CSC 261/461 – Database Systems
Lecture 16

Fall 2017

Announcement

- CIRC account
- Quiz 6
  - Due: Monday at 11:59 pm)
- Project 1 Milepost 3
  - Due: Nov 10
- Project 2 Part 2 (Optional)
  - Due: Nov 15
- For graduate students:
  - We will provide:
    - Term paper feedback (yes/no)

The IO Model & External Sorting

Today's Lecture

1. Chapter 16 (Disk Storage, File Structure and Hashing)
2. Chapter 17 (Indexing)

Simplified Database System Environment

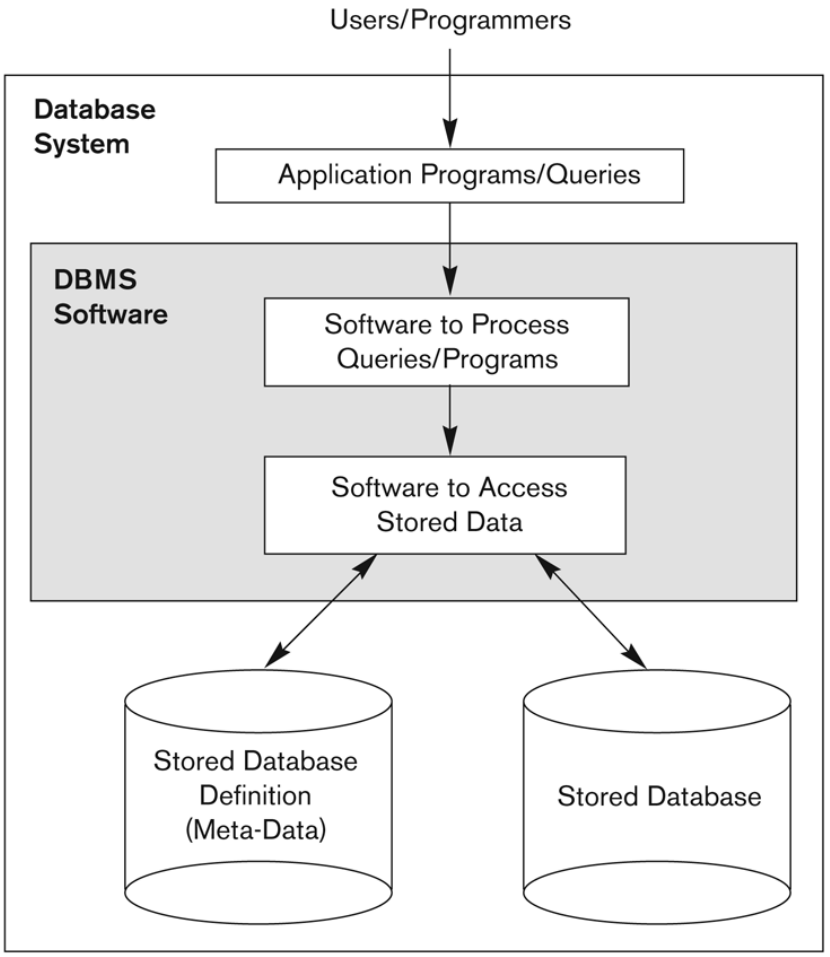


Figure 1.1
A simplified database system environment.

What you will learn about in this section

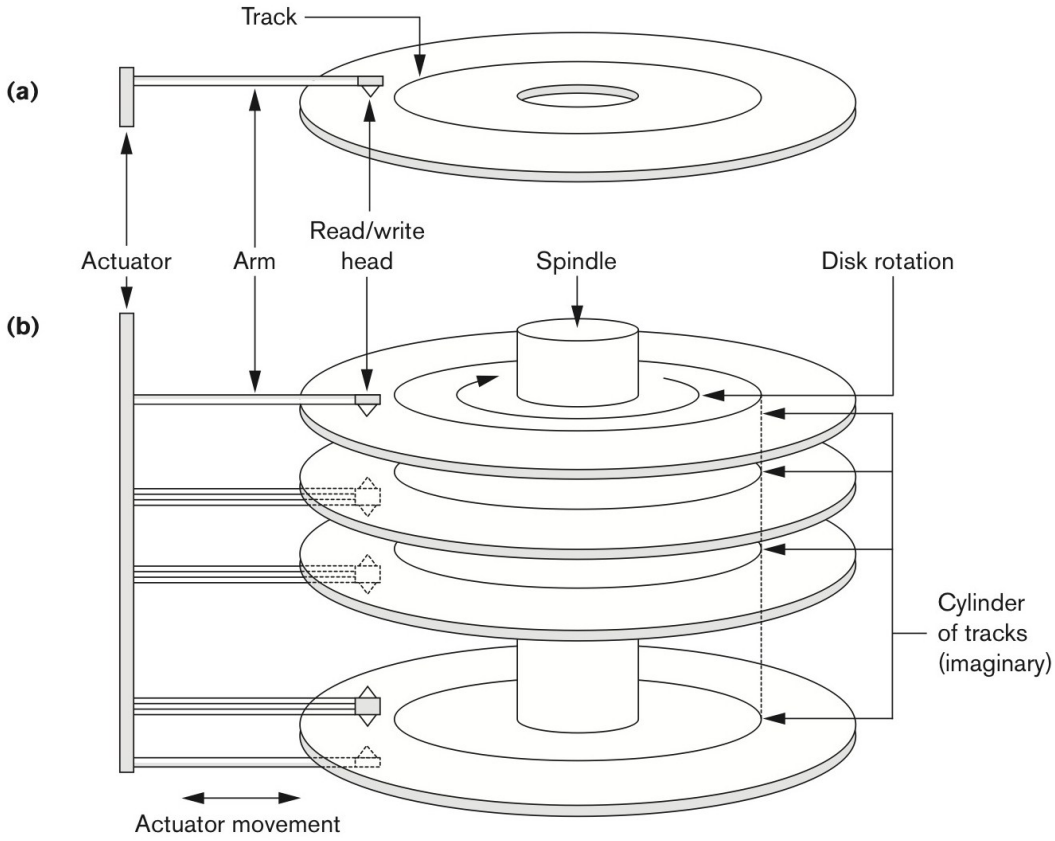
1. Storage and memory model

2. Buffer

1. THE BUFFER

High-level: Disk vs. Main Memory

- **Disk**:
  - Slow
    - Sequential access
      - (although fast sequential reads)
  - Durable
    - We will assume that once on disk, data is safe!
  - Cheap



High-level: Disk vs. Main Memory

- Random Access Memory (RAM) or **Main Memory**:
  - Fast
    - Random access, byte addressable
      - ~10x faster for sequential access
      - ~100,000x faster for random access!
  - *Volatile*
    - Data can be lost if e.g. crash occurs, power goes o
  - Expensive
    - For $100, get 16GB of RAM vs. 2TB of disk!

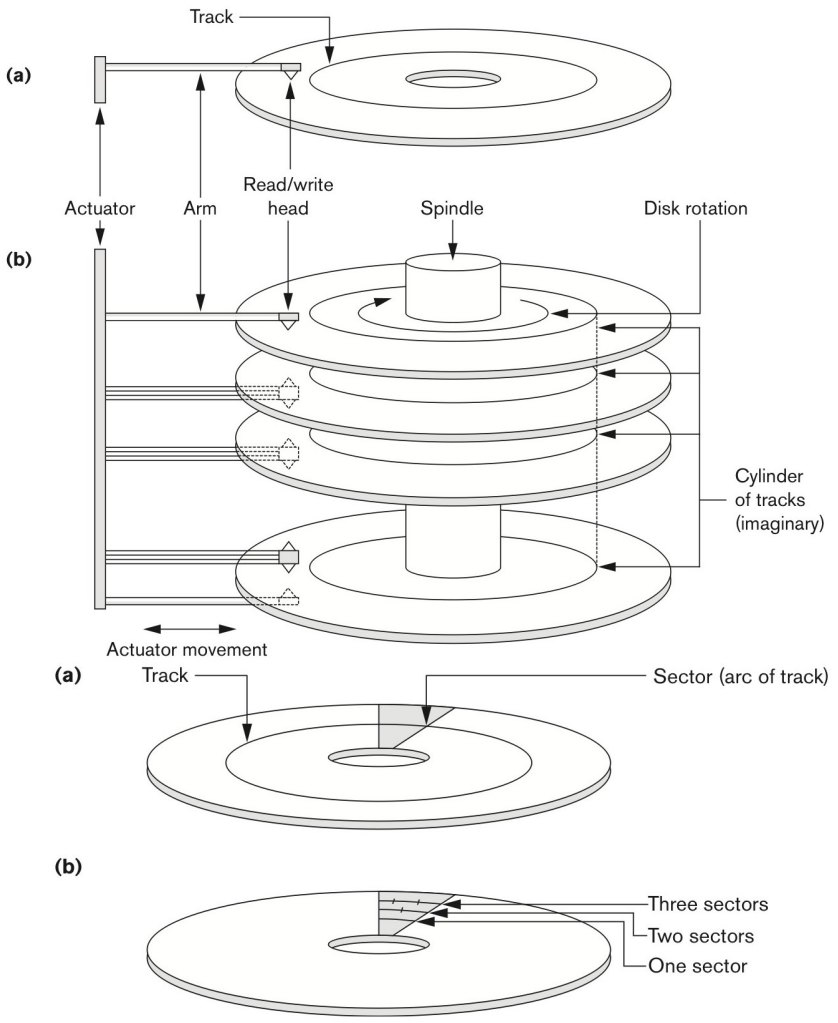
High-level: Disk vs. Main Memory

- Keep in mind the tradeoffs here as motivation for the mechanisms we introduce
  - Main memory: fast but limited capacity, volatile
    - Vs.
  - Disk: slow but large capacity, durable

How do we effectively utilize ***both*** ensuring certain critical guarantees?

Hardware Description of Disk Devices

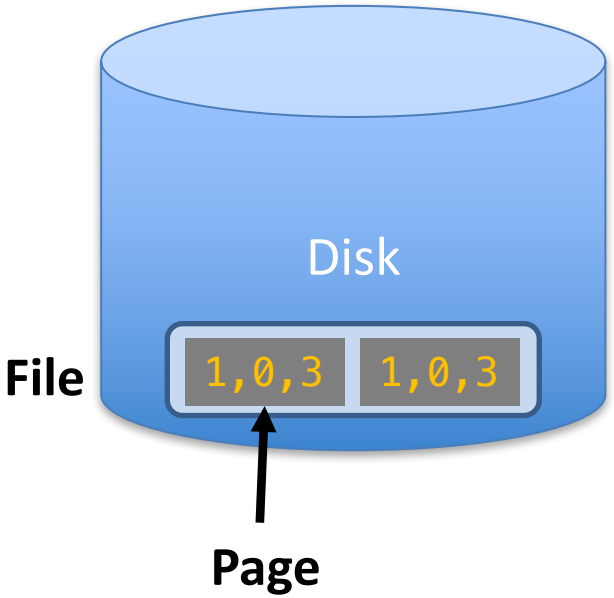
- Information is stored on a disk surface in concentric circles (Track)
- Tracks with same diameter on various surfaces is called cylinder
- Tracks are divided into sectors
- OS divides a track into equal sized disk blocks (pages)
  - One page = one or more sectors



A Simplified Filesystem Model

- For us, a **page** is a ***fixed-sized array*** of memory
  - One (or more) disk block (blocks)
  - Interface:
    - write to an entry (called a **slot**) or set to "None"

- And a **file** is a *variable-length list* of pages
  - Interface: create / open / close; next_page(); etc.



