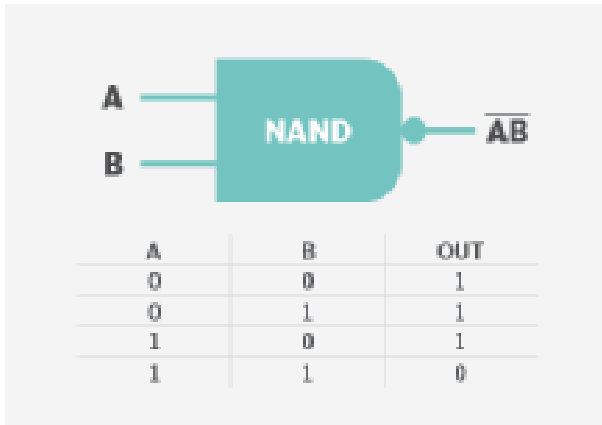
- **File organization**
  - optimize block access time by organizing the blocks to correspond to how data will be accessed
    - Ex: Store related information on the same or nearby cylinders.
  - Files may get **fragmented** over time
    - E.x: if data is inserted to/deleted from the file
    - Or free 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

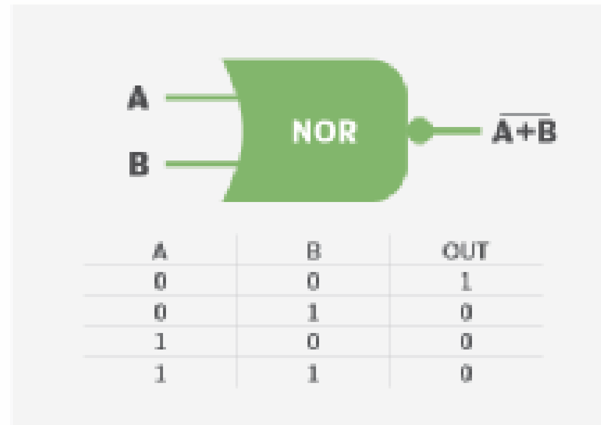
Optimization of Disk Block Access

- **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
  - *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

Flash Storage

- NOR flash vs NAND flash



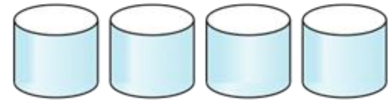


- NAND flash
  - used widely for storage, since it is much cheaper than NOR flash
  - erase is very slow
  - **solid state disks**:
    - use multiple flash storage devices to provide higher transfer rate of 100 to 200 MB/sec

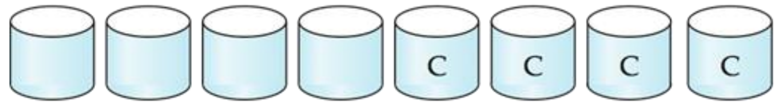
RAID

- **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
- Originally a cost-effective alternative to large, expensive disks
  - I in RAID originally stood for ``inexpensive"
  - Today RAIDs are used for their higher reliability and bandwidth.
    - The "I" is interpreted as independent

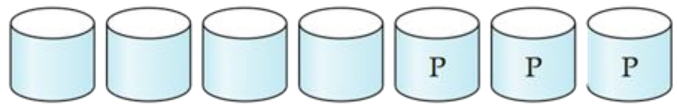
RAID



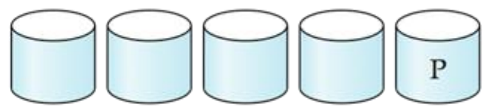
(a) RAID 0: nonredundant striping



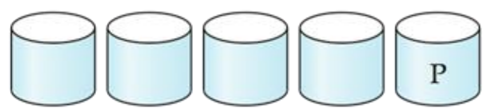
(b) RAID 1: mirrored disks



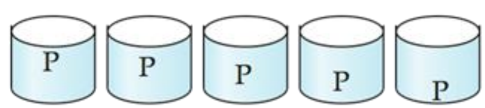
(c) RAID 2: memory-style error-correcting codes



