Chapter 2

Indexing Structures for Files

Adapted from the slides of “Fundamentals of Database Systems” (Elmasri et al., 2003)

Chapter outline

- Types of Single-level Ordered Indexes
  - Primary Indexes
  - Clustering Indexes
  - Secondary Indexes
- Multilevel Indexes
- Dynamic Multilevel Indexes Using B-Trees and B+-Trees
- Indexes on Multiple Keys

Indexes as Access Paths

- A single-level index is an auxiliary file that makes it more efficient to search for a record in the data file.
- The index is usually specified on one field of the file (although it could be specified on several fields)
- One form of an index is a file of entries **<field value, pointer to record>,** which is ordered by field value

- The index is called an *access path* on the field.

Indexes as Access Paths (cont.)

- The index file usually occupies considerably less disk blocks than the data file because its entries are much smaller
- A binary search on the index yields a pointer to the file record
- Indexes can also be characterized as dense or sparse.
  - A **dense index** has an index entry for *every search key value* (and hence every record) in the data file.
  - A **sparse** (or **nondense**) **index**, on the other hand, has index entries for only some of the search values

Example: Given the following data file:
EMPLOYEE(NAME, SSN, ADDRESS, JOB, SAL, ...)
Suppose that:
record size R=150 bytes
block size B=512 bytes
r=30000 records

Then, we get:
blocking factor Bfr= B div R= $\lfloor B/R \rfloor$ = 512 div 150= 3 records/block
number of file blocks b= $\lceil r/Bfr \rceil$= (30000/3)= 10000 blocks

For an index on the SSN field, assume the field size V_{SSN}=9 bytes,
assume the record pointer size P_R=7 bytes. Then:
index entry size R_I=(V_{SSN}+ P_R)=(9+7)=16 bytes
index blocking factor Bfr_I= B div R_I= 512 div 16= 32 entries/block
number of index blocks b_i= (r/ Bfr_I)= (30000/32)= 938 blocks
binary search needs $\log_2 b_i$= $\log_2$938= 10 block accesses

