Chapter 11: Indexing and Storage

Modified from:

Database System Concepts, 6th Ed.

Chapter 11: Indexing and Storage

- **DBMS Storage**
  - Memory hierarchy
  - File Organization
  - Buffering
- **Indexing**
  - Basic Concepts
  - B$^+$-Trees
  - Static Hashing
  - Index Definition in SQL
 - Multiple-Key Access

Memory Hierarchy

DBMS Storage

- Modern Computers have different types of memory

  - Cache, Main Memory, Harddisk, SSD, …

- Memory types have different characteristics in terms of

  - Persistent vs. volatile

  - Speed (random vs. sequential access)

  - Size

  - Price – this usually determines size

- Database systems are designed to be use these different memory types effectively

  - Need for persistent storage: the state of the database needs to be written to persistent storage

    - guarantee database content is not lost when the computer is shutdown

  - Moving data between different types of memory

    - Want to use fast memory to speed-up operations

    - Need slower memory for the size

Storage Hierarchy

Persistent storage

cache

main memory

flash memory

magnetic disk

optical disk

magnetic tapes

Size

Speed
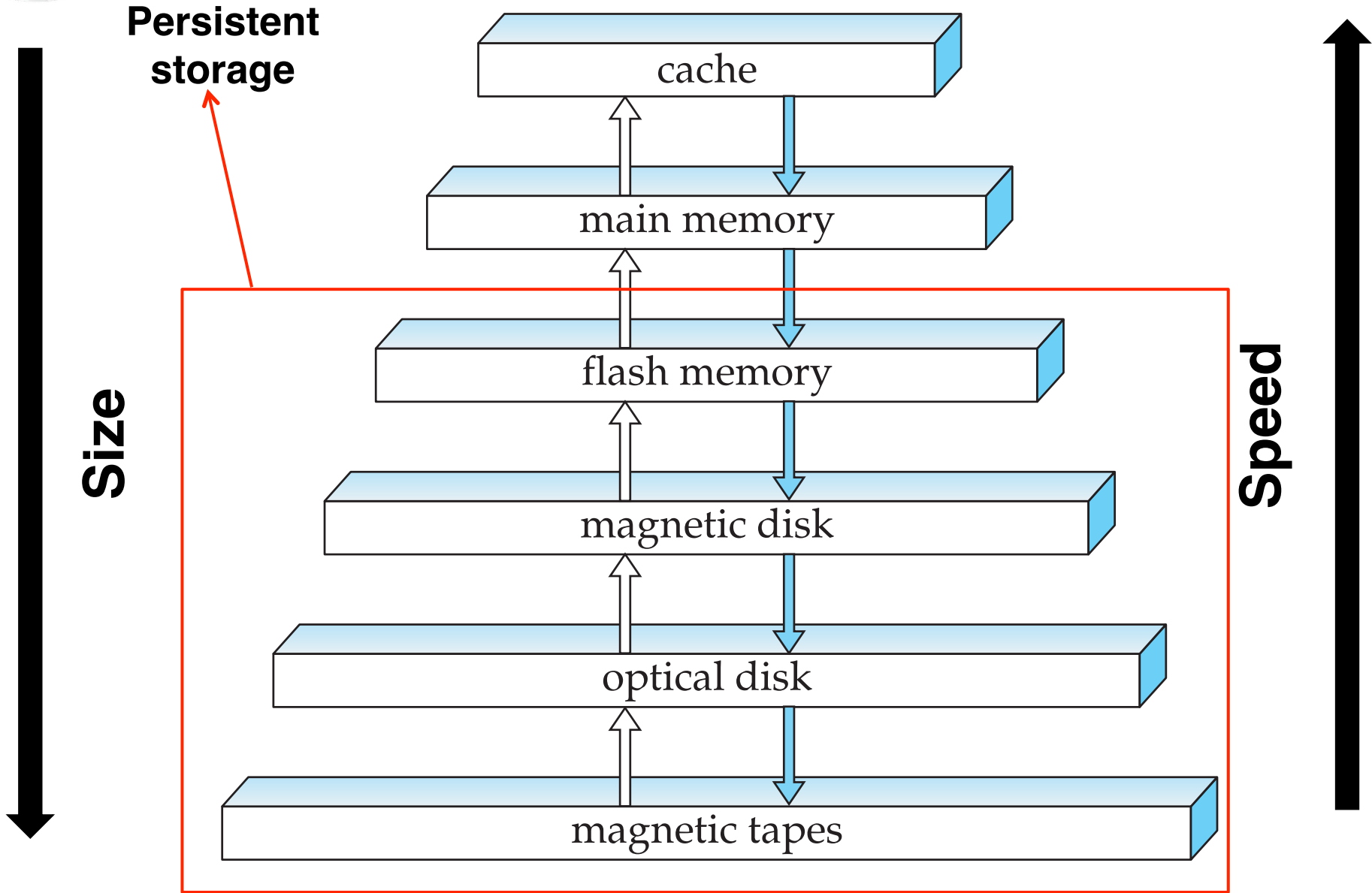
Main Memory vs. Disk

- Why do we not only use main memory
  - What if database does not fit into main memory?
  - Main memory is volatile
- Main memory vs. disk
  - Given available main memory when do we keep which part of the database in main memory
    - **Buffer manager**: Component of DBMS that decides when to move data between disk and main memory
  - How do we ensure transaction property durability
    - Buffer manager needs to make sure data written by committed transactions is written to disk to ensure durability

Magnetic Hard Disk Mechanism

track *t*

spindle

sector *s*

arm assembly

cylinder *c* →

read–write head

platter

arm

rotation
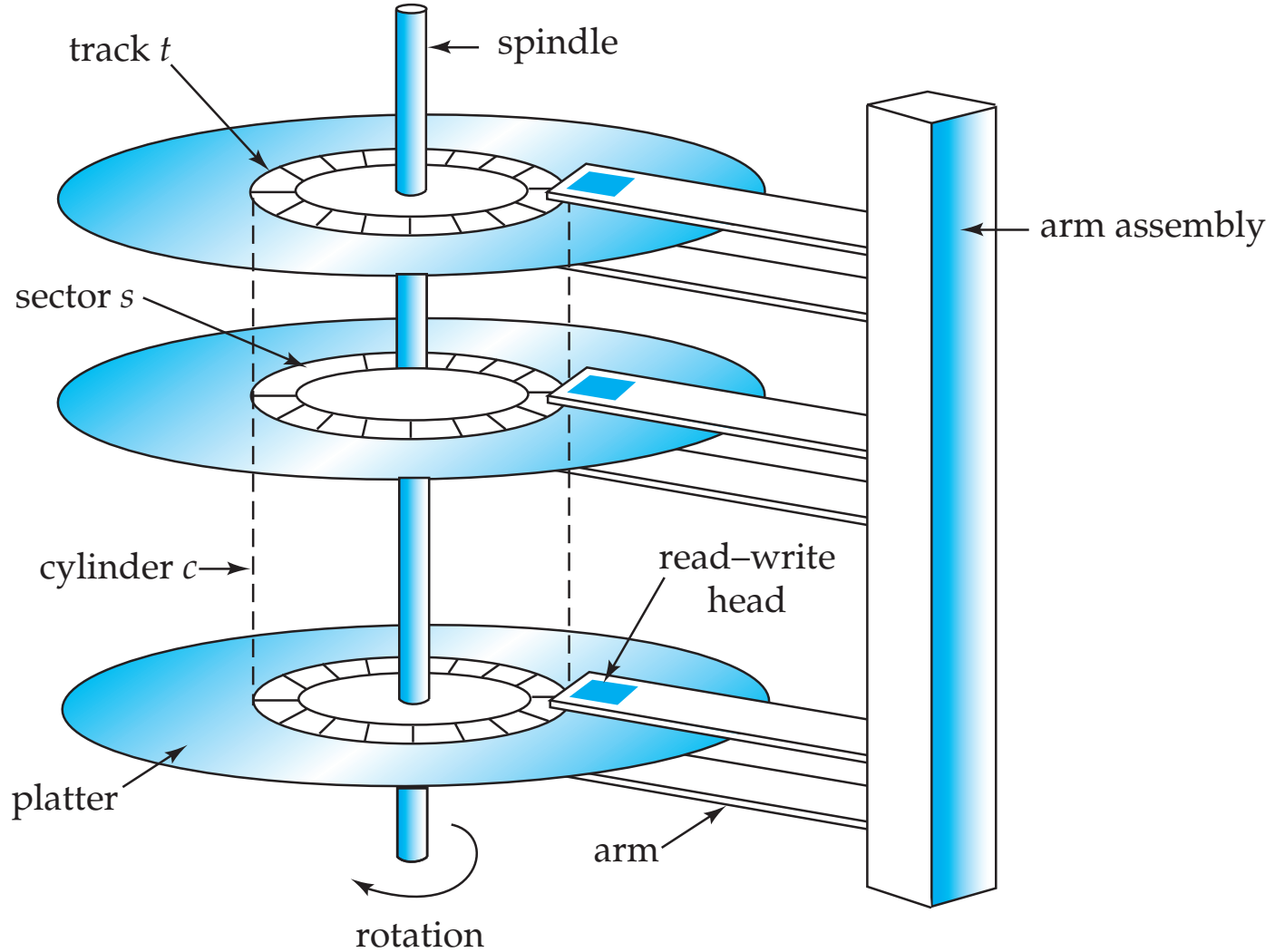
NOTE: Diagram is schematic, and simplifies the structure of actual disk drives

Performance Measures of Disks

- **Access time** – the time it takes from when a read or write request is issued to when data transfer begins. Consists of:
  - **Seek time** – time it takes to reposition the arm over the correct track.
    - Average seek time is 1/2 the worst case seek time.
      - Would be 1/3 if all tracks had the same number of sectors, and we ignore the time to start and stop arm movement
    - 4 to 10 milliseconds on typical disks
  - **Rotational latency** – time it takes for the sector to be accessed to appear under the head.
    - Average latency is 1/2 of the worst case latency.
    - 4 to 11 milliseconds on typical disks (5400 to 15000 r.p.m.)
- **Data-transfer rate** – the rate at which data can be retrieved from or stored to the disk.
  - 25 to 100 MB per second max rate, lower for inner tracks
  - Multiple disks may share a controller, so rate that controller can handle is also important
    - E.g. SATA: 150 MB/sec, SATA-II 3Gb (300 MB/sec)
    - Ultra 320 SCSI: 320 MB/s, SAS (3 to 6 Gb/sec)
    - Fiber Channel (FC2Gb or 4Gb): 256 to 512 MB/s

Random vs. Sequential Access

- Transfer of data from disk has a minimal size = 1 block

  - Reading 1 byte is as fast as reading one block (e.g., 4KB)

- **Random Access**

  - Read data from anywhere on the disk

  - Need to get to the right track (**seek time**)

  - Need to wait until the right sector is under the arm (on avg ½ time for one rotation) (**rotational delay**)

  - Then can transfer data at ~ **transfer rate**

- **Sequential Access**

  - Read data that is on the current track + sector

  - can transfer data at ~ **transfer rate**

- Reading large number of small pieces of data randomly is very slow compared to sequential access

  - Thus, try layout data on disk in a way that enables sequential access

File Organization

File Organization

- The database is stored as a collection of **files**. Each file stores **records** (tuples from a table). A record is a sequence of **fields** (the attributes of a tuple).

- Reading one record of a time from disk would be very slow (random access)

  - Organize our database files in pages (size of block or larger)

  - Read/write data in units of pages

  - One page will usually contain several records

- One approach:

  - assume record size is fixed

  - each file has records of one particular type only

  - different files are used for different relations

  This case is easiest to implement; will consider variable length records later.

Fixed-Length Records

- Simple approach:

  - Store record i starting from byte $n * (i - 1)$, where n is the size of each record. Put maximal P / n records on each page.

