Datenbanksysteme II: Overview and General Architecture

Ulf Leser

Table of Content

- Storage Hierarchy
- 5-Layer Architecture
- Overview: Layer-by-Layer

2010: Price versus speed

Really expensive	Register	1-10 ns/byte
Very expensive	Cache	10-60 ns/cache line
~ 200 € / GB	Main Memory	100-300 ns/block
~ 1 € / GB	Disk	10-[illegible] ms/block
< 1€/GB	Tape	

Difference ~10^5

Difference ~10^4

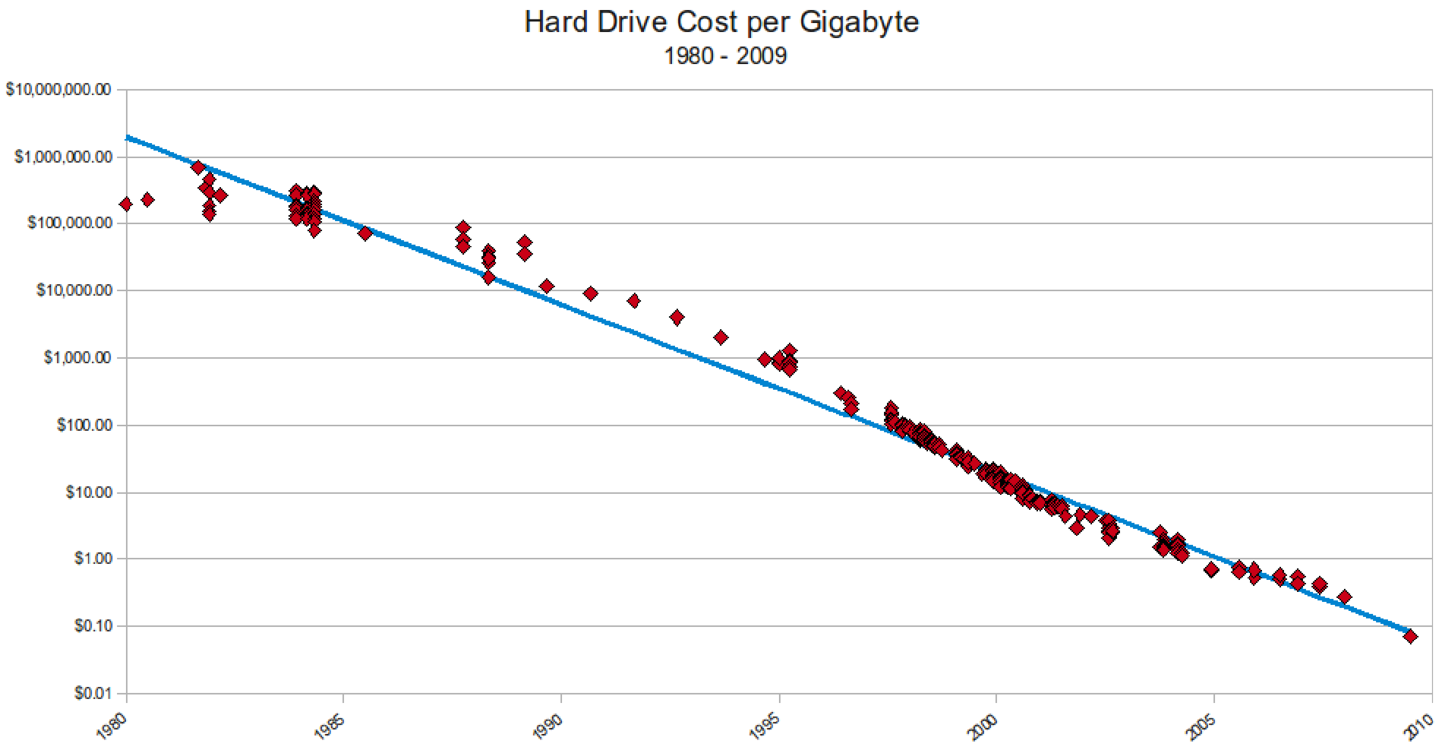
2010: Storage Hierarchy

Really expensive	Register	1-4 byte
Very expensive	Cache	1-4 MB
~ 200 € / GB	Main Memory	1-16 GB
~ 1 € / GB	Disk	512GB – 1TB discs
< 1€/GB	Tape	"Infinite" tape robots

2016: Storage Hierarchy

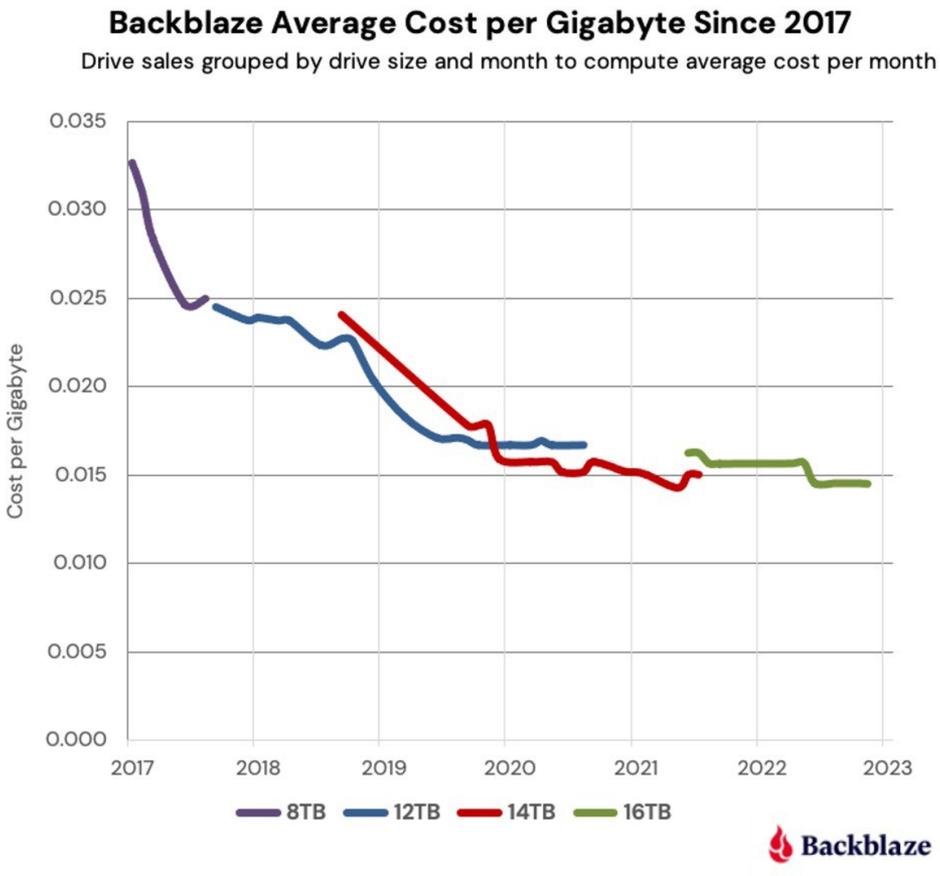
Really expensive	Register	1 – 32 byte
Very expensive	Cache	1-16 MB
~ 7 € / GB	Main Memory	16-256 GB
~ 0,04 € / GB	Disk	1-16 TB
	Tape	“Infinite” tape robots

