Overview of Storage and Indexing

(Chapter 9, 3rd edition)

"How index-learning turns no student pale
Yet holds the eel of science by the tail."
-- Alexander Pope (1688-1744)

Data on External Storage

- Disks: Can retrieve random page at fixed cost
  - But reading several consecutive pages is much cheaper than reading them in random order
- Tapes: Can only read pages in sequence
  - Way Cheaper than disks; used for archival storage
- File organization: Method of arranging a file of records on external storage.
  - Record id (rid) is sufficient to physically locate record
  - Indexes are data structures that allow us to find the record ids of records with given values in index search key fields
- Architecture: Buffer manager stages pages from external storage to main memory buffer pool. File and index layers make calls to the buffer manager.

Alternative File Types

Many alternatives exist, *each ideal for some situations, and not so good for others:*

- Heap (random order) files: *pages that are linked and recordes are in MO particular order*
- Sorted Files: contiguous pages on the disk is used and records are sorted. Can apply "divide and conquer"
- Clustered Files: Pages that are linked and records are stored in sorted order, but NOT with contiguous pages
- Indexes: Data structures to organize records via trees or hashing.

Impact of Alternative File Organizations

Many alternatives exist, *each ideal for some situations, and not so good for others:*

- Heap (random order) files: Suitable when typical access is a file scan retrieving all records.
- Sorted Files: Best if records must be retrieved in some order, or only a `range' of records is needed. Binary search can be applied
- Clustered Files: suited for range search with index. Binary search cannot be applied
- Indexes:
  - Like sorted/clustered files, they speed up searches for a subset of records, based on values in certain ("search key") fields
  - Updates are much faster than in sorted files. Why?

Indexes

- An *index* on a file speeds up selections on the *search key fields* for the index.
  - Any subset of the fields of a relation can be the search key for an index on the relation.
  - *Search key (search field or search attribute)* is not the same as a *key* (minimal set of fields that uniquely identify a record in a relation).
- An index contains a collection of *data entries*, and supports efficient retrieval of all data entries **k*** with a given key value **k**.

*Alternatives for Data Entry **k**** in Index*

- Three alternatives:
  - Data record with key value **k**
  - **<k**, rid of data record with search key value **k>**
  - **<k**, list of rids of data records with search key **k>**
- Choice of alternative for data entries is orthogonal to the indexing technique used to locate data entries with a given key value **k**.
  - Examples of indexing techniques: B+ trees, hash-based structures
  - Typically, index contains auxiliary information that directs searches to the desired data entries

Alternatives for Data Entries (Contd.)

- Alternative 1:
  - If this is used, index structure is a file organization for data records (instead of a Heap file or sorted file).
  - At most one index on a given collection of data records can use Alternative 1. (Otherwise, data records are duplicated, leading to redundant storage and potential inconsistency.)
  - If data records are very large, # of pages containing data entries is high. Implies size of auxiliary information in the index is also large, typically.

Alternatives for Data Entries (Contd.)

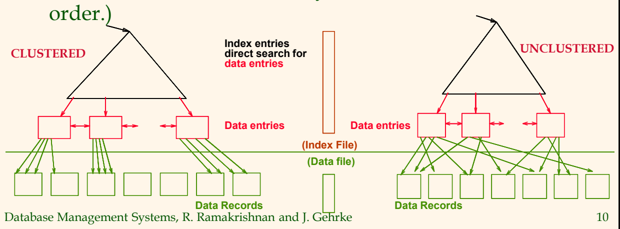
- Alternatives 2 (widely used) and 3:
  - Key entries are typically much smaller than data records. So, better than Alternative 1 with large data records, especially if search keys are small. (Portion of index structure used to direct search, which depends on size of data entries, is much smaller than with Alternative 1.)
  - Alternative 3 is more compact than Alternative 2, but leads to variable sized data entries even if search keys are of fixed length.
    - Under what conditions do we need alternative 3?