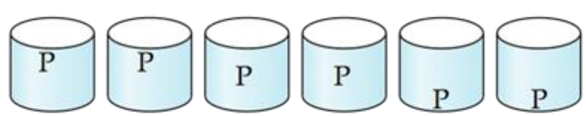
(d) RAID 3: bit-interleaved parity



(e) RAID 4: block-interleaved parity



(f) RAID 5: block-interleaved distributed parity



(g) RAID 6: P + Q redundancy

RAID

LEVEL 6	Independent Data Disks with Double Parity
LEVEL 5	Block Interleaved Distributed Parity
LEVEL 4	Dedicated Parity Drive
LEVEL 3	Bit-Interleaved Parity
LEVEL 2	Error-Correcting Coding
LEVEL 1	Mirroring and Duplexing
LEVEL 0	Striped Disk Array without Fault Tolerance

Choice of RAID Level

- Factors in choosing RAID level
  - Monetary cost
  - Performance:
  - Performance during failure
  - Performance during rebuild of failed disk
- RAID 0 is used only when data safety is not important
- Level 2 and 4 never used since they are subsumed by 3 and 5
- Level 3 is not used anymore
- Level 6 is rarely used since levels 1 and 5 offer adequate safety for most applications

Hardware Issues

- **Software RAID**
  - RAID implementations done entirely in software, with no special hardware support
- **Hardware RAID**
  - RAID implementations with special hardware
  - Use non-volatile RAM to record writes that are being executed
  - Power failure during write can result in corrupted disk

Hardware Issues

- **Latent failures**
  - data successfully written earlier gets damaged
  - can result in data loss even if only one disk fails
- **Data scrubbing**
  - continually scan for latent failures, and recover from copy/parity
- **Hot swapping**
  - replacement of disk while system is running, without power down
- Many systems maintain spare disks which are kept online, and used as replacements for failed disks immediately on detection of failure
- Many hardware RAID systems ensure that a single point of failure will not stop the functioning of the system by using

Homework

- Study more about RAID technology

File Organization

- The database is stored as a collection of *files*.
- Each file is a sequence of *records*.
- A record is a sequence of fields.
- One approach:
  - assume record size is fixed
  - each file has records of one particular type only
  - different files are used for different relations

Fixed-Length Records

- Simple approach:

record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

Deleting record 3 and compacting

record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

Deleting record 3 and moving last record

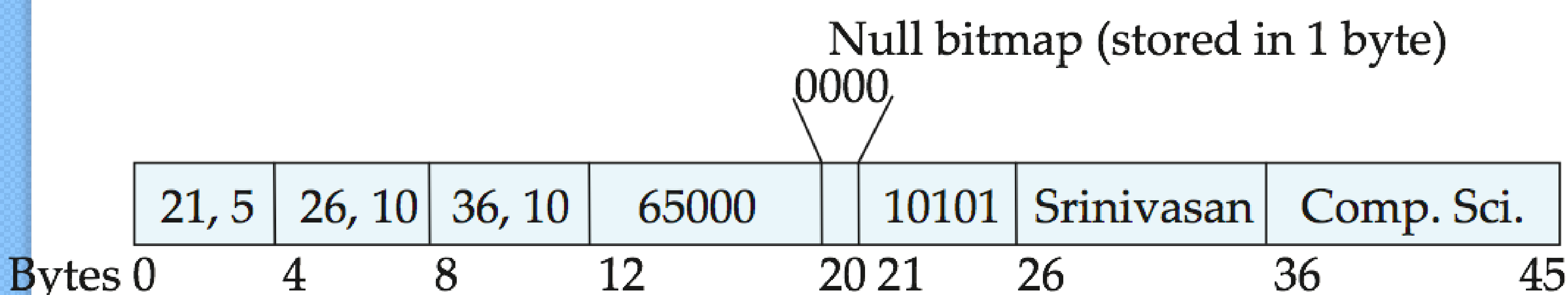
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 11	98345	Kim	Elec. Eng.	80000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000