  - Record access is simple but records may cross blocks

    - Modification: do not allow records to cross block boundaries

- Deletion of record i: alternatives:

  - move records $i + 1, \ldots, n$ to $i, \ldots, n - 1$

  - move record n to i

  - do not move records, but link all free records on a *free list*

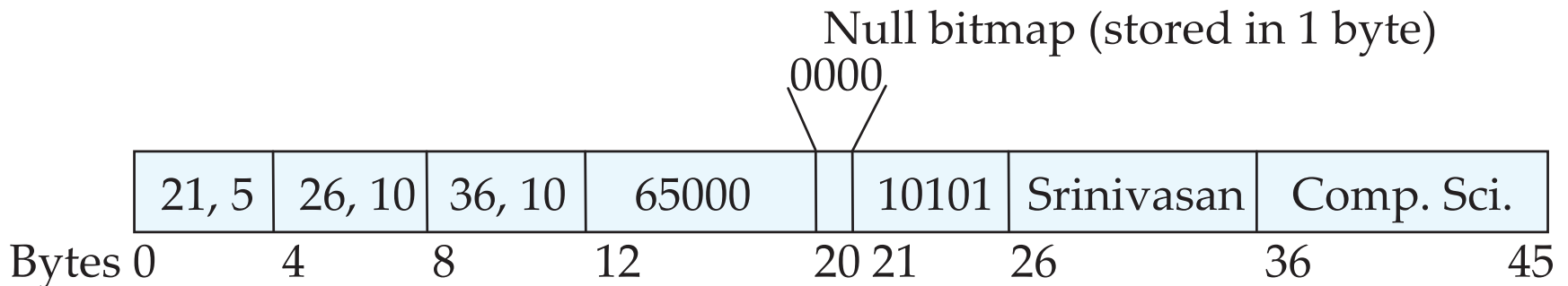|				
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1	12121	Wu	Finance	90000
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4	32343	El Said	History	60000
record 5	33456	Gold	Physics	87000
record 6	45565	Katz	Comp. Sci.	75000
record 7	58583	Califieri	History	62000
record 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

Free Lists

- Store the address of the first deleted record in the file header.

- Use this first record to store the address of the second deleted record, and so on

- Can think of these stored addresses as pointers since they "point" to the location of a record.

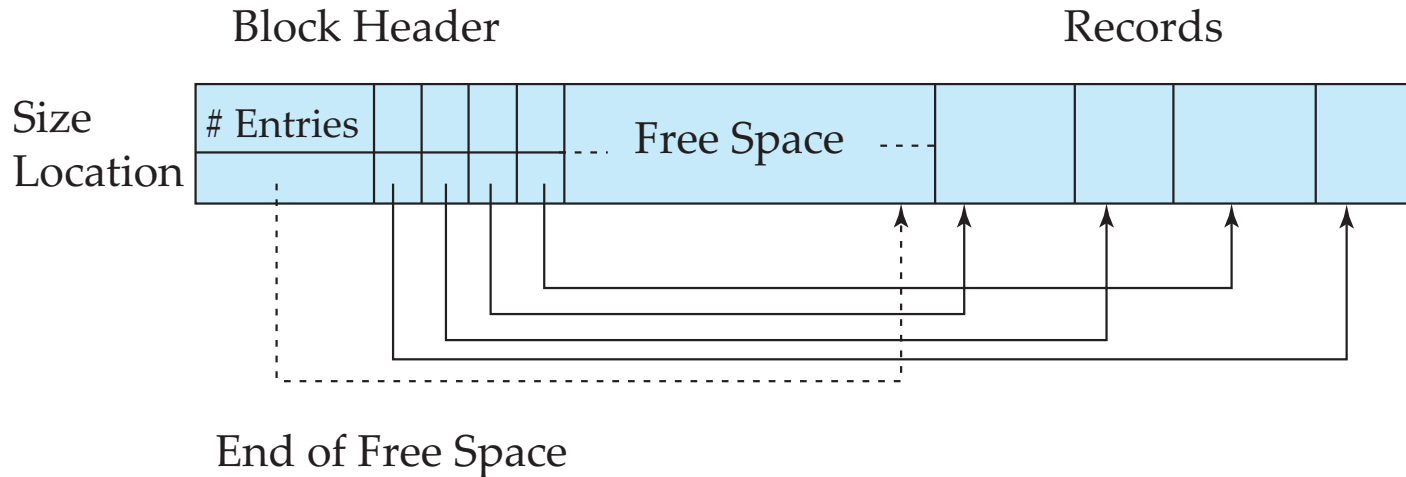|				
header				
record 0	10101	Srinivasan	Comp. Sci.	65000
record 1				
record 2	15151	Mozart	Music	40000
record 3	22222	Einstein	Physics	95000
record 4				
record 5	33456	Gold	Physics	87000
record 6				
record 7	58583	Califieri	History	62000
record 8	76543	Singh	Finance	80000
record 9	76766	Crick	Biology	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 (used in some older data models).
- 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

Null bitmap (stored in 1 byte)

0000

21, 5	26, 10	36, 10	65000		10101	Srinivasan	Comp. Sci.

Bytes 0   4   8   12   20 21   26   36   45

Variable-Length Records: Slotted Page Structure

Block Header   Records

End of Free Space

- **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

  - Deletion efficient

  - Insertion efficient

  - Search is expensive

    - Example: Get instructor with name Glavic

      - Have to search through all instructors

- **Sequential** – store records in sequential order, based on the value of some search key of each record

  - Deletion expensive and/or waste of space

  - Insertion expensive and/or waste of space

  - Search is efficient (e.g., binary search)

    - As long as the search is on the search key we are ordering on

Buffering

Modified from:

Database System Concepts, 6th Ed.

Buffer Manager

■ Buffer Manager

● Responsible for loading pages from disk and writing modified pages back to disk

■ Handling blocks

1. If the block is already in the buffer, the buffer manager returns the address of the block in main memory

2. If the block is not in the buffer, the buffer manager

    1. Allocates space in the buffer for the block

        1. Replacing (throwing out) some other block, if required, to make space for the new block.

        2. Replaced block written back to disk only if it was modified since the most recent time that it was written to/fetched from the disk.

    2. Reads the block from the disk to the buffer, and returns the address of the block in main memory to requester.

Buffer-Replacement Policies

■ Most operating systems replace the block **least recently used** (LRU strategy)

■ Idea behind LRU – use past pattern of block references as a predictor of future references

■ Queries have well-defined access patterns (such as sequential scans), and a database system can use the information in a user's query to predict future references

  ● LRU can be a bad strategy for certain access patterns involving repeated scans of data

    ‣ For example: when computing the join of 2 relations r and s by a nested loops
      for each tuple *tr* of *r* do
        for each tuple *ts* of *s* do
          if the tuples *tr* and *ts* match …

  ● Mixed strategy with hints on replacement strategy provided by the query optimizer is preferable

Buffer-Replacement Policies (Cont.)

■ **Pinned block** – memory block that is not allowed to be written back to disk. E.g., an operation still needs this block.

■ **Toss-immediate** strategy – frees the space occupied by a block as soon as the final tuple of that block has been processed

■ **Most recently used (MRU) strategy** – system must pin the block currently being processed. After the final tuple of that block has been processed, the block is unpinned, and it becomes the most recently used block.

■ Buffer manager can use statistical information regarding the probability that a request will reference a particular relation

 ● E.g., the data dictionary is frequently accessed. Heuristic: keep data-dictionary blocks in main memory buffer

■ Buffer managers also support **forced output** of blocks for the purpose of recovery (more in Chapter 16 in the textbook)

Indexing and Hashing

Basic Concepts

■ Indexing mechanisms used to speed up access to desired data.

● E.g., author catalog in library

■ **Search Key** - attribute or set of attributes used to look up records in a file.

■ An **index file** consists of records (called **index entries**) of the form

search-key	pointer

■ Index files are typically much smaller than the original file

■ Two basic kinds of indices:

● **Ordered indices:** search keys are stored in some sorted order

● **Hash indices:** search keys are distributed uniformly across "buckets" using a "hash function".

Index Evaluation Metrics

- Access types supported efficiently. E.g.,
  - records with a specified value in the attribute
  - or records with an attribute value falling in a specified range of values.
- Access time
- Insertion time
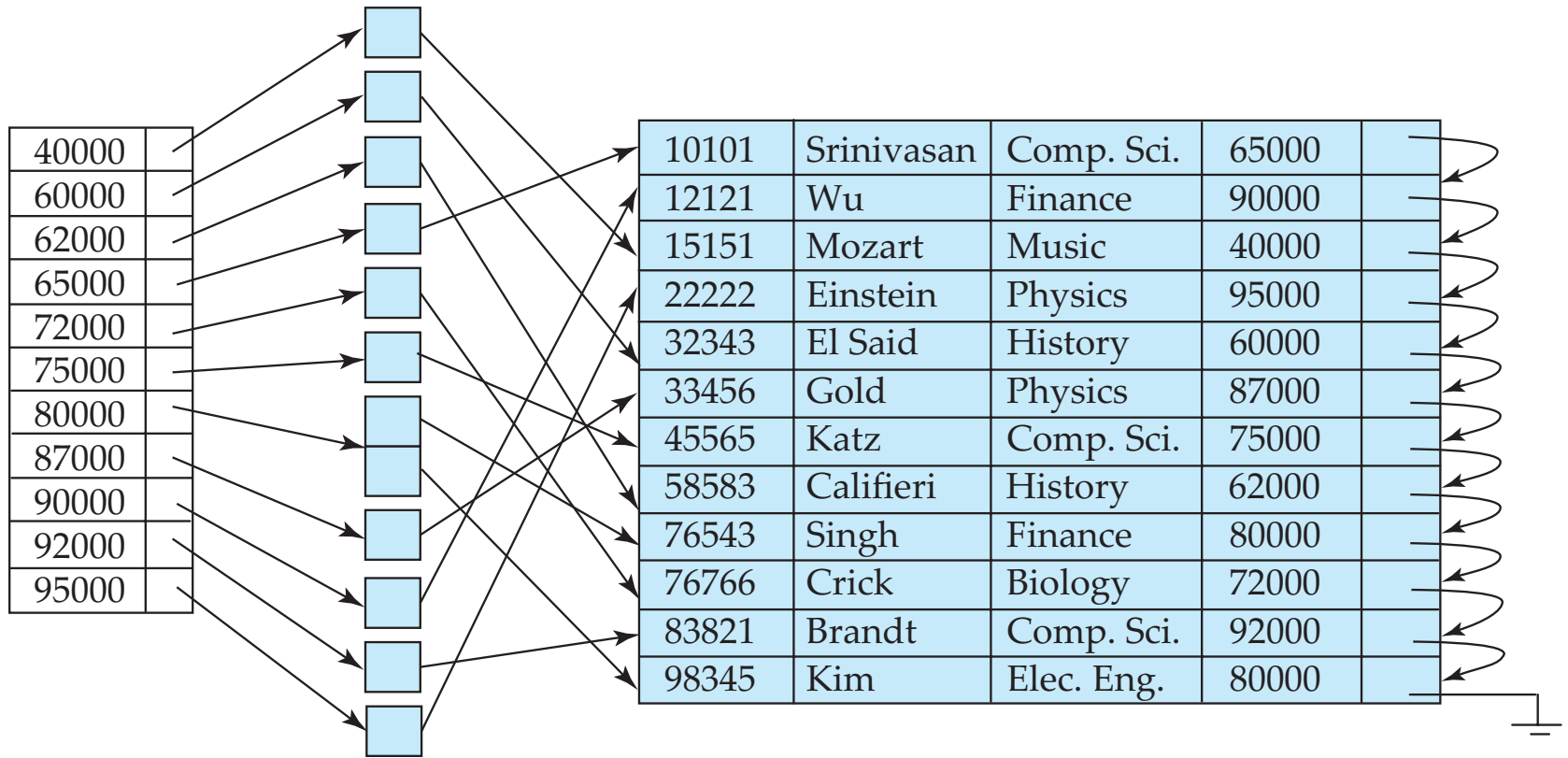- Deletion time
- Space overhead

Ordered Indices

- In an **ordered index,** index entries are stored sorted on the search key value. E.g., author catalog in library.