Index Classification (or types)

- *Primary* vs. *secondary*: If search key contains primary key, then called primary index.
  - *Unique* index: Search key contains a candidate key. Alt 2 is fine! (why?)
- *Clustered* vs. *unclustered*: If order of data records is the same as, or `close to', order of data (or key) entries, then called clustered index.
  - Alternative 1 implies clustered; in practice, clustered also implies Alternative 1 (since sorted files are rare).
  - A file can be clustered (or sorted) on at most one search key.
  - Cost of retrieving data records through index varies *greatly* based on whether index is clustered or not!
  - If there are multiple indexes, only one can be clustered! Why?
  - Need to be chosen judiciously.

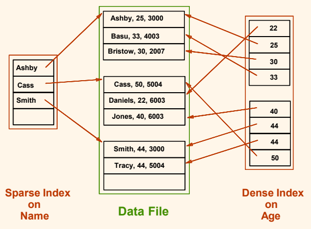
Clustered vs. Unclustered Index

- Suppose that Alternative (2) is used for data entries, and that the data records are stored in a Heap file.
  - To build clustered index, first sort the Heap file (with some free space on each page for future inserts).
  - Overflow pages may be needed for inserts. (Thus, order of data recs is `close to', but may not be identical to, the sort order.)



Index Classification (Contd.)

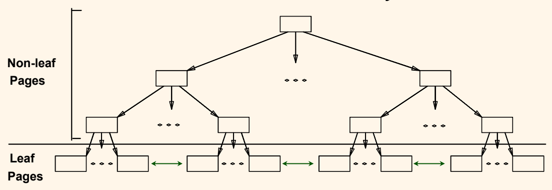
- *Dense* vs. *Sparse* index
- Dense: an index record appears for **every** search key in the file!
- Sparse: index records are created only for some of the records
  - Alternative 1 always leads to dense index. Why?
  - Every sparse index is clustered! Why?
  - Sparse indexes are smaller. why?
  - However, some useful optimizations are based on dense indexes (e.g., count)



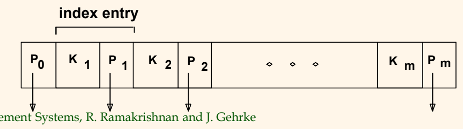
Hash-Based Indexes

- Good for equality selections. Why?
  - Index is a collection of *buckets*. Bucket = *primary* page plus zero or more *overflow* pages.
  - *Hashing function* **h**: **h**(*r*) = bucket in which record *r* belongs. **h** looks at the *search key* fields of *r*.
- If Alternative (1) is used, the buckets contain the data records; otherwise, they contain <key, rid> or <key, rid-list> pairs.

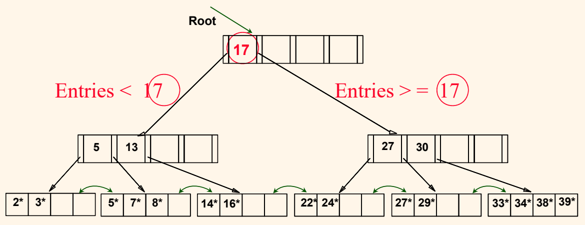
B+ Tree Indexes (B stands for ?)



- ❖ Leaf pages contain *data (search key) entries*, and are
- ❖chained (prev & next). Non-leaf pages contain *index entries*
- ❖to guide the search:
- ❖Data pages (not shown!) are separate from Leaf index pages



Example B+ Tree



- Find 28*? (not present) 29*? (present)
- All > 15* and < 30* (range condition)
- Insert/delete: Find data entry in leaf, then change it. Need to adjust parent sometimes.
  - And change sometimes bubbles up the tree

Tree-Structured Indexes

Chapter 10

Introduction

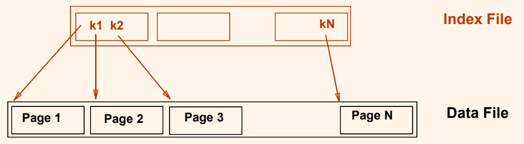
- Tree-structured indexing techniques support both *range searches* and *equality searches*.
  - ISAM: static structure (pre 1970)
  - B tree: dynamic structure (around 1971)
  - *B+ tree*: dynamic, adjusts gracefully under inserts and deletes. (improved B Tree, after 1975)