Costs Drop Faster than you Think

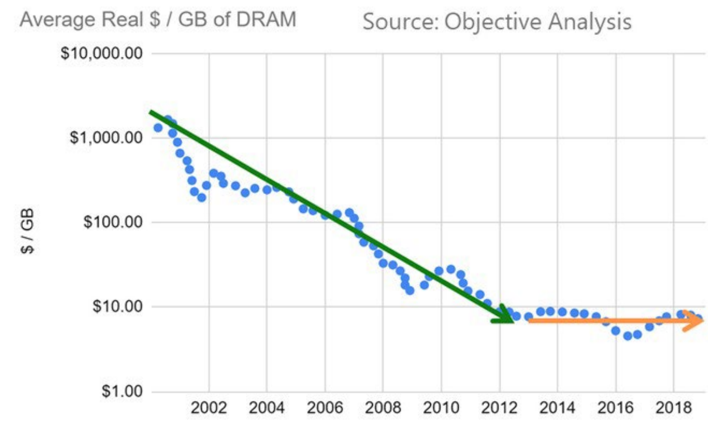


Source: http://analystfundamentals.com/?p=88

But now Trends Level out

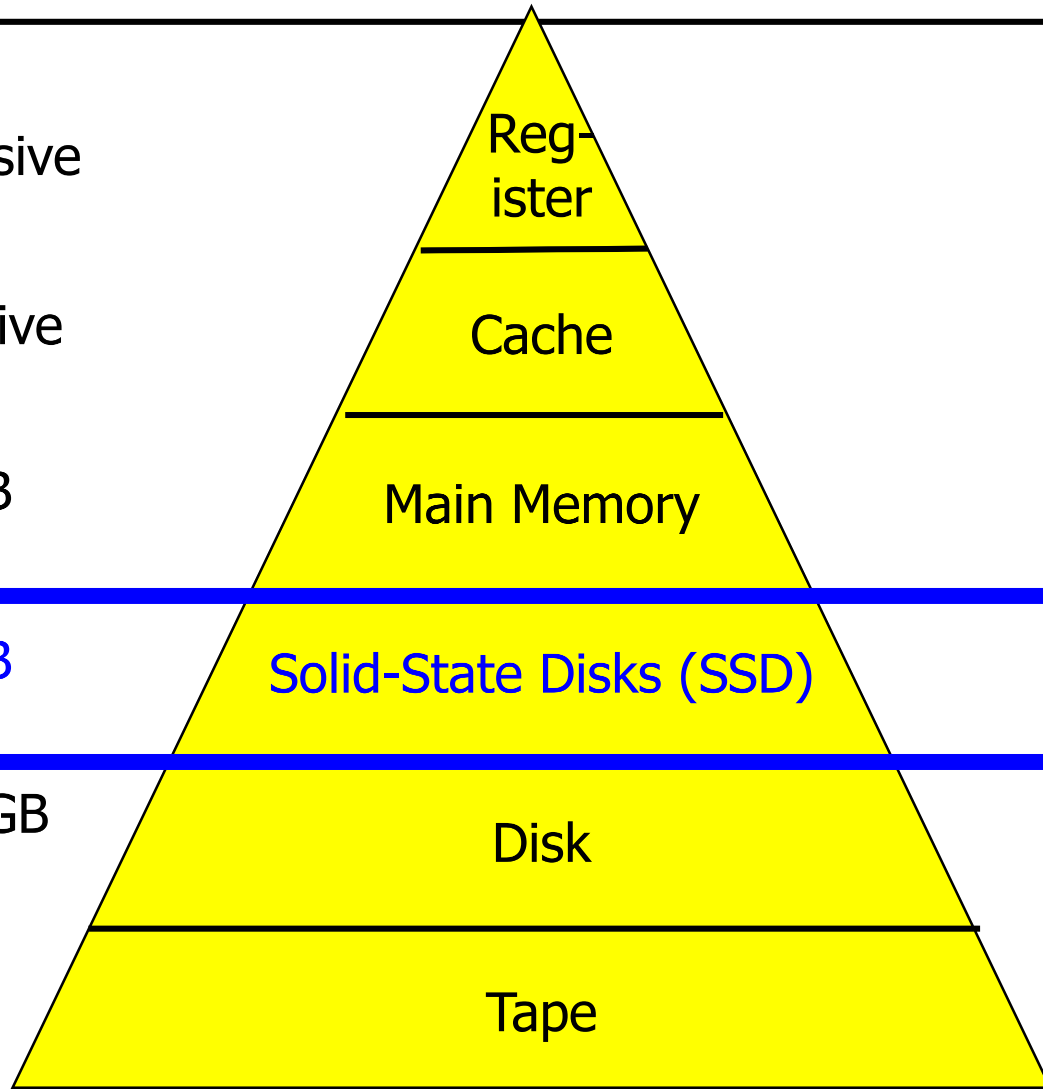


Source: https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/



Source: https://www.nextplatform.com/2023/01/18/what-do-we-do-when-compute-and-memory-stop-getting-cheaper/

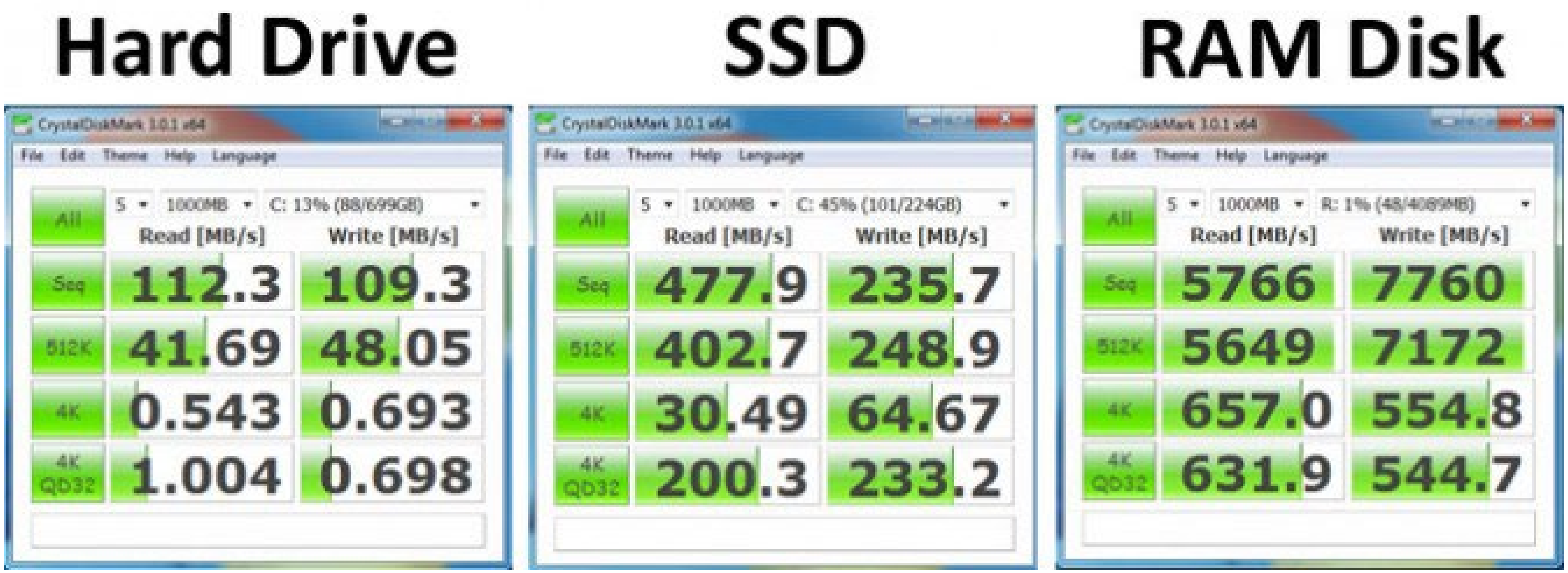
New Players



Really expensive	Register	1-10ns / byte
Very expensive	Cache	10-100ns / cache line
~ 7 € / GB	Main Memory	60-300ns / block
~ 1 € / GB	Solid-State Disks (SSD)	1 ms / block
~ 0,04 € / GB	Disk	10-20 ms / block
	Tape	sec – min

Characteristics

random access != sequential



	Hard Drive Read [MB/s]	Hard Drive Write [MB/s]	SSD Read [MB/s]	SSD Write [MB/s]	RAM Disk Read [MB/s]	RAM Disk Write [MB/s]
Seq	112.3	109.3	477.9	235.7	5766	7760
512K	41.69	48.05	402.7	248.9	5649	7172
4K	0.543	0.693	30.49	64.67	657.0	554.8
4K QD32	1.004	0.698	200.3	233.2	631.9	544.7

read != write

Quelle: http://blog.laptopmag.com/faster-than-an-ssd-how-to-turn-extra-memory-into-a-ram-disk

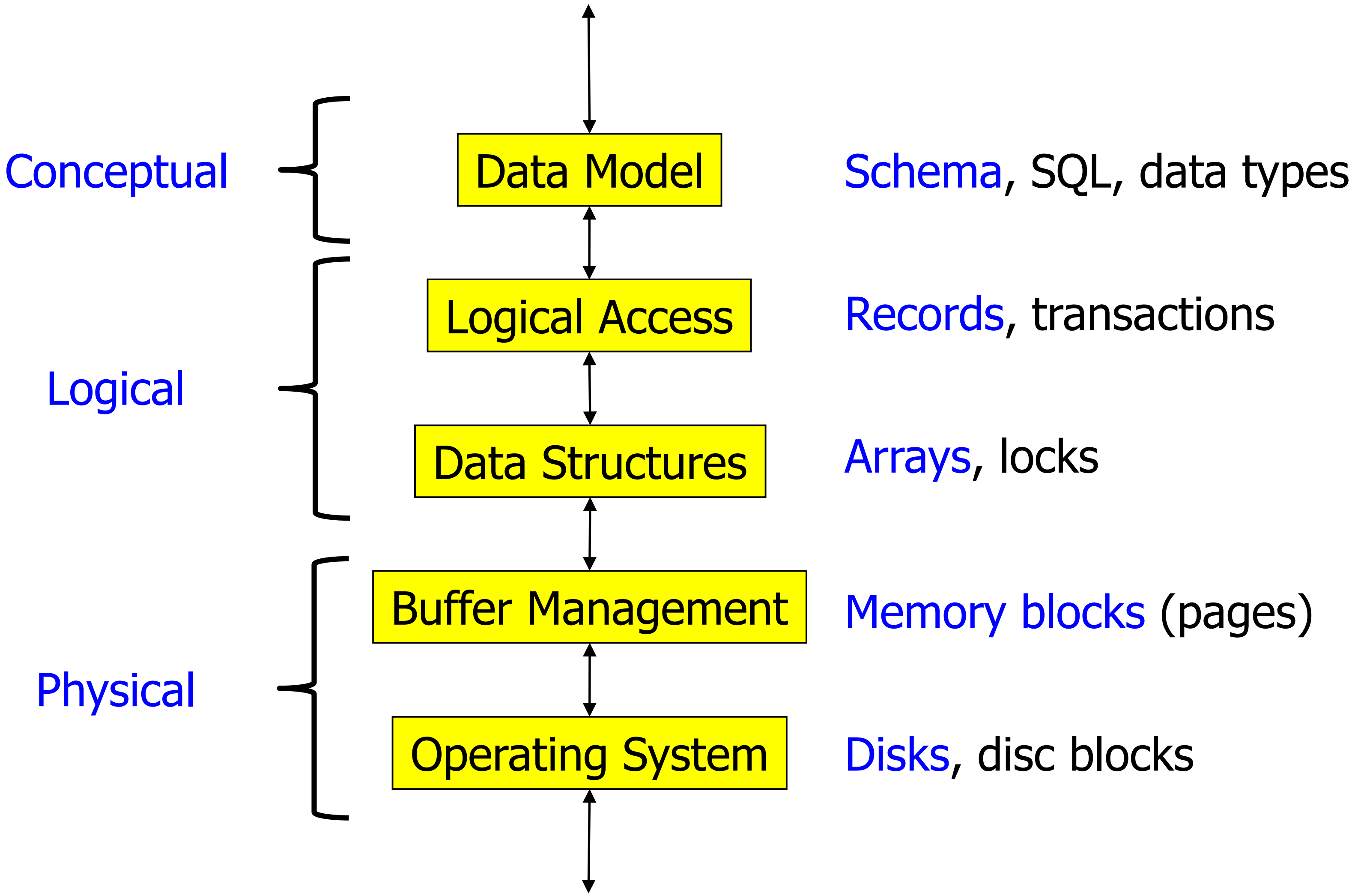
Consequences

- Dealing with memory hierarchy is core concern of DBMS
  - Another issue is multi-core
- Many differences between storage media
  - Speed, durability, size, cost
  - Block sizes
  - Read/write, random-access/sequential
  - Error rates, longevity
- My Guess: 99% of all databases are <100GB
  - Main memory databases
- This lecture will mostly focus on disk versus RAM
  - But highly similar problems for cache-RAM, disk-SSD, …
  - The principle is important, not the concrete parameters
  - … that change all the time anyway