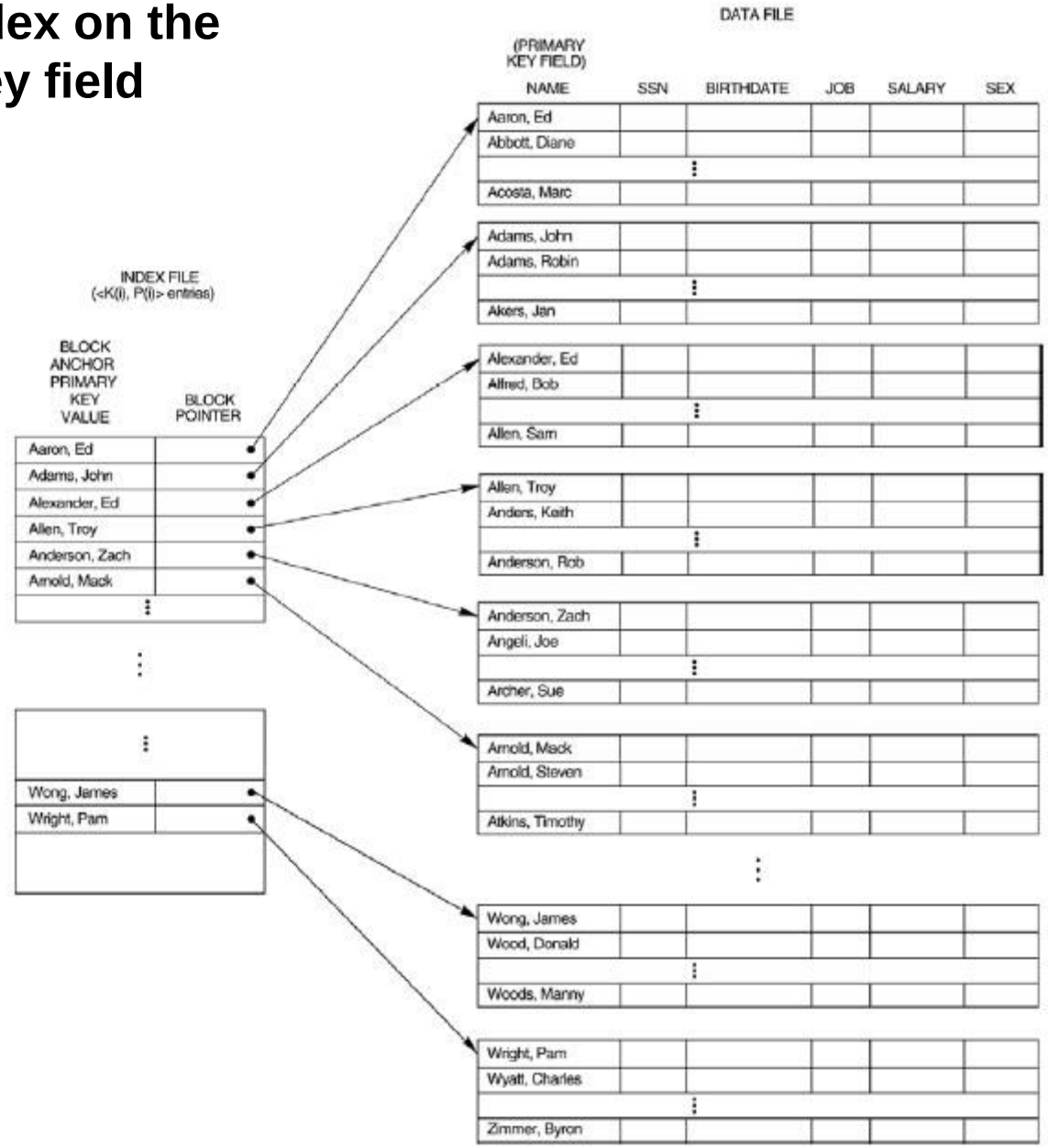
This is compared to an average linear search cost of:
(b/2)= 10000/2 = 5000 block accesses
If the file records are ordered, the binary search cost would be:
$\log_2 b$= $\log_2$10000 = 13 block accesses

Types of Single-Level Indexes

Primary Index

- Defined on an ordered data file
- The data file is ordered on a *key field,* includes one index entry *for each block* in the data file; the index entry has the key field value for the *first record* in the block, which is called the *block anchor*
- A similar scheme can use the *last record* in a block.
- A primary index is a nondense (sparse) index, since it includes an entry for each disk block of the data file and the keys of its anchor record rather than for every search value.