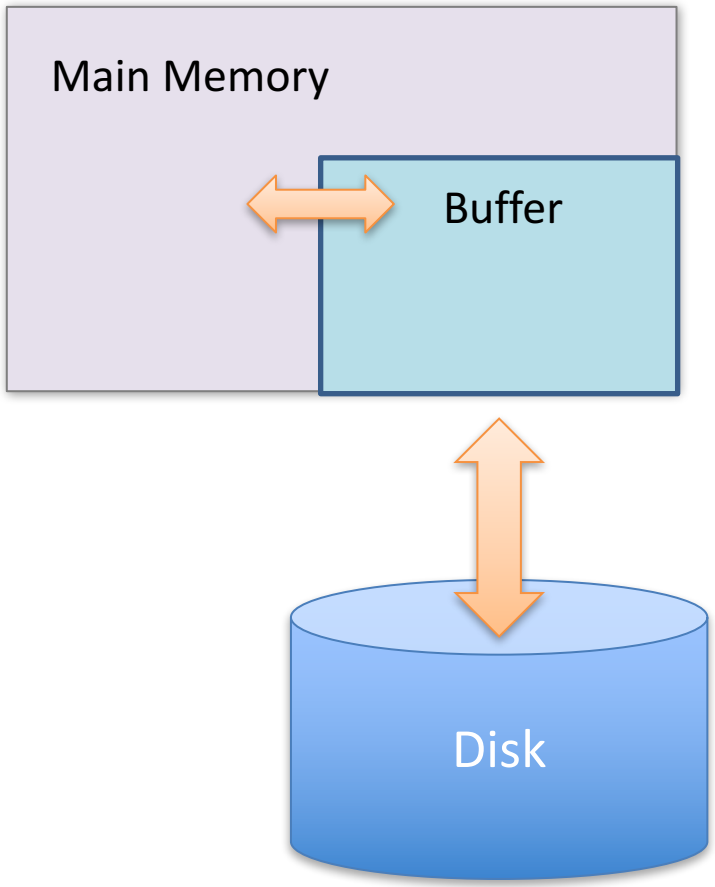
The Buffer

- Transfer of data between main memory and disk takes place in units of disk blocks.
- The hardware address of a block is a combination of a cylinder number, track number, and block number.

- A **buffer** is a region of physical memory used to store a single block.

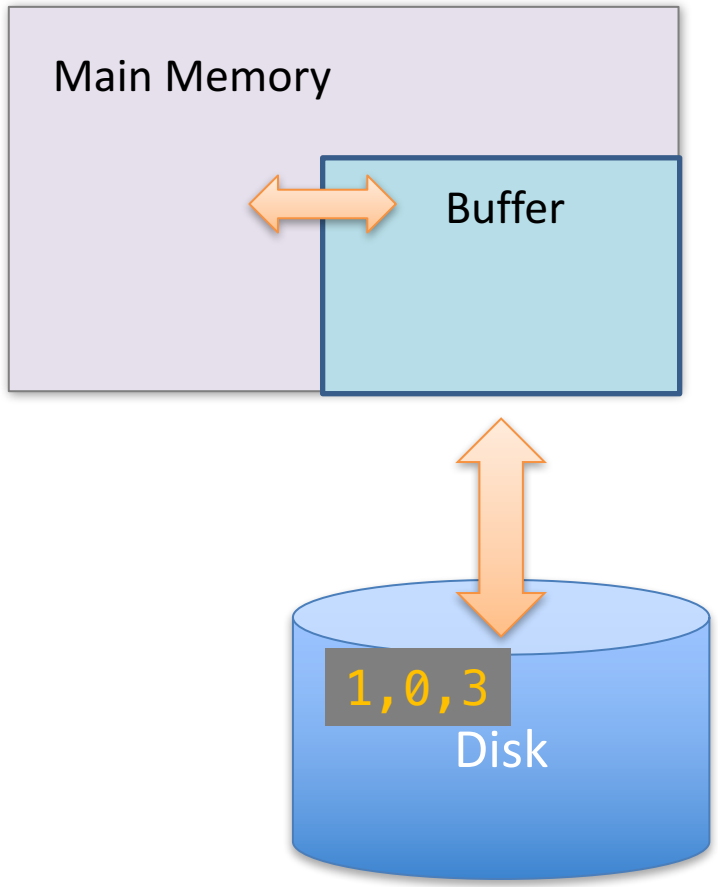
- Sometimes, several contiguous blocks can be copied into a cluster
  - In this lecture: We will mostly not distinguish between a buffer and a cluster.

- *Key idea:* Reading / writing to disk is slow- need to cache data!



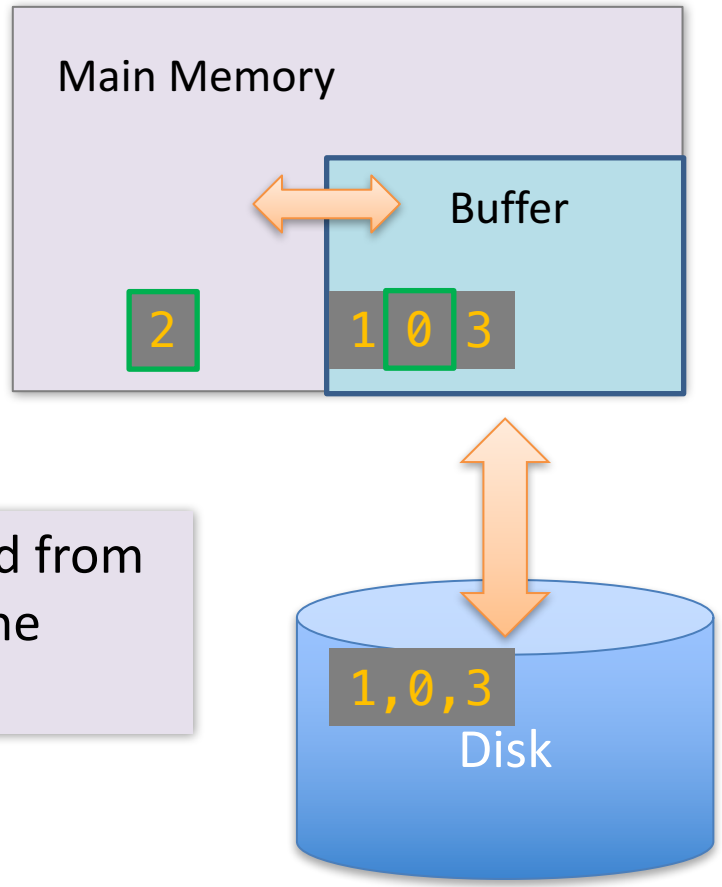
The (Simplified) Buffer

- In this class: We'll consider a buffer located in **main memory** that operates over **pages** and **files**:
  - **Read(page):** Read page from disk -> buffer *if not already in buffer*



The (Simplified) Buffer

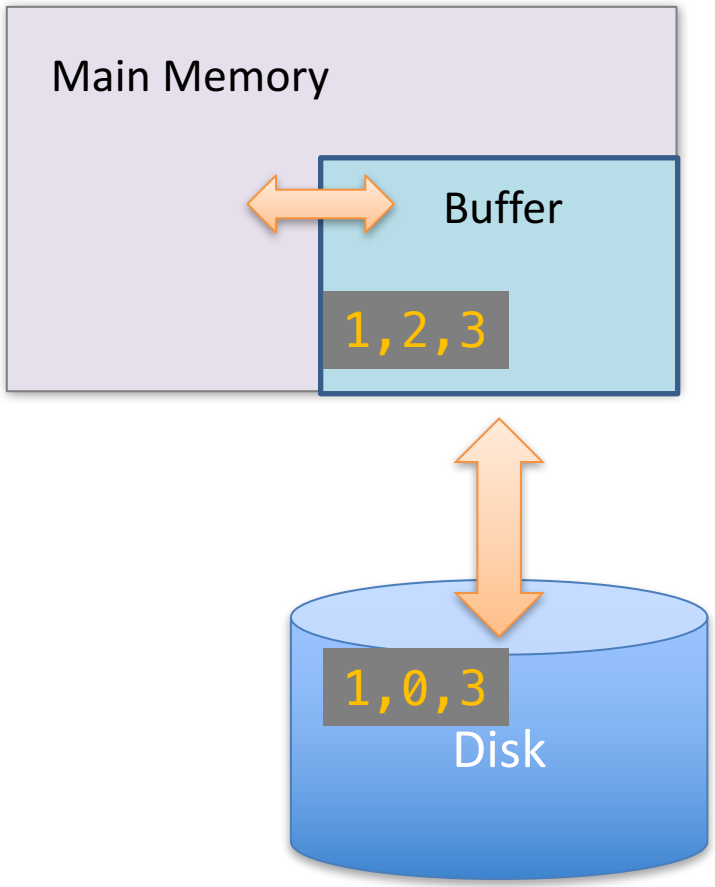
- In this class: We'll consider a buffer located in **main memory** that operates over **pages** and **files**:
  - **Read(page):** Read page from disk -> buffer *if not already in buffer*



Processes can then read from / write to the page in the buffer

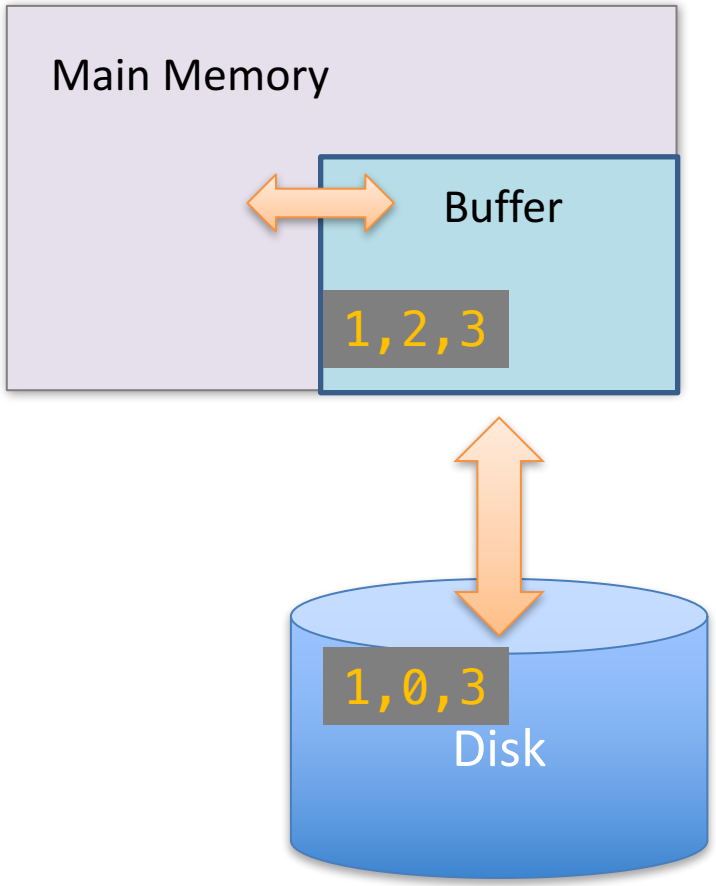
The (Simplified) Buffer

- In this class: We'll consider a buffer located in **main memory** that operates over **pages** and **files**:
  - **Read(page):** Read page from disk -> buffer *if not already in buffer*
  - **Flush(page):** Evict page from buffer & write to disk



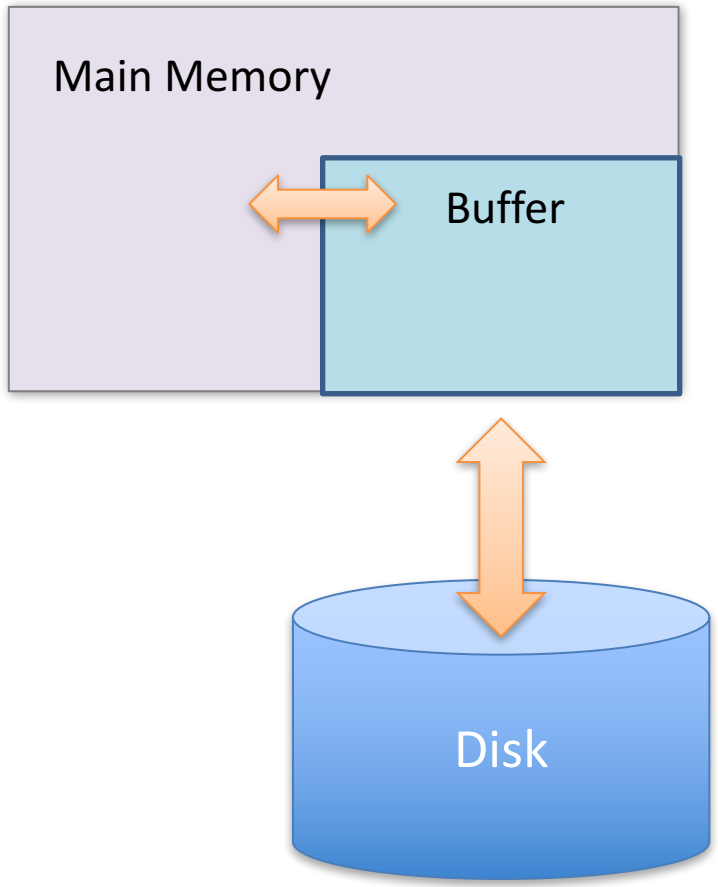
The (Simplified) Buffer

- In this class: We'll consider a buffer located in **main memory** that operates over **pages** and **files**:
  - **Read(page):** Read page from disk -> buffer *if not already in buffer*
  - **Flush(page):** Evict page from buffer & write to disk
  - **Release(page):** Evict page from buffer *without* writing to disk



Managing Disk: The DBMS Buffer

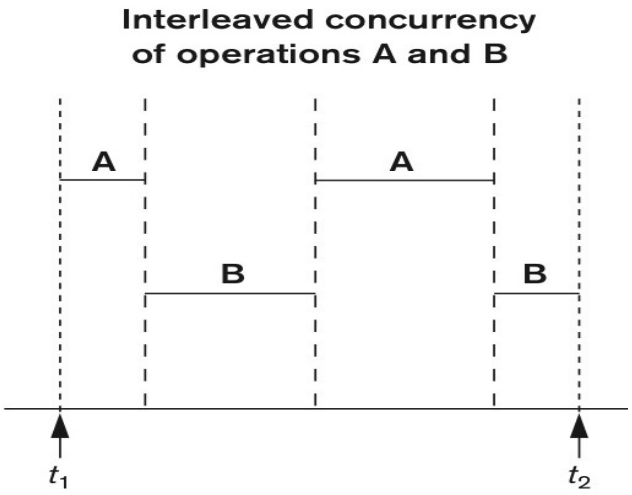
- Database maintains its own buffer
  - Why? The OS already does this…
  - DB knows more about access patterns.
  - Recovery and logging require ability to **flush** to disk.