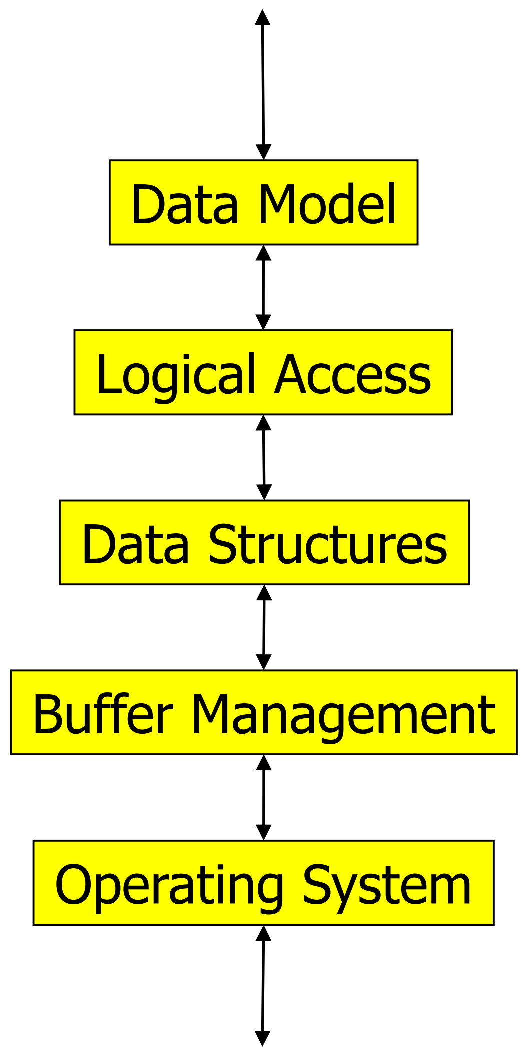
Table of Content

- Storage Hierarchy
- 5-Layer Architecture
- Overview: Layer-by-Layer

Five Layer Architecture



Tasks



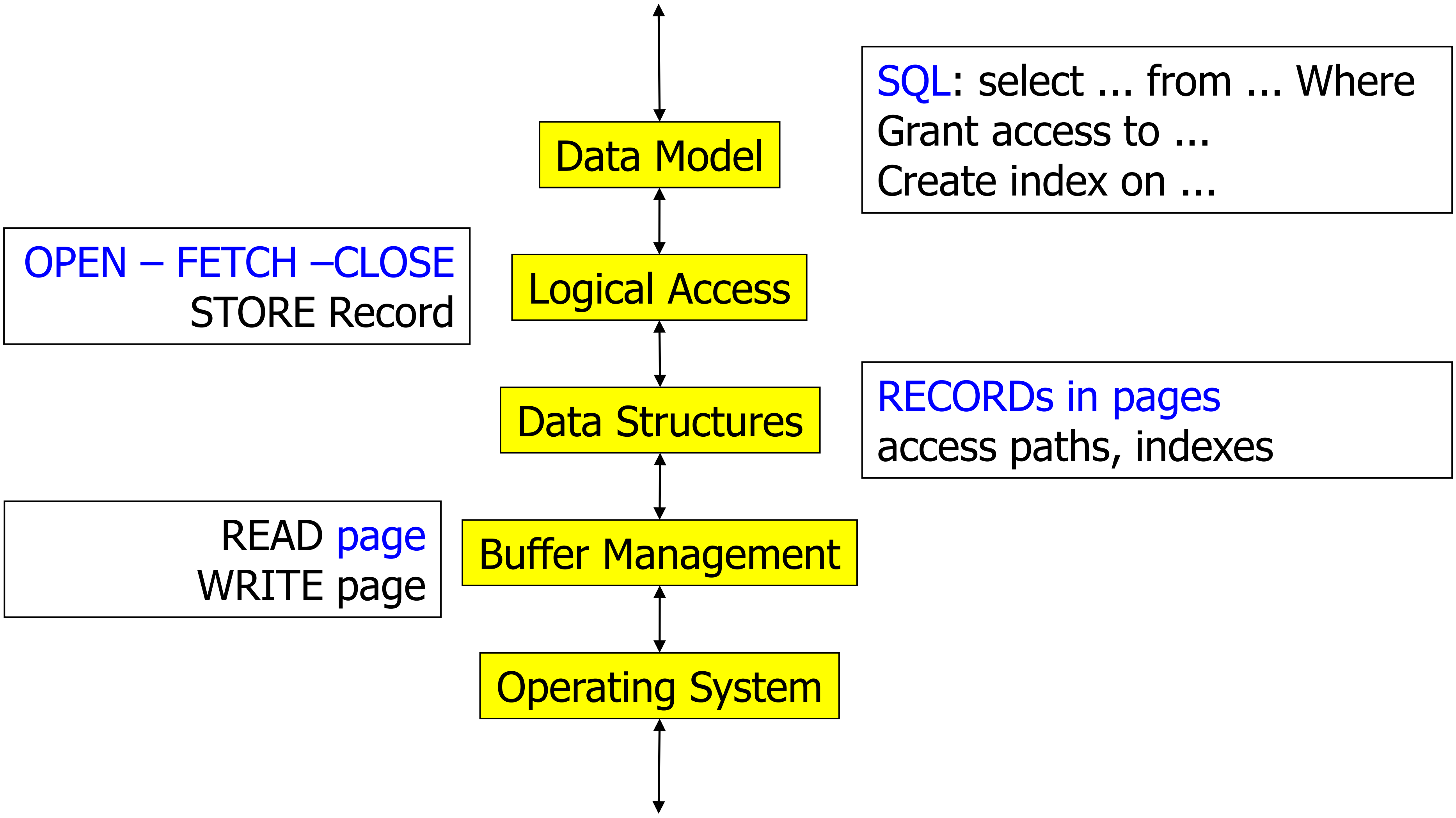
Query optimization
Access control
Integrity constraints

Sorting
Transaction processing
Cursor management

Physical record manager
Index manager
Lock manager
Log / Recovery

Block management
Caching

Operations


SQL: select ... from ... Where
Grant access to ...
Create index on ...
Data Model
OPEN – FETCH –CLOSE
STORE Record
Logical Access
RECORDs in pages
access paths, indexes
Data Structures
READ page
WRITE page
Buffer Management
Operating System

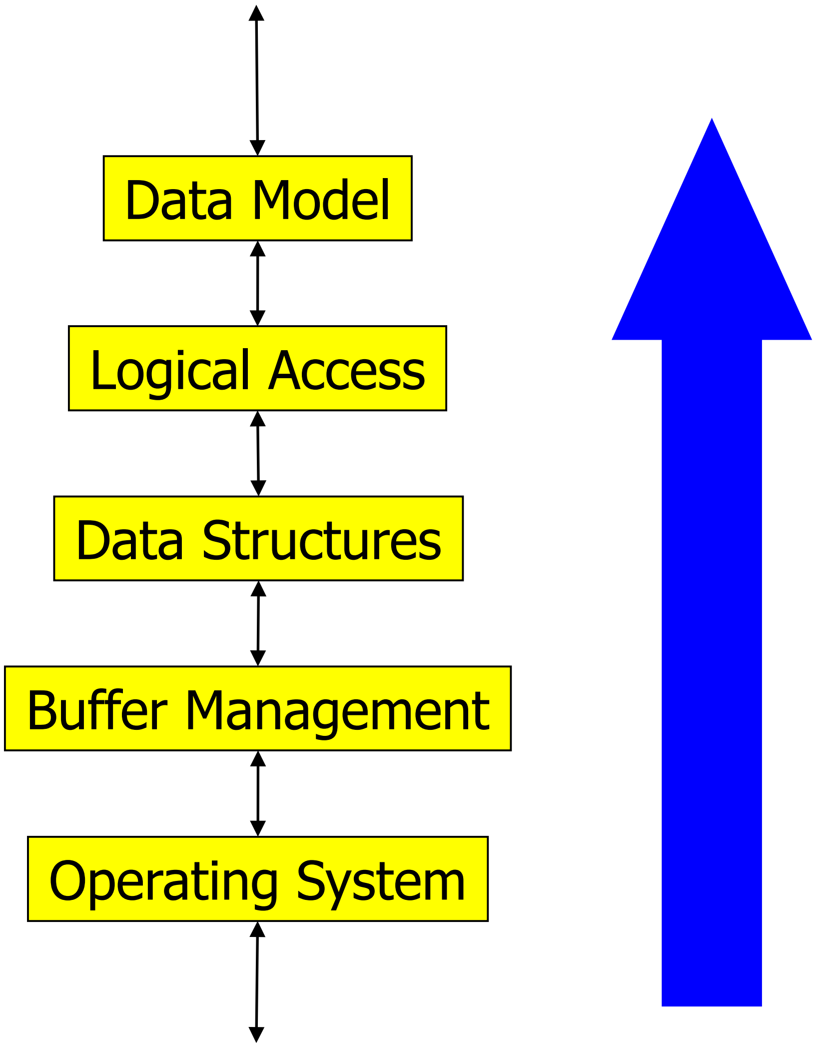
Note: Idealized Representation

- Layers may be merged
  - E.g. logical and internal record-based layers
- Not all functionality can be assigned to exactly one layer
  - E.g. recovery, optimization
- Layers sometimes must access non-neighboring layers
  - Prefetching needs to know the query
    - Layer 4 to Layer 1/2
  - Optimizer needs to know about physical data layout
    - Layer 1 to layer 4/5
  - Breaks information hiding principle

