Database Systems

Data Storage and Buffer Management

何明昕 HE Mingxin, Max

Send your email to **c.max@yeah.net** with

a subject like: **DBS***345-Andy: On What …*

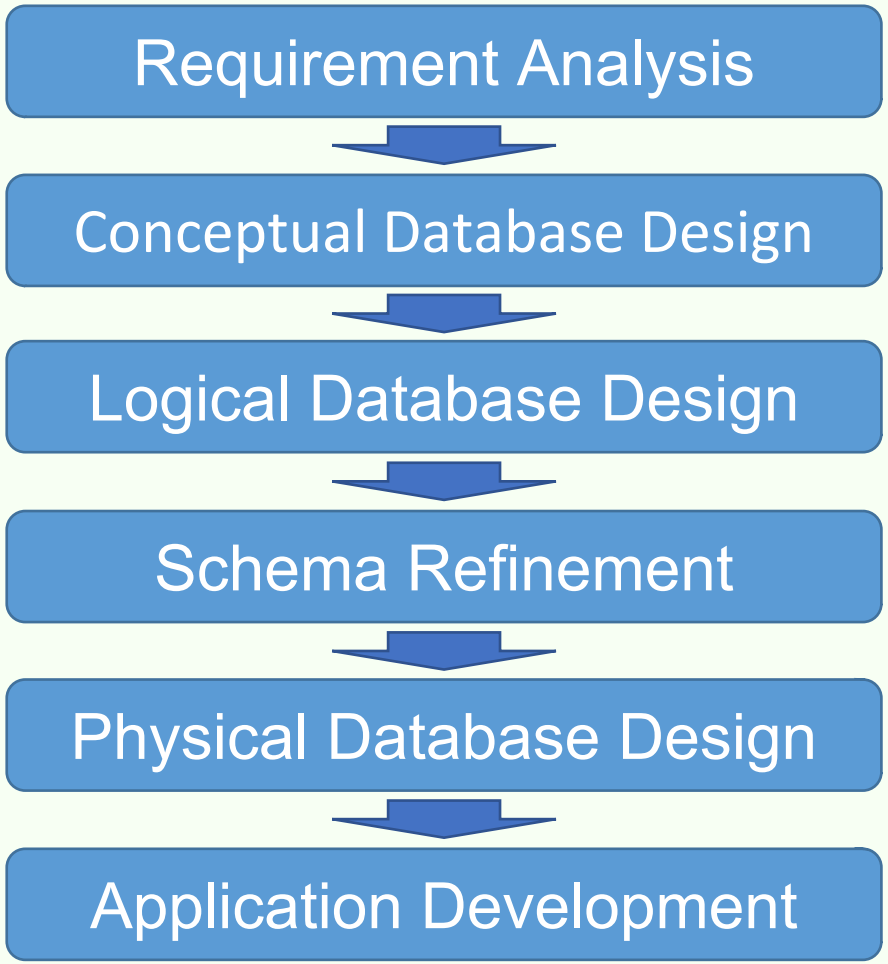
Download from **c.program@yeah.net**

/文件中心/网盘/DatabaseSystems2021

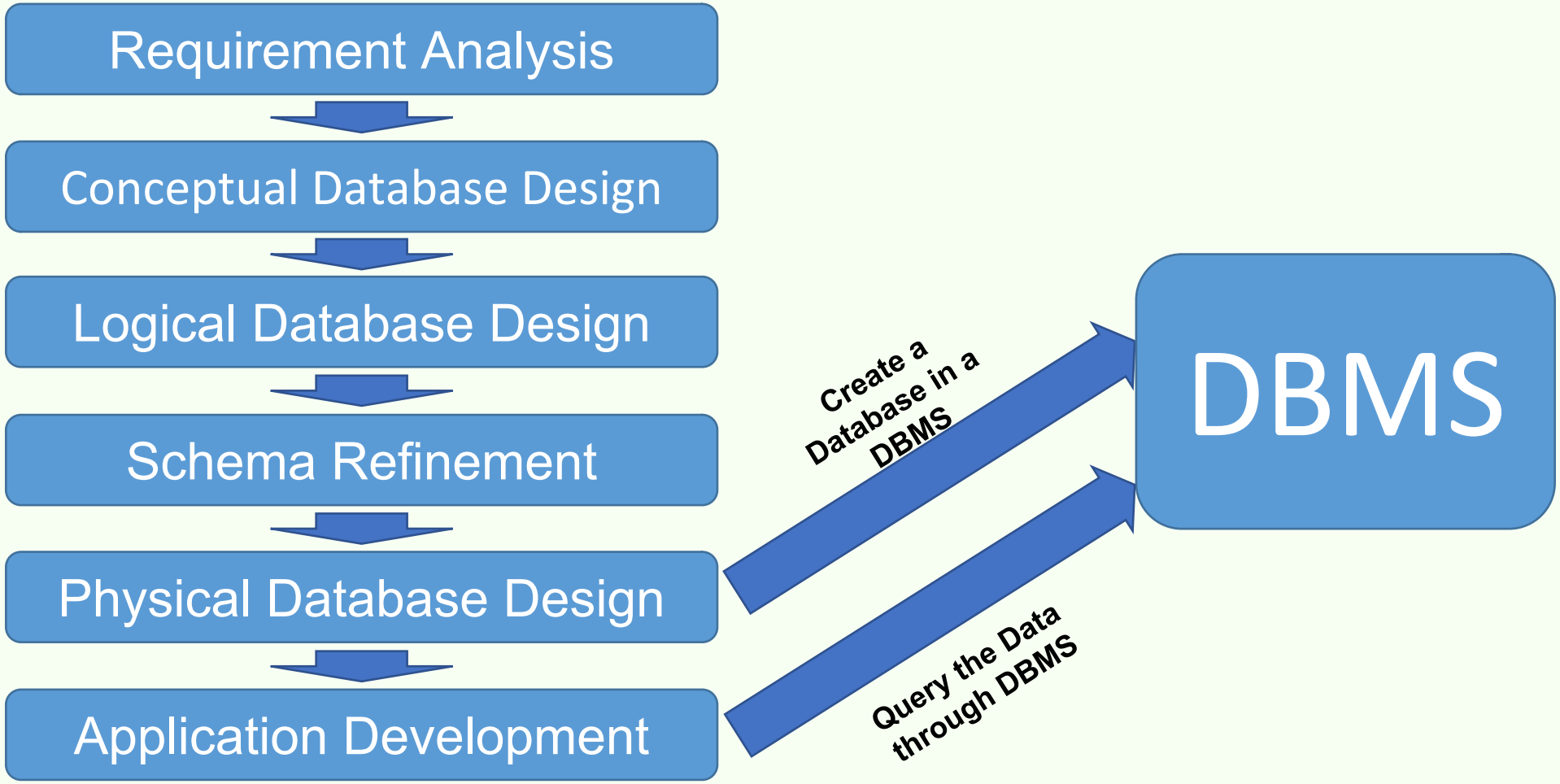
Data Storage and Buffer Management