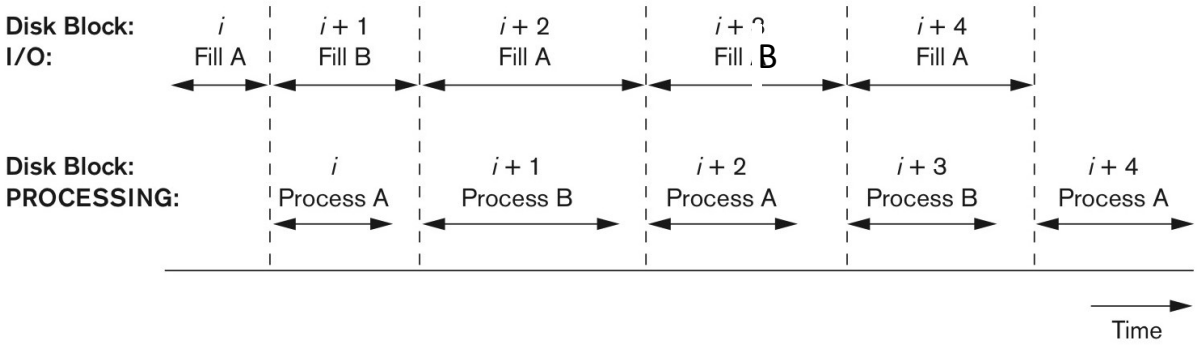


The Buffer Manager

- A **buffer manager** handles supporting operations for the buffer:
  - Primarily, handles & executes the "replacement policy"
    - i.e. finds a page in buffer to flush/release if buffer is full and a new page needs to be read in
  - DBMSs typically implement their own buffer management routines

Use of Two Buffer





Buffer Replacement Strategies

- Least recently used (LRU)
- Clock policy
- First-in-first-out (FIFO)

- Refer 16.3..2 for details

Records and Files

- Data is usually stored in the form of records
- Each record consists of a collection of related data values or items.
  - Record usually describe entities

File Types

- Unordered Records (Heap Files)
- Ordered Records (Sorted Files)

Heap Files

- Insertion (of a record):
  - Very efficient.
  - Last disk block is copied into a buffer
  - New record is added
  - Block is rewritten back to disk
- Searching:
  - Linear search
- Deletion:
  - Rewrite empty block after deleting record. (or)
  - Use deletion marker

Sorted Files

- Physically sort the records of a file
  - Based on the values of one of the fields (ordering fields)
  - Ordered and sequential file

- Searching:
  - Can perform Binary Search.
- Insertion and Deletion:
  - Expensive

Average Access Times for a File of *b* Blocks under Basic File Organizations

Table 16.3 Average Access Times for a File of *b* Blocks under Basic 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	$\log_2 b$

2. EXTERNAL MERGE & SORT

Challenge: Merging Big Files with Small Memory

How do we *efficiently* merge two sorted files when both are much larger than our main memory buffer?

External Merge Algorithm

- **Input**: 2 sorted lists of length M and N
- **Output:** 1 sorted list of length M + N
- **Required:** At least 3 Buffer Pages
- **IOs**: 2(M+N)

Key (Simple) Idea

To find an element that is no larger than all elements in two lists, one only needs to compare minimum elements from each list.

If:

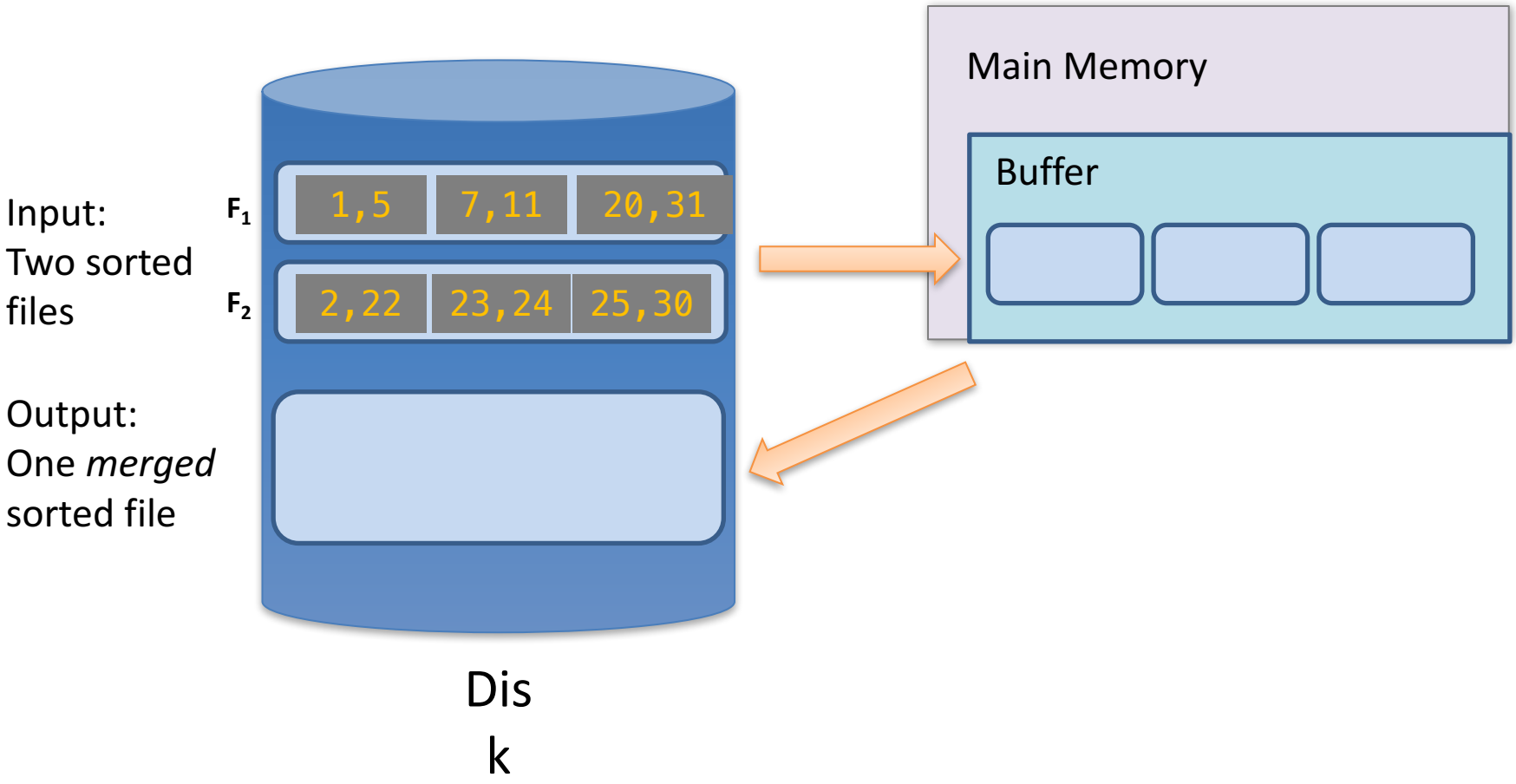
$$A_1 \leq A_2 \leq \cdots \leq A_N$$
$$B_1 \leq B_2 \leq \cdots \leq B_M$$

Then:

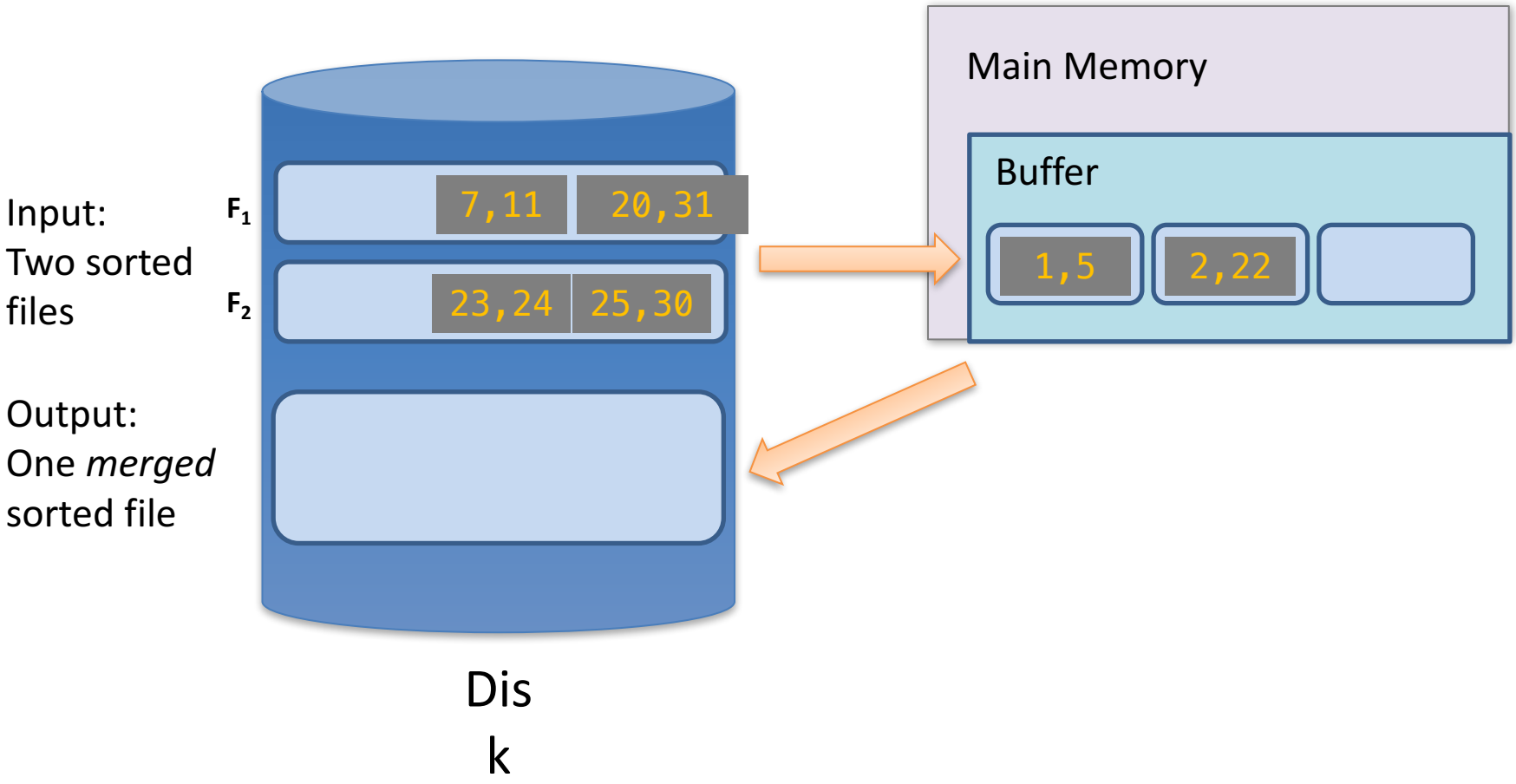
$$Min(A_1, B_1) \leq A_i$$
$$Min(A_1, B_1) \leq B_j$$