Table of Content

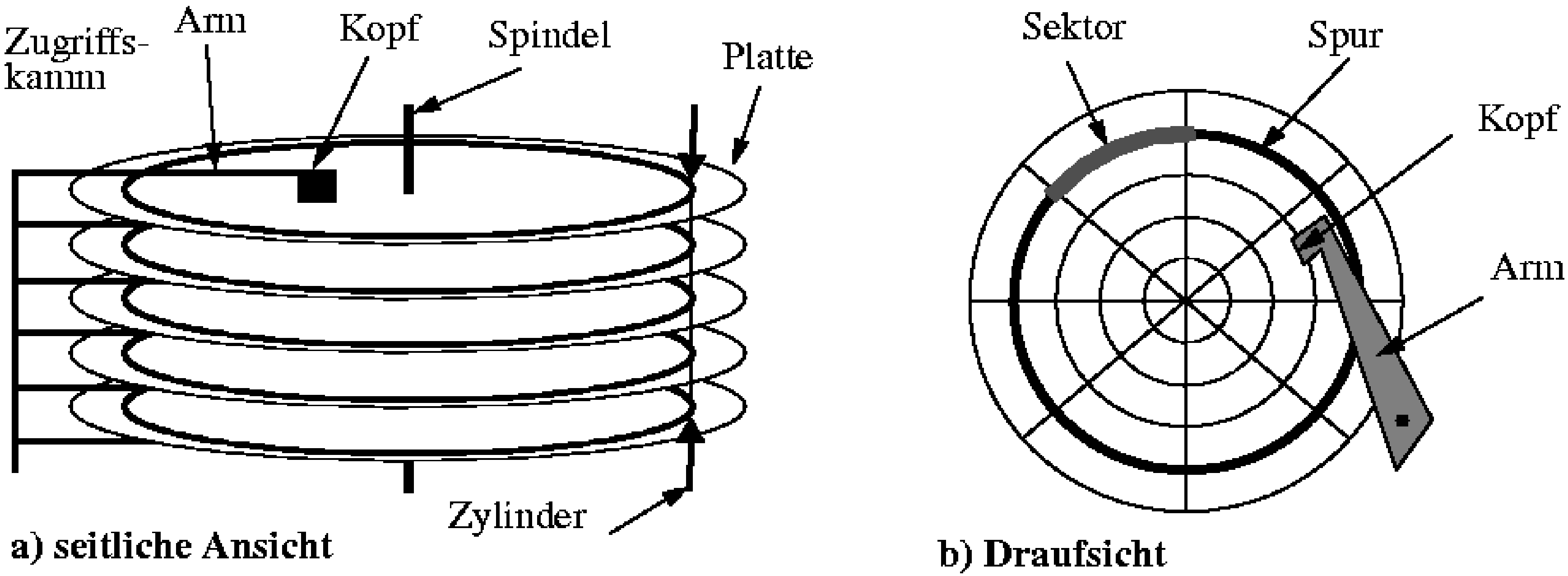
- Storage Hierarchy
- 5-Layer Architecture
- Overview: Layer-by-Layer

Bottom-Up



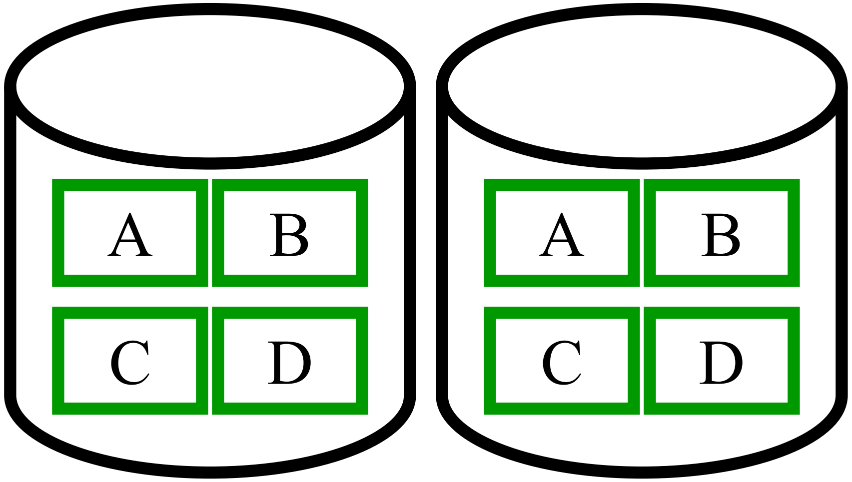
Data Model
Logical Access
Data Structures
Buffer Management
Operating System

Classical Discs



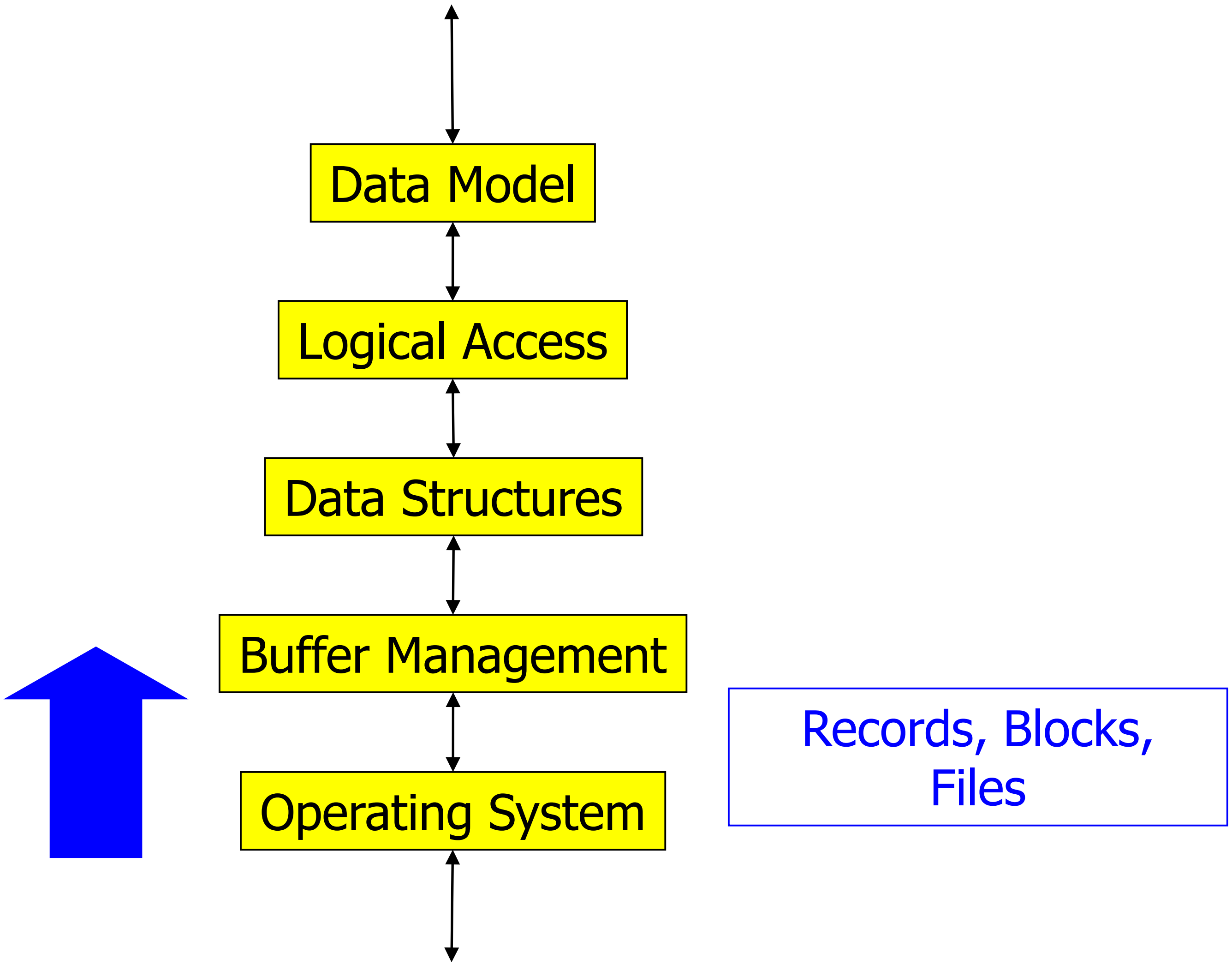
- Durable, slow, cheap, large, not very robust
- In principle: Same read/write speed
- Much difference between random-access / sequential

RAID 1: Mirroring



- Redundancy: Fail-safety and access speed
  - Increased read performance, write perf. not affected (parallel write)
  - Disc crash (one) can be tolerated
  - Be careful about dependent components (controller, power, …)
- Drawbacks
  - Which value is correct in case of divergence in the two copies?
  - Space consumption doubles

Bottom-Up


Data Model
Logical Access
Data Structures
Buffer Management
Operating System
Records, Blocks, Files

Access Methods: Sequential Unsorted Files

- Access to records by record/tuple identifier (RID or TID)

1522	Bond	...
123	Mason	...
...	...	...
1754	Miller	...

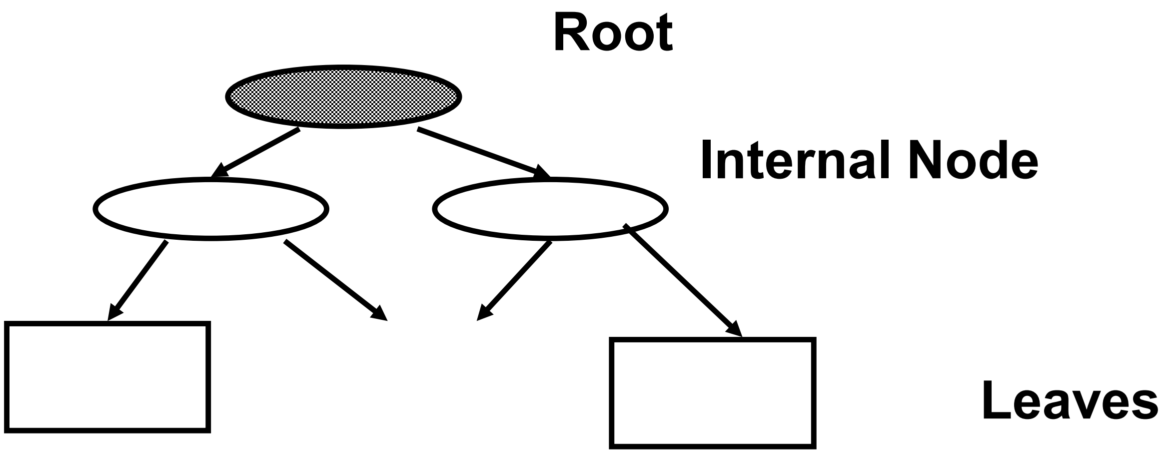
- Operations
  - INSERT(Record): Move to end of file and add, O(1)
  - SEEK(TID): Sequential scan, O(n)
    - FIRST (File): O(1)
    - NEXT(File): O(1)
    - EOF (File): O(1)
  - DELETE(TID): Seek TID; flag as deleted, O(n)
  - REPLACE(TID, Record): Seek TID; write record, O(n)
    - What happens if records have variable size?