Introduction

- *As for any index, 3 alternatives for data entries* **k***:
  - Data record with key value **k**
  - <**k**, rid of data record with search key value **k**>
  - <**k**, list of rids of data records with search key **k**>
- Choice is orthogonal to the *indexing technique* used to locate data entries **k***.
- Tree-structured indexing techniques support both *range searches* and *equality searches*.
- *ISAM*: static structure; *B+ tree*: dynamic, adjusts gracefully under inserts and deletes.

Range Searches

- *``Find all students with gpa > 3.0''*
  - If data is in sorted file, do binary search to find first such student, then scan to find others.
  - Cost of binary search can be quite high.
- Simple idea: Create an `index' file.



- *Can do binary search on (smaller) index file!*

Example (sorted vs. tree indexing)

- 10000 pages sorted (contiguous)
- Search on the data pages takes $\log_2 10000$ or 14 page accesses ($\log_2 10000$ is 14 + 1 more for data access)
- Let the record size be 200 bytes; key size be 10 bytes, 40 records/page (8K page size)
- Now the data pages and leaf nodes of the index for the above are as follows:
  - 10000*40 = 400000 keys
  - Key, ptr pairs per page is 8000/20 = 400 (10 bytes for pointer and 10 bytes for the key)
  - Fan out of the index tree is ~ 400
  - # of index leaf pages is 400000/400 (key, ptr per page)= 1000 pages
- We reduce page access from 14+1 to 2+1 (why?) with the index
- Log_{400} 1000 is 2 (+ 1 more access for the data)

Indexes

- Binary Trees
- AVL Trees
- ISAM (used by DBMSs before B and B+ trees)
- B Trees (height Balanced)
- B+ trees (also height balanced)

Index Characteristics

- Static Vs. Dynamic
- Top Down Vs. Bottom up
- Fixed number Vs. dynamic number of index pages
- Balanced Vs. Unbalanced

Binary and AVL trees

- Binary trees
  - At most 2 descendents (or children) per node.
  - No constraint on the length of the paths from the root.
- AVL trees
  - An AVL tree is a binary tree in which the difference between the height of the right and left sub-trees (or the root node) is never more than one.

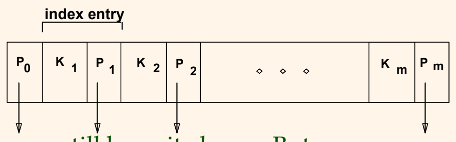
Binary and Avl trees

- Binary and AVL are top-down tree constructions. Once the wrong key is placed in the root of the tree (or in the root of any sub tree), it is difficult to balance the tree without significant overhead (reorganization)
- How can we guarantee that each of the pages contain at least some minimum number of keys (important for large page sizes)
- How can we guarantee that the heights of different paths are the same (or are not very different)

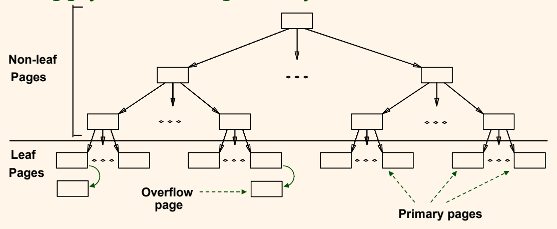
Balanced (B and B+) trees

- A balanced tree builds the tree upward from the bottom instead of downward (like AVL and binary trees) from the top.
- Rather than finding ways to undo a bad situation, the problem is avoided altogether from the very beginning.
- With balanced trees, you allow the root to split and merge, rather than set it up and find ways to change it.

ISAM

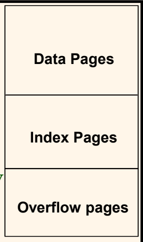


- Index file may still be quite large. But we can apply the idea repeatedly!