Where it hits the metal

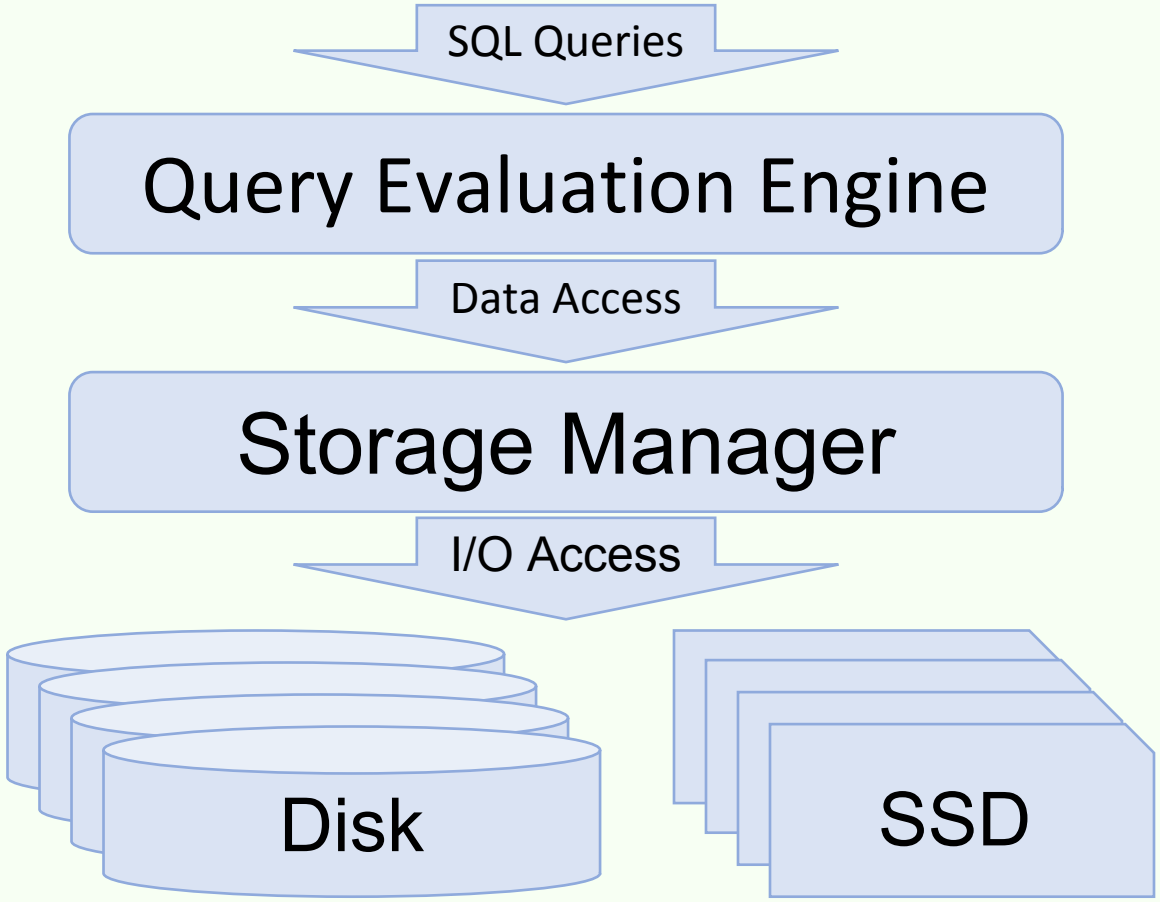
Building a Data-Driven Application



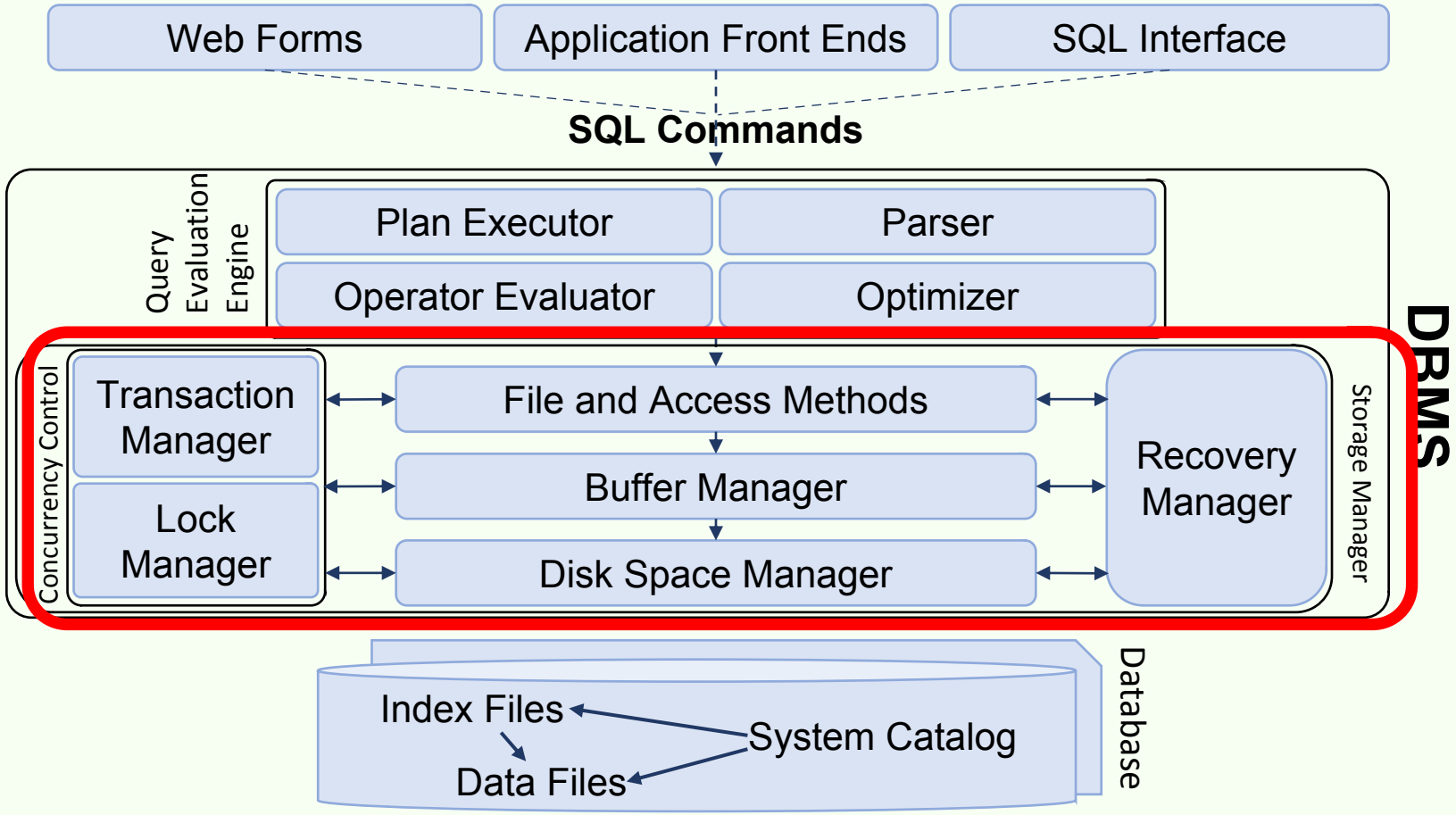
Building a Data-Driven Application, Augmented