Free Lists

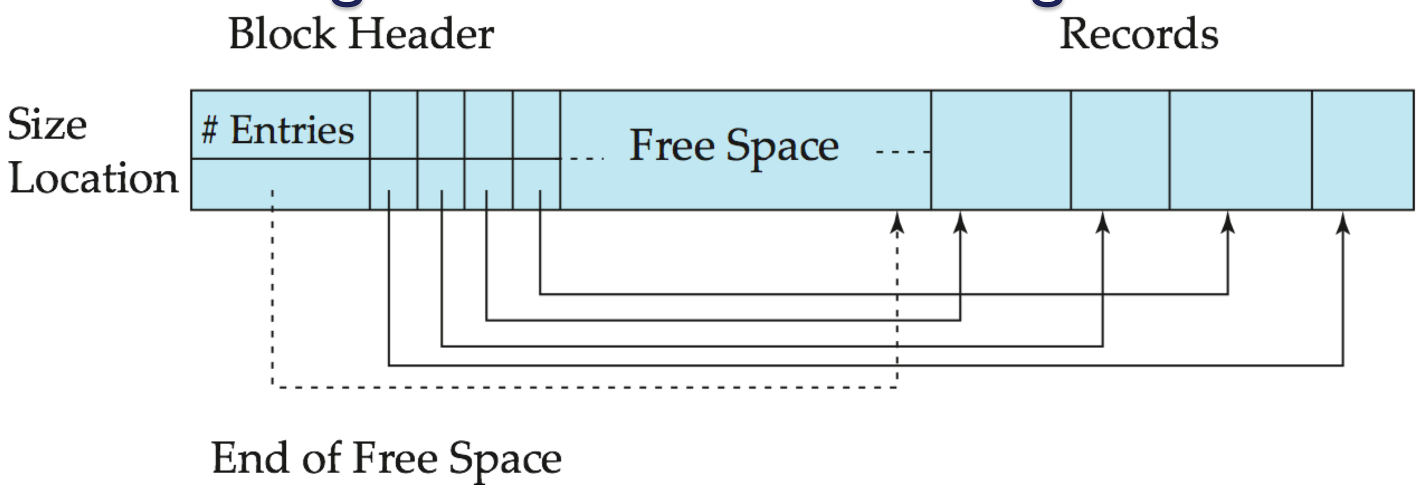
header				
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1				
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4				
record 5	33456	Gold	Physics	87000
record 6				
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	72000
record 10	83821	Brandt	Comp. Sci.	92000
record 11	98345	Kim	Elec. Eng.	80000

Variable-Length Records

- Variable-length records arise in database systems in several ways:
  - Storage of multiple record types in a file.
  - Record types that allow variable lengths for one or more fields such as strings (**varchar**)
  - Record types that allow repeating fields
- Attributes are stored in order
- Variable length attributes represented by fixed size (offset, length), with actual data stored after all fixed length attributes
- Null values represented by null-value bitmap



Variable-Length Records: Slotted Page Structure



- **Slotted page** header contains:
  - number of record entries
  - end of free space in the block
  - location and size of each record
- Records can be moved around within a page to keep them contiguous with no empty space between them; entry in the header must be updated.
- Pointers should not point directly to record — instead they should point to the entry for the record in header.

