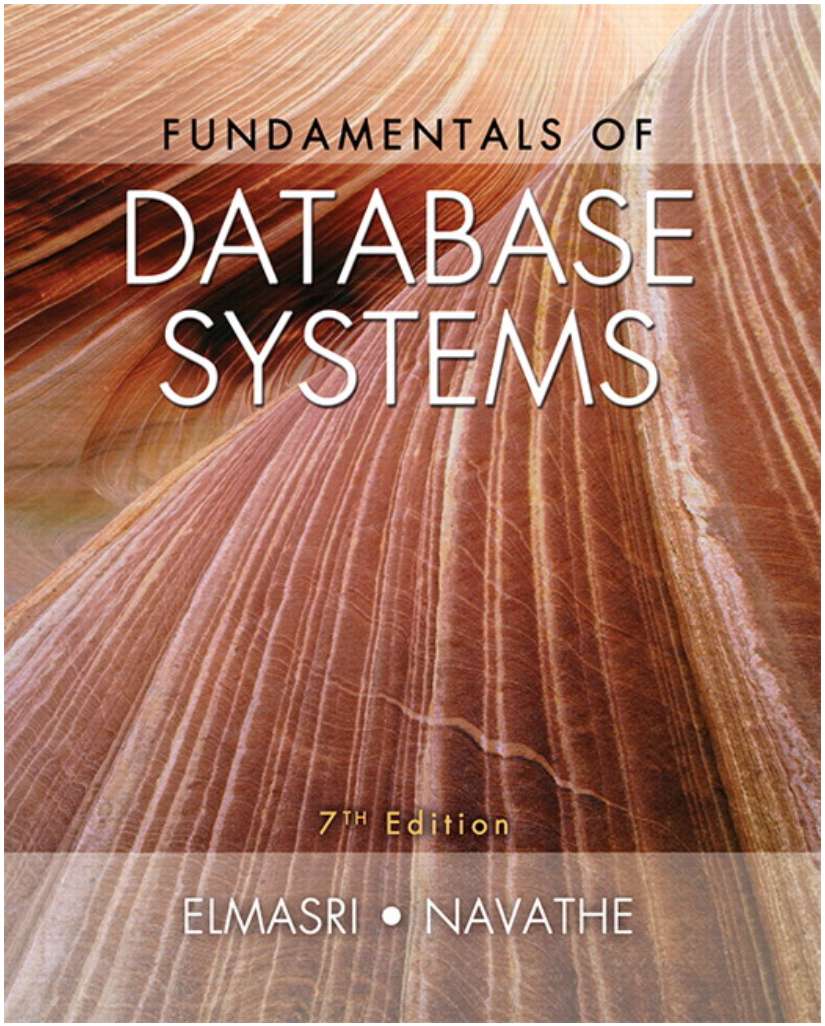

CHAPTER 16

Disk Storage, Basic File Structures, Hashing, and Modern Storage Architectures

16.1 Introduction

- Databases typically stored on magnetic disks
  - Accessed using physical database file structures
- Storage hierarchy
  - Primary storage
    - CPU main memory, cache memory
  - Secondary storage
    - Magnetic disks, flash memory, solid-state drives
  - Tertiary storage
    - Removable media

Memory Hierarchies and Storage Devices

- Cache memory
  - Static RAM
  - DRAM
- Mass storage
  - Magnetic disks
    - CD-ROM, DVD, tape drives
- Flash memory
  - Nonvolatile

Storage Types and Characteristics

Type	Capacity*	Access Time	Max Bandwidth	Commodity Prices (2014)**
Main Memory- RAM	4GB–1TB	30ns	35GB/sec	$100–$20K
Flash Memory- SSD	64 GB–1TB	50µs	750MB/sec	$50–$600
Flash Memory- USB stick	4GB–512GB	100µs	50MB/sec	$2–$200
Magnetic Disk	400 GB–8TB	10ms	200MB/sec	$70–$500
Optical Storage	50GB–100GB	180ms	72MB/sec	$100
Magnetic Tape	2.5TB–8.5TB	10s–80s	40–250MB/sec	$2.5K–$30K
Tape jukebox	25TB–2,100,000TB	10s–80s	250MB/sec–1.2PB/sec	$3K–$1M+

*Capacities are based on commercially available popular units in 2014.