Access Methods: Sequential sorted Files

123	Mason	...
1522	Bond	...
...	...	...
1754	Miller	...

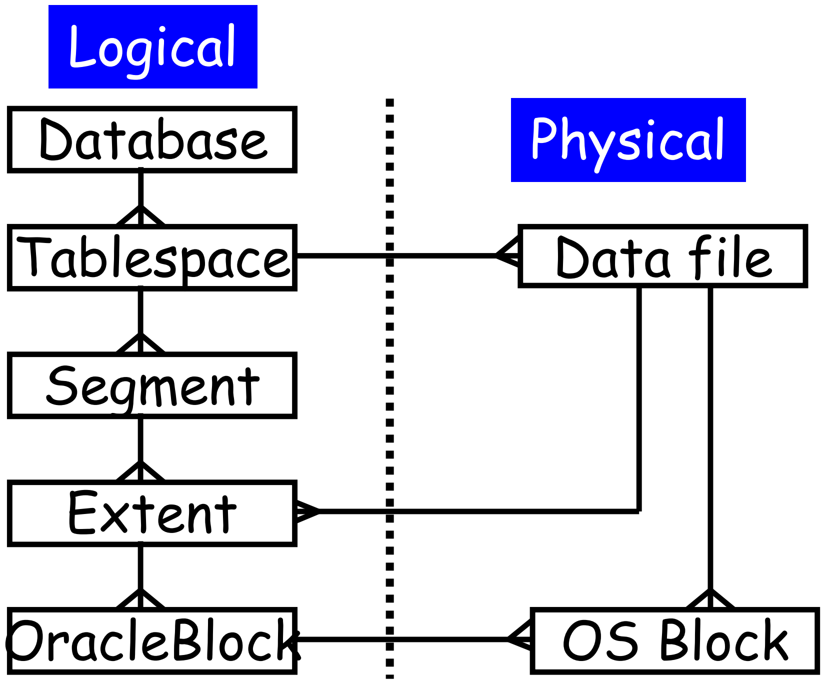
- Operations
  - SEEK(TID): Bin search, O(log(n))
    - But a lot of random access
    - Might be slower than scanning the file
  - INSERT(Record): Seek(TID), move records by one, O(n)
    - This is terribly expensive – avoid!
  - …

Indexed Files



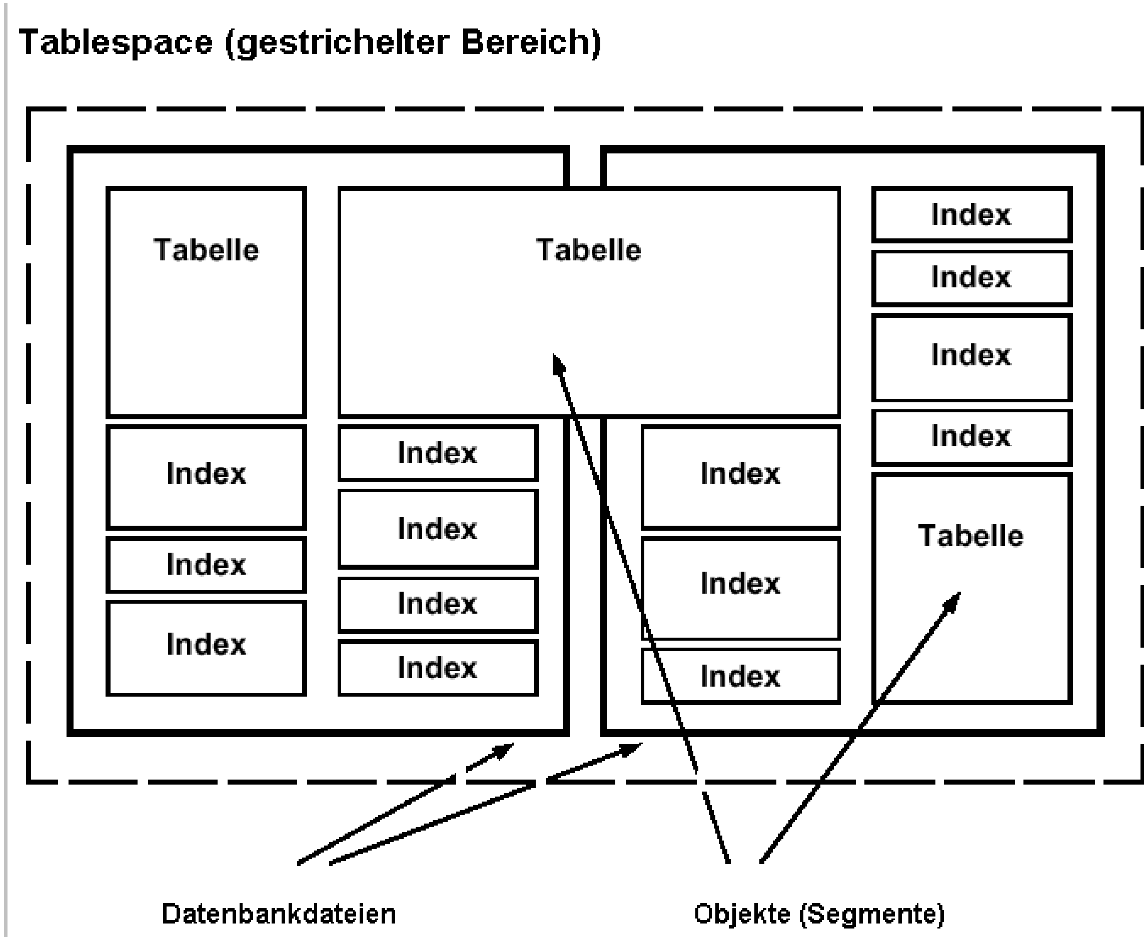
- Operations
  - SEEK(TID): Using order in TIDs: O(log(n))
    - Only if tree is balanced
    - Only if tree is ordered by the search attribute
  - INSERT(TID): Seek TID and insert; possibly restructuring
  - …

Storage in Oracle

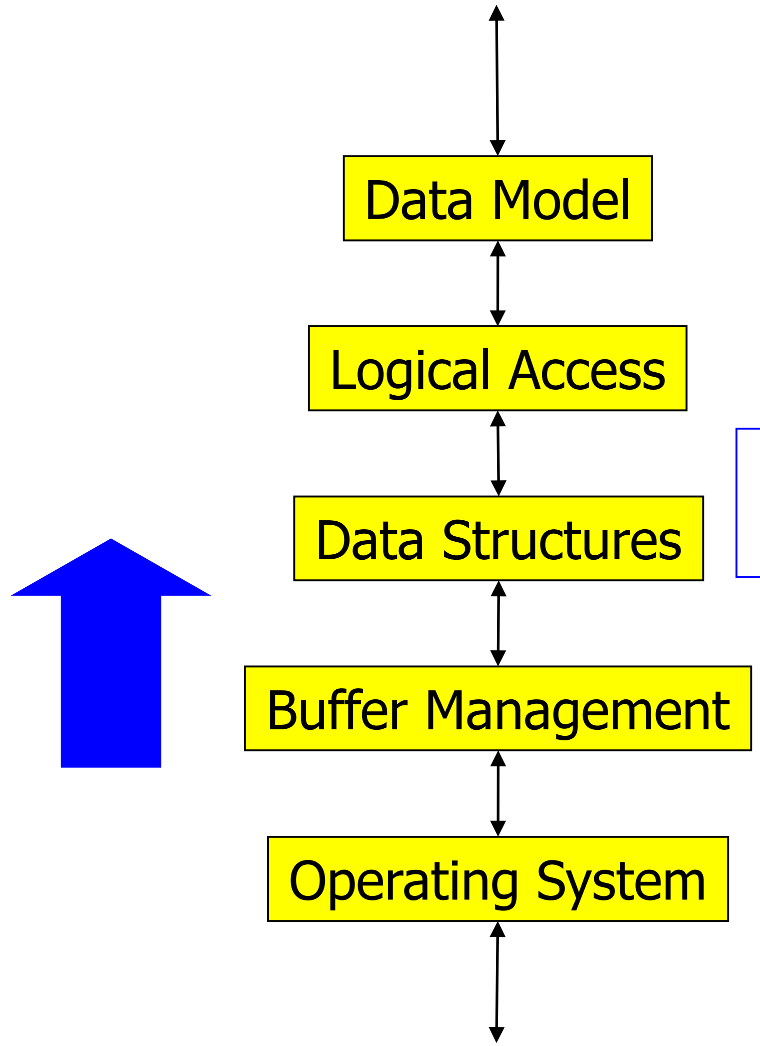


- Data files are assigned to tablespaces
  - May consist of multiple files
  - All data from one object (table, index) are in one tablespace
    - But table and index can be in different ones
  - Backup, quotas, access, …
- Extents: Continuous sequences of blocks on disc
- Space is allocated in extents (min, next, max, ...)
- Segments logically group all extents of an object