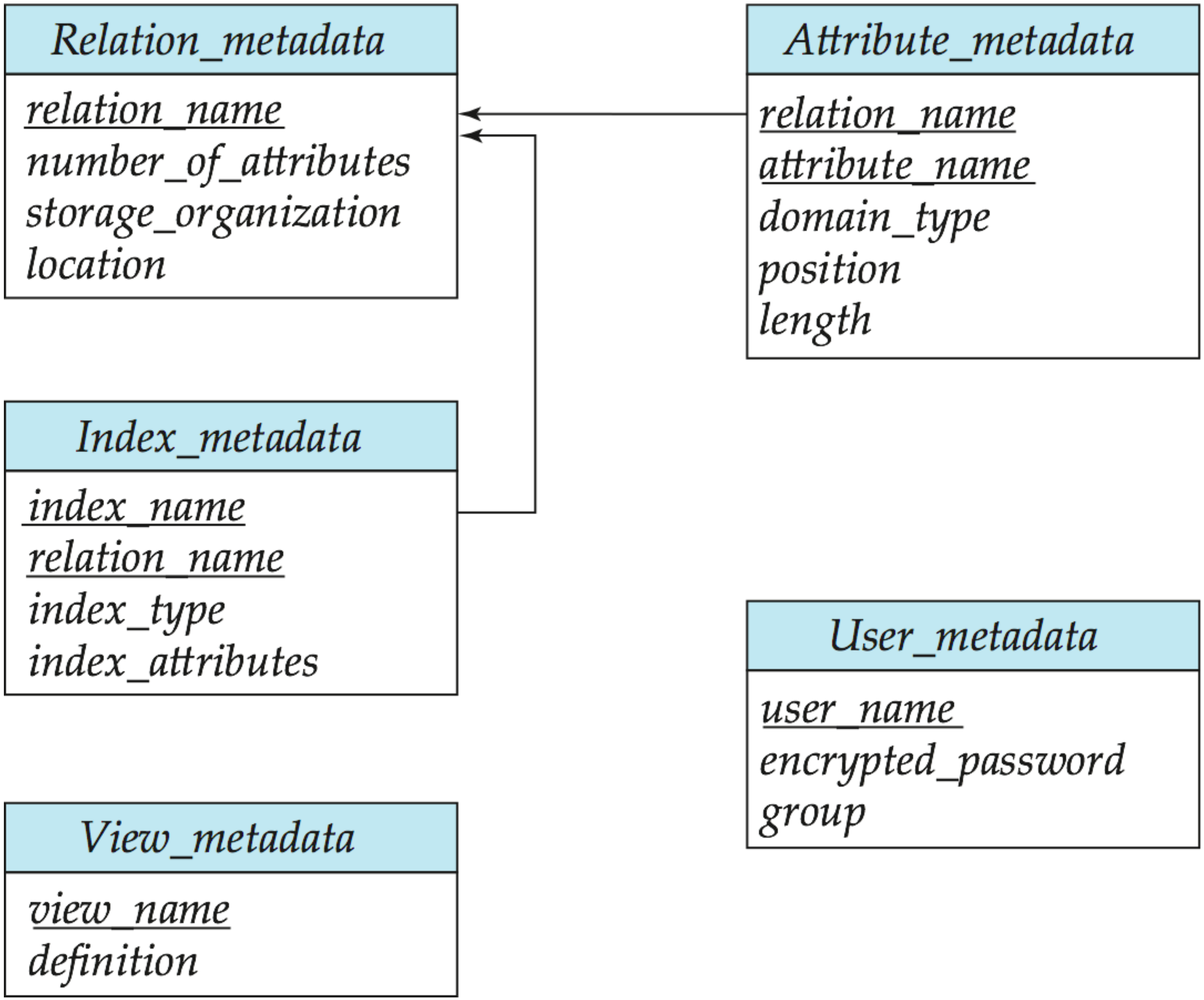
Organization of Records in Files

- **Heap**
  - a record can be placed anywhere in the file where there is space
- **Sequential**
  - store records in sequential order, based on the value of the search key of each record
- **Hashing**
  - a hash function computed on some attribute of each record; the result specifies in which block of the file the record should be placed

Data Dictionary Storage

- **Data dictionary**
  - also called **system catalog**
- stores **metadata** such as
  - Information about relations
    - names of relations
    - names, types and lengths of attributes of each relation
    - names and definitions of views
    - integrity constraints
  - User and accounting information, including passwords
  - Statistical and descriptive data
    - number of tuples in each relation
  - Physical file organization information
    - How relation is stored (sequential/hash/…)
    - Physical location of relation
  - Information about indices

Relational Representation of System Metadata



Storage Access

- A database file is partitioned into fixed-length storage units called **blocks**.
  - Blocks are units of both storage allocation and data transfer.
- Database system seeks to minimize the number of block transfers between the disk and memory.
  - can reduce the number of disk accesses by keeping as many blocks as possible in main memory
- **Buffer**
  - portion of main memory available to store copies of disk blocks.
- **Buffer manager**
  - subsystem responsible for allocating buffer space in main memory.

Blocking

- Blocking: refers to storing a number of records in one blo ck on the disk.
- Blocking factor (*bfr*) refers to the number of records per block.
- There may be empty space in a block if an integral number of records do not fit in one block.
- *Spanned Records*: refer to records that exceed the size of one or more blocks and hence span a number of blocks.

Exercise

Consider a disk with a sector size of 512 bytes, 2000 tracks per surface, 50 sectors per track, five double-sided platters, and average seek time of 10 msec.

1. What is the capacity of a track in bytes? What is the capacity of each surface? What is the capacity of the disk?

2. How many cylinders does the disk have?

3. Give examples of valid block sizes. Is 256 bytes a valid block size? 2048? 51,200?

4. If the disk platters rotate at 5400 rpm (revolutions per minute), what is the maximum rotational delay?

5. If one track of data can be transferred per revolution, what is the transfer rate?

Summary

- Overview of Physical Storage Media
- Magnetic Disks
- RAID Technology
- Tertiary Storage
- Storage Access
- File Organization
- Organization of Records in Files
- Data-Dictionary Storage

References

- *Fundamentals of Database Systems*

(6th Edition) By Remez Elmasri & Shamkant B. Navathe