**Costs are based on commodity online marketplaces.

Table 16.1 Types of Storage with Capacity, Access Time, Max Bandwidth (Transfer Speed), and Commodity Cost

Storage Organization of Databases

- Persistent data
  - Most databases
- Transient data
  - Exists only during program execution
- File organization
  - Determines how records are physically placed on the disk
  - Determines how records are accessed

16.2 Secondary Storage Devices

- Hard disk drive
- Bits (ones and zeros)
  - Grouped into bytes or characters
- Disk capacity measures storage size
- Disks may be single or double-sided
- Concentric circles called tracks
  - Tracks divided into blocks or sectors
- Disk packs
  - Cylinder

Single-Sided Disk and Disk Pack

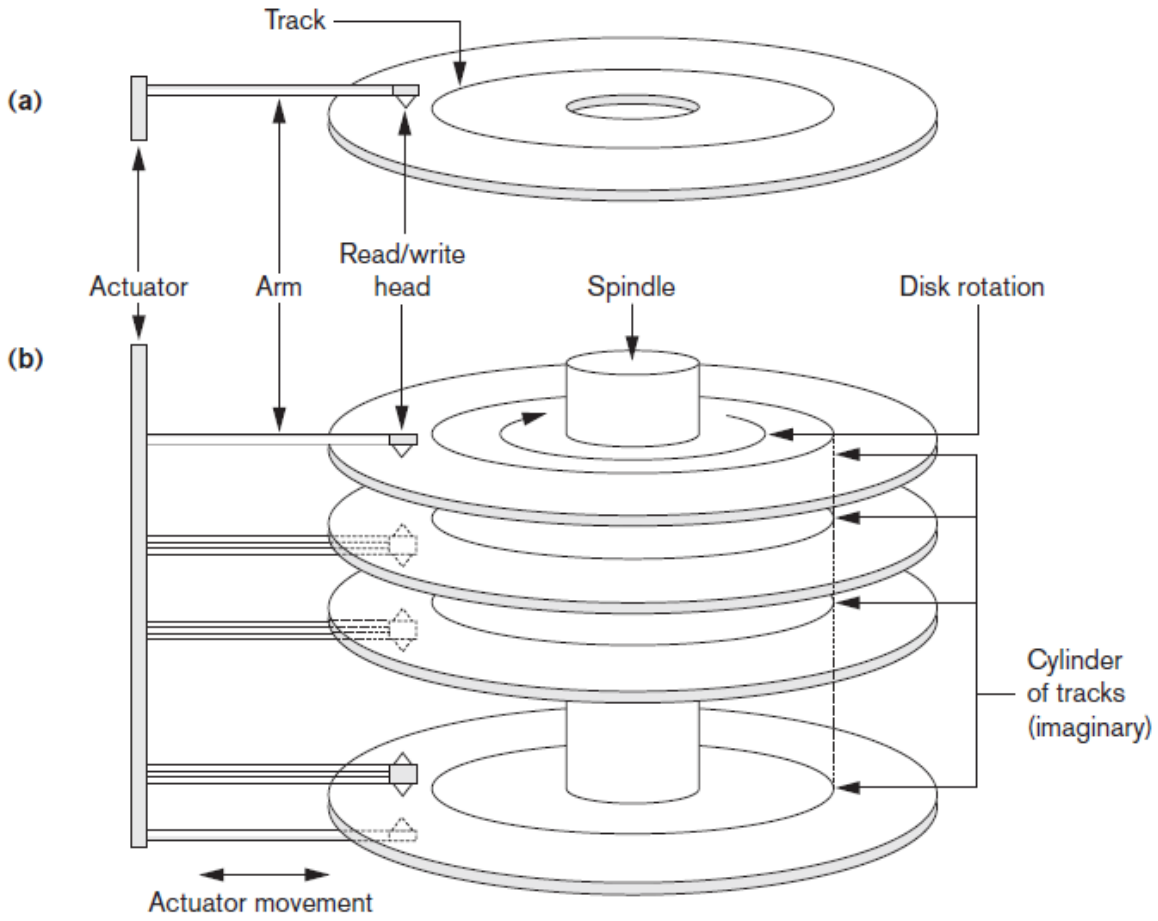

Figure 16.1 (a) A single-sided disk with read/write hardware
(b) A disk pack with read/write hardware