Managing space in Oracle

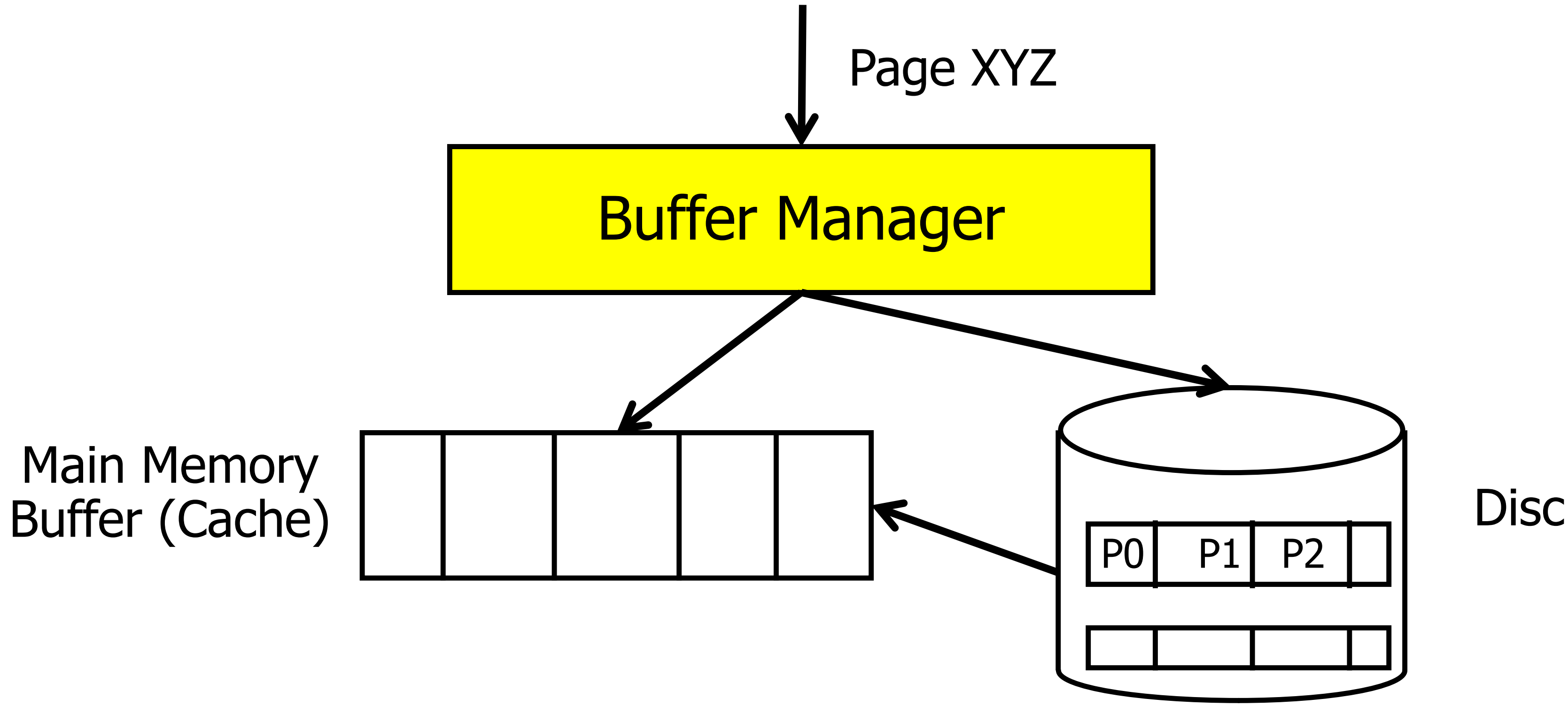


Bottom-Up



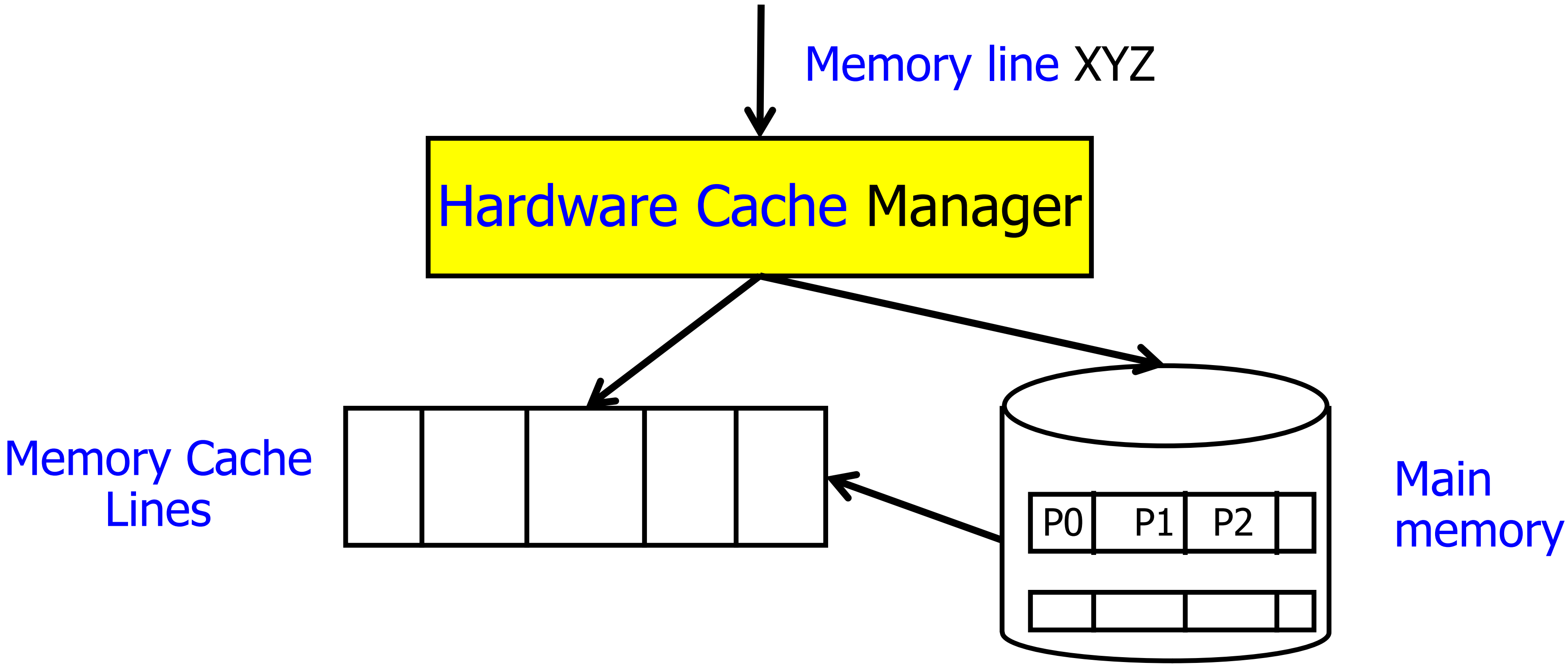
Virtual – physical blocks, access paths

Caching = Buffer Management



- Which blocks should be cached – for how long?
- Caching data blocks? Index blocks?
- Competition: Intermediate data, data buffers, sort buffer, …

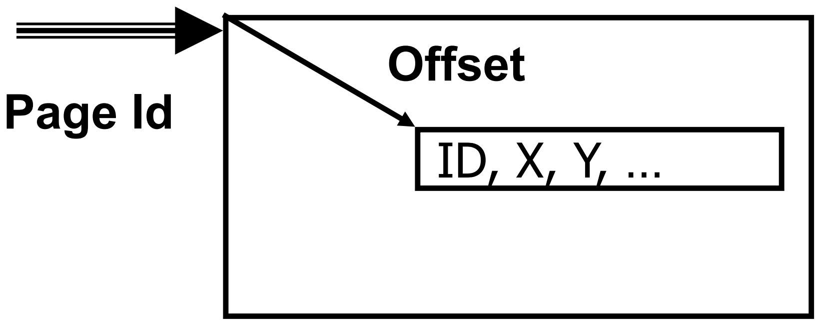
Buffer Management at Main Memory Level



- Which lines should be cached – for how long?
- Cache Hierarchy: Cache Level L1 (core), L2 (CPU / board), L3 (memory manager)

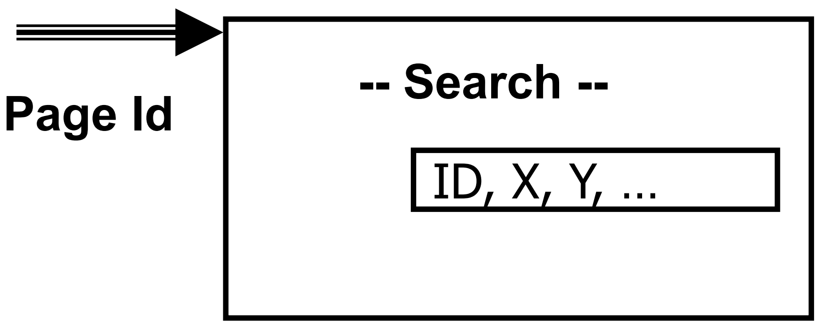
From Buffers to Records

- Absolute addressing: TID = <PageId, Offset, ID>



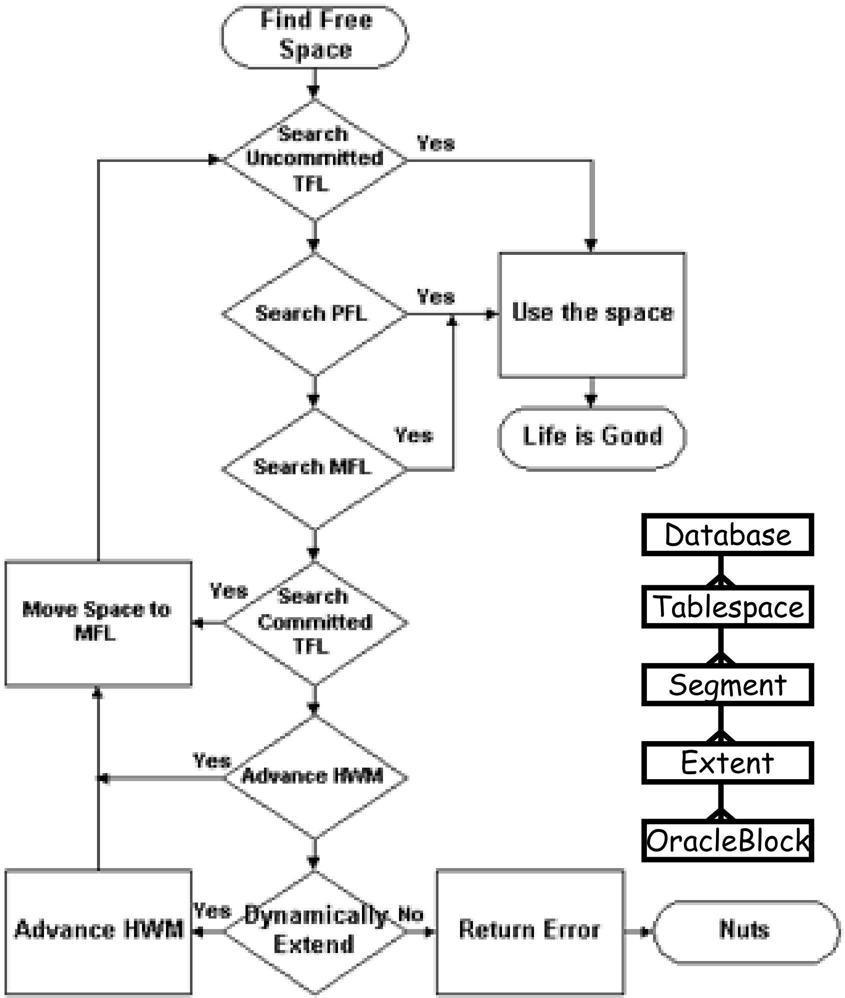
- Pro: Fast access
- Con: Records cannot be moved