for i=1….N and j=1….M

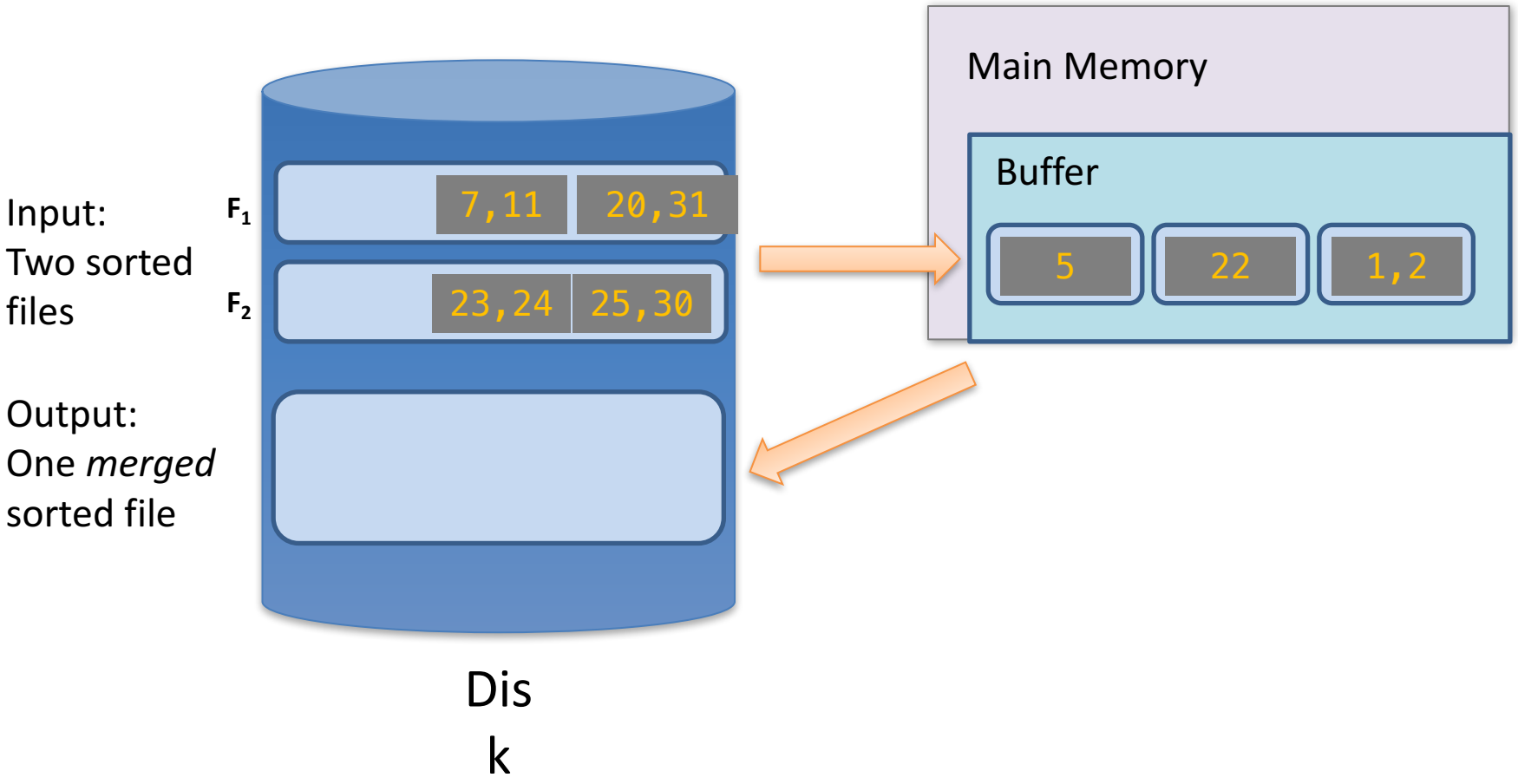
External Merge Algorithm



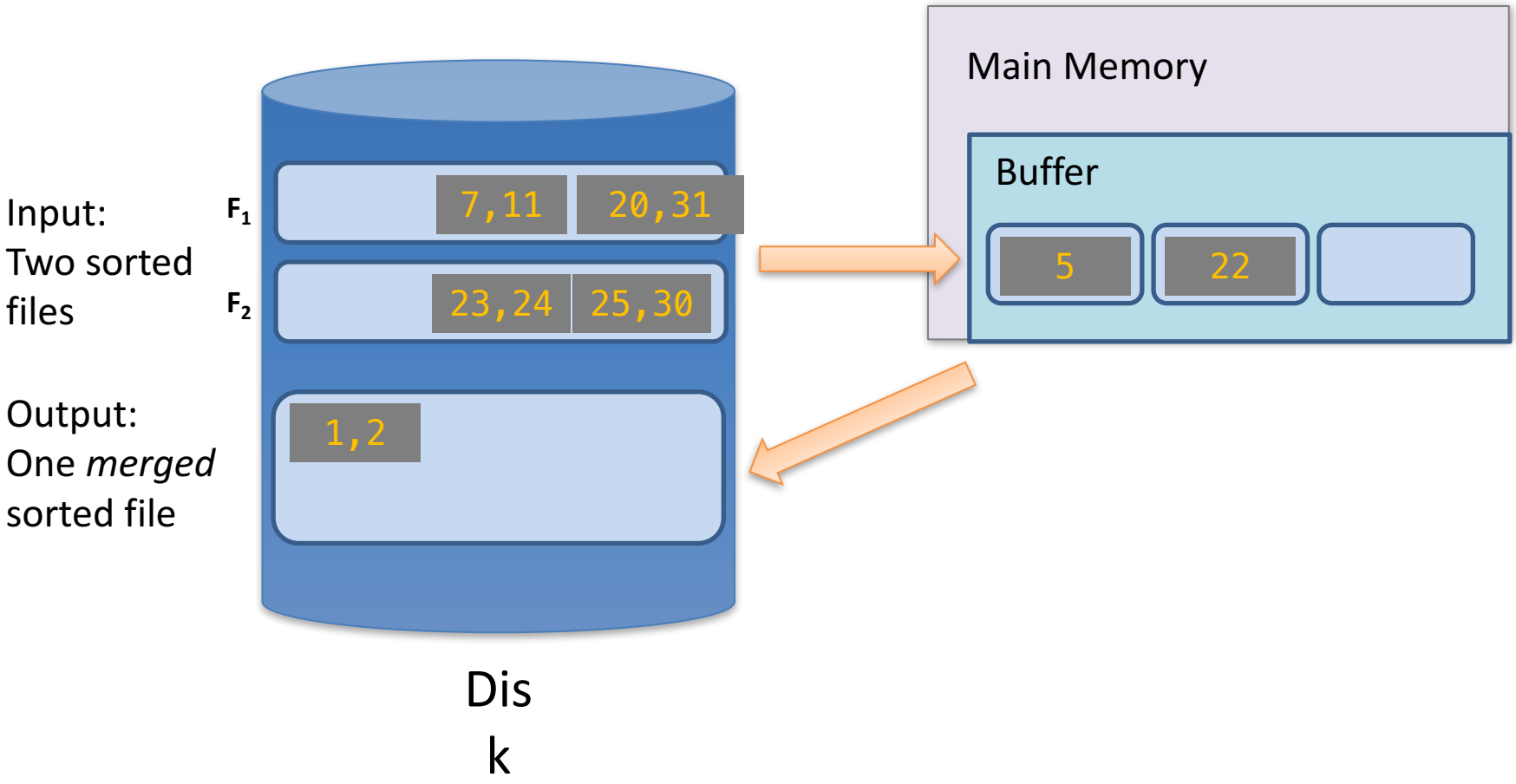
External Merge Algorithm



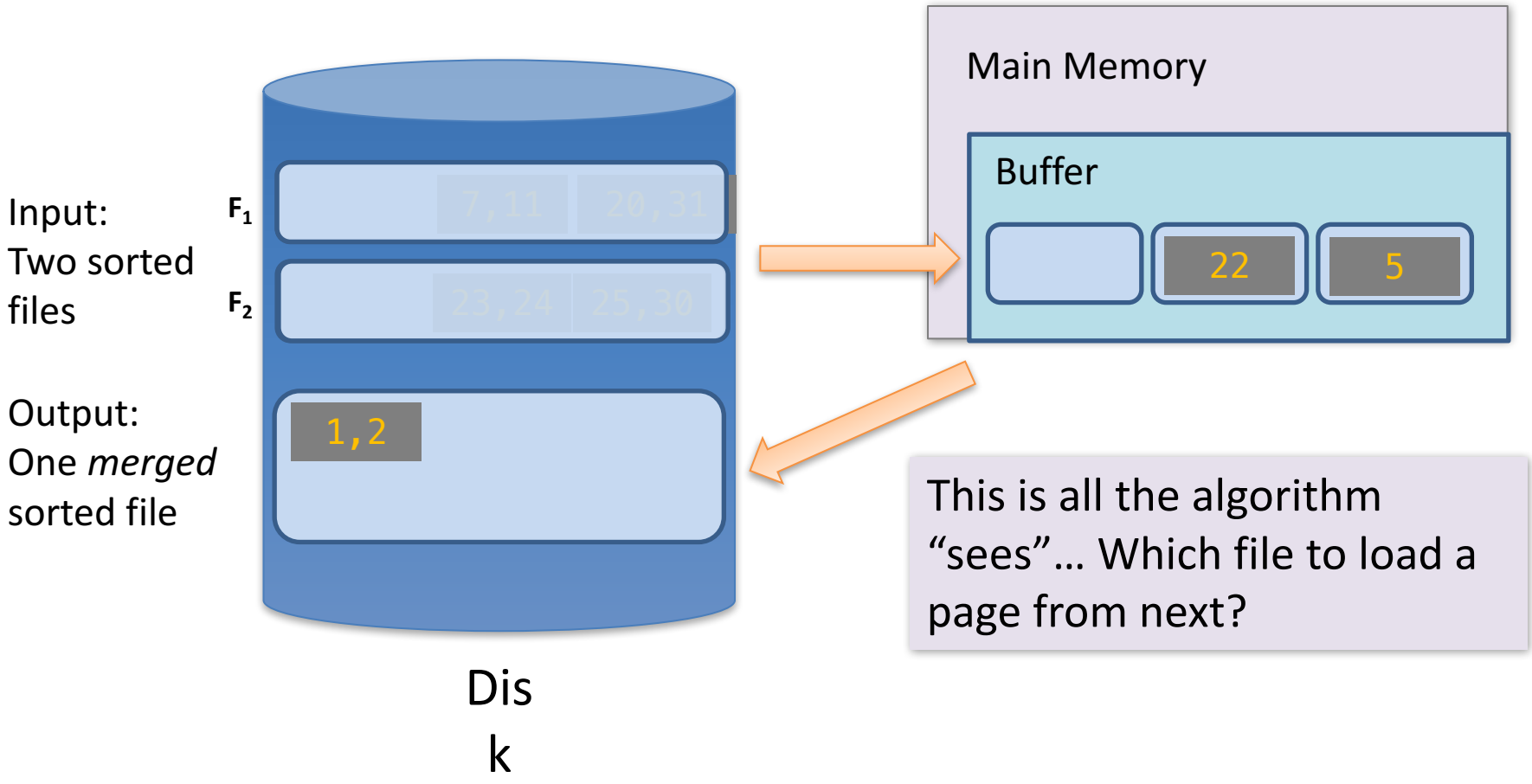
External Merge Algorithm



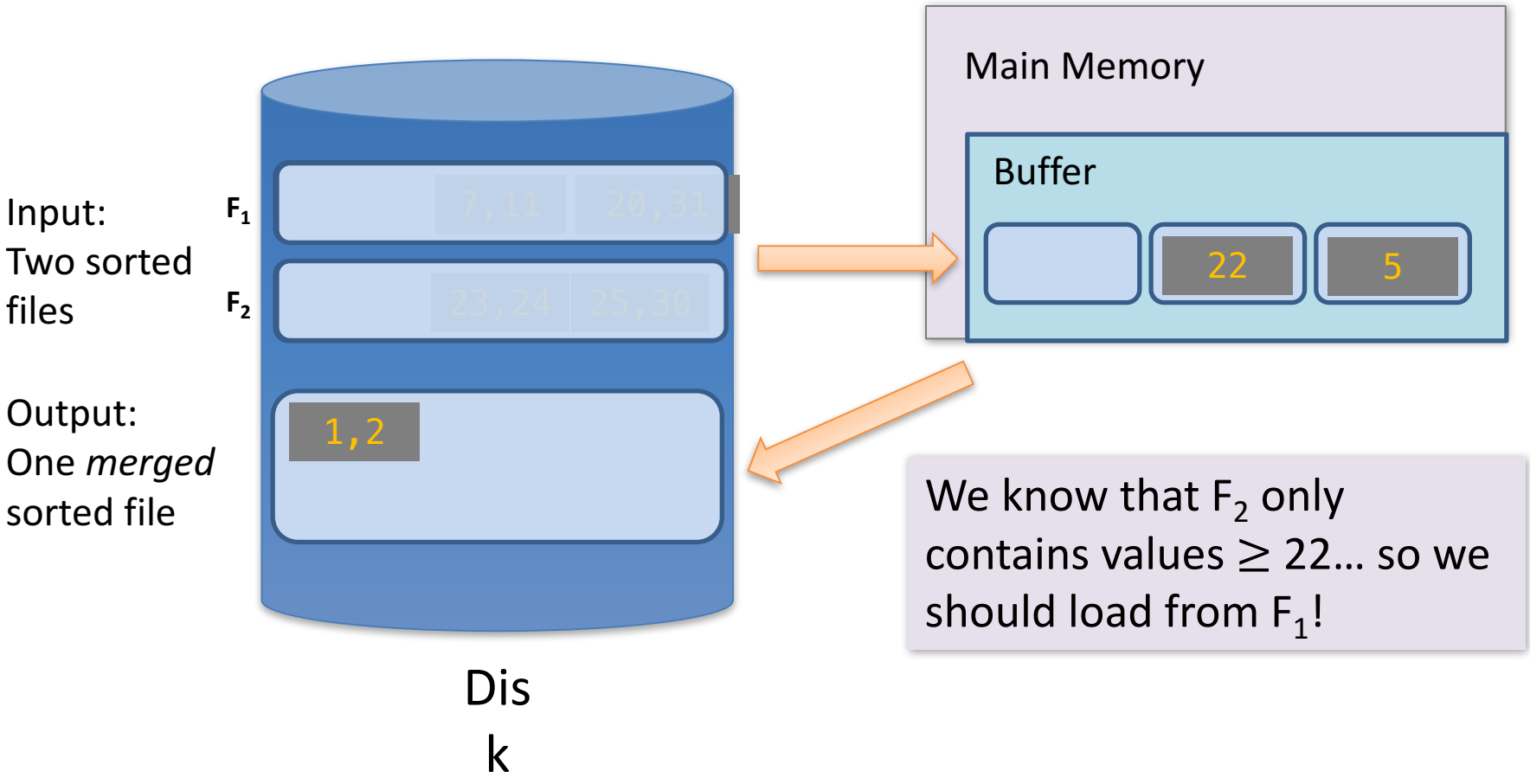
External Merge Algorithm



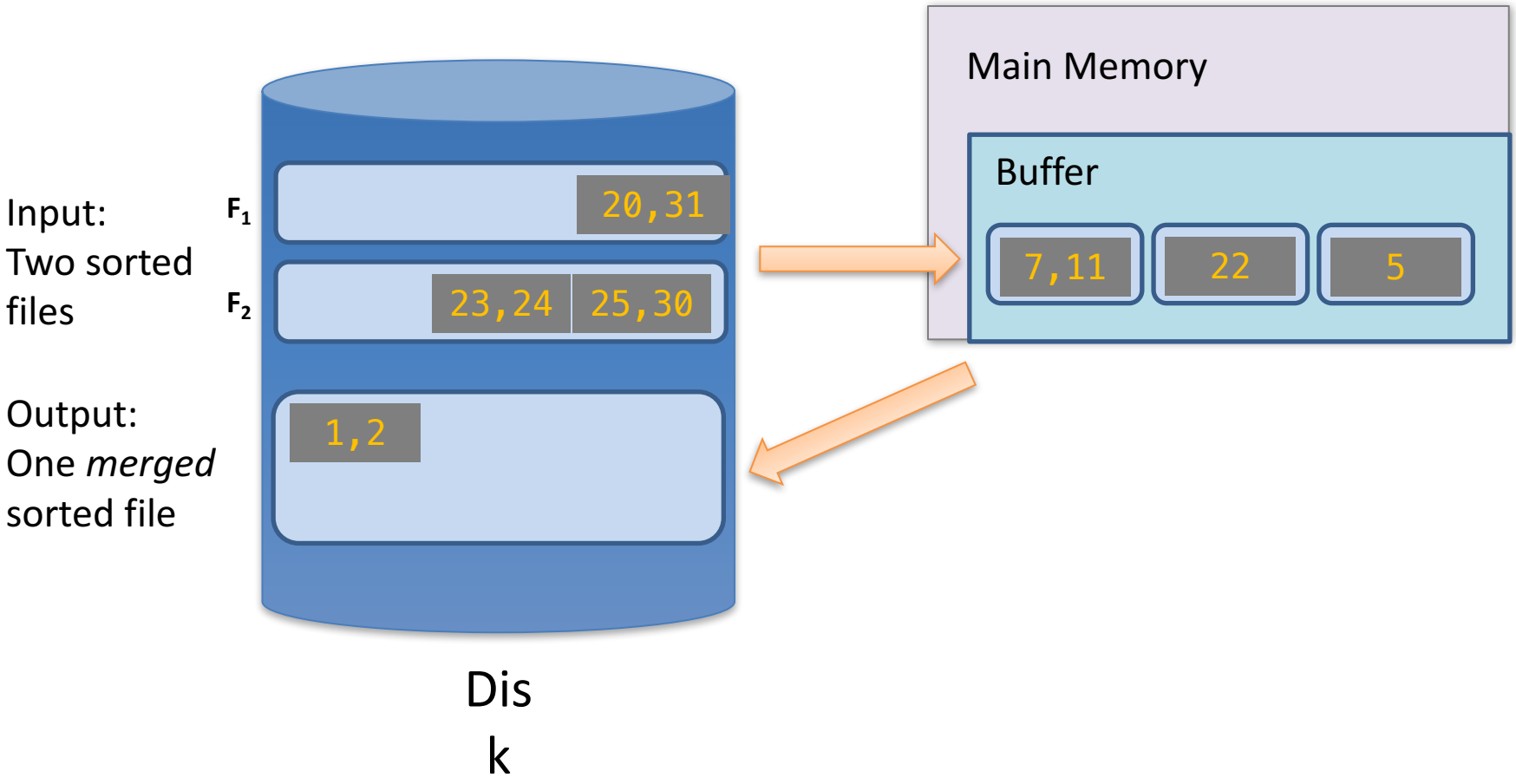
External Merge Algorithm



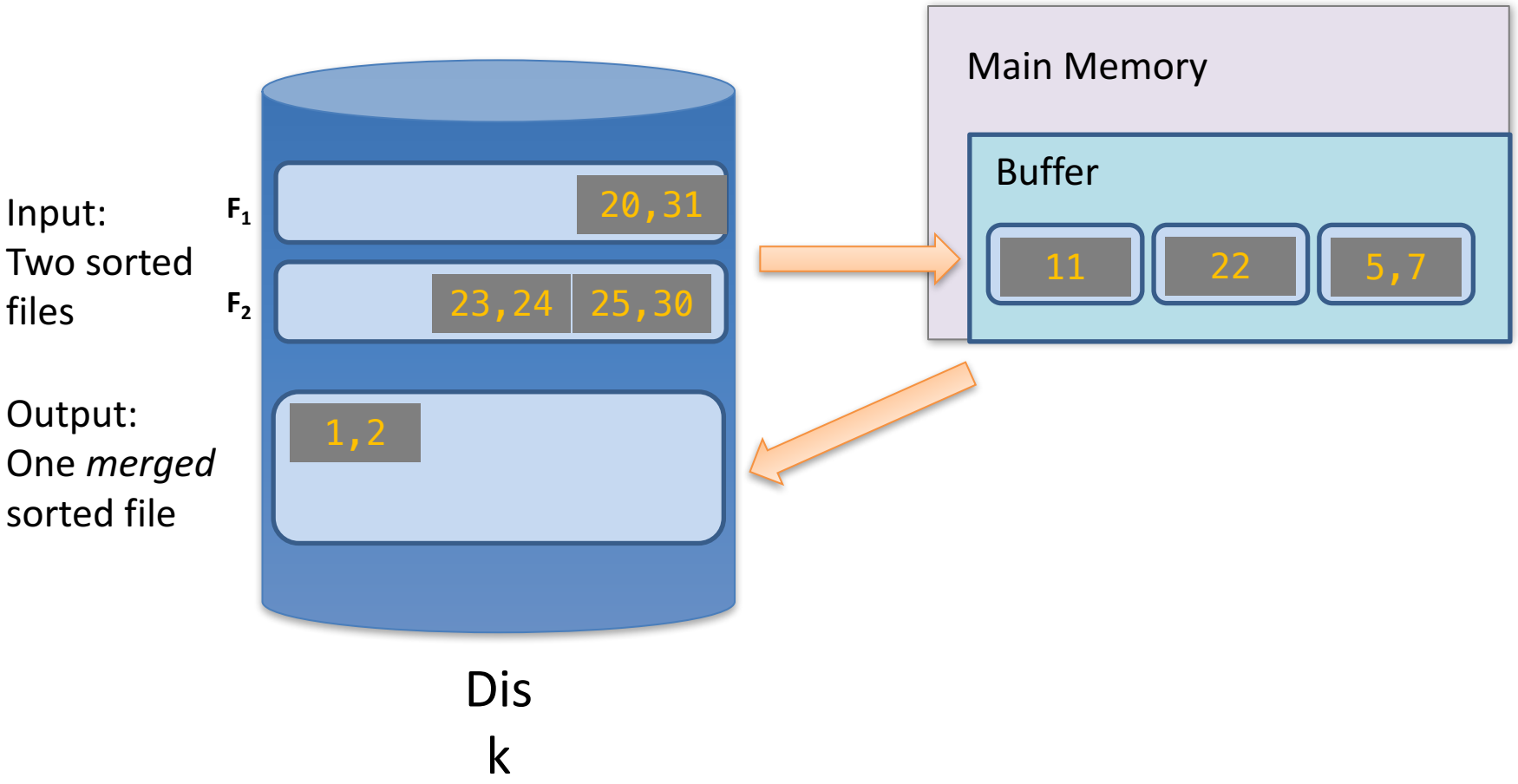
External Merge Algorithm



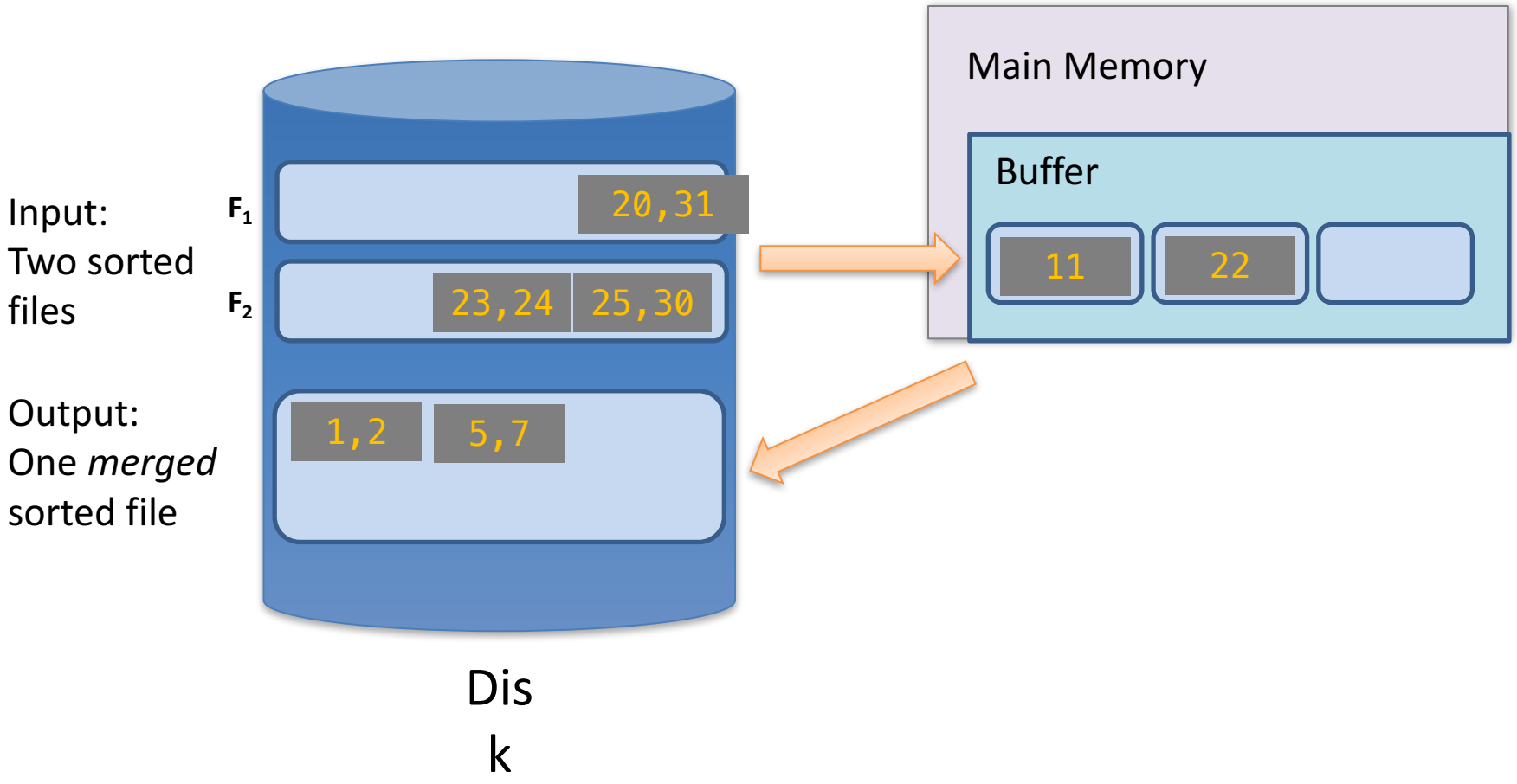