- *Leaf pages contain data entries.*

Comments on ISAM

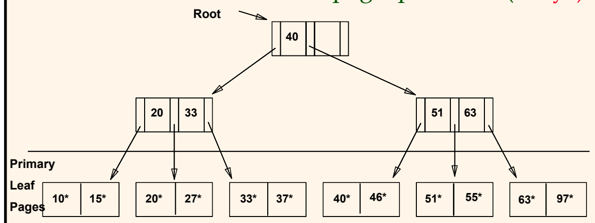


- *File creation*: Leaf (data) pages allocated sequentially (or contiguously), sorted by search key; then index pages allocated, then space for overflow pages.
- *Index entries*: <search key value, page id>; they `direct' search for *data entries*, which are in leaf pages.
- Search: Start at root; use key comparisons to go to leaf. Cost $\propto \log_F N$; F = # entries/index pg, N = # leaf pgs
- Insert: Find leaf where data entry belongs to, and put it there.
- Delete: Find and remove from leaf; if empty overflow page, de-allocate.

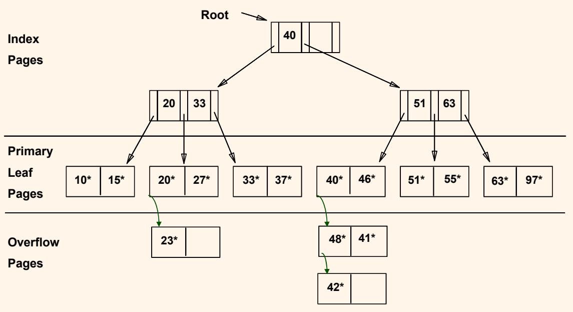
- **Static tree structure**: *inserts/deletes affect only leaf pages.*

Example ISAM Tree

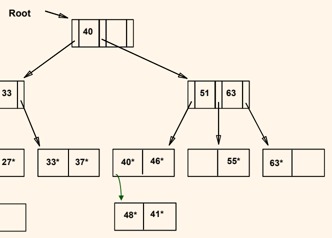
- Each node can hold 2 entries;
- no need for `next-leaf-page' pointers. (Why?)



After Inserting 23*, 48*, 41*, 42* ...



... Then Deleting 42*, 51*, 97*



- *Note that 51* appears in index levels, but not in leaf!*

Summary of ISAM

- ISAM structure is created for a given file
- Number of index pages (entries) do not change
- Overflow pages are added and deleted as needed
- The number of primary leaf pages do not change
- Leaf pages are allocated sequentially (hence no need for pointers!)
- No need to lock index pages (more concurrency) why?
- The index tree is NOT balanced dynamically (balanced statically at creation time)
- Index value exists, but may not be a record for that value
- Potentially long overflow lists
- May need to re-create ISAM index to overcome the above

B and B+ trees

- The common theme of all index structures is that they associatively map some attribute of a data object to some locator information which can be used to retrieve the actual data object.
- Typically, index scans are separated from record lookup:
  - Allows to scan an index without retrieving data from the data file (reduces I/O for count and other computations)
  - For B+ trees, leaves (of index) can be accessed sequentially
  - Joining of indices using record id (rid/tid)

B and B+ trees

- Both are balanced trees; i.e., the length of the path from the root to any leaf node is the same.
- Both store pointers to data (in Alt 2 and 3) in the index nodes. Alt 1 is typically not used!!
- The data records are stored on separate pages.
- Both are constructed in a bottom-up manner
- However, there are some fundamental differences between them

FIGURE 14.10
B-tree structures. (a) A node in a B-tree with q – 1 search values. (b) A B-tree of order p = 3. The values were inserted in the order 8, 5, 1, 7, 3, 12, 9, 6.