- Absolute addressing + search: TID = <PageId,ID>



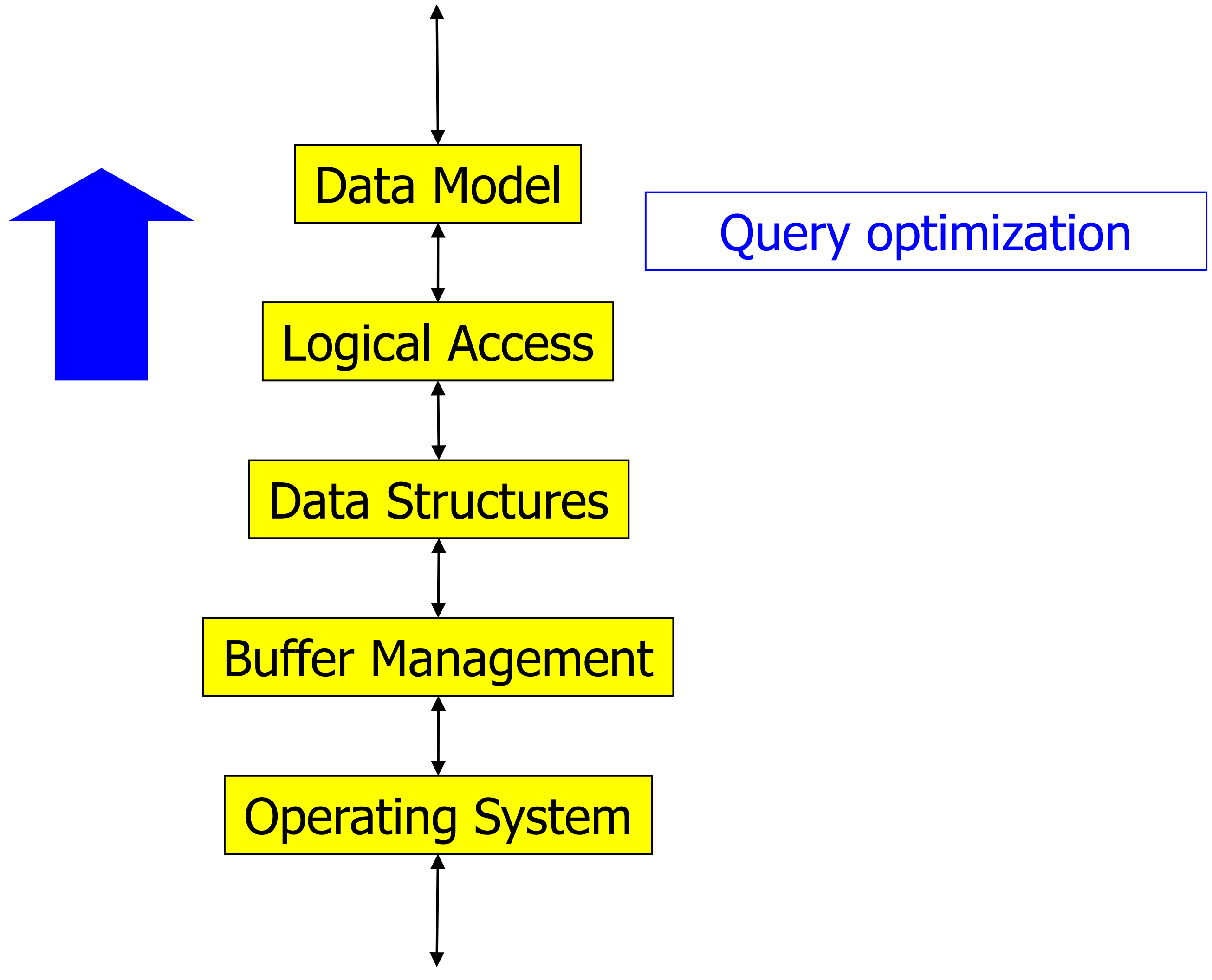
- Pro: Records can be moved within page
- Con: Slower access

Free Space, TX, and Concurrent Processes

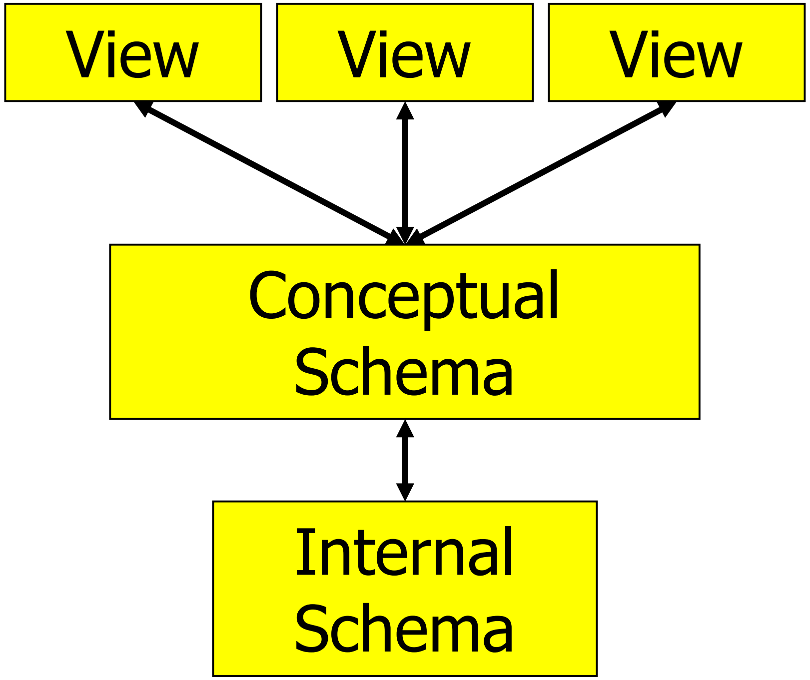


- Oracle procedure for finding free space
- Free space managed at the level of segments
  - Logical database objects
- Explanation
  - TFL: transaction free list
  - PFL: process free list
  - MFL: master free list
  - HWM: High water mark

Bottom-Up


Data Model
Query optimization
Logical Access
Data Structures
Buffer Management
Operating System

The ANSI/SPARC Three Layer-Model



Query rewriting, view expansion

Query execution plan generation and optimization: Access paths, join order, …

Execution of operators, pipelining

Query Processing

- Declarative query

```
SELECT Name, Address, Checking, Balance
FROM   customer C, account A
WHERE Name = "Bond" and C.Account# = A.Account#
```

- Translated in procedural Query Execution Plan (QEP)

```
FOR EACH c in CUSTOMER DO
  IF c.Name = "Bond" THEN
    FOR EACH  a IN ACCOUNT DO
      IF a.Account# = c.Account# THEN
        Output ("Bond", c.Address, a.Checking, a.Balance)
```

One Query – Many QEPs

```
SELECT    Name, Address, Checking, Balance
FROM      customer C, account A
WHERE     Name = "Bond" and C.Acco# = A.Acco#
```

```
FOR EACH c in CUSTOMER DO
  IF c.Name = "Bond" THEN
    FOR EACH  a IN ACCOUNT DO
      IF a.Acco# = c.Acco# THEN Output ("Bond", c.Address, a.Checking, a.Balance)

FOR EACH a in ACCOUNT DO
  FOR EACH  c IN CUSTOMER DO
    IF a.Acco# = c.Acco# THEN
      IF c.Name = "BOND" THEN Output ("Bond", c.Address, a.Checking, a. Balance)

FOR EACH c in CUSTOMER WITH Name="Bond" BY INDEX DO
  FOR EACH  a IN ACCOUNT DO
    IF a.Acco# = c.Acco# THEN Output ("Bond", c.Address, a.Checking, a. Balance)

FOR EACH c in CUSTOMER WITH Name="Bond" BY INDEX DO
  FOR EACH  a IN ACCOUNT with a.Acco#=c.Acco# BY INDEX DO
    Output ("Bond", c.Address, a.Checking, a. Balance)

…
```

Query optimization

- Task: Find the (hopefully) fastest QEP
- Two interdependent levels: Best plan, best implementation
  - Different QEPs by algebraic rewriting
    - P1: $\sigma_{Name=Bond}(Account \bowtie Customer)$
    - P2: $Account \bowtie \sigma_{Name=Bond}(Customer)$
  - Different QEPs by different operator implementations
    - P1': Access by scan, hash-join
    - P1'': Access by index, nested-loop-join
- Plan space: Enumerate and evaluate (some? all?) QEPs
- Optimization goal: Minimize size of intermediate results
  - Heuristic that proved helpful over four decades
  - Predicting actual runtime still is infeasible (and costly)