External Merge Algorithm



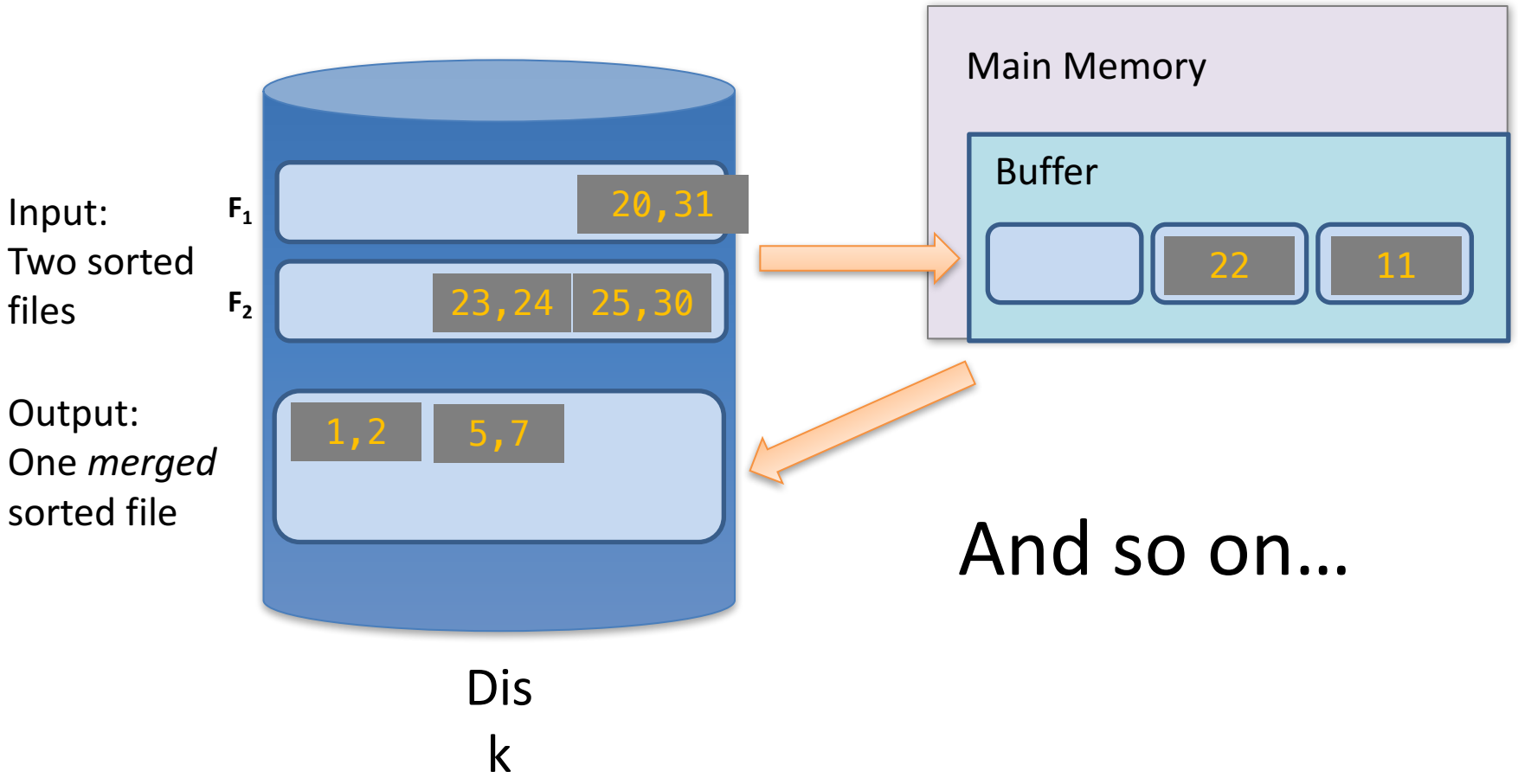
External Merge Algorithm



External Merge Algorithm



External Merge Algorithm



We can merge lists of **arbitrary length** with *only* 3 buffer pages.

If lists of size M and N, then

Cost: 2(M+N) IOs

Each page is read once, written once

2. HASHING TECHNIQUES

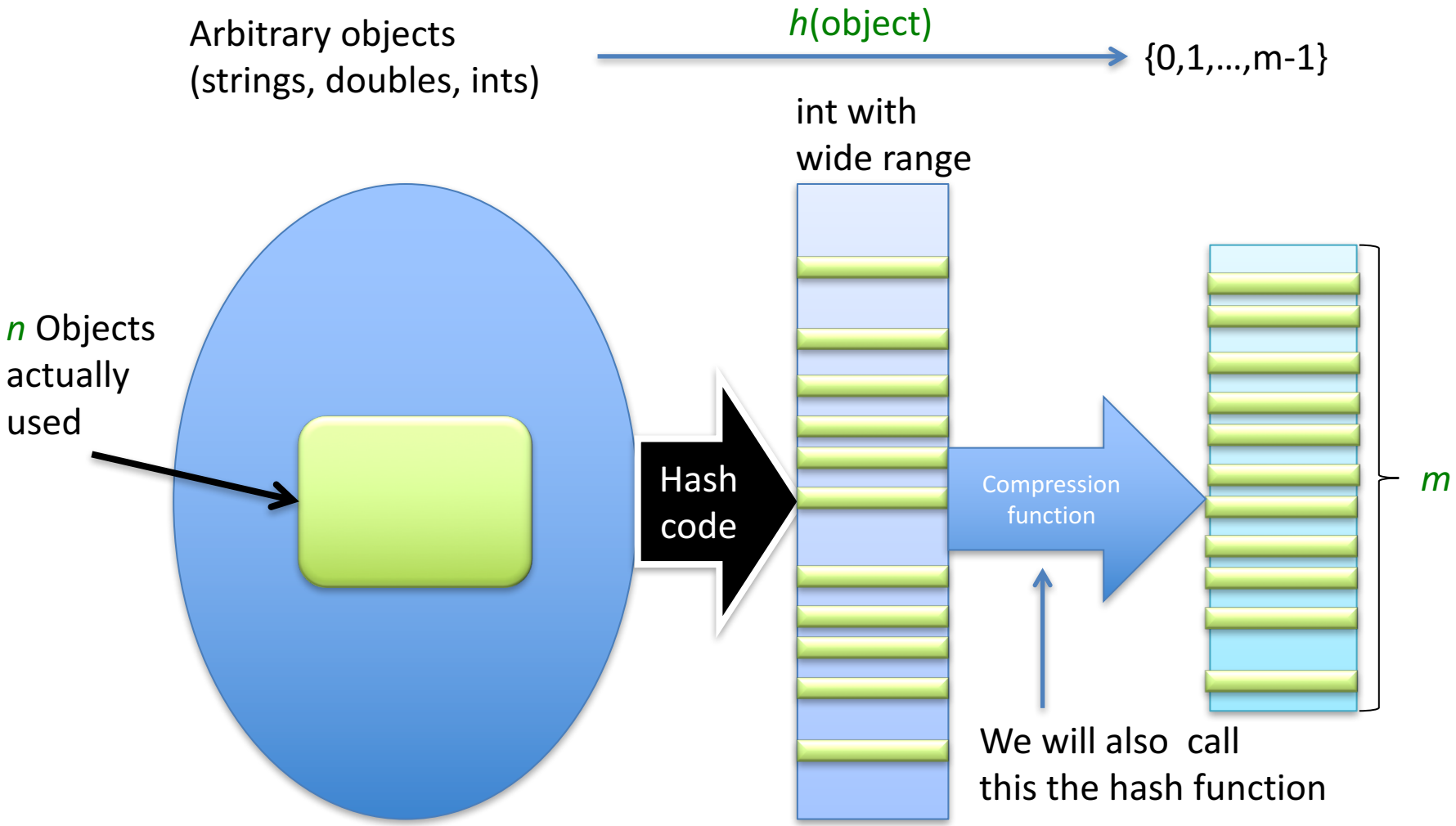
- Sears, Roebuck and Co., Consumers Guide, 1897

“IF YOU DON’T FIND IT IN THE INDEX, LOOK VERY CAREFULLY THROUGH THE ENTIRE CATALOG”

HASING – GENERAL IDEAS

- Hashing first proposed by Arnold Dumey (1956)
- Hash codes
- Chaining
- Open addressing

Top level view



Good Hash Function

- If $\text{key}_1 \neq \text{key}_2$, then it's extremely unlikely that $h(\text{key}_1) = h(\text{key}_2)$
  - Collision problem!

- Pigeonhole principle
  - K+1 pigeons, K holes → at least one hole with ≥ 2 pigeons

Division method

$$h(s) = s \mod m$$

- How does this function perform for different m?