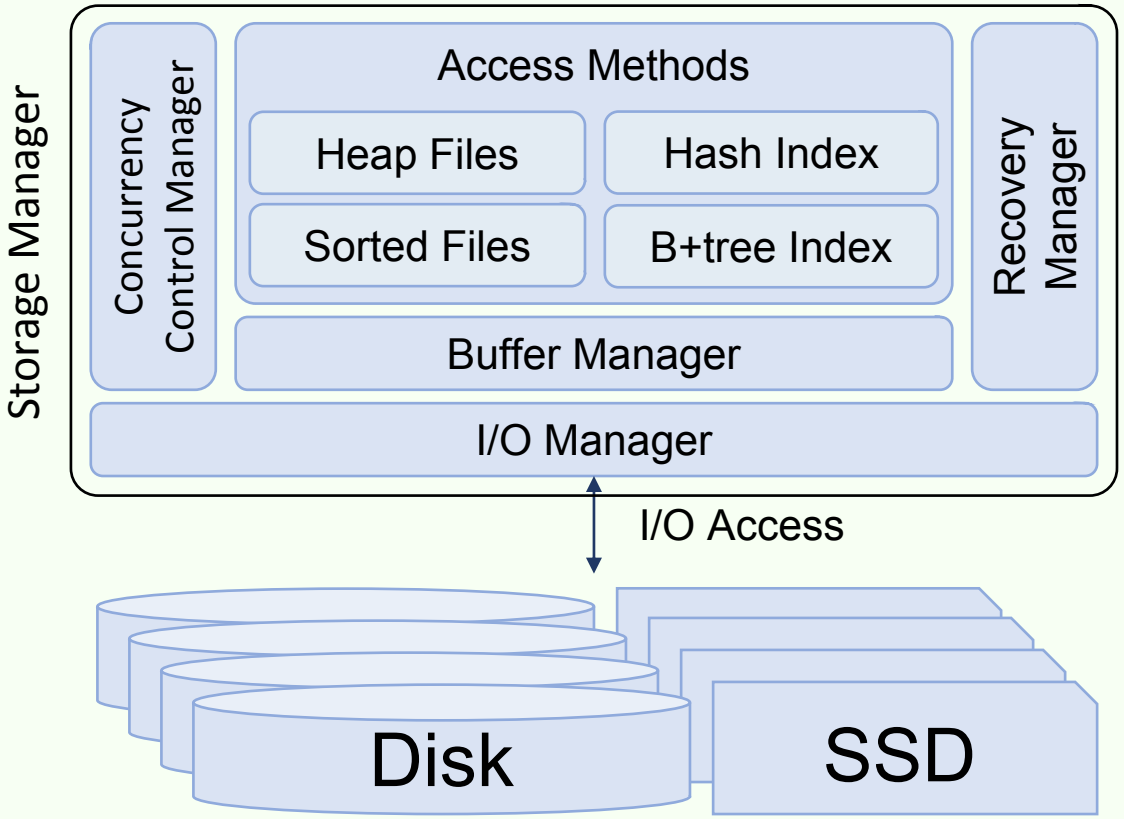
Simplified DBMS Architecture



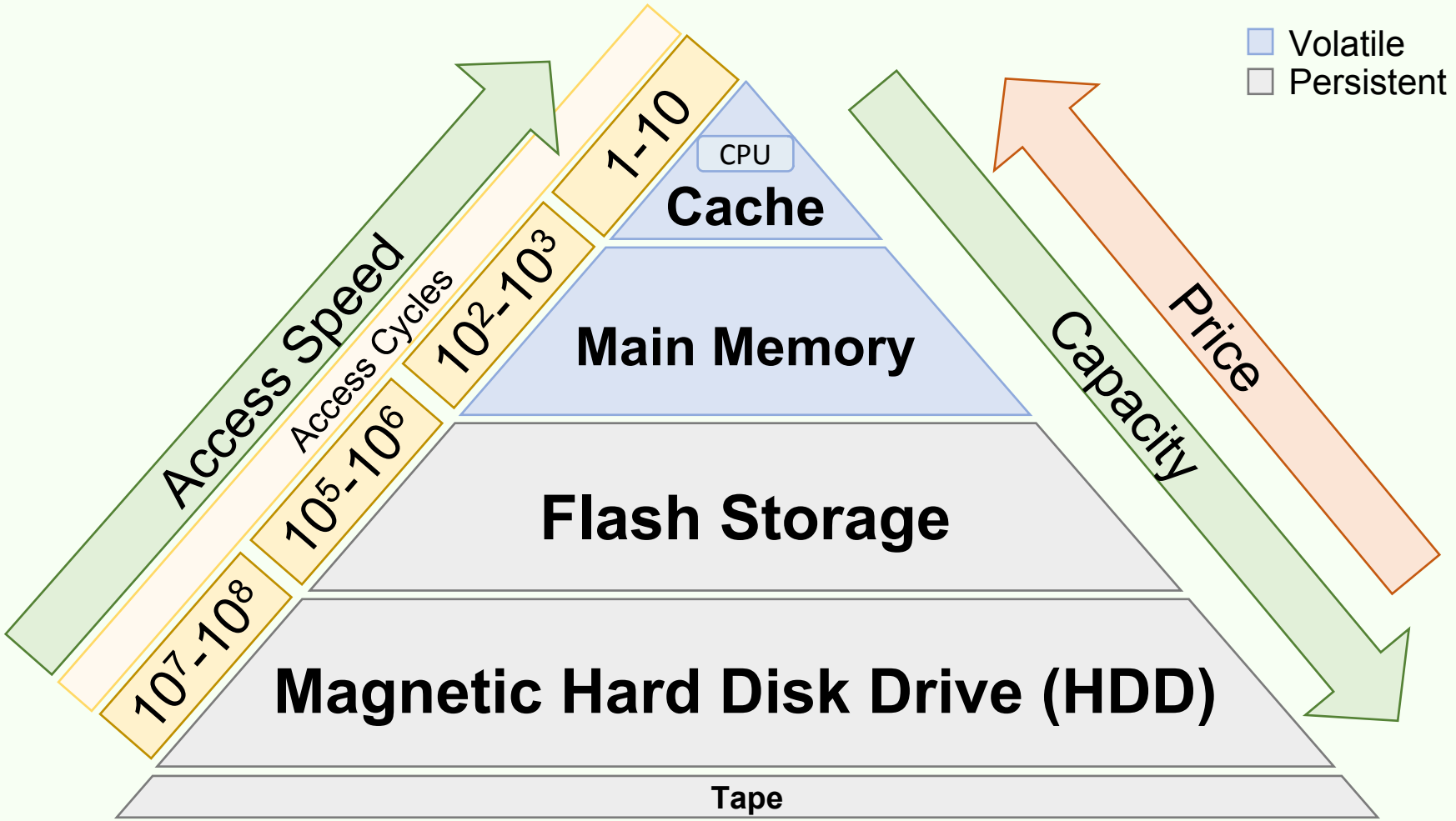
Detailed DBMS Architecture



Storage Manager



Memory Hierarchy



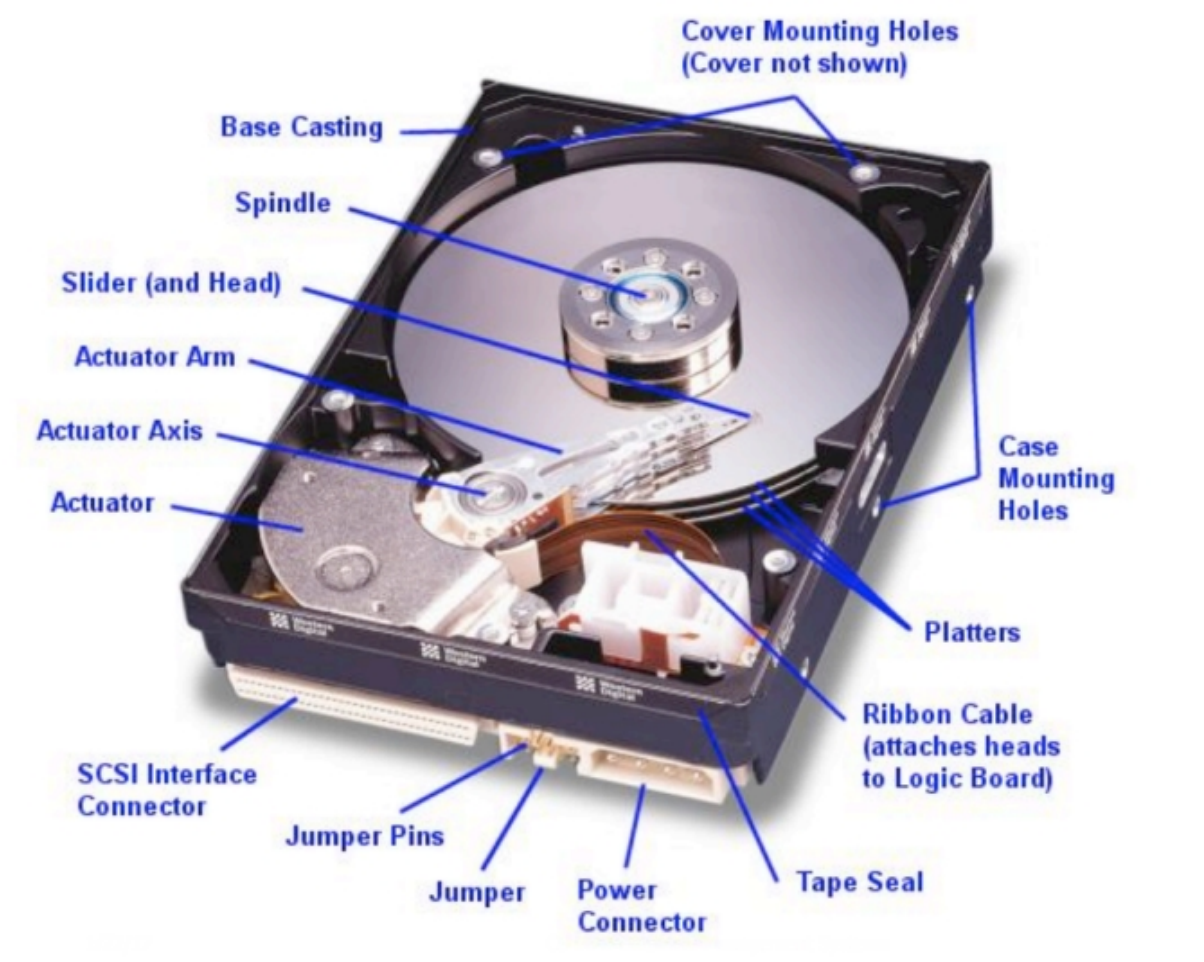
Memory Hierarchy (Cont.)

- Primary storage: main memory (RAM) for currently-used data
- Secondary storage: disk for the main database
  - Increasingly replaced by flash storage
- Tertiary storage: tape for archiving older versions of the data
  - Increasingly replaced by disk