- **Primary index:** in a sequentially ordered file, the index whose search key specifies the sequential order of the file.

  - Also called **clustering index**

  - The search key of a primary index is usually but not necessarily the primary key.

- **Secondary index**: an index whose search key specifies an order different from the sequential order of the file. Also called non-clustering index**.**

- Index-sequential file**:** ordered sequential file with a primary index.

Secondary Indices Example

10101	Srinivasan	Comp. Sci.	65000	
12121	Wu	Finance	90000	
15151	Mozart	Music	40000	
22222	Einstein	Physics	95000	
32343	El Said	History	60000	
33456	Gold	Physics	87000	
45565	Katz	Comp. Sci.	75000	
58583	Califieri	History	62000	
76543	Singh	Finance	80000	
76766	Crick	Biology	72000	
83821	Brandt	Comp. Sci.	92000	
98345	Kim	Elec. Eng.	80000	

Secondary index on *salary* field of *instructor*

- Index record points to a bucket that contains pointers to all the actual records with that particular search-key value.

- Secondary indices have to be dense

Primary and Secondary Indices

- Indices offer substantial benefits when searching for records.

- BUT: Updating indices imposes overhead on database modification --when a file is modified, every index on the file must be updated,

- Sequential scan using primary index is efficient, but a sequential scan using a secondary index is expensive
  - Each record access may fetch a new block from disk
  - Block fetch requires about 5 to 10 milliseconds, versus about 100 nanoseconds for memory access
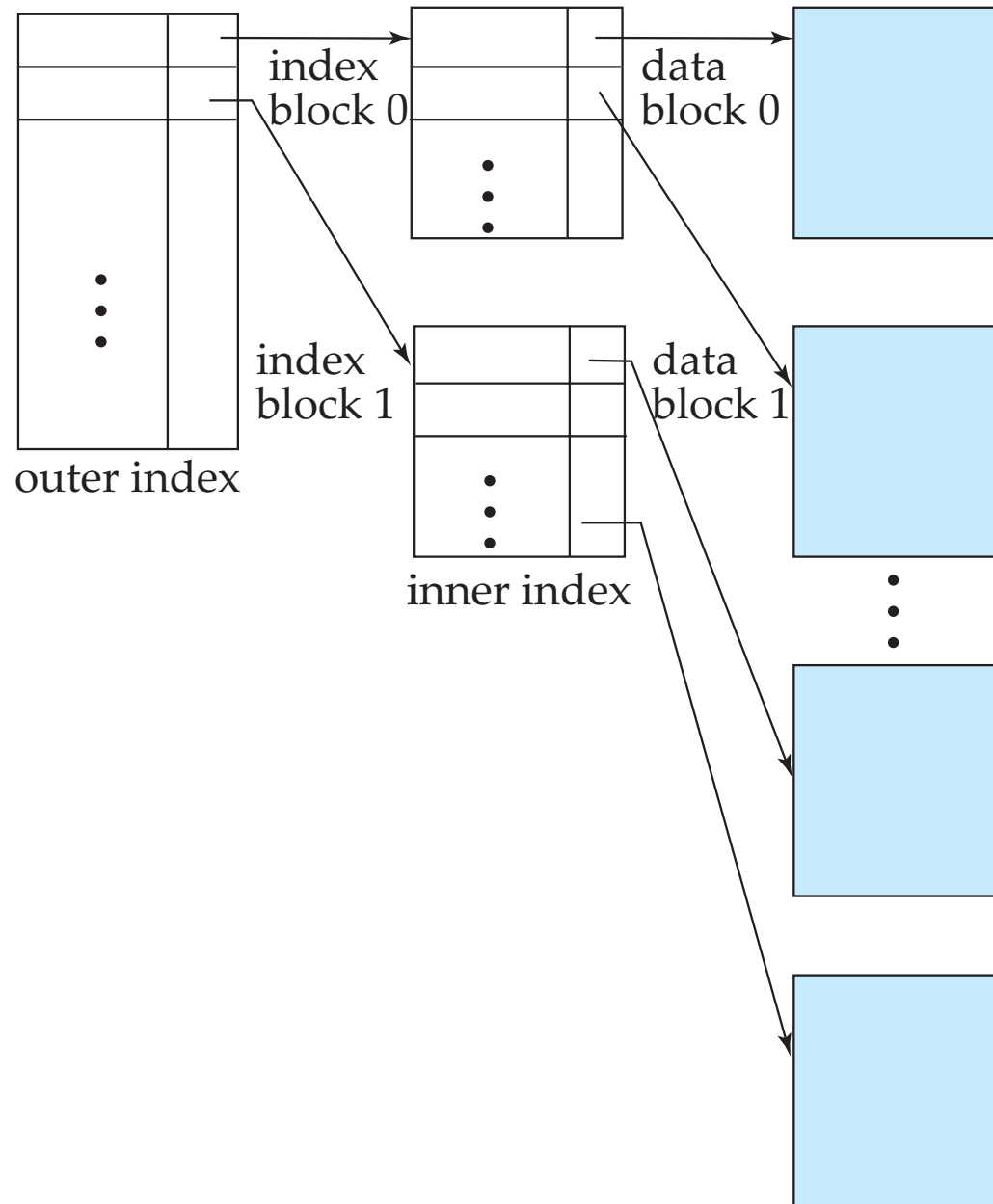
Multilevel Index

- If primary index does not fit in memory, access becomes expensive.

- Solution: treat primary index kept on disk as a sequential file and construct a sparse index on it.

  - outer index – a sparse index of primary index
  - inner index – the primary index file

- If even outer index is too large to fit in main memory, yet another level of index can be created, and so on.

- Indices at all levels must be updated on insertion or deletion from the file.
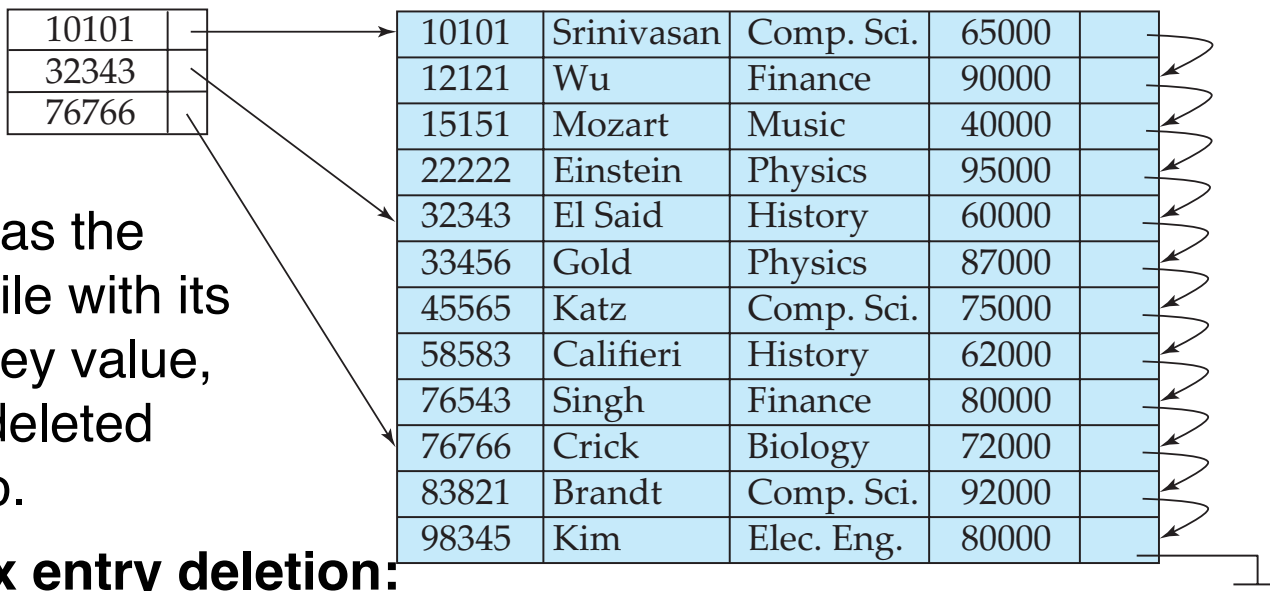
outer index

index block 0

index block 1

inner index

data block 0

data block 1

Index Update: Deletion

10101	→
32343	→
76766	→

10101	Srinivasan	Comp. Sci.	65000	
12121	Wu	Finance	90000	
15151	Mozart	Music	40000	
22222	Einstein	Physics	95000	
32343	El Said	History	60000	
33456	Gold	Physics	87000	
45565	Katz	Comp. Sci.	75000	
58583	Califieri	History	62000	
76543	Singh	Finance	80000	
76766	Crick	Biology	72000	
83821	Brandt	Comp. Sci.	92000	
98345	Kim	Elec. Eng.	80000	

- ■ **If deleted record was the only record in the file with its particular search-key value, the search-key is deleted from the index also.**

- ■ **Single-level index entry deletion:**

  - ● **Dense indices** – deletion of search-key is similar to file record deletion.

  - ● **Sparse indices** –
    - ‣ if an entry for the search key exists in the index, it is deleted by replacing the entry in the index with the next search-key value in the file (in search-key order).
    - ‣ If the next search-key value already has an index entry, the entry is deleted instead of being replaced.

Index Update: Insertion

- **Single-level index insertion:**
  - Perform a lookup using the search-key value appearing in the record to be inserted.
  - **Dense indices** – if the search-key value does not appear in the index, insert it.
  - **Sparse indices** – if index stores an entry for each block of the file, no change needs to be made to the index unless a new block is created.
    - ▸ If a new block is created, the first search-key value appearing in the new block is inserted into the index.
- **Multilevel insertion and deletion:** algorithms are simple extensions of the single-level algorithms

Secondary Indices

■ Frequently, one wants to find all the records whose values in a certain field (which is not the search-key of the primary index) satisfy some condition.