Sectors on a Disk

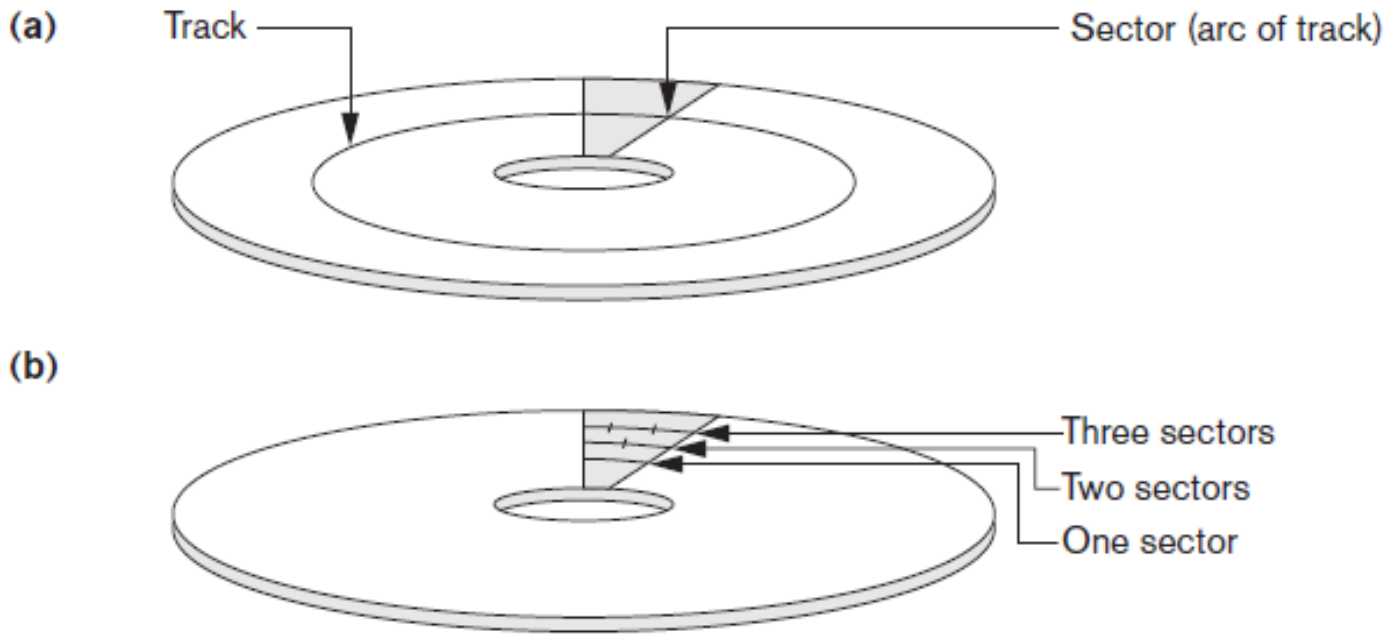

Figure 16.2 Different sector organizations on disk (a) Sectors subtending a fixed angle (b) Sectors maintaining a uniform recording density

Secondary Storage Devices (cont'd.)

- Formatting
  - Divides tracks into equal-sized disk blocks
  - Blocks separated by interblock gaps
- Data transfer in units of disk blocks
  - Hardware address supplied to disk I/O hardware
- Buffer
  - Used in read and write operations
- Read/write head
  - Hardware mechanism for read and write operations

Secondary Storage Devices (cont'd.)