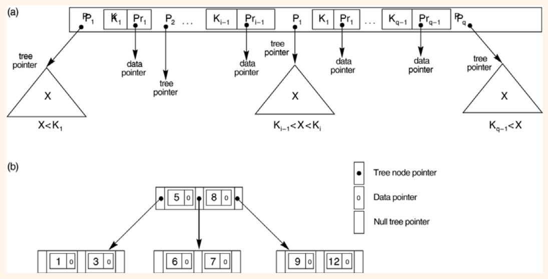
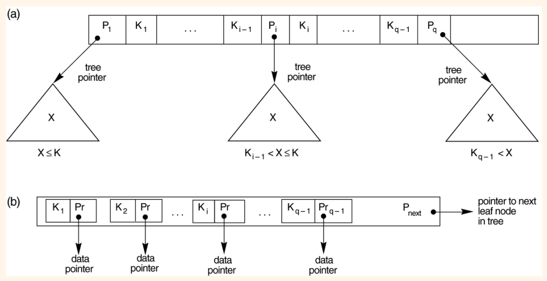


FIGURE 14.11
The nodes of a B+-tree. (a) Internal node of a B+-tree with q –1 search values. (b) Leaf node of a B+-tree with q – 1 search values and q – 1 data pointers.



B-Tree

- B tree index nodes/pages (both leaf and non-leaf) contain
  <key, data ptr, child ptr> whereas a
- B+ tree non-leaf index node/page contains
  <key, child ptr> pairs and
- B+ tree index leaf node/page contains
  <key, data ptr> pairs
- B+ tree leaf index pages/nodes are doubly linked
- Data pages are separate in both the cases

B Tree -- Implications

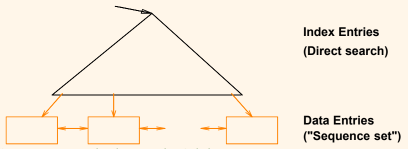
- In a B tree (as compared to a B+ tree):
  - Packing density of an index page is less (why?)
  - Key values are NOT repeated in the index (why?)
  - Leaf index pages cannot be traversed for scan (why?)
  - For a key that exists in the file, search may be stopped before reaching the leaf node (why?)

B Tree -- Implications

- Sequential access of the entire file requires touching every node in the index (why?)
  - What kind of traversal is needed?
- Perhaps good (why?) when the key forms most of the data record.
  - i.e., the ratio of data record size to key size is closer to 1!

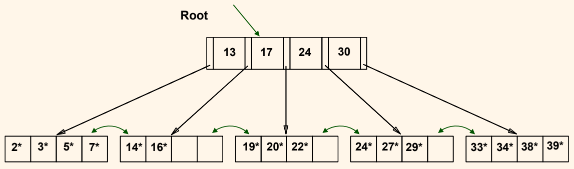
B+ Tree: Most Widely Used Index

- Insert/delete at log $_F$ N cost; keep tree *height-balanced*. (F = fanout, N = # leaf pages)
- Minimum 50% occupancy (except for root). Each node contains **d** <= *m* <= 2**d** entries. The parameter **d** is called the *order* of the tree.
- Supports equality and range-searches efficiently.



Example B+ Tree

- Search begins at root, and key comparisons direct it to a leaf (as in ISAM).
- Search for 5*, 15*, all data entries >= 24* ...



- *Based on the search for 15*, we <u>know</u> it is not in the tree!*

B+ Trees in Practice

- Typical order: 100+. Typical fill-factor: 67%.
  - average fanout = 133
- Typical capacities (not including root):
  - Height 1: 133^1 = 133 nodes (can access 133^2 data recs)
  - Height 2: 133^2 = 17689 nodes
  - Height 3: 133^3 = 2,352,637 nodes
  - Height 4: 133^4 = 312,900,700 nodes
    - (can access $4.16^{10 \text{ records}}$)
- Remember that each leaf index node holds 133 keys and 133 record pointers
- Can often hold top levels in buffer pool:
  - Level 1 = 1 page = 8 Kbytes
  - Level 2 = 133 pages = 1 Mbyte
  - Level 3 = 17,689 pages = 133 MBytes

Inserting a Data Entry into a B+ Tree

- Find correct leaf node *L*.
- Put data entry onto *L*.
  - If *L* has enough space, *done*!
  - Else, either (overflow case)
    - i) Redistribute entries evenly **with a sibling**, **copy up (not insert)** middle key, or
    - ii) *split* *L (into L and a new node L2) and* Insert index entry pointing to *L2* into parent of *L*.