- Example 1: In the *instructor* relation stored sequentially by ID, we may want to find all instructors in a particular department

- Example 2: as above, but where we want to find all instructors with a specified salary or with salary in a specified range of values
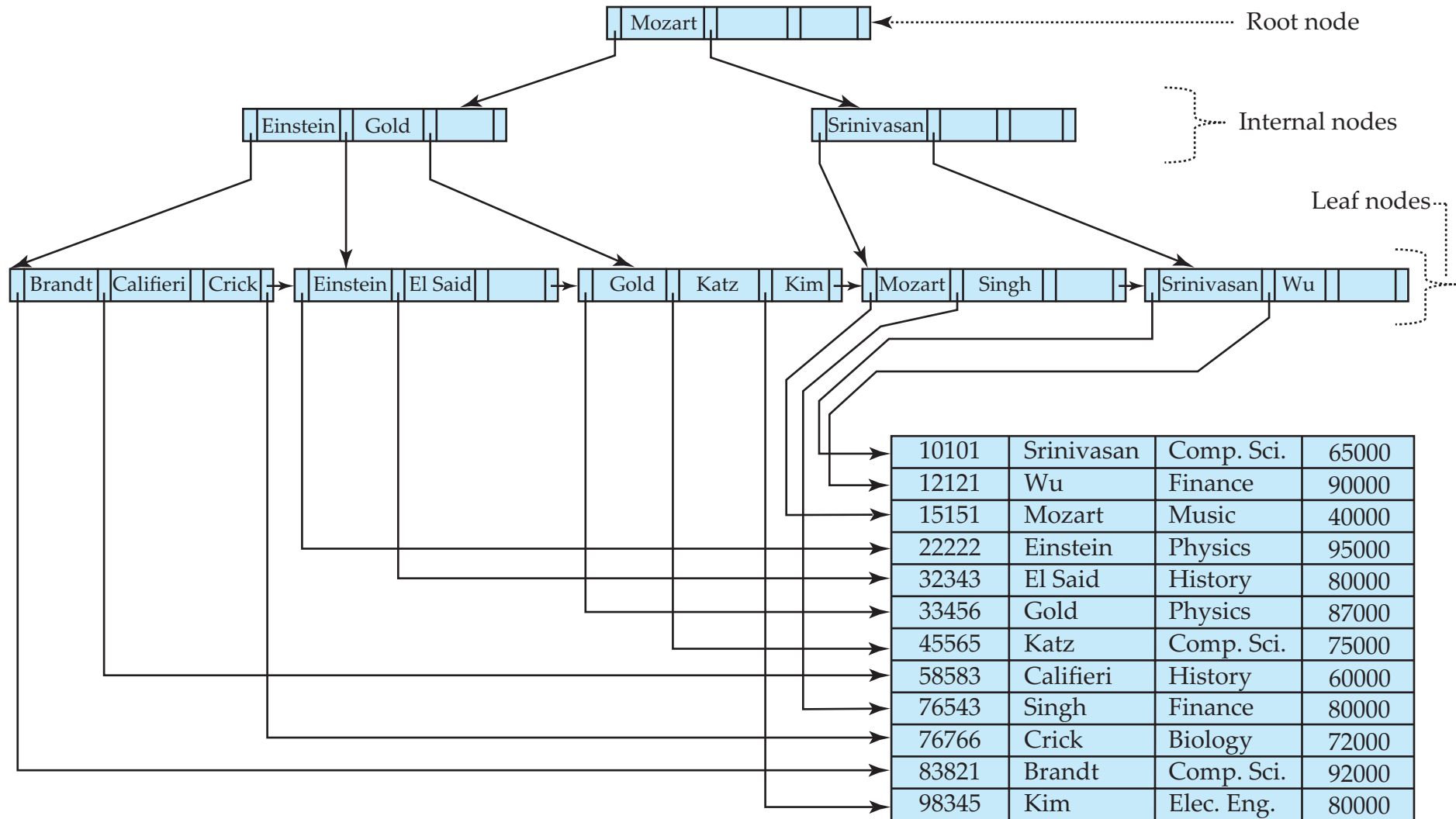
■ We can have a secondary index with an index record for each search-key value

B⁺-Tree Index

B⁺-tree indices are an alternative to indexed-sequential files.

- **Disadvantage of indexed-sequential files**
  - performance degrades as file grows, since many overflow blocks get created.
  - Periodic reorganization of entire file is required.
- **Advantage of B⁺-tree index files:**
  - automatically reorganizes itself with small, local, changes, in the face of insertions and deletions.
  - Reorganization of entire file is not required to maintain performance.
- **(Minor) disadvantage of B⁺-trees:**
  - extra insertion and deletion overhead, space overhead.
- **Advantages of B⁺-trees outweigh disadvantages**
  - B⁺-trees are used extensively

Example of B+-Tree

Root node

Internal nodes

Leaf nodes

| Mozart | | | |

| Einstein | Gold | | | Srinivasan | | | |

| Brandt | Califieri | Crick | | Einstein | El Said | | | Gold | Katz | Kim | | Mozart | Singh | | | Srinivasan | Wu | | |

10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	80000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	60000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

B⁺-Tree Index Files (Cont.)

A B⁺-tree is a rooted tree satisfying the following properties:

- All paths from root to leaf are of the same length

- Each node that is not a root or a leaf has between $\lceil n/2 \rceil$ and n children.

- A leaf node has between $\lceil (n–1)/2 \rceil$ and $n–1$ values

- Special cases:

  - If the root is not a leaf, it has at least 2 children.

  - If the root is a leaf (that is, there are no other nodes in the tree), it can have between 0 and ($n–1$) values.

B+-Tree Node Structure

- Typical node

P_1	K_1	P_2	...	P_{n-1}	K_{n-1}	P_n

- K_i are the search-key values

- P_i are pointers to children (for non-leaf nodes) or pointers to records or buckets of records (for leaf nodes).

- The search-keys in a node are ordered

$$K_1 < K_2 < K_3 < \ldots < K_{n-1}$$

(Initially assume no duplicate keys, address duplicates later)

Leaf Nodes in B⁺-Trees

Properties of a leaf node:

- For $i = 1, 2, \ldots, n-1$, pointer P_i points to a file record with search-key value K_i,

- If L_i, L_j are leaf nodes and $i < j$, L_i's search-key values are less than or equal to L_j's search-key values

- P_n points to next leaf node in search-key order

leaf node

Pointer to next leaf node

10101	Srinivasan	Comp. Sci.	65000
12121	Wu	Finance	90000
15151	Mozart	Music	40000
22222	Einstein	Physics	95000
32343	El Said	History	80000
33456	Gold	Physics	87000
45565	Katz	Comp. Sci.	75000
58583	Califieri	History	60000
76543	Singh	Finance	80000
76766	Crick	Biology	72000
83821	Brandt	Comp. Sci.	92000
98345	Kim	Elec. Eng.	80000

The leaf node contains: Brandt, Califieri, Crick

Non-Leaf Nodes in B⁺-Trees

■ Non leaf nodes form a multi-level sparse index on the leaf nodes. For a non-leaf node with m pointers:

● All the search-keys in the subtree to which P_1 points are less than K_1

● For $2 \leq i \leq n - 1$, all the search-keys in the subtree to which P_i points have values greater than or equal to K_{i-1} and less than K_i

● All the search-keys in the subtree to which P_n points have values greater than or equal to K_{n-1}

P_1	K_1	P_2	...	P_{n-1}	K_{n-1}	P_n

Example of B+-tree
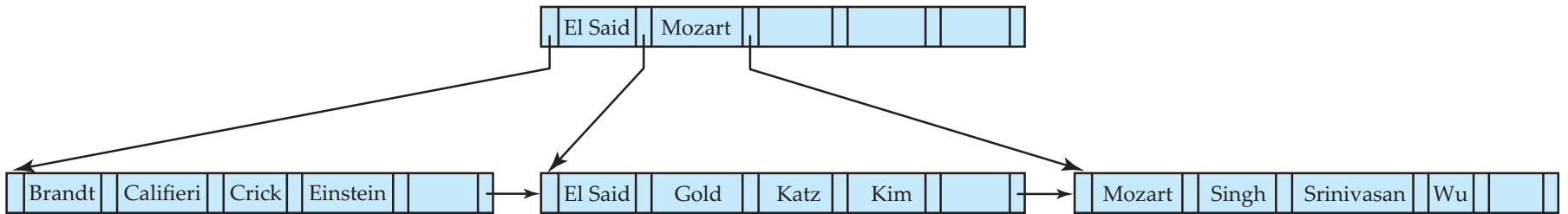
B+-tree for *instructor* file (*n* = 6)

- Leaf nodes must have between 3 and 5 values ($\lceil (n-1)/2 \rceil$ and $n-1$, with $n = 6$).

- Non-leaf nodes other than root must have between 3 and 6 children ($\lceil (n/2) \rceil$ and n with $n = 6$).

- Root must have at least 2 children.

Observations about B⁺-trees

- Since the inter-node connections are done by pointers, "logically" close blocks need not be "physically" close.

- The non-leaf levels of the B⁺-tree form a hierarchy of sparse indices.

- The B⁺-tree contains a relatively small number of levels

  ▸ Level below root has at least 2* $\lceil n/2 \rceil$ values

  ▸ Next level has at least 2* $\lceil n/2 \rceil$ * $\lceil n/2 \rceil$ values

  ▸ .. etc.

  - If there are K search-key values in the file, the tree height is no more than $\lceil \log_{\lceil n/2 \rceil}(K) \rceil$

  - thus searches can be conducted efficiently.

- Insertions and deletions to the main file can be handled efficiently, as the index can be restructured in logarithmic time (as we shall see).

Queries on B⁺-Trees

- Find record with search-key value *V.*

  1. *C=root*
  2. While C is not a leaf node {
     1. Let *i* be least value s.t. $V \le K_i$.
     2. If no such exists, set *C = last non-null pointer in C*
     3. Else { if (V= K_i) Set C = P_{i+1} else set C = P_i}
     }
  3. Let *i* be least value s.t. $K_i = V$
  4. If there is such a value *i*, follow pointer P_i to the desired record.
  5. Else no record with search-key value *k* exists.
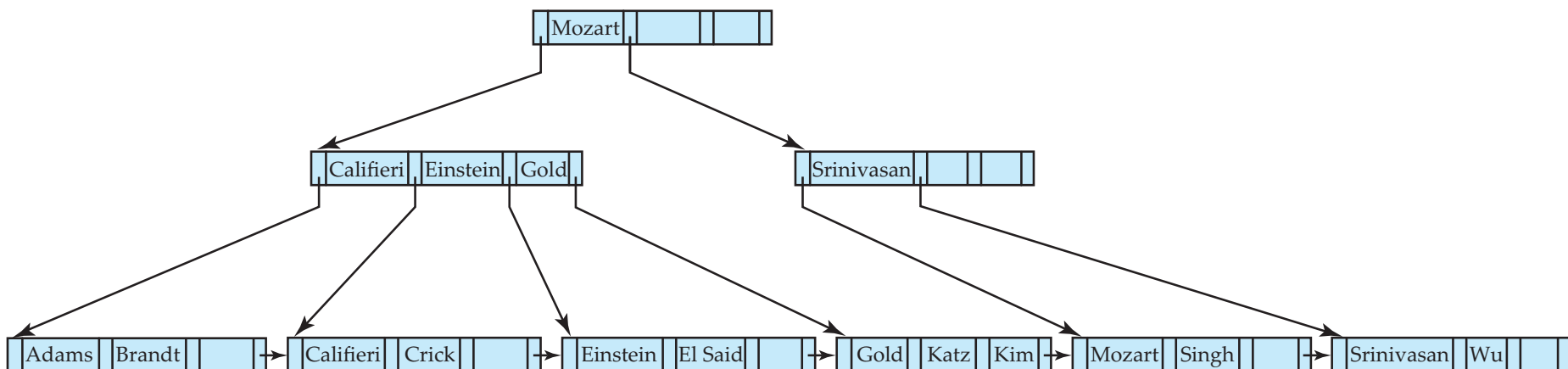
Handling Duplicates

■ With duplicate search keys

- In both leaf and internal nodes,
    ▸ we cannot guarantee that $K_1 < K_2 < K_3 < \ldots < K_{n-1}$
    ▸ but can guarantee $K_1 \leq K_2 \leq K_3 \leq \ldots \leq K_{n-1}$
- Search-keys in the subtree to which P_i points
    ▸ are $\leq K_i$, but not necessarily $< K_i$
    ▸ To see why, suppose same search key value V is present in two leaf node L_i and L_{i+1}. Then in parent node K_i must be equal to V

Handling Duplicates

- We modify find procedure as follows
  - traverse P_i even if $V = K_i$
  - As soon as we reach a leaf node C check if C has only search key values less than V
    - ▸ if so set C = right sibling of C before checking whether C contains V
- Procedure printAll
  - uses modified find procedure to find first occurrence of V
  - Traverse through consecutive leaves to find all occurrences of V

** Errata note: modified find procedure missing in first printing of 6th edition

Queries on B+-Trees (Cont.)

- If there are *K* search-key values in the file, the height of the tree is no more than $\lceil \log_{\lceil n/2 \rceil}(K) \rceil$.

- A node is generally the same size as a disk block, typically 4 kilobytes

  - and *n* is typically around 100 (40 bytes per index entry).

- With 1 million search key values and *n* = 100

  - at most $\log_{50}(1{,}000{,}000) = 4$ nodes are accessed in a lookup.