- Disk controller
  - Interfaces disk drive to computer system
  - Standard interfaces
    - SCSI
    - SATA
    - SAS

Secondary Storage Devices (cont'd.)

- Techniques for efficient data access
  - Data buffering
  - Proper organization of data on disk
  - Reading data ahead of request
  - Proper scheduling of I/O requests
  - Use of log disks to temporarily hold writes
  - Use of SSDs or flash memory for recovery purposes

Solid State Device Storage

- Sometimes called flash storage
- Main component: controller
- Set of interconnected flash memory cards
- No moving parts
- Data less likely to be fragmented
- More costly than HDDs
- DRAM-based SSDs available
  - Faster access times compared with flash

Magnetic Tape Storage Devices

- Sequential access
  - Must scan preceding blocks
- Tape is mounted and scanned until required block is under read/write head
- Important functions
  - Backup
  - Archive

16.3 Buffering of Blocks

- Buffering most useful when processes can run concurrently in parallel

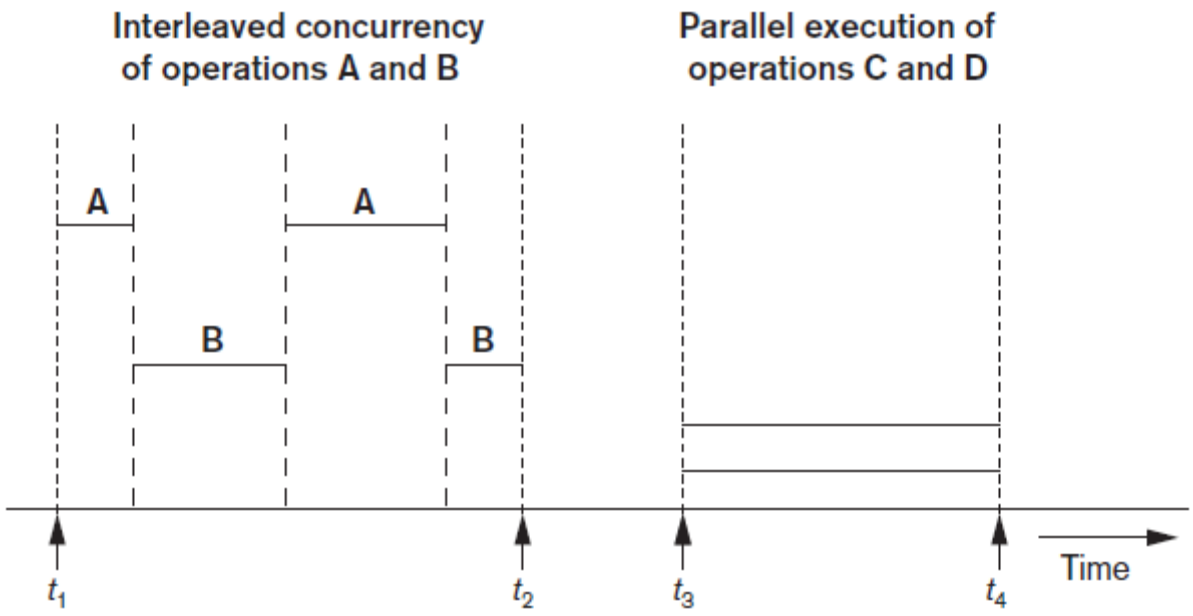

Figure 16.3 Interleaved concurrency versus parallel execution

Buffering of Blocks (cont'd.)