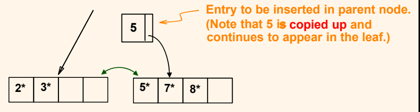
    // Can avoid split sometimes by redistributing

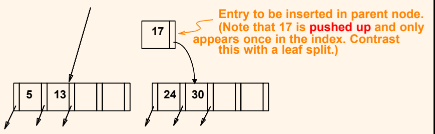
    //when can you NOT redistribute?

    // if there is a proper sibling, and update parent node!
- This can happen recursively
  - For an index node, you can either redistribute evenly or split, and **push up** middle key. (Contrast with leaf splits.)
- Splits "grow" tree; root split increases height.
  - Tree growth: gets *wider* or *one level taller at top.*

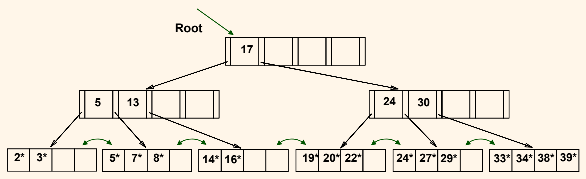
Inserting 8 into Example B+ Tree*

- Observe how minimum occupancy is guaranteed in both leaf and index pg splits.
- Note the difference between *copy-up* and *push-up;* be sure you understand the reasons for this.


5
Entry to be inserted in parent node.
(Note that 5 is copied up and continues to appear in the leaf.)
2* 3*
5* 7* 8*


17
Entry to be inserted in parent node.
(Note that 17 is pushed up and only appears once in the index. Contrast this with a leaf split.)
5 13
24 30

*Example B+ Tree After Inserting 8**

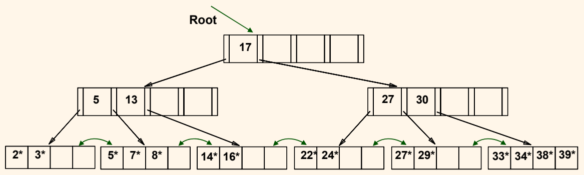


- Notice that root was split, leading to increase in height.
- In this example, we can avoid split by re-distributing entries; however, this is usually not done in practice.

Deleting a Data Entry from a B+ Tree

- Start at root, find leaf *L* where entry belongs.
  - If no entry, *done!*
- Remove the entry.
  - If L is at least half-full, *done!*
  - If L has only **d-1** entries, (underflow)
    - Try to re-distribute, borrowing from *sibling (adjacent node with same parent as L)*. Middle key is copied up
    - If re-distribution fails, *merge* *L* and sibling.
- If merge occurred, must delete entry (pointing to *L* or sibling) from parent of *L*.
- Merge could propagate to root, decreasing height.
- If re-distribution, update the parent node to reflect this (copy up and update)! Not move up!
- Recursive algorithm

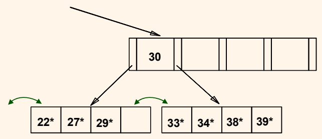
Example Tree After (Inserting 8, Then) Deleting 19* and 20* ...*

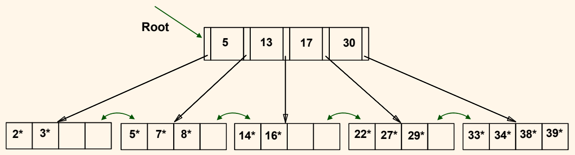


- Deleting 19* is easy.
- Deleting 20* is done with re-distribution. Notice how middle key is *copied up for update*.

*... And Then Deleting 24**

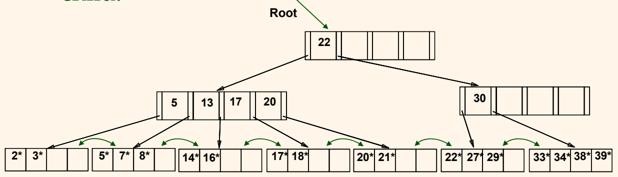
- Must merge.
- Observe *\`toss'* of index entry (on right), and *\`pull down'* of index entry (below).