Disk

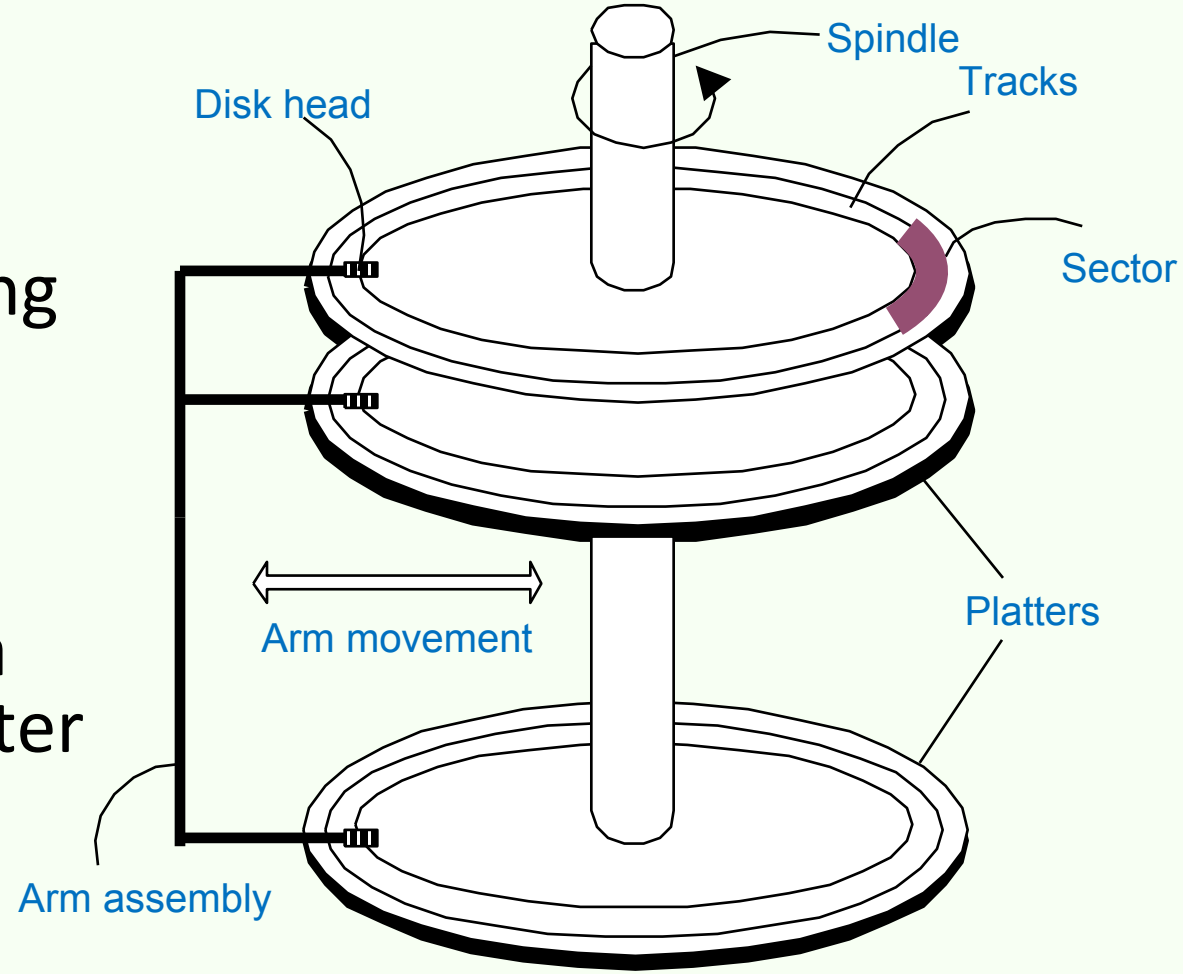
- Secondary storage device of choice
- Data is stored and retrieved in units called *disk blocks* or *pages*
- Unlike RAM, time to retrieve a disk page varies *depending upon its location* on disk
  - Therefore, relative placement of pages on disk has major impact on DBMS performance!

Disk Anatomy



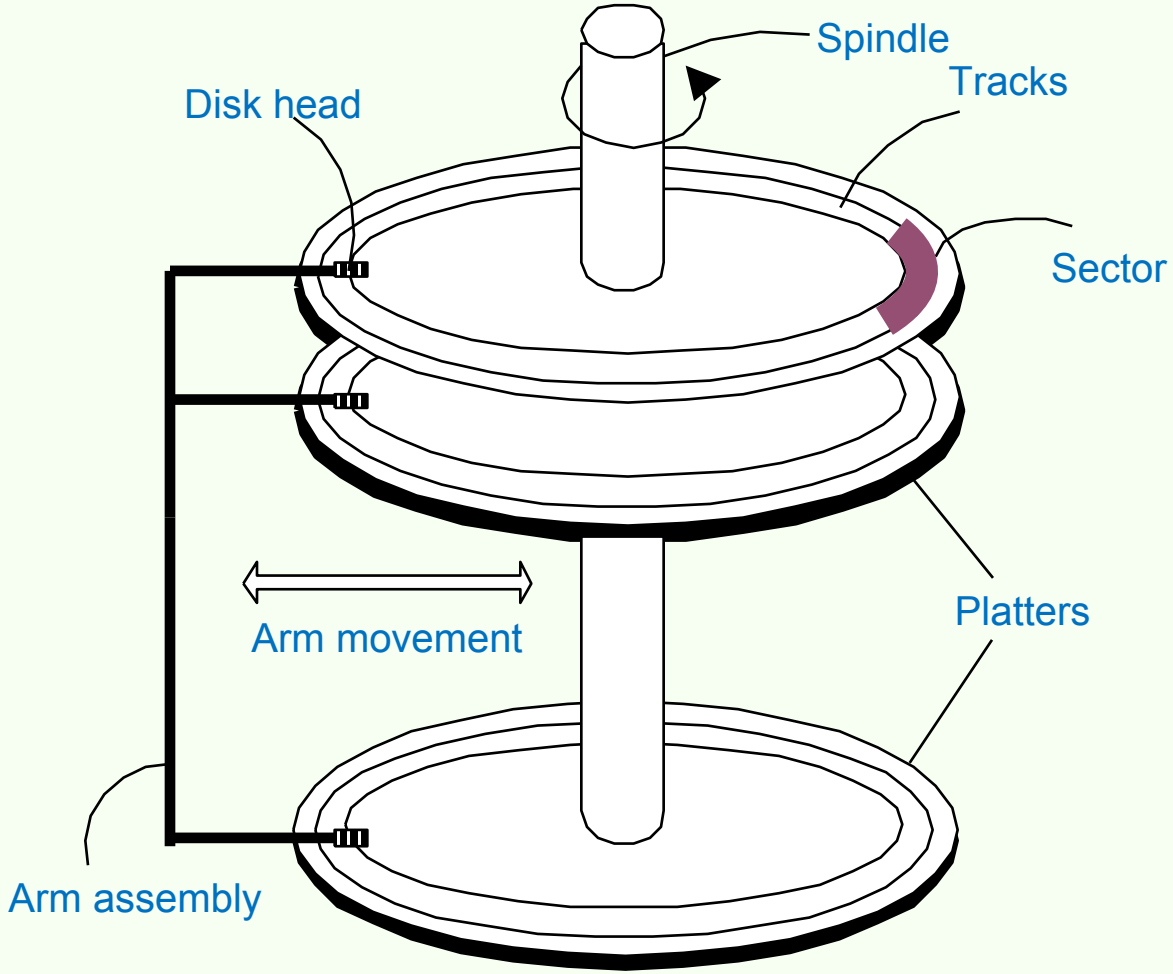
Disk Anatomy (Cont.)