- Contrast this with a balanced binary tree with 1 million search key values — around 20 nodes are accessed in a lookup

  - above difference is significant since every node access may need a disk I/O, costing around 20 milliseconds

Updates on B⁺-Trees: Insertion

1. Find the leaf node in which the search-key value would appear

2. If the search-key value is already present in the leaf node

   1. Add record to the file

   2. If necessary add a pointer to the bucket.

3. If the search-key value is not present, then

   1. add the record to the main file (and create a bucket if necessary)

   2. If there is room in the leaf node, insert (key-value, pointer) pair in the leaf node

   3. Otherwise, split the node (along with the new (key-value, pointer) entry) as discussed in the next slide.

Updates on B⁺-Trees: Insertion (Cont.)

- Splitting a leaf node:
  - take the n (search-key value, pointer) pairs (including the one being inserted) in sorted order. Place the first $\lceil n/2 \rceil$ in the original node, and the rest in a new node.
  - let the new node be p, and let k be the least key value in p. Insert (k,p) in the parent of the node being split.
  - If the parent is full, split it and **propagate** the split further up.
- Splitting of nodes proceeds upwards till a node that is not full is found.
  - In the worst case the root node may be split increasing the height of the tree by 1.
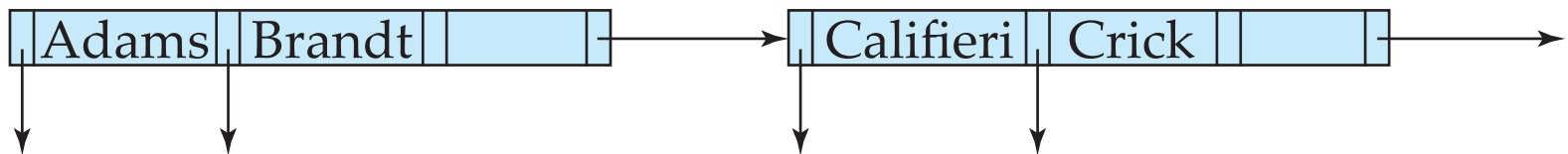
Result of splitting node containing Brandt, Califieri and Crick on inserting Adams
Next step: insert entry with (Califieri,pointer-to-new-node) into parent

B+-Tree Insertion
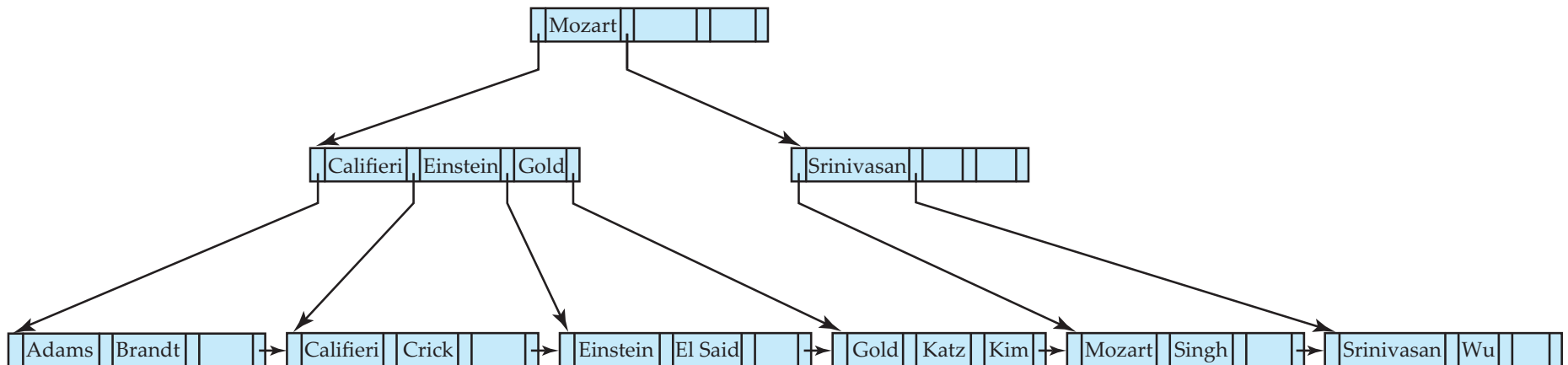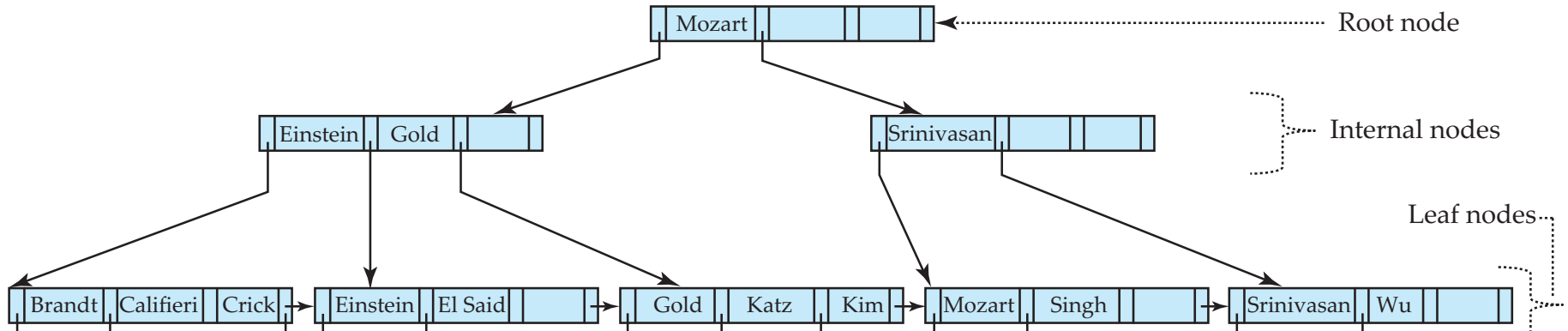
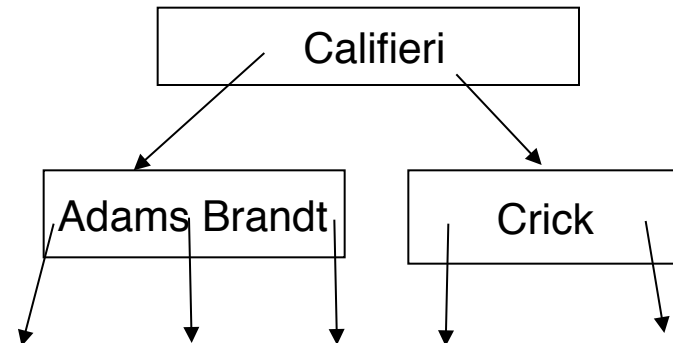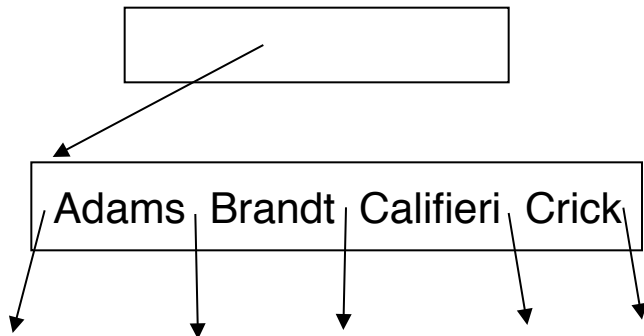B+-Tree before and after insertion of "Adams"

B+-Tree Insertion
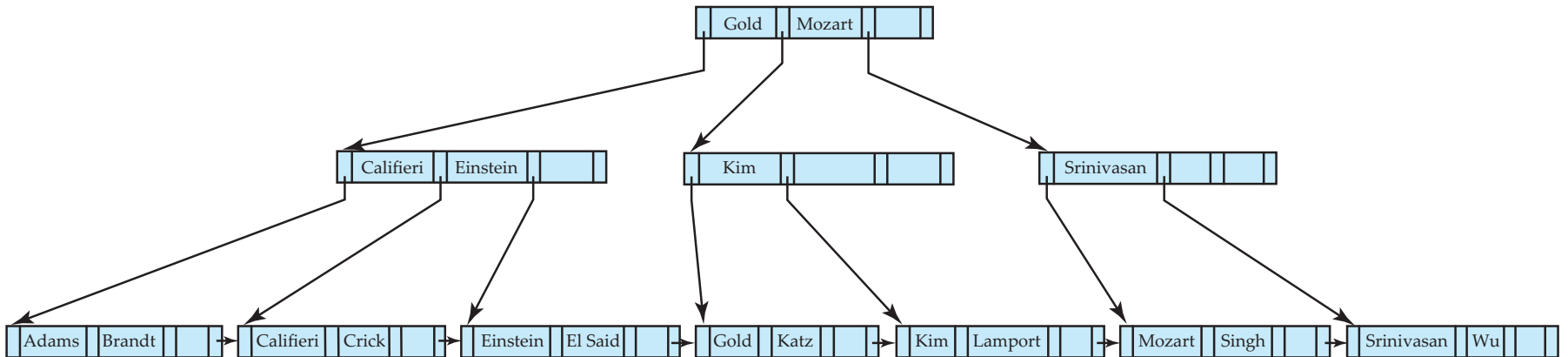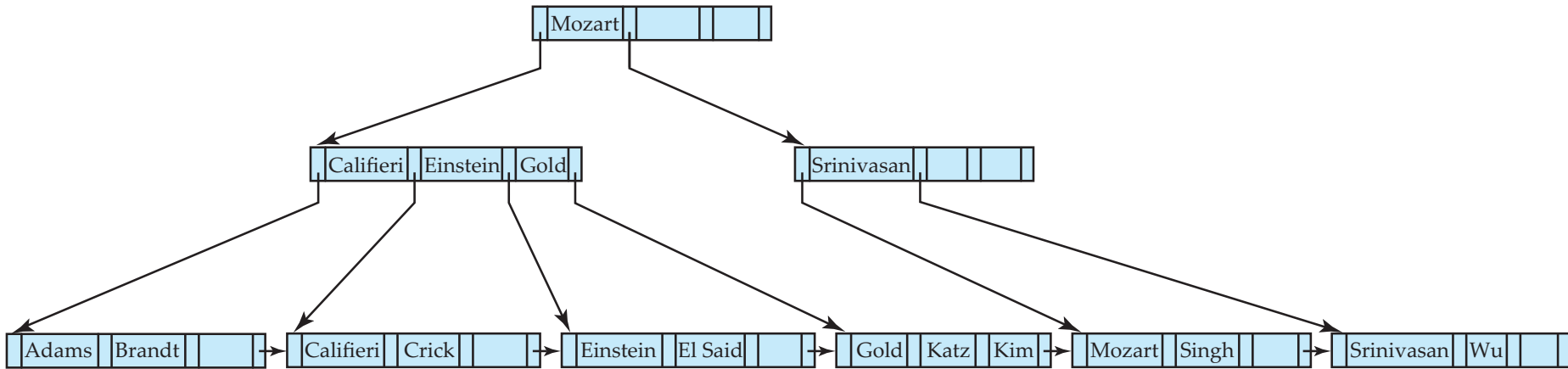
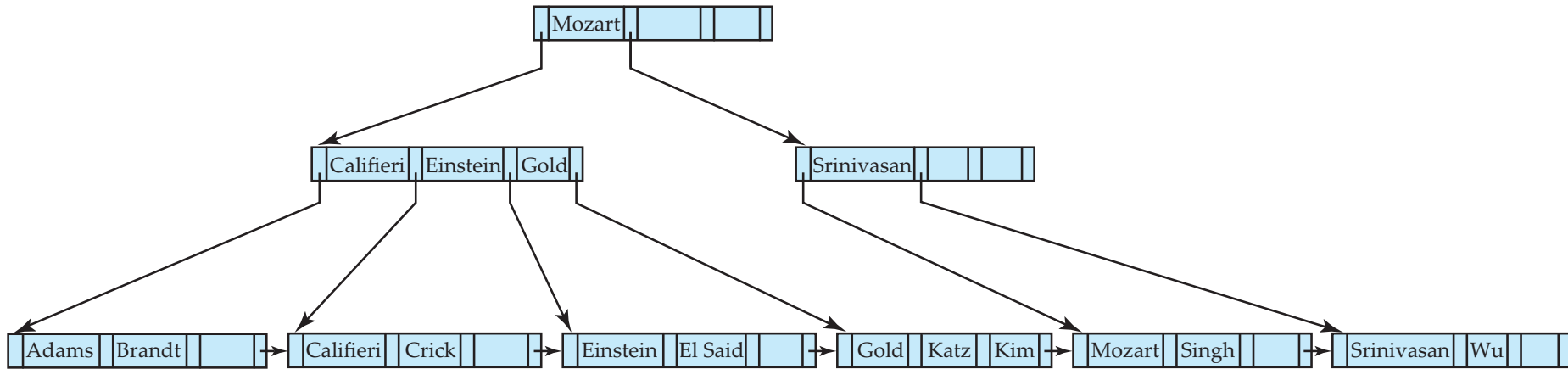B+-Tree before and after insertion of "Lamport"

Insertion in B⁺-Trees (Cont.)