- Double buffering can be used to read continuous stream of blocks

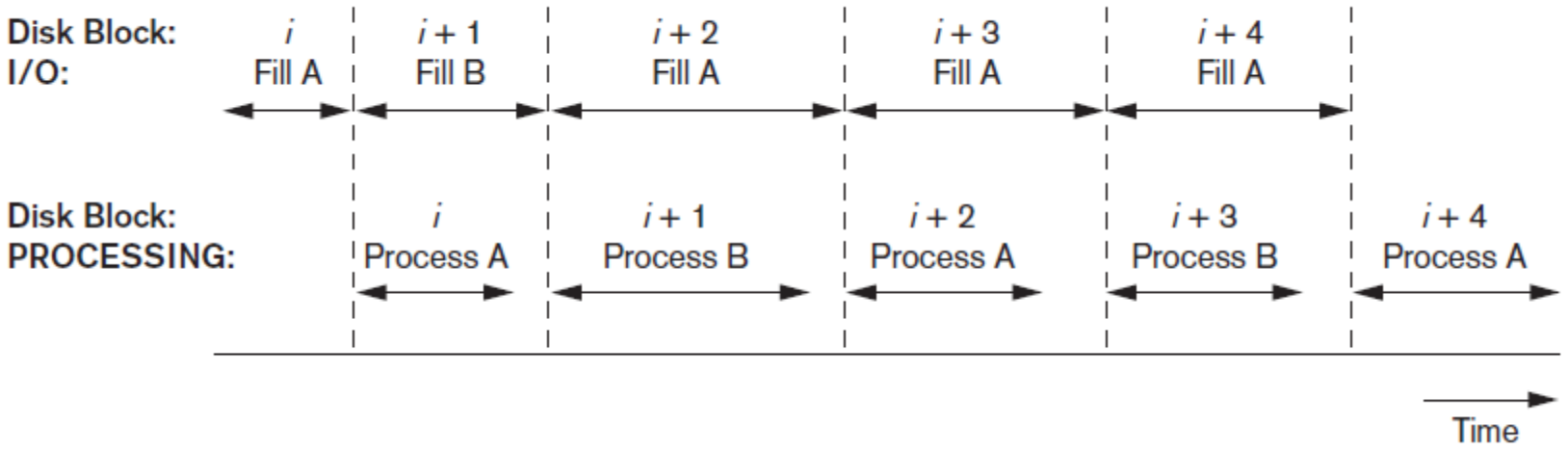

Figure 16.4 Use of two buffers, A and B, for reading from disk

Buffer Management and Replacement Strategies

- Buffer management information
  - Pin count
  - Dirty bit
- Buffer replacement strategies
  - Least recently used (LRU)
  - Clock policy
  - First-in-first-out (FIFO)

16.4 Placing File Records on Disk

- Record: collection of related data values or items
  - Values correspond to record field
- Data types
  - Numeric
  - String
  - Boolean
  - Date/time
- Binary large objects (BLOBs)
  - Unstructured objects

Placing File Records on Disk (cont'd.)

- Reasons for variable-length records
  - One or more fields have variable length
  - One or more fields are repeating
  - One or more fields are optional
  - File contains records of different types

Record Blocking and Spanned Versus Unspanned Records

- File records allocated to disk blocks
- Spanned records
  - Larger than a single block
  - Pointer at end of first block points to block containing remainder of record
- Unspanned
  - Records not allowed to cross block boundaries

Record Blocking and Spanned Versus Unspanned Records (cont'd.)

- Blocking factor
  - Average number of records per block for the file

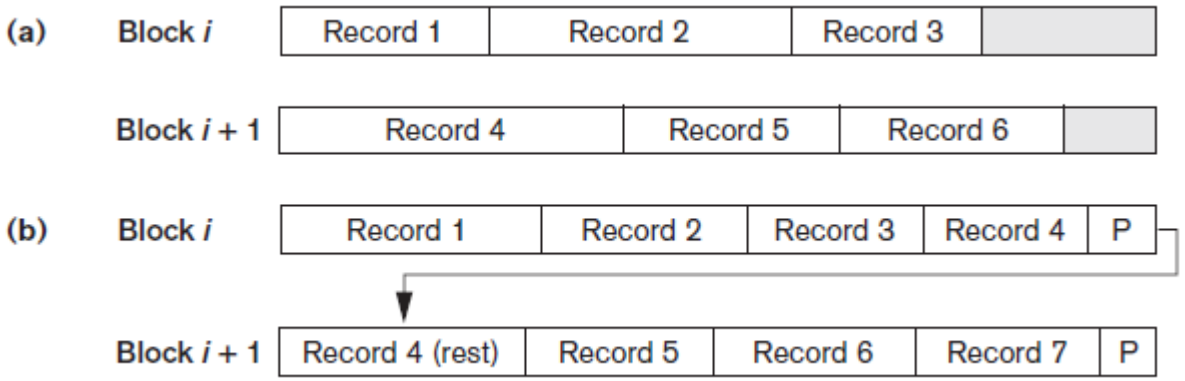

Figure 16.6 Types of record organization (a) Unspanned (b) Spanned

Record Blocking and Spanned Versus Unspanned Records (cont'd.)