- Platter: circular hard surface on which data is stored by inducing magnetic changes
- Spindle: axis responsible for rotating the platters
- Disk head: mechanism to read or write data
- Arm assembly: moves to position a head on a desired track of the platter
- RPM (Rotations Per Minute)
  - 7200 RPM – 15000 RPM



Disk Anatomy (Cont.)

- Data is encoded in concentric circles of sectors called tracks
  - Sector size is fixed, a characteristic of the disk
- Block (page) size: multiple of sector size
- At any time, exactly one head can read/write



Accessing the Disk

Disk Access Time

=

Seek Time

+

Rotational Delay

+

Data Transfer Time

Dominated by seek time and rotational delay

- Time to move the arm to position disk head on the right track
- Typical seek time: ~ 9 ms,
  - ~ 4 ms for high-end disks

- Time to wait for sector to rotate under the disk head
- Typical delay: 0–10 ms
  - Maximum delay = 1 full rotation
  - Average delay ~ half rotation

- Time to move the data to/from the disk surface
- Typical rates: ~100 MB/s

RPM	Average delay (ms)
5,400	5.56
7,200	4.17
10,000	3.00
15,000	2.00

Example of HDD Specs

- I/O rates
  - Random access workload (~0.3 MB/s)
  - Sequential workload (~210 MB/s)

	Seagate HDD
Capacity	3 TB
RPM	7,200
Average Seek Time	9 ms
Max Transfer Rate	210 MB/s
# Platters	3

Accessing the Disk (Cont.)

- Key to low disk access time: reduce seek time and/or rotational delay
  - Through optimizing the sequential arrangement of blocks
- "Next" block concept: for each block, load
  - blocks on the same track, followed by
  - blocks on the same cylinder, followed by
  - blocks on adjacent cylinders
- For a sequential read, *pre-fetching* several pages at a time is a big win
  - Since you don't need to seek and rotate per page

Reminder:
disk fragmentation

Nice overview of disk architecture and history at http://www.storagereview.com/guide/index.html

Managing Disk Space

- Database IO layer works with the disk device in one of the two ways
  - OS exports a “raw” device interface, which essentially looks like one big file that is a large byte array
  - OR, the DBMS grabs a big file/directory space in the OS and then uses the OS file as a container for the database
- Either way, disk is organized as files, pages and records

Tables on Disk: A Birds Eye View

- Data is stored in tables
- Each table is stored in a file on disk
- Each file consists of multiple pages
- Each page consists of multiple records (i.e. tuples)
- Each records consists of multiple fields
- Data is allocated/deallocated in increments of pages
- Logically-close pages should be nearby in the disk

Solid-state Drive (SSD)

- Another secondary storage technology
- Uses flash memory
  - No moving parts (i.e. no rotate or seek motors)
  - Eliminates seek time and rotational delay
  - Low power consumption and lightweight
- Data transfer rate: 300-600 MB/s
- Fast sequential **and** random access

SSDs (Cont.)

- Limitation (vanishing)
  - Small storage capacity (~0.1-0.5x of HDD)
  - Expensive (~7-20x of HDD)
  - Writes are much more expensive (~10x) than reads
  - Limited lifetime
    - 1-10k writes per page
    - 6 year average failure rate

Recap

- Architecture of a typical DBMS
- Memory hierarchy
  - CPU cache, main memory, SSD, disk, tape
- Disk
  - Anatomy
  - Accessing the disk
    - Seek time, rotational delay, data transfer time
- SSD