Primary index on the ordering key field

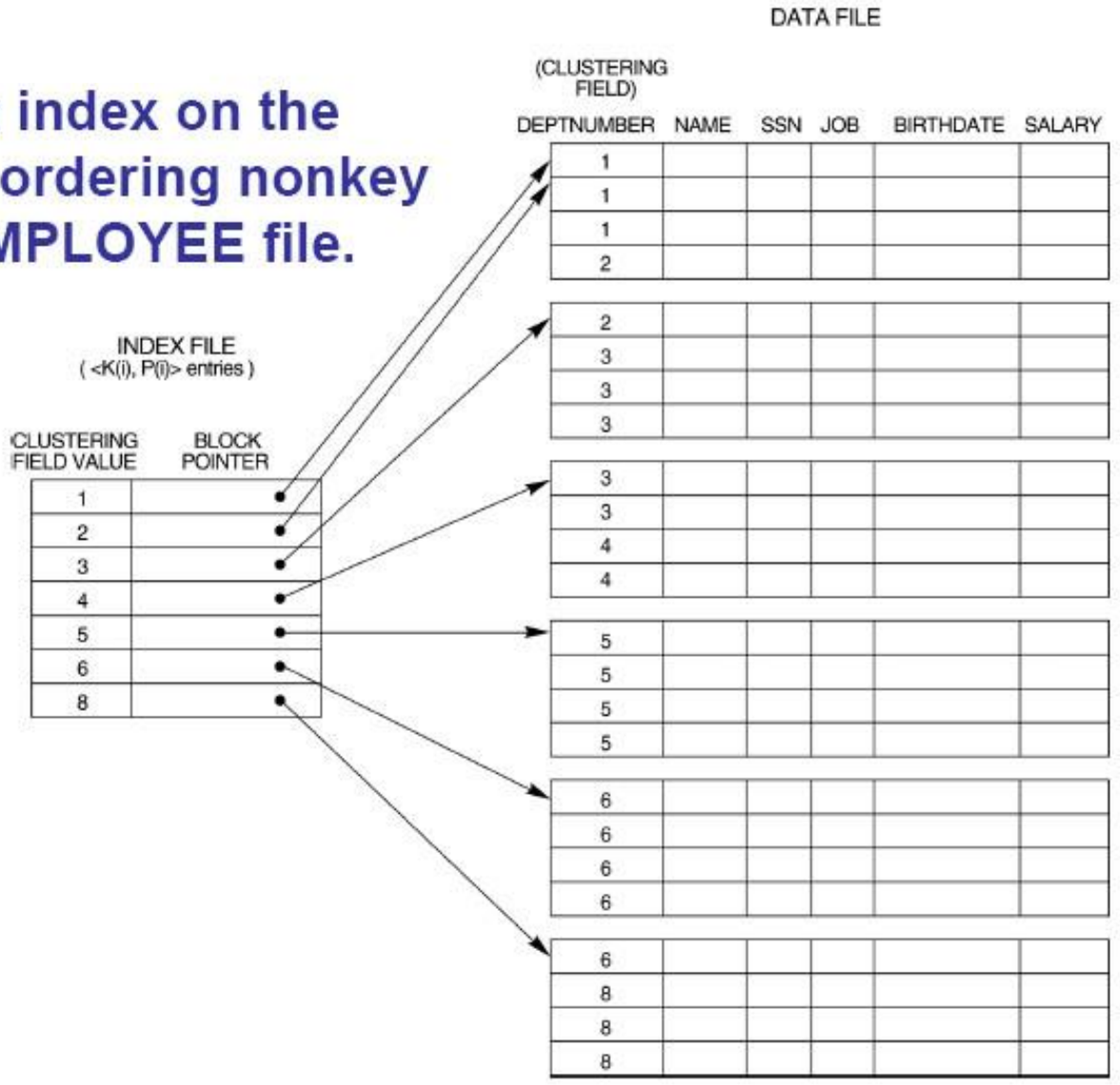


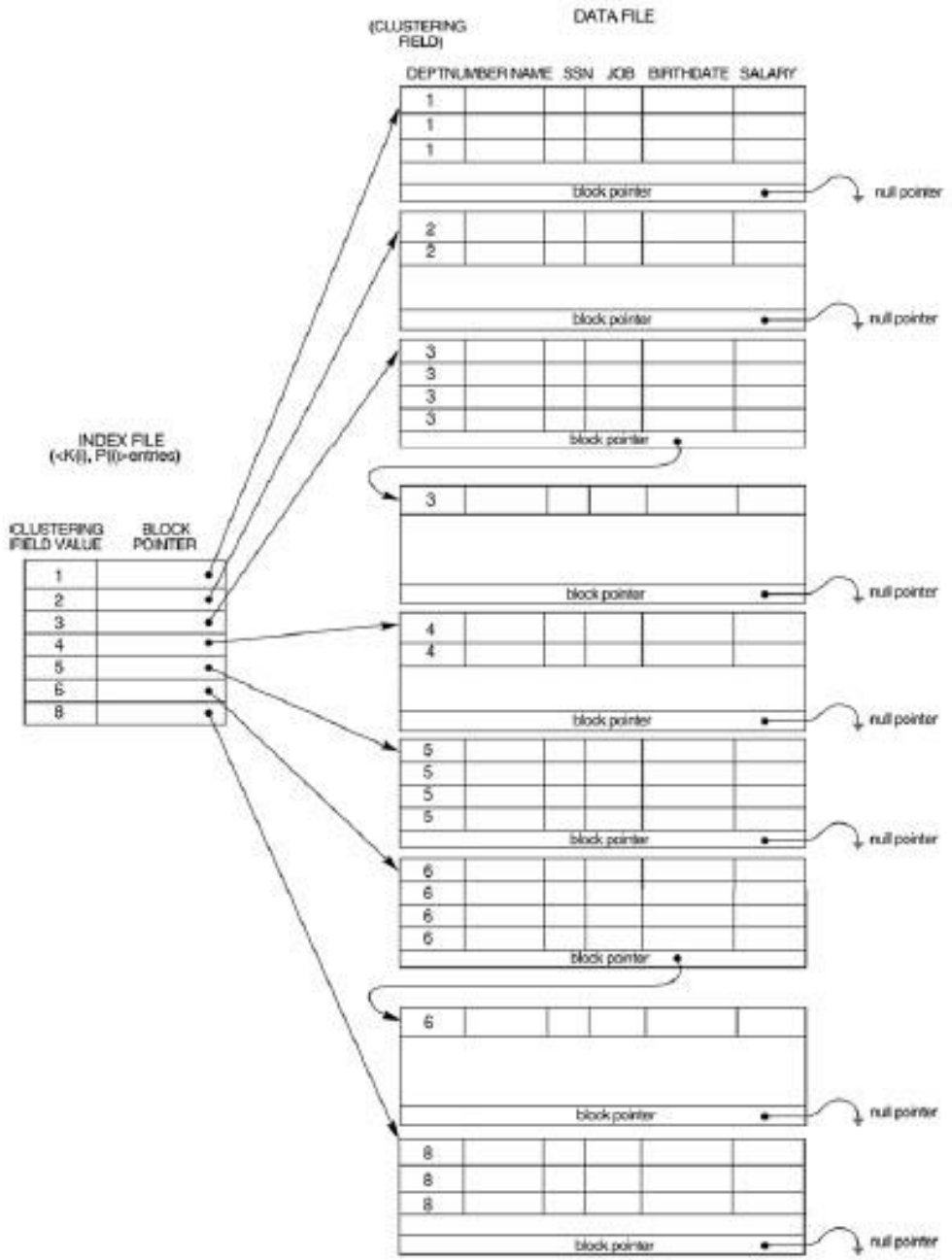
Types of Single-Level Indexes

Clustering Index

- Defined on an ordered data file
- The data file is ordered on a *non-key field* unlike primary index, which requires that the ordering field of the data file have a distinct value for each record.
- Includes one index entry *for each distinct value* of the field; the index entry points to the first data block that contains records with that field value.
- It is another example of *nondense* index where insertion and deletion is relatively straightforward with a clustering index.

A clustering index on the DEPTNUMBER ordering nonkey field of an EMPLOYEE file.