- Allocating file blocks on disk
  - Contiguous allocation
  - Linked allocation
  - Indexed allocation
- File header (file descriptor)
  - Contains file information needed by system programs
    - Disk addresses
    - Format descriptions

16.5 Operations on Files

- Retrieval operations
  - No change to file data
- Update operations
  - File change by insertion, deletion, or modification
- Records selected based on selection condition

Operations on Files (cont'd.)

- Examples of operations for accessing file records
  - Open
  - Find
  - Read
  - FindNext
  - Delete
  - Insert
  - Close
  - Scan

16.6 Files of Unordered Records (Heap Files)

- Heap (or pile) file
  - Records placed in file in order of insertion
- Inserting a new record is very efficient
- Searching for a record requires linear search
- Deletion techniques
  - Rewrite the block
  - Use deletion marker

16.7 Files of Ordered Records (Sorted Files)

- Ordered (sequential) file
  - Records sorted by ordering field
    - Called ordering key if ordering field is a key field
- Advantages
  - Reading records in order of ordering key value is extremely efficient
  - Finding next record
  - Binary search technique

Access Times for Various File Organizations

Type of Organization	Access/Search Method	Average Blocks to Access a Specific Record
Heap (unordered)	Sequential scan (linear search)	$b/2$
Ordered	Sequential scan	$b/2$
Ordered	Binary search	$\log_2 b$

Table 16.3 Average access times for a file of b blocks under basic file organizations

16.8 Hashing Techniques

- Hash function (randomizing function)
  - Applied to hash field value of a record
  - Yields address of the disk block of stored record
- Organization called hash file
  - Search condition is equality condition on the hash field
  - Hash field typically key field
- Hashing also internal search structure
  - Used when group of records accessed exclusively by one field value

Hashing Techniques (cont'd.)

- Internal hashing
  - Hash table
- Collision
  - Hash field value for inserted record hashes to address already containing a different record
- Collision resolution
  - Open addressing
  - Chaining
  - Multiple hashing

Hashing Techniques (cont'd.)

- External hashing for disk files
  - Target address space made of buckets
  - Bucket: one disk block or contiguous blocks
- Hashing function maps a key into relative bucket
  - Table in file header converts bucket number to disk block address
- Collision problem less severe with buckets
- Static hashing
  - Fixed number of buckets allocated

Hashing Techniques (cont'd.)

- Hashing techniques that allow dynamic file expansion
  - Extendible hashing
    - File performance does not degrade as file grows
  - Dynamic hashing
    - Maintains tree-structured directory
  - Linear hashing
    - Allows hash file to expand and shrink buckets without needing a directory

16.9 Other Primary File Organizations

- Files of mixed records
  - Relationships implemented by logical field references
  - Physical clustering
- B-tree data structure
- Column-based data storage

16.10 Parallelizing Disk Access Using RAID Technology

- Redundant arrays of independent disks (RAID)
  - Goal: improve disk speed and access time
- Set of RAID architectures (0 through 6)
- Data striping
  - Bit-level striping
  - Block-level striping
- Improving Performance with RAID
  - Data striping achieves higher transfer rates

Parallelizing Disk Access Using RAID Technology (cont'd.)

- Improving reliability with RAID
  - Redundancy techniques: mirroring and shadowing
- RAID organizations and levels
  - Level 0
    - Data striping, no redundant data
    - Spits data evenly across two or more disks
  - Level 1
    - Uses mirrored disks

Parallelizing Disk Access Using RAID Technology (cont'd.)