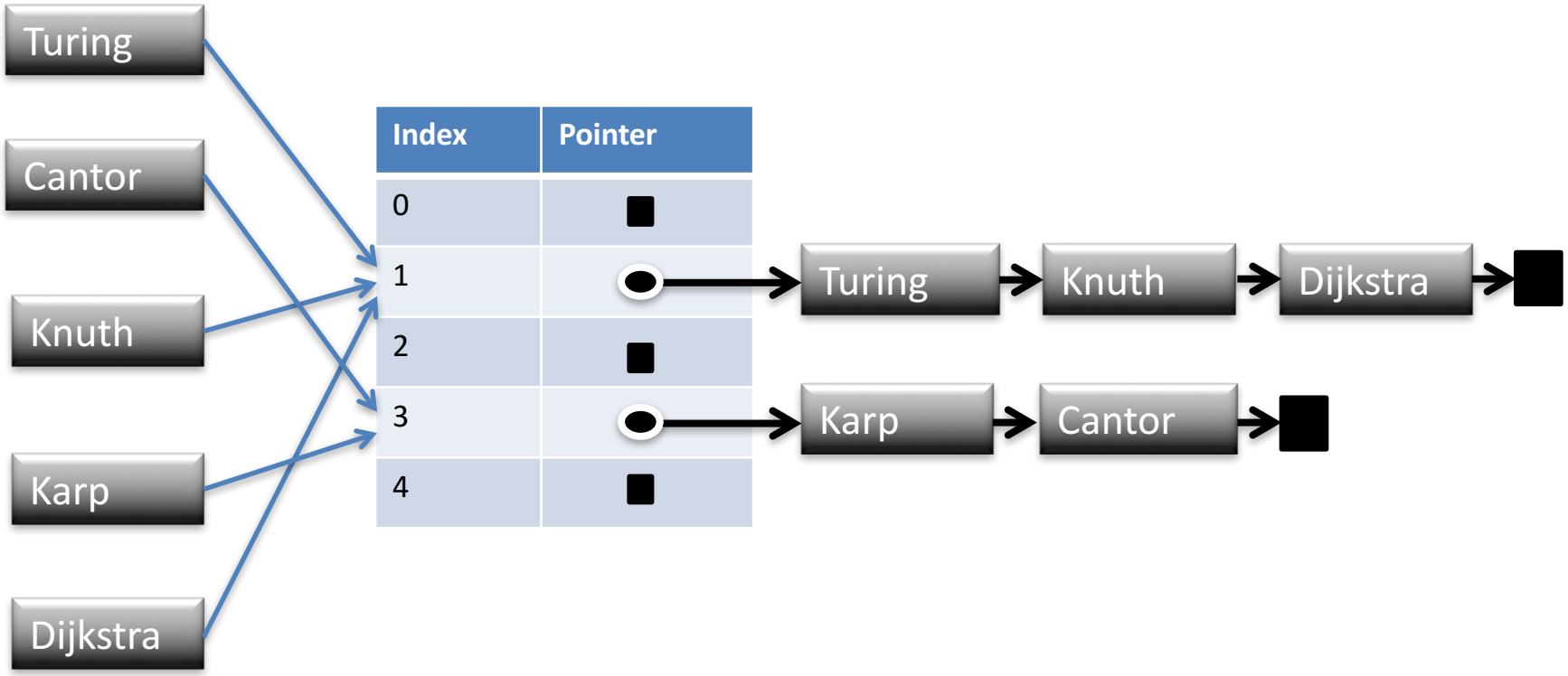
Separate chaining

Open addressing

Cuckoo hashing

COLLISION RESOLUTION

Separate Chaining



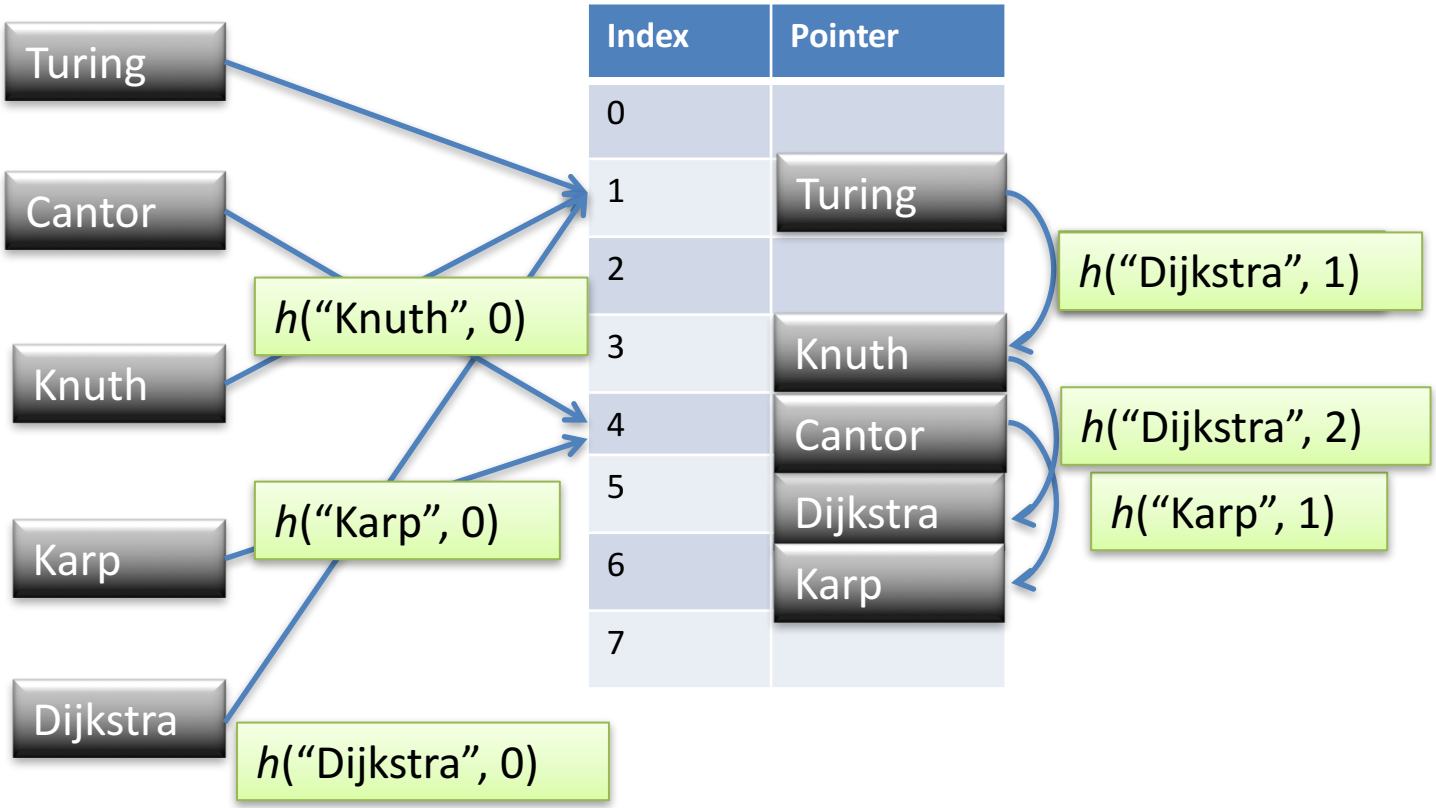
Open Addressing

- Store all entries in the hash table itself, no pointer to the "outside"

- Advantage
  - Less space waste
  - Perhaps good cache usage

- Disadvantage
  - More complex collision resolution
  - Slower operations

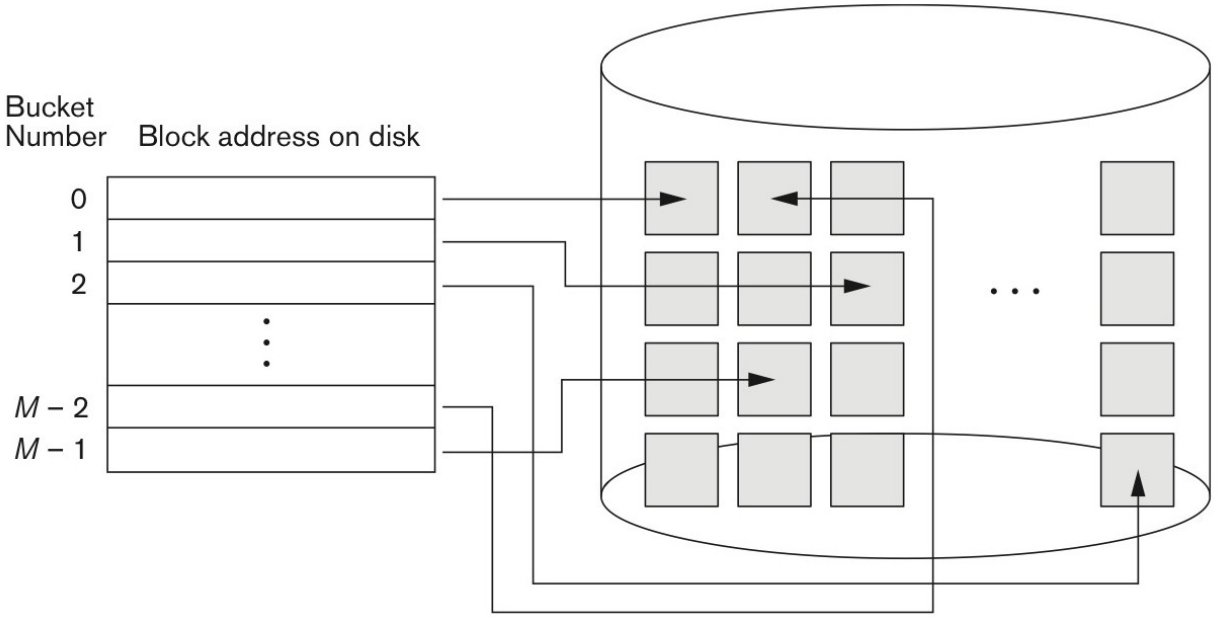
Open Addressing



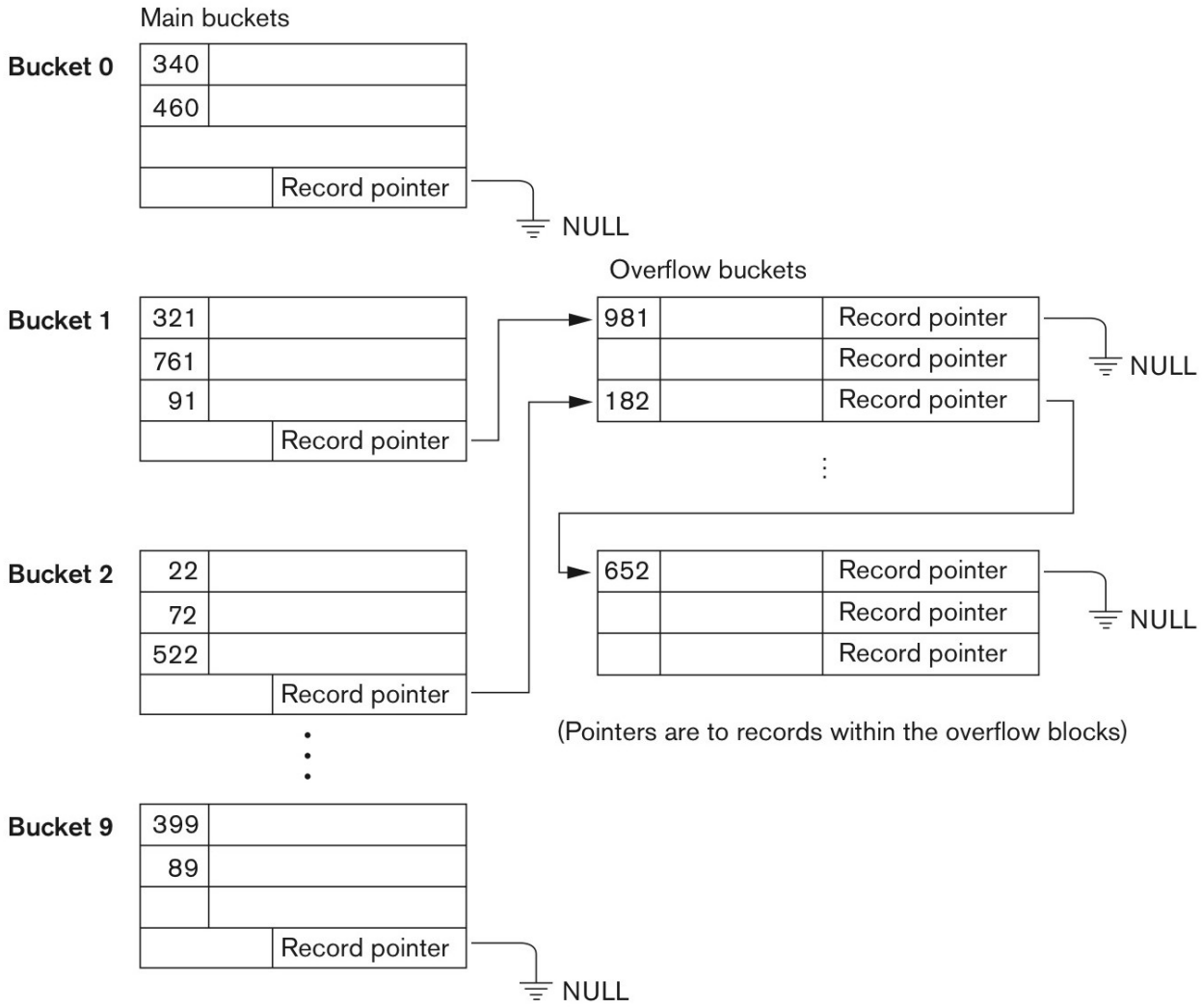
External Hashing

- Hashing for disk files
- Target address space is made of buckets
- Hashing function maps a key into relative bucket number
- Convert the bucket number into corresponding disk block address