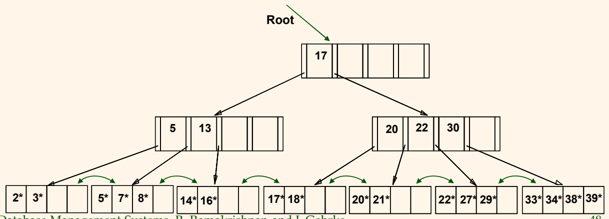
Example of Non-leaf Re-distribution

- Tree is shown below *during deletion* of 24*. (What could be a possible initial tree?)
- In contrast to previous example, we can re-distribute entry from left child of root to right child.



After Re-distribution

- Intuitively, entries are re-distributed by *\`pushing through'* the splitting entry in the parent node.
- It suffices to re-distribute index entry with key 20; we've re-distributed 17 as well for illustration.

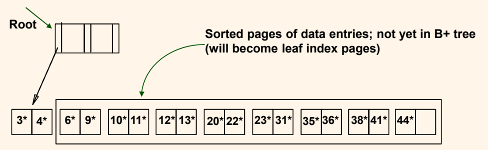


Prefix Key Compression

- Important to increase fan-out. (Why?)
- Key values in index entries only `direct traffic'; can often compress them.
  - E.g., If we have adjacent index entries with search key values *Dannon Yogurt*, *David Smith* and *Devarakonda Murthy*, we can abbreviate *David Smith* to *Dav*. (The other keys can be compressed too ...)
    - Is this correct? Not quite! What if there is a data entry *Davey Jones*? (Can only compress *David Smith* to *Davi*)
    - In general, while compressing, must leave each index entry greater than every key value (in any subtree) to its left.
- Insert/delete must be suitably modified.

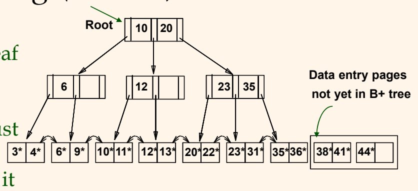
Bulk Loading of a B+ Tree

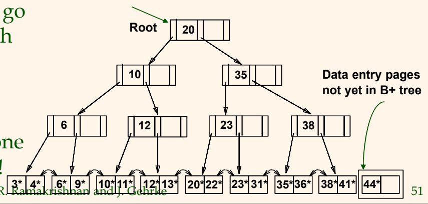
- If we have a large collection of records, and we want to create a B+ tree on some field, doing so by repeatedly inserting records is very slow.
- Bulk Loading can be done much more efficiently.
- *Initialization*: Sort all data entries, insert pointer to first (leaf index) page in a new (root) page.


Root
Sorted pages of data entries; not yet in B+ tree (will become leaf index pages)
3* 4*
6* 9*
10* 11*
12* 13*
20* 22*
23* 31*
35* 36*
38* 41*
44*

Bulk Loading (Contd.)

- Index entries for leaf pages always entered into right-most index page just above leaf level. When this fills up, it splits. (Split may go up right-most path to the root.)
- Much faster than repeated inserts, especially when one considers locking!





Cost of bulk loading (creating an index)

1) Creating the data entries to insert in the index
2) Sort the data entries
3) Building the index

Cost of 1) is (R+E), R #pages containing records and E is the # of pages with data entries; access each page of R to get the data entries

Sort cost: approximately 4E

Cost of writing out all the index pages

Summary of Bulk Loading

- Option 1: multiple inserts.
  - Slow.
  - Does not give sequential storage of leaves.
- Option 2: *Bulk Loading*
  - Has advantages for concurrency control.
  - Fewer I/Os during build.
  - Leaves will be stored sequentially (and linked, of course).
  - Can control "fill factor" on pages.