- Splitting a non-leaf node: when inserting (k,p) into an already full internal node N

  - Copy N to an in-memory area M with space for n+1 pointers and n keys

  - Insert (k,p) into M

  - Copy $P_1, K_1, \ldots, K_{\lceil n/2 \rceil -1}, P_{\lceil n/2 \rceil}$ from M back into node N

  - Copy $P_{\lceil n/2 \rceil +1}, K_{\lceil n/2 \rceil +1}, \ldots, K_n, P_{n+1}$ from M into newly allocated node N'

  - Insert $(K_{\lceil n/2 \rceil}, N')$ into parent N
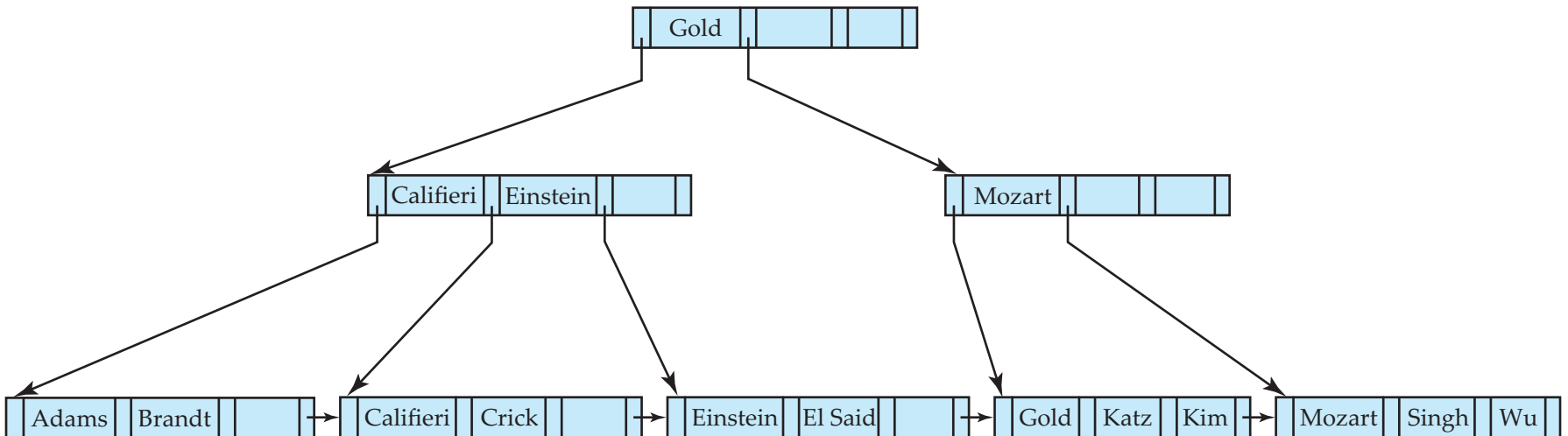
- **Read pseudocode in book!**

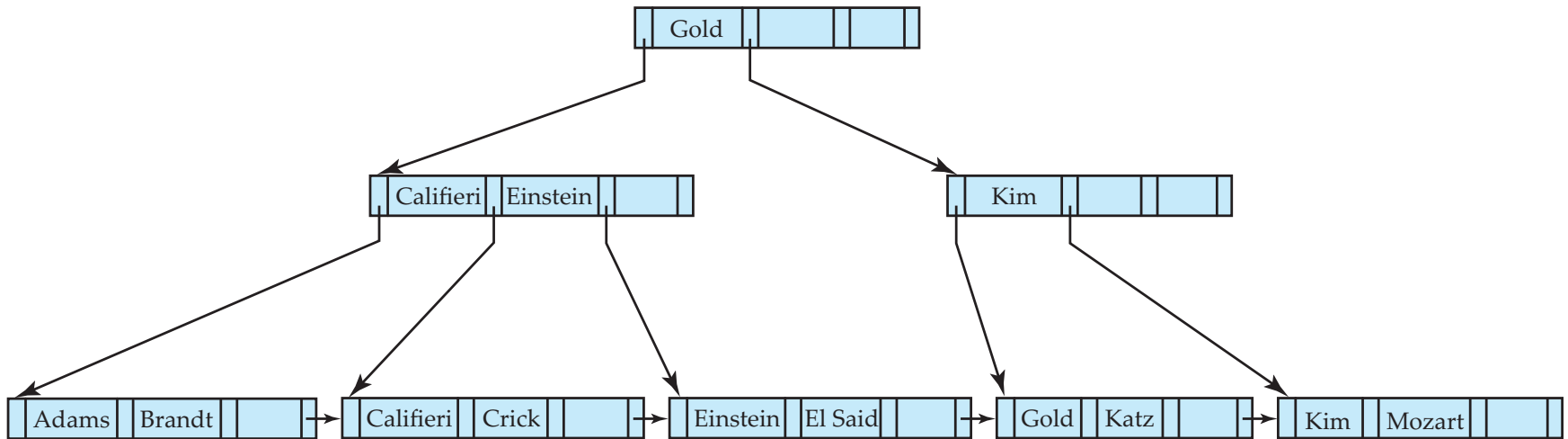# Examples of B+-Tree Deletion

Mozart

Califieri | Einstein | Gold

Srinivasan

Adams | Brandt → Califieri | Crick → Einstein | El Said → Gold | Katz | Kim → Mozart | Singh → Srinivasan | Wu

Before and after deleting "Srinivasan"

Gold

Califieri | Einstein

Mozart

Adams | Brandt → Califieri | Crick → Einstein | El Said → Gold | Katz | Kim → Mozart | Singh | Wu

- Deleting "Srinivasan" causes merging of under-full leaves

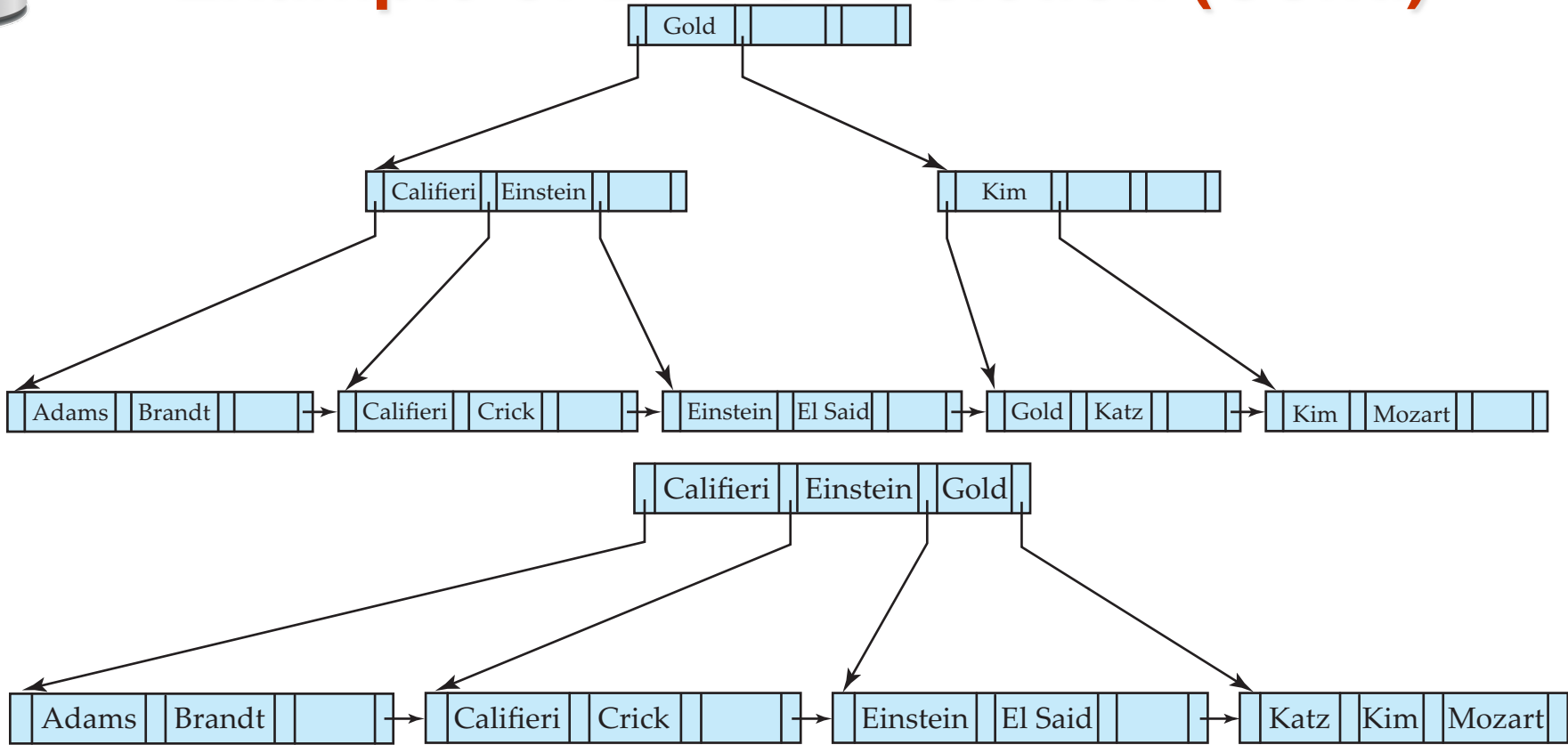# Examples of B+-Tree Deletion (Cont.)

Deletion of "Singh" and "Wu" from result of previous example

- Leaf containing Singh and Wu became underfull, and borrowed a value Kim from its left sibling

- Search-key value in the parent changes as a result

Example of B⁺-tree Deletion (Cont.)

Before and after deletion of "Gold" from earlier example

- Node with Gold and Katz became underfull, and was merged with its sibling
- Parent node becomes underfull, and is merged with its sibling
  - Value separating two nodes (at the parent) is pulled down when merging
- Root node then has only one child, and is deleted

Updates on B⁺-Trees: Deletion

- Find the record to be deleted, and remove it from the main file and from the bucket (if present)

- Remove (search-key value, pointer) from the leaf node if there is no bucket or if the bucket has become empty

- If the node has too few entries due to the removal, and the entries in the node and a sibling fit into a single node, then ***merge siblings***:

    - Insert all the search-key values in the two nodes into a single node (the one on the left), and delete the other node.

    - Delete the pair (K_{i-1}, P_i), where P_i is the pointer to the deleted node, from its parent, recursively using the above procedure.

Updates on B⁺-Trees: Deletion

- Otherwise, if the node has too few entries due to the removal, but the entries in the node and a sibling do not fit into a single node, then **redistribute pointers**:

  - Redistribute the pointers between the node and a sibling such that both have more than the minimum number of entries.