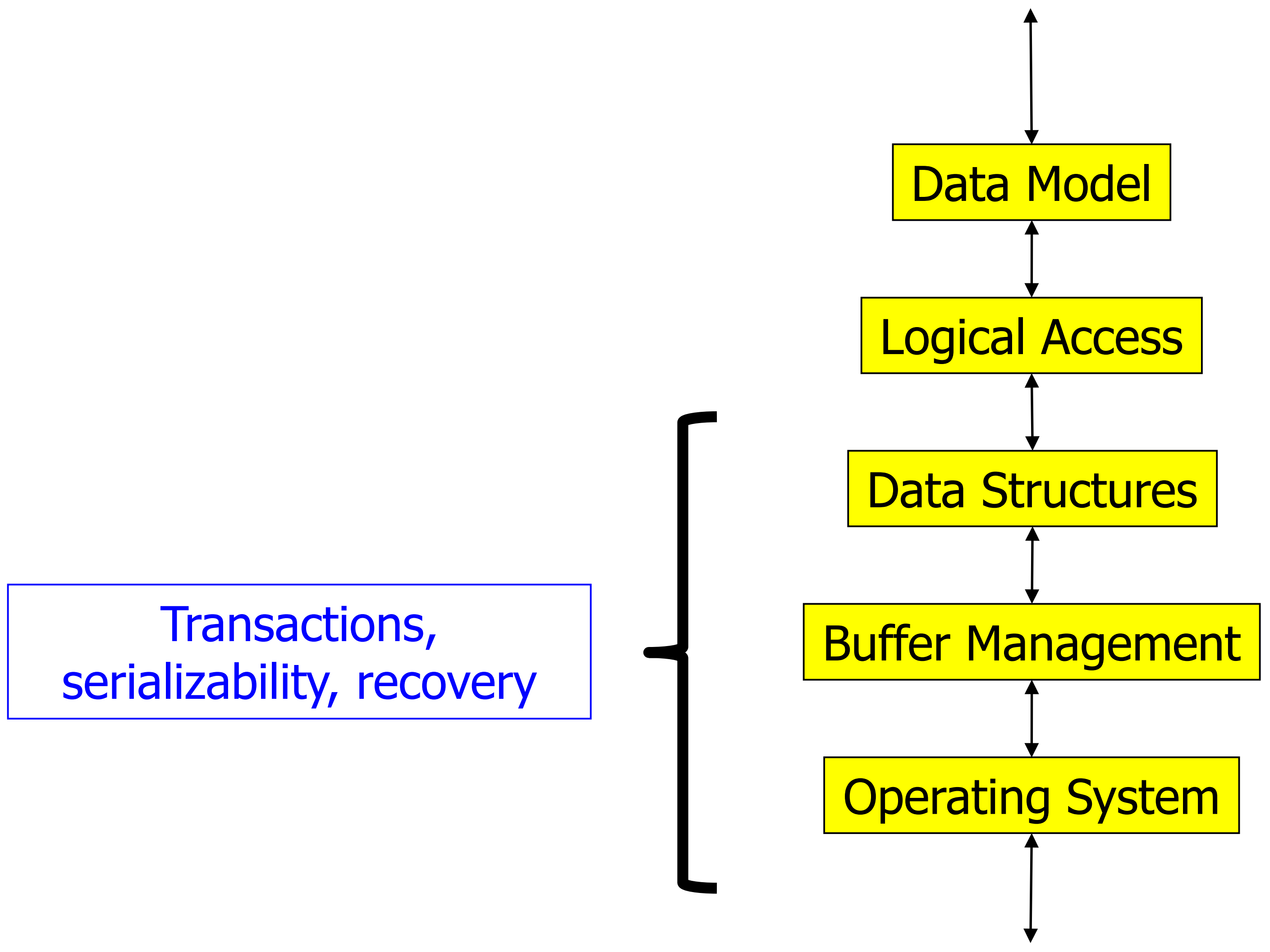
Cost-Based Optimizer

- Use statistics on current state of relations
  - Size, value distribution, fragmentation, cluster factors, …

```
FOR EACH a in ACCOUNT DO
  FOR EACH  c IN CUSTOMER DO
    IF a.Account# = c.Account# THEN
      IF c.Name = "BOND" THEN …
```

  - Let selectivity of $\sigma_{Name=Bond}$ be 1%, |Customer|=10.000, |Account|=12.000, Customer/Account evenly distributed
  - Performs …
    - Join: 10.000 * 12.000 = 120M comparisons
    - Produces ~12.000 intermediate result tuples
    - Filters down to ~120 results

Bottom-Up


Data Model
Logical Access
Data Structures
Buffer Management
Operating System
Transactions, serializability, recovery

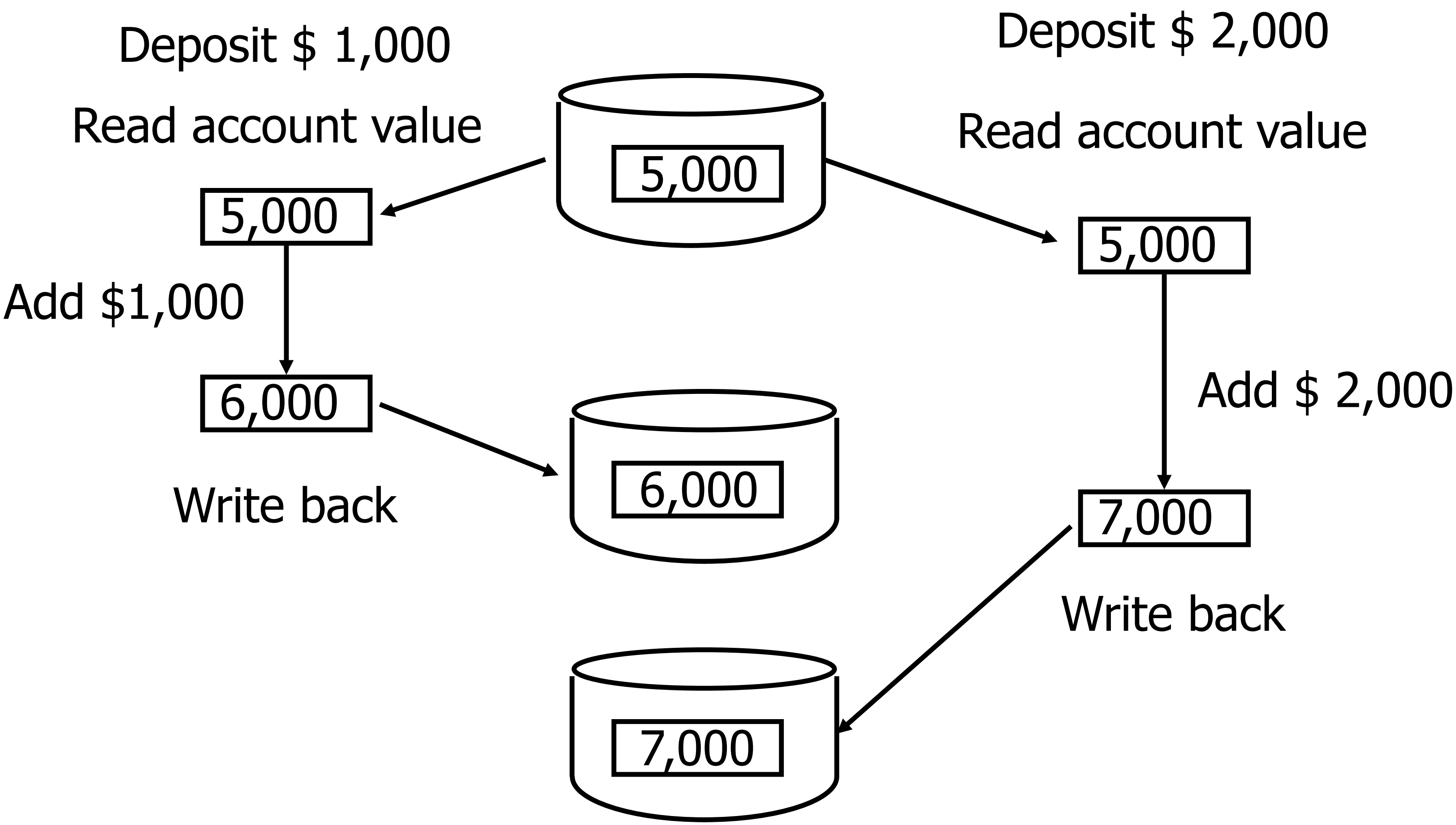
Transactions (TX)

- Transaction: "Logical unit of work"

```
Begin_Transaction
    UPDATE ACCOUNT
        SET Savings = Savings + 1M
        SET Checking = Checking - 1M
    WHERE Account# = 007;
    INSERT JOURNAL <007, NNN, "Transfer", ...>
End_Transaction
```

- ACID properties
  - Atomic execution
  - Consistent DB state after commits
  - Isolation: No influence on result by concurrent TX
  - Durability: After commit, changes are reflected in the database

Lost Update Problem



Synchronization and schedules

```
T1: read A;           T2: read B;
    A := A − 10;          B := B − 20;
    write A;              write B;
    read B;               read C;
    B := B + 10;          C := C + 20;
    write B;              write C;
```

Schedule S_1		Schedule S_2		Schedule S_3	
T_1	T_2	T_1	T_2	T_1	T_2
read A		**read** A		**read** A	
$A - 10$			**read** B	$A - 10$	
write A		$A - 10$			**read** B
read B			$B - 20$	**write** A	
$B + 10$		**write** A			$B - 20$
write B			**write** B	**read** B	
	read B	**read** B			**write** B
	$B - 20$		**read** C	$B + 10$	
	write B	$B + 10$			**read** C
	read C		$C + 20$	**write** B	
	$C + 20$	**write** B			$C + 20$
	write C		**write** C		**write** C

?

Synchronization and locks

- When is a schedule „fine"?
  - When it is serializable
  - I.e., when it is equivalent to a serial schedule
  - Proof serializability of schedules
- Strategy: Blocking everything is dreadful
- Strategy: Checking after execution is wasteful
- Synchronization protocols
  - Guarantee to produce only serializable schedules
  - Require certain well-behavior of transactions
    - Two phase locking, multi-version synchronization, timestamp synchronization, …
- Be careful with deadlocks