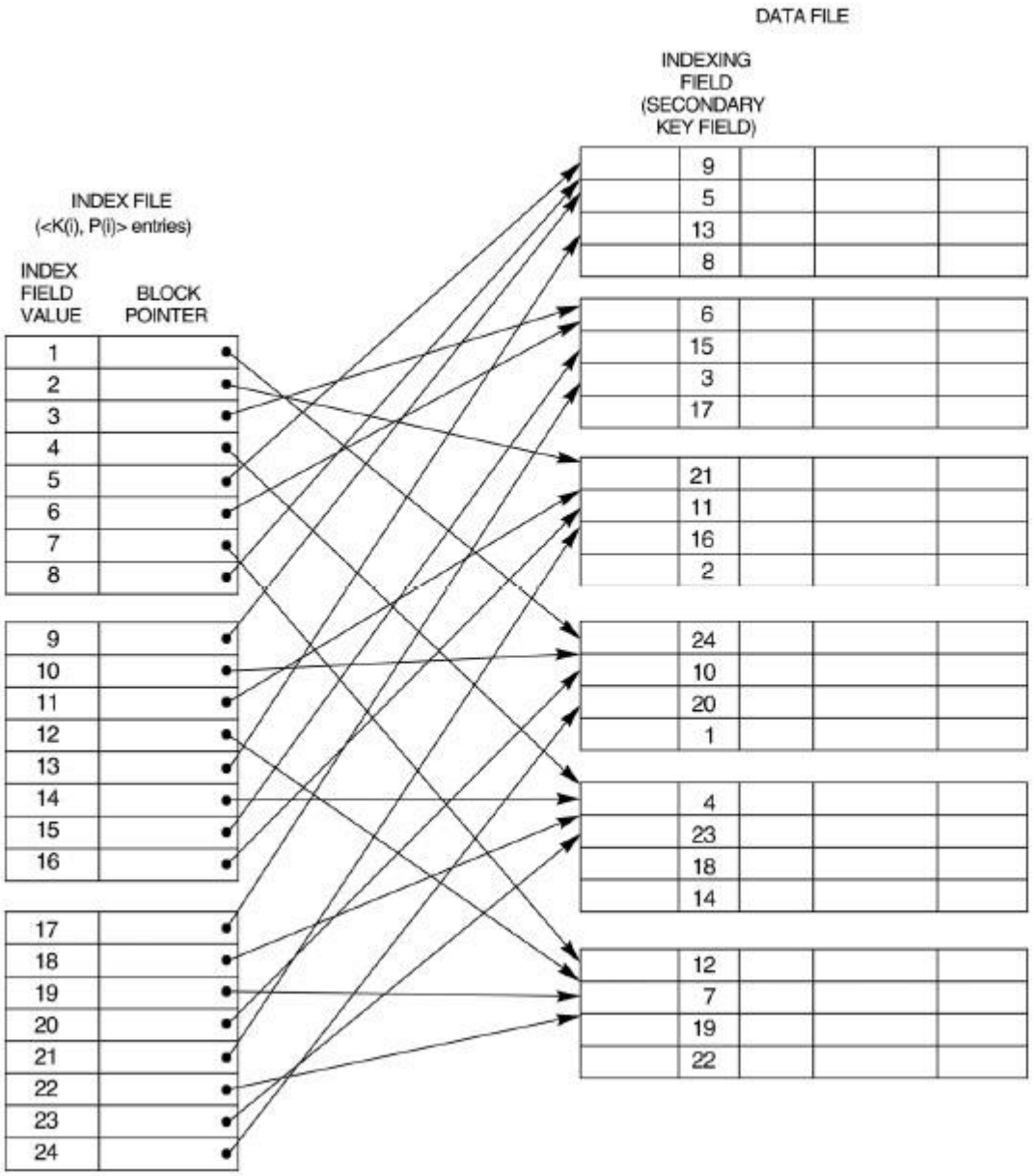


Clustering index with a separate block cluster for each group of records that share the same value for the clustering field.

Types of Single-Level Indexes

Secondary Index

- A secondary index provides a secondary means of accessing a file for which some primary access already exists.
- The secondary index may be on a field which is a candidate key and has a unique value in every record, or a nonkey with duplicate values.
- The index is an ordered file with two fields.
  - The first field is of the same data type as some *nonordering field* of the data file that is an *indexing field.*
  - The second field is either a *block* pointer or a *record* pointer. There can be *many* secondary indexes (and hence, indexing fields) for the same file.
- Includes one entry *for each record* in the data file; hence, it is a *dense index*



A dense secondary index (with block pointers) on a nonordering key field of a file.

A secondary index (with recored pointers) on a nonkey field implemented using one level of indirection so that index entries are of fixed length and have unique field values.