  - Update the corresponding search-key value in the parent of the node.

- The node deletions may cascade upwards till a node which has $\lceil n/2 \rceil$ or more pointers is found.

- If the root node has only one pointer after deletion, it is deleted and the sole child becomes the root.

Non-Unique Search Keys

■ Alternatives to scheme described earlier

- Buckets on separate block (bad idea)
- List of tuple pointers with each key
  ▸ Extra code to handle long lists
  ▸ Deletion of a tuple can be expensive if there are many duplicates on search key (why?)
  ▸ Low space overhead, no extra cost for queries
- Make search key unique by adding a record-identifier
  ▸ Extra storage overhead for keys
  ▸ Simpler code for insertion/deletion
 ▸ Widely used

B+-Tree File Organization

- Index file degradation problem is solved by using B+-Tree indices.

- Data file degradation problem is solved by using B+-Tree File Organization.

- The leaf nodes in a B+-tree file organization store records, instead of pointers.

- Leaf nodes are still required to be half full
  - Since records are larger than pointers, the maximum number of records that can be stored in a leaf node is less than the number of pointers in a nonleaf node.

- Insertion and deletion are handled in the same way as insertion and deletion of entries in a B+-tree index.

B⁺-Tree File Organization (Cont.)
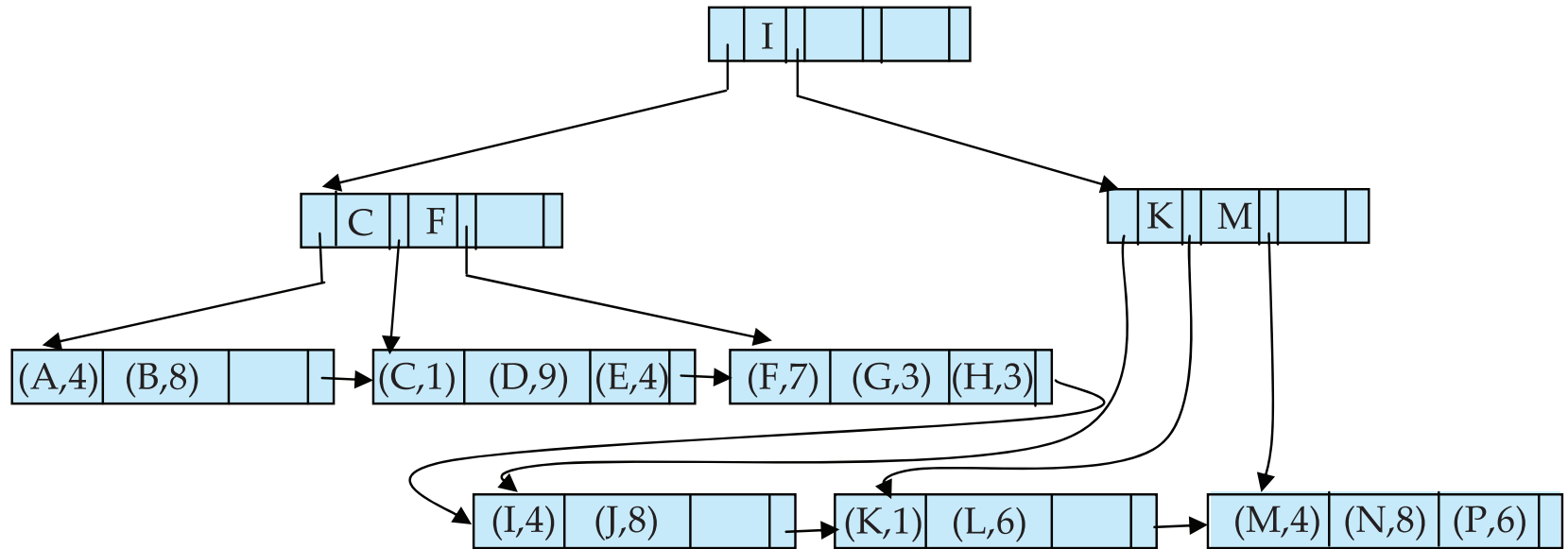
Example of B⁺-tree File Organization

- Good space utilization important since records use more space than pointers.

- To improve space utilization, involve more sibling nodes in redistribution during splits and merges

  - Involving 2 siblings in redistribution (to avoid split / merge where possible) results in each node having at least $\lfloor 2n/3 \rfloor$ entries

Hashing

Static Hashing

- A **bucket** is a unit of storage containing one or more records (a bucket is typically a disk block).

- In a **hash file organization** we obtain the bucket of a record directly from its search-key value using a **hash function.**

- Hash function h is a function from the set of all search-key values K to the set of all bucket addresses B.

- Hash function is used to locate records for access, insertion as well as deletion.

- Records with different search-key values may be mapped to the same bucket; thus entire bucket has to be searched sequentially to locate a record.

Example of Hash File Organization

Hash file organization of *instructor* file, using *dept_name* as key
(See figure in next slide.)

- There are 10 buckets,

- The binary representation of the ith character is assumed to be the integer i.

- The hash function returns the sum of the binary representations of the characters modulo 10

  - E.g. h(Music) = 1   h(History) = 2
    h(Physics) = 3 h(Elec. Eng.) = 3

Example of Hash File Organization

bucket 0

bucket 1

15151	Mozart	Music	40000

bucket 2

32343	El Said	History	80000
58583	Califieri	History	60000

bucket 3

22222	Einstein	Physics	95000
33456	Gold	Physics	87000
98345	Kim	Elec. Eng.	80000

bucket 4

12121	Wu	Finance	90000
76543	Singh	Finance	80000

bucket 5

76766	Crick	Biology	72000

bucket 6

10101	Srinivasan	Comp. Sci.	65000
45565	Katz	Comp. Sci.	75000
83821	Brandt	Comp. Sci.	92000

bucket 7

Hash file organization of *instructor* file, using *dept_name* as key
(see previous slide for details).

Hash Functions

- Worst hash function maps all search-key values to the same bucket; this makes access time proportional to the number of search-key values in the file.

- An ideal hash function is **uniform**, i.e., each bucket is assigned the same number of search-key values from the set of *all* possible values.

- Ideal hash function is **random**, so each bucket will have the same number of records assigned to it irrespective of the *actual distribution* of search-key values in the file.

- Typical hash functions perform computation on the internal binary representation of the search-key.

  - For example, for a string search-key, the binary representations of all the characters in the string could be added and the sum modulo the number of buckets could be returned. .

Handling of Bucket Overflows

- Bucket overflow can occur because of
  - Insufficient buckets
  - Skew in distribution of records. This can occur due to two reasons:
    - multiple records have same search-key value
    - chosen hash function produces non-uniform distribution of key values
- Although the probability of bucket overflow can be reduced, it cannot be eliminated; it is handled by using *overflow buckets*.

Handling of Bucket Overflows (Cont.)

■ **Overflow chaining** – the overflow buckets of a given bucket are chained together in a linked list.

■ Above scheme is called **closed hashing.**

● An alternative, called **open hashing**, which does not use overflow buckets, is not suitable for database applications.