A Note on `Order'

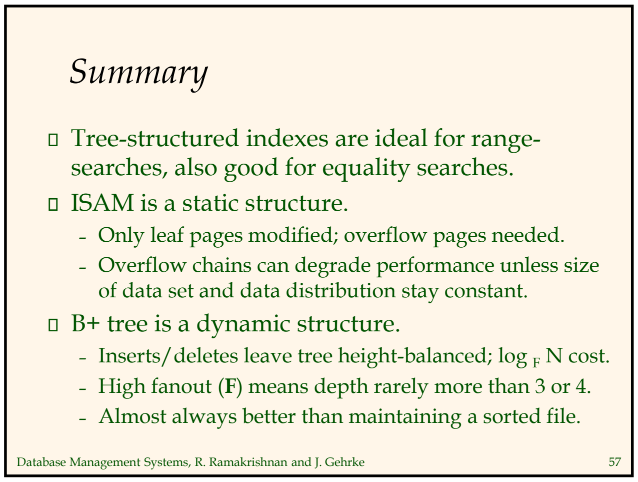
- *Order* (**d**) concept replaced by physical space criterion in practice (`*at least half-full*').
  - Index pages can typically hold many more entries than leaf pages.
  - Variable sized records and search keys mean different nodes will contain different numbers of entries.
  - Even with fixed length fields, multiple records with the same search key value (*duplicates*) can lead to variable-sized data entries (if we use Alternative (3)).

A Note on Alt 3

- *Storing multiple data pointers is not desirable as it gives rise to variable size records in leaf index pages!*
- *For clustered B+ tree index, this can be done relatively easily by scanning additional data records on the same page as it is clustered on key value*
- *For non-clustered B+ tree index, the above cannot be done. (why?)*
  - *One possibility is to add overflow pages from the leaf index page that points to multiple data records!*

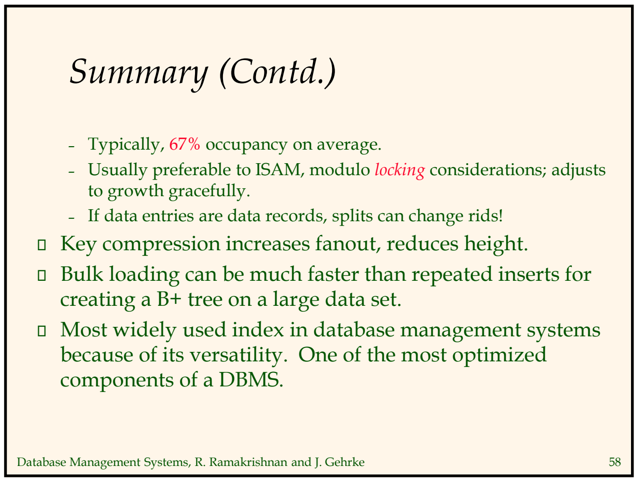
Summary

- B and B+ trees can be viewed as multi-level indexes or height balanced trees.
- B and B+ trees overcome the problems associated with binary and AVL trees
  - Binary requires too many i/os (log 2)
  - May be expensive to keep the index sorted
  - Very sensitive to order of inserts
  - Avl is a height balanced 1-tree or HB(1) tree
- B and B+ trees are height-balanced 0-tree or H(0) or completely balanced



Summary

- Tree-structured indexes are ideal for range-searches, also good for equality searches.
- ISAM is a static structure.
  - Only leaf pages modified; overflow pages needed.
  - Overflow chains can degrade performance unless size of data set and data distribution stay constant.
- B+ tree is a dynamic structure.
  - Inserts/deletes leave tree height-balanced; log $_F$ N cost.
  - High fanout (**F**) means depth rarely more than 3 or 4.
  - Almost always better than maintaining a sorted file.



Summary (Contd.)

- Typically, 67% occupancy on average.
- Usually preferable to ISAM, modulo *locking* considerations; adjusts to growth gracefully.
- If data entries are data records, splits can change rids!

- Key compression increases fanout, reduces height.
- Bulk loading can be much faster than repeated inserts for creating a B+ tree on a large data set.
- Most widely used index in database management systems because of its versatility. One of the most optimized components of a DBMS.

Thank You !