- RAID organizations and levels (cont'd.)
  - Level 2
    - Hamming codes for memory-style redundancy
    - Error detection and correction
  - Level 3
    - Single parity disk relying on disk controller
  - Levels 4 and 5
    - Block-level data striping
    - Data distribution across all disks (level 5)

Parallelizing Disk Access Using RAID Technology (cont'd.)

- RAID organizations and levels (cont'd.)
  - Level 6
    - Applies P+Q redundancy scheme
    - Protects against up to two disk failures by using just two redundant disks
- Rebuilding easiest for RAID level 1
  - Other levels require reconstruction by reading multiple disks
- RAID levels 3 and 5 preferred for large volume storage

RAID Levels

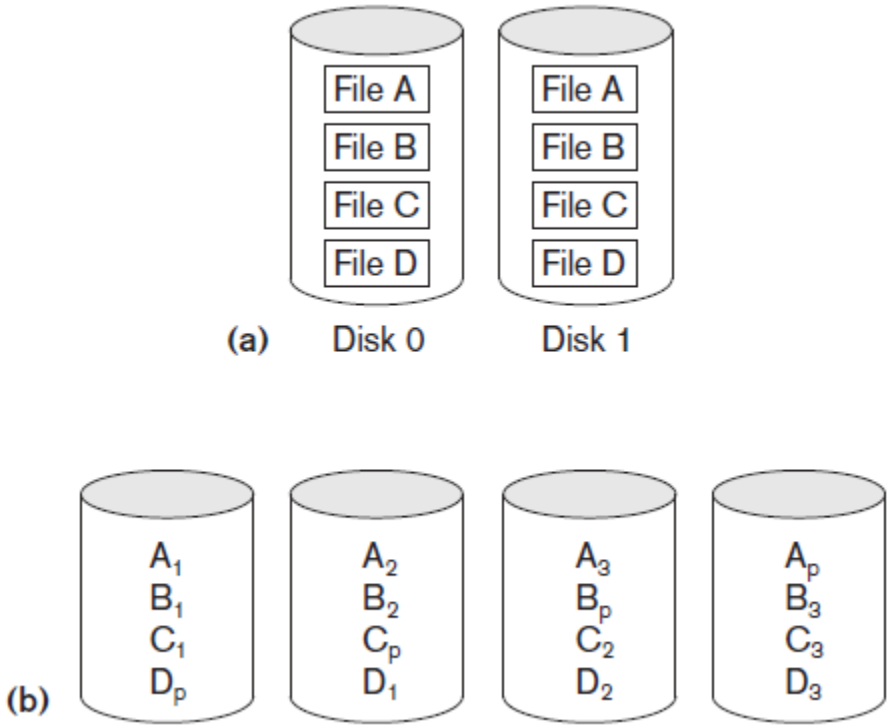

Figure 16.14 Some popular levels of RAID (a) RAID level 1: Mirroring of data on two disks (b) RAID level 5: Striping of data with distributed parity across four disks

16.11 Modern Storage Architectures

- Storage area networks
  - Online storage peripherals configured as nodes on high-speed network
- Network-attached storage
  - Servers used for file sharing
  - High degree of scalability, reliability, flexibility, performance
- iSCSI
  - Clients send SCSI commands to SCSI storage devices on remote channels

Modern Storage Architectures (cont'd.)

- Fibre Channel over IP (FCIP)
  - Fibre Channel control codes and data translated into IP packets
  - Transmitted between geographically distant Fibre Channel SANs
- Fibre Channel over Ethernet (FCoE)
  - Similar to iSCSI without the IP

Modern Storage Architectures (cont'd.)

- Automated storage tiering
  - Automatically moves data between different storage types depending on need
    - Frequently-used data moved to solid-state drives
- Object-based storage
  - Data managed in form of objects rather than files made of blocks
  - Objects carry metadata and global identifier
  - Ideally suited for scalable storage of unstructured data

16.12 Summary

- Magnetic disks
  - Accessing a disk block is expensive
- Commands for accessing file records
- File organizations: unordered, ordered, hashed
- RAID
- Modern storage trends