bucket 0

bucket 1

overflow buckets for bucket 1
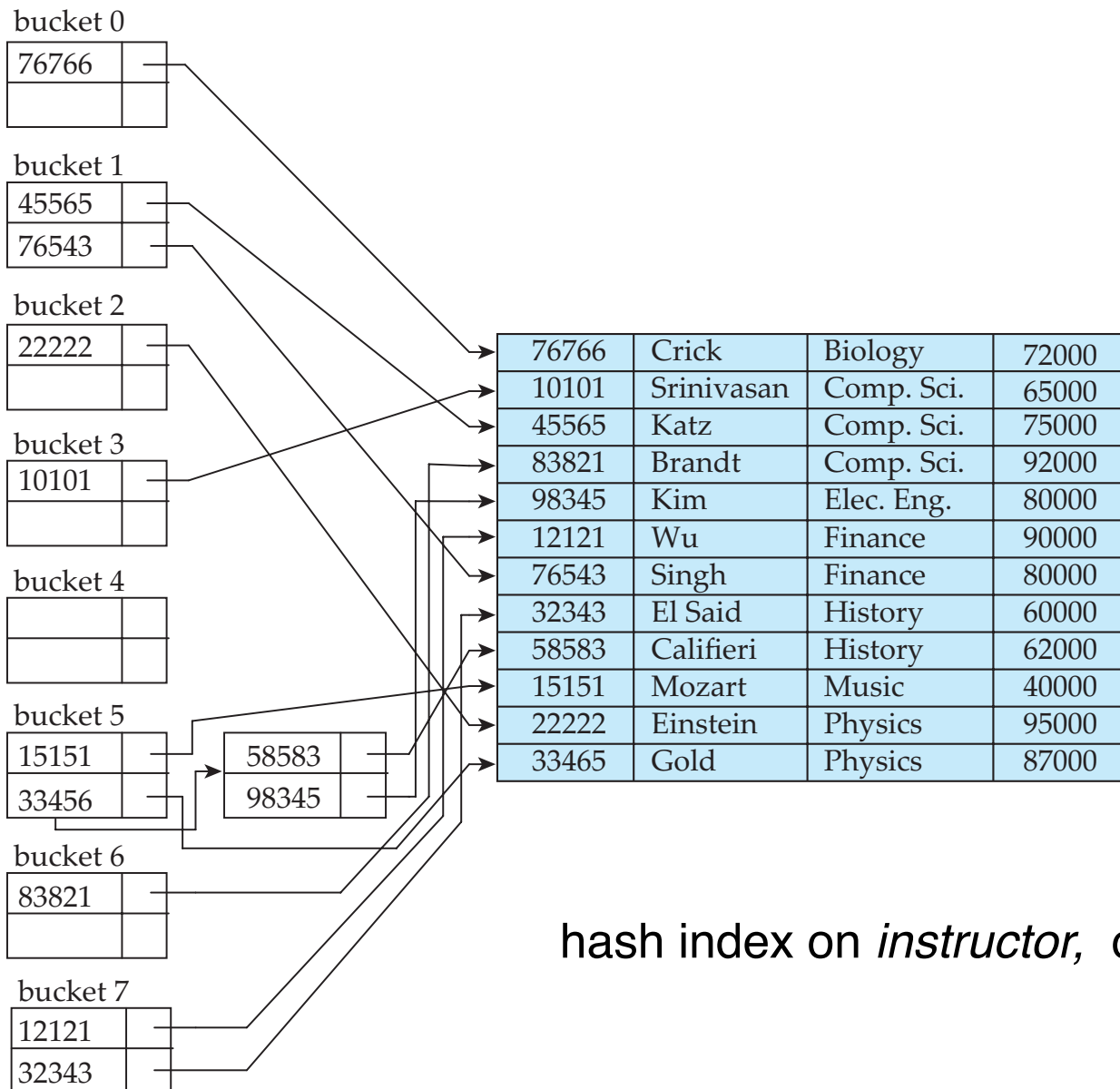
bucket 2

bucket 3
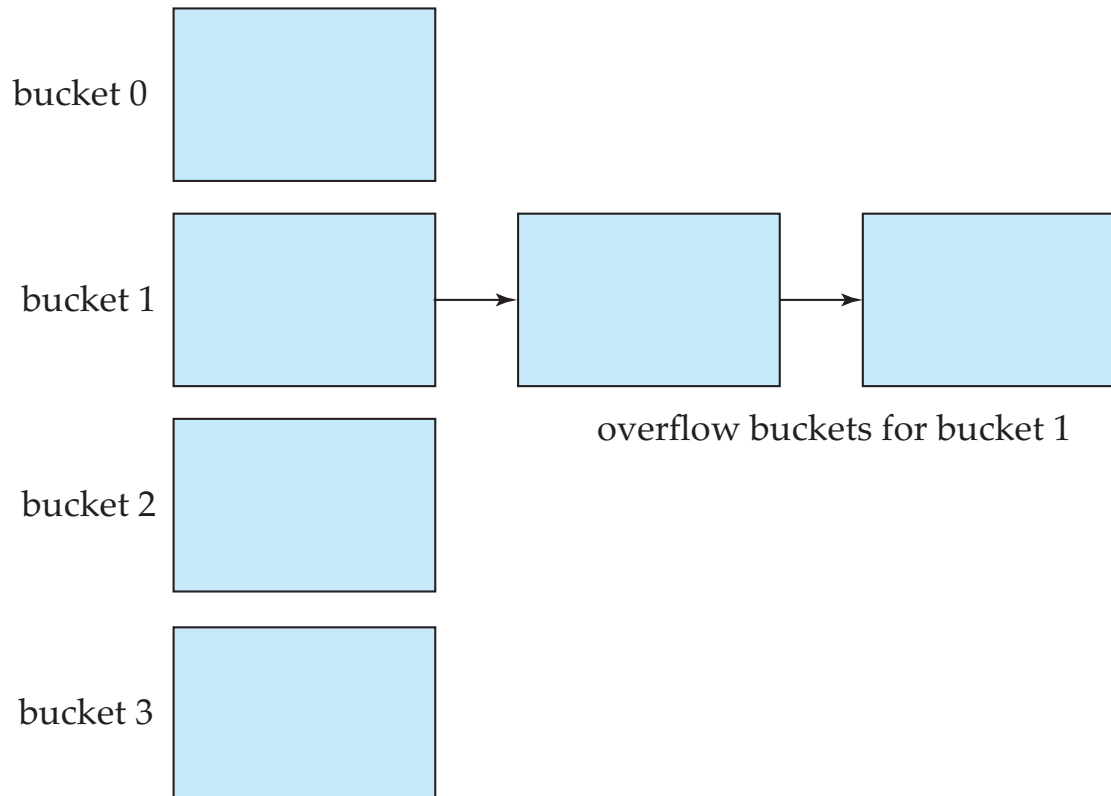
Hash Indices

- Hashing can be used not only for file organization, but also for index-structure creation.

- A **hash index** organizes the search keys, with their associated record pointers, into a hash file structure.

- Strictly speaking, hash indices are always secondary indices

  - if the file itself is organized using hashing, a separate primary hash index on it using the same search-key is unnecessary.

  - However, we use the term hash index to refer to both secondary index structures and hash organized files.

Example of Hash Index

hash index on *instructor,* on attribute *ID*

Deficiencies of Static Hashing

- In static hashing, function *h* maps search-key values to a fixed set of *B* of bucket addresses. Databases grow or shrink with time.

  - If initial number of buckets is too small, and file grows, performance will degrade due to too much overflows.

  - If space is allocated for anticipated growth, a significant amount of space will be wasted initially (and buckets will be underfull).

  - If database shrinks, again space will be wasted.

- One solution: periodic re-organization of the file with a new hash function

  - Expensive, disrupts normal operations

- Better solution: allow the number of buckets to be modified dynamically.

Index Definition in SQL

- Create an index

  create index <index-name> **on** <relation-name>
  (<attribute-list>)

  E.g.: **create index** *b-index* **on** *branch(branch_name)*

- Use **create unique index** to indirectly specify and enforce the condition that the search key is a candidate key is a candidate key.

  - Not really required if SQL **unique** integrity constraint is supported

- To drop an index

  drop index <index-name>

- Most database systems allow specification of type of index, and clustering.

End of Chapter

Figure 11.01

10101	Srinivasan	Comp. Sci.	65000	
12121	Wu	Finance	90000	
15151	Mozart	Music	40000	
22222	Einstein	Physics	95000	
32343	El Said	History	60000	
33456	Gold	Physics	87000	
45565	Katz	Comp. Sci.	75000	
58583	Califieri	History	62000	
76543	Singh	Finance	80000	
76766	Crick	Biology	72000	
83821	Brandt	Comp. Sci.	92000	
98345	Kim	Elec. Eng.	80000	

Figure 11.15
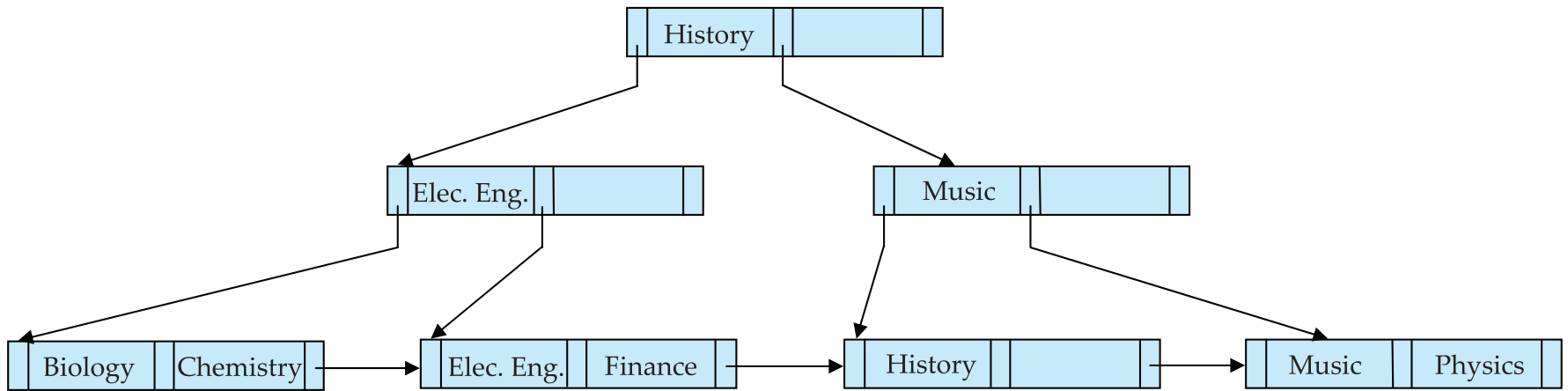
Partitioned Hashing

- Hash values are split into segments that depend on each attribute of the search-key.

$$(A_1, A_2, \ldots, A_n) \text{ for } n \text{ attribute search-key}$$

- Example: $n = 2$, for *customer,* search-key being (*customer-street, customer-city*)

search-key value	hash value
(Main, Harrison)	101 111
(Main, Brooklyn)	101 001
(Park, Palo Alto)	010 010
(Spring, Brooklyn)	001 001
(Alma, Palo Alto)	110 010

- To answer equality query on single attribute, need to look up multiple buckets. Similar in effect to grid files.
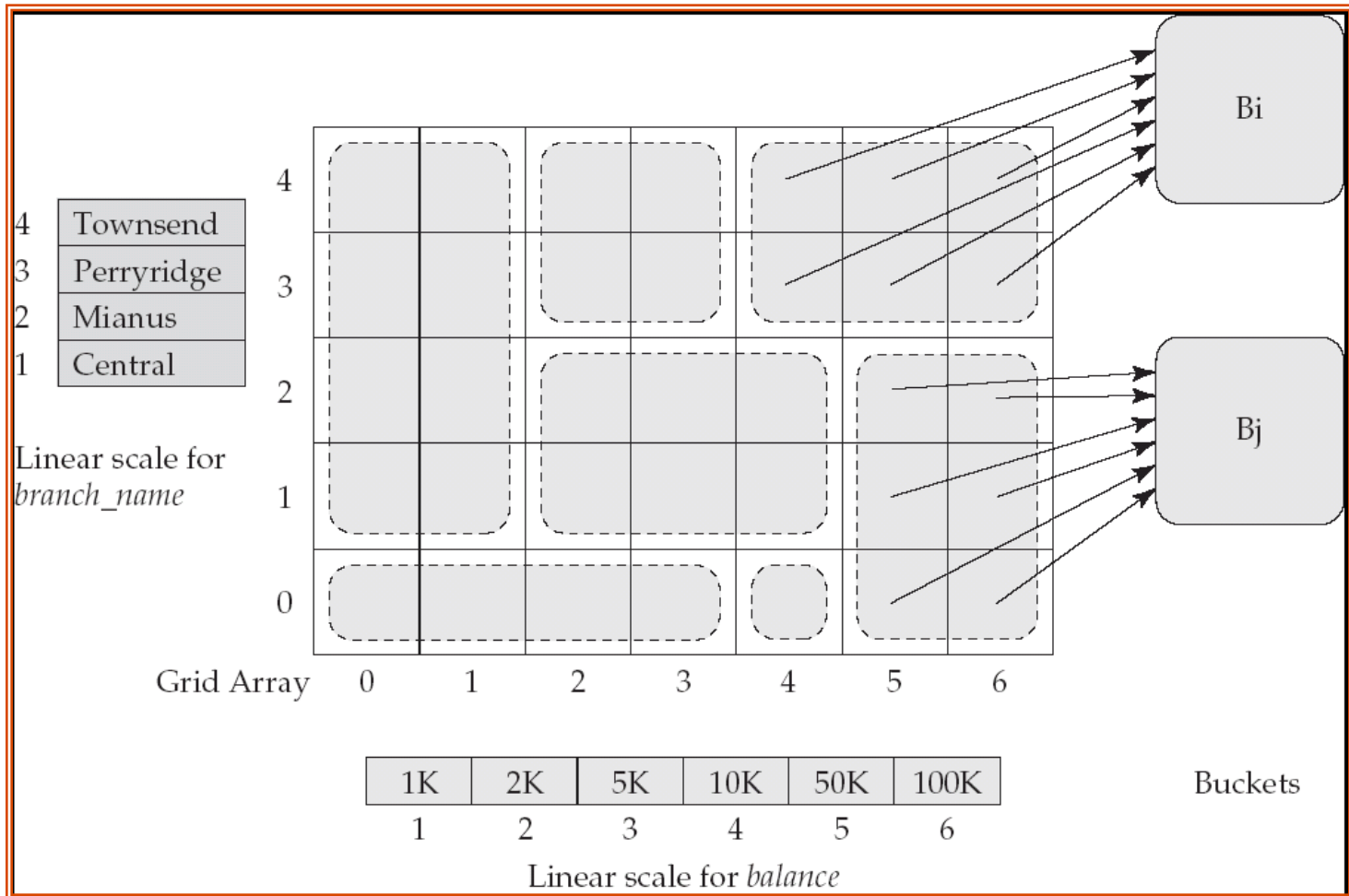
Grid Files

- Structure used to speed the processing of general multiple search-key queries involving one or more comparison operators.

- The grid file has a single grid array and one linear scale for each search-key attribute. The grid array has number of dimensions equal to number of search-key attributes.

- Multiple cells of grid array can point to same bucket

- To find the bucket for a search-key value, locate the row and column of its cell using the linear scales and follow pointer

Example Grid File for *account*

Queries on a Grid File

■ A grid file on two attributes A and B can handle queries of all following forms with reasonable efficiency

- $(a_1 \leq A \leq a_2)$
- $(b_1 \leq B \leq b_2)$
- $(a_1 \leq A \leq a_2 \ \wedge \ b_1 \leq B \leq b_2)$,.

■ E.g., to answer $(a_1 \leq A \leq a_2 \ \wedge \ b_1 \leq B \leq b_2)$, use linear scales to find corresponding candidate grid array cells, and look up all the buckets pointed to from those cells.

Grid Files (Cont.)

- During insertion, if a bucket becomes full, new bucket can be created if more than one cell points to it.

  - Idea similar to extendable hashing, but on multiple dimensions

  - If only one cell points to it, either an overflow bucket must be created or the grid size must be increased

- Linear scales must be chosen to uniformly distribute records across cells.

  - Otherwise there will be too many overflow buckets.

- Periodic re-organization to increase grid size will help.

  - But reorganization can be very expensive.

- Space overhead of grid array can be high.

- R-trees (Chapter 23) are an alternative