Bucket Number to Disk Block address



Bucket Number to Disk Block address

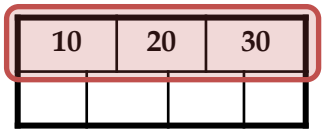


1. B+ TREES

B+ Trees

- Search trees
  - B does not mean binary!
- Idea in B Trees:
  - make 1 node = 1 physical page
  - Balanced, height adjusted tree (not the B either)
- Idea in B+ Trees:
  - Make leaves into a linked list (for range queries)

B+ Tree Basics

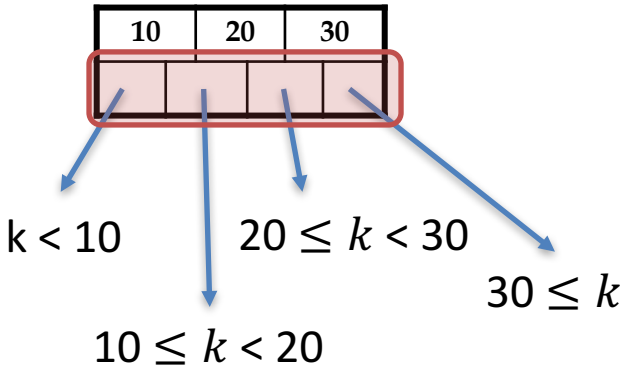


Parameter ***d*** = the degree

Each *non-leaf ("interior")* ***node*** has ≥ d and ≤ 2d ***keys****

except for root node, which can have between **1 and 2d keys*

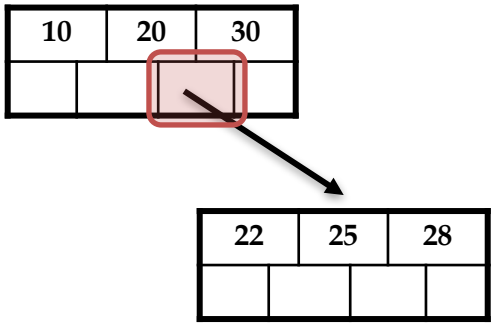
B+ Tree Basics



The *n* keys in a node define *n+1* ranges

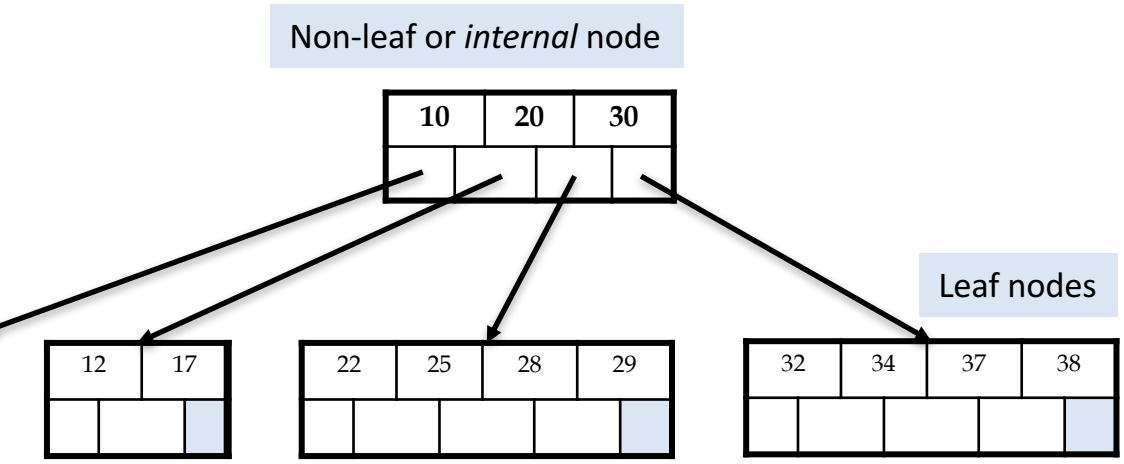
B+ Tree Basics

Non-leaf or *internal* node



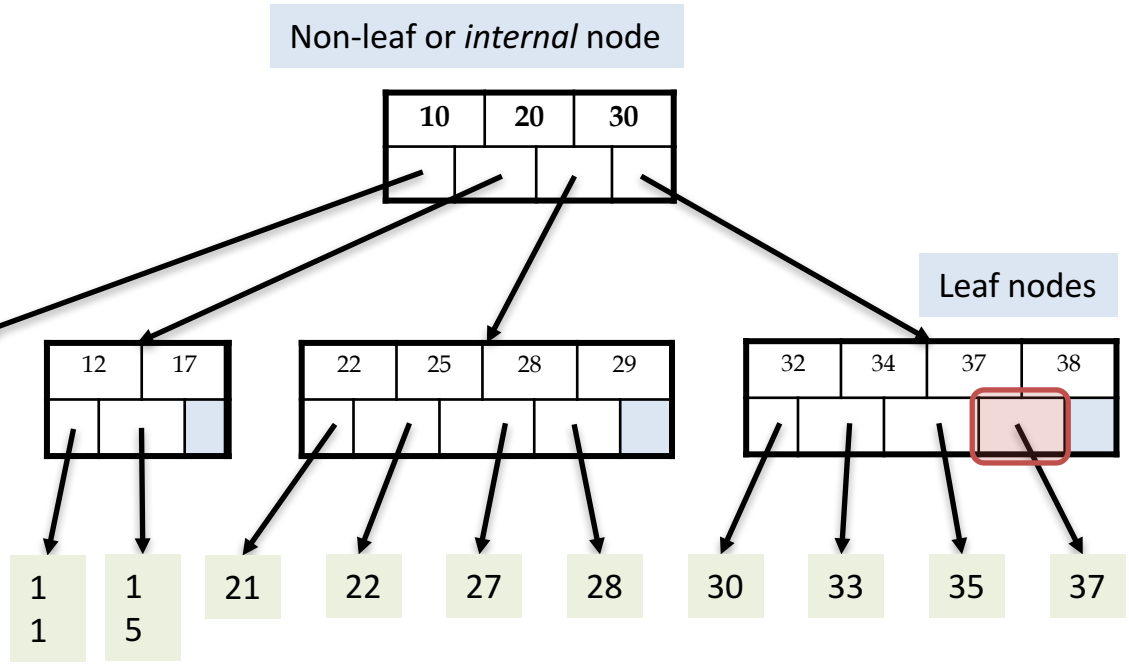
For each range, in a *non-leaf* node, there is a **pointer** to another node with keys in that range

B+ Tree Basics



Leaf nodes also have between *d* and *2d* keys, and are different in that:

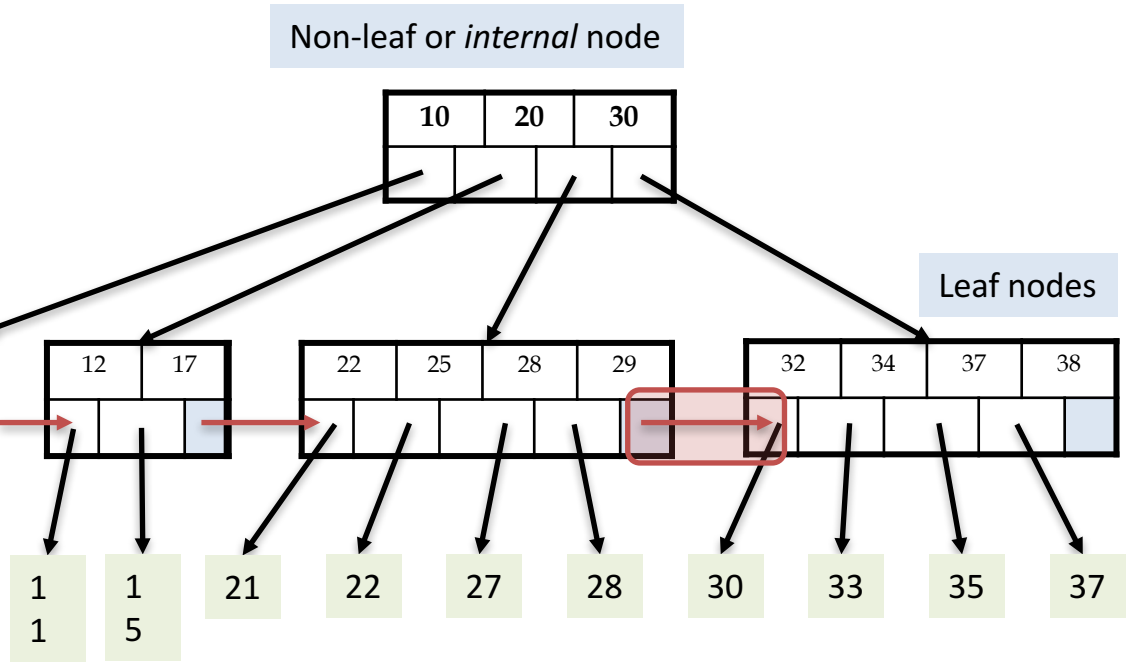
B+ Tree Basics



Leaf nodes also have between *d* and *2d* keys, and are different in that:

Their key slots contain pointers to data records

B+ Tree Basics

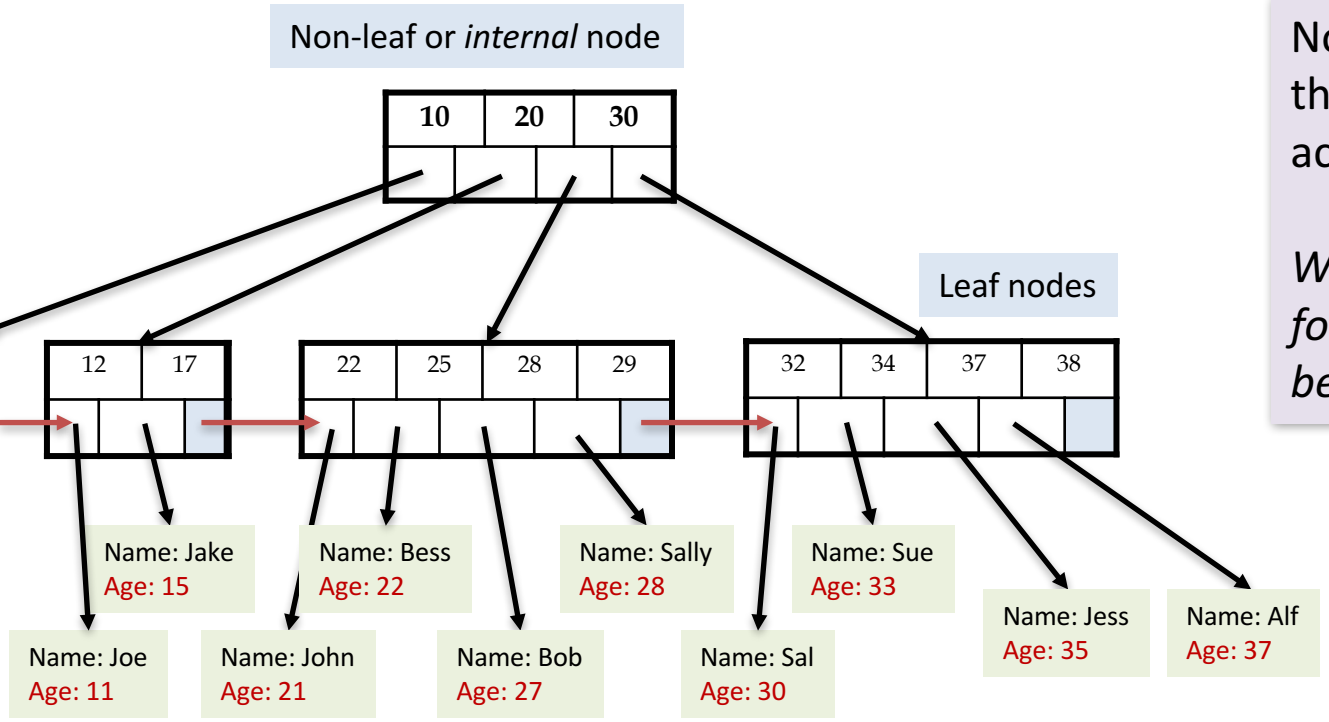


Leaf nodes also have between *d* and *2d* keys, and are different in that:

Their key slots contain pointers to data records

They contain a pointer to the next leaf node as well, ***for faster sequential traversal***

B+ Tree Basics



Note that the pointers at the leaf level will be to the actual data records (rows).

We might truncate these for simpler display (as before)…

Acknowledgement

- Some of the slides in this presentation are taken from the slides provided by the authors.
- Many of these slides are taken from cs145 course offered by Stanford University.