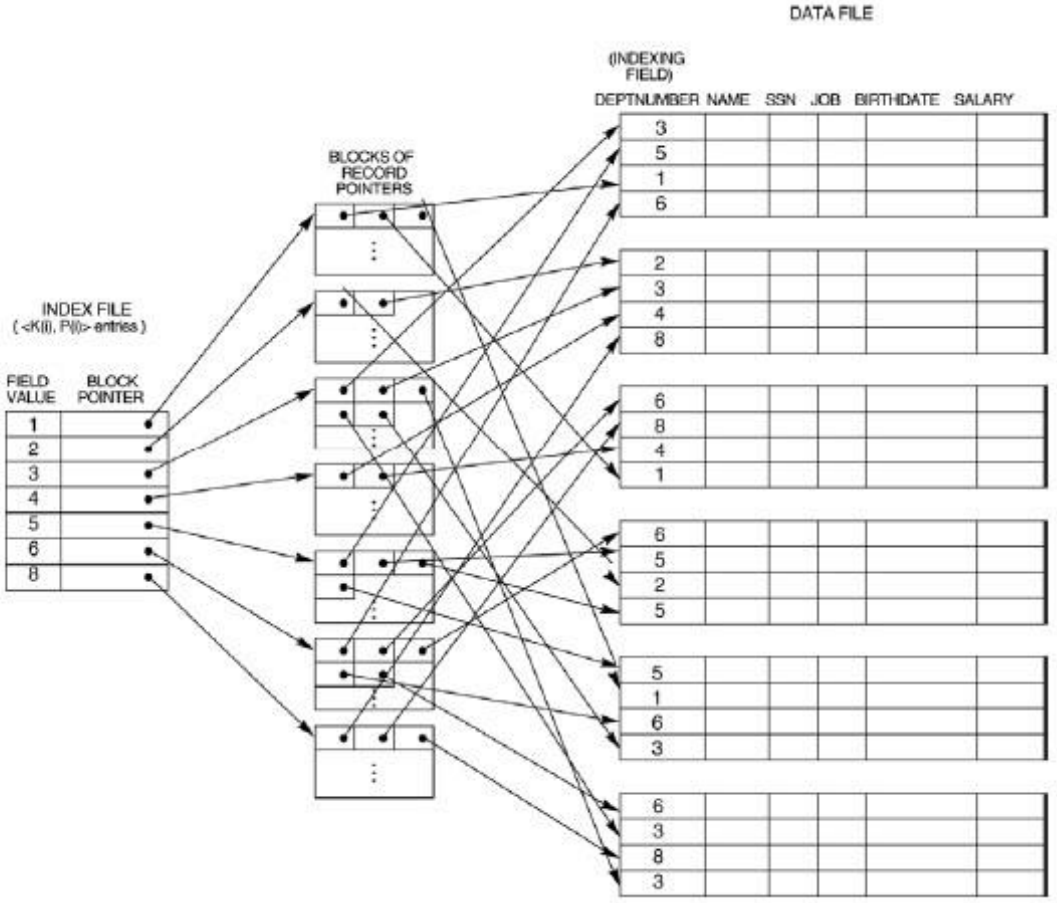


TABLE 14.1 TYPES OF INDEXES BASED ON THE PROPERTIES OF THE INDEXING FIELD

	INDEX FIELD USED FOR ORDERING THE FILE	INDEX FIELD NOT USED FOR ORDERING THE FILE
Indexing field is key	Primary index	Secondary index (Key)
Indexing field is nonkey	Clustering index	Secondary index (NonKey)

TABLE 14.2 PROPERTIES OF INDEX TYPES

TYPE OF INDEX	NUMBER OF (FIRST-LEVEL) INDEX ENTRIES	DENSE OR NONDENSE	BLOCK ANCHORING ON THE DATA FILE
Primary	Number of blocks in data file	Nondense	Yes
Clustering	Number of distinct index field values	Nondense	Yes/no[a]
Secondary (key)	Number of records in data file	Dense	No
Secondary (nonkey)	Number of records[b] or Number of distinct index field values[c]	Dense or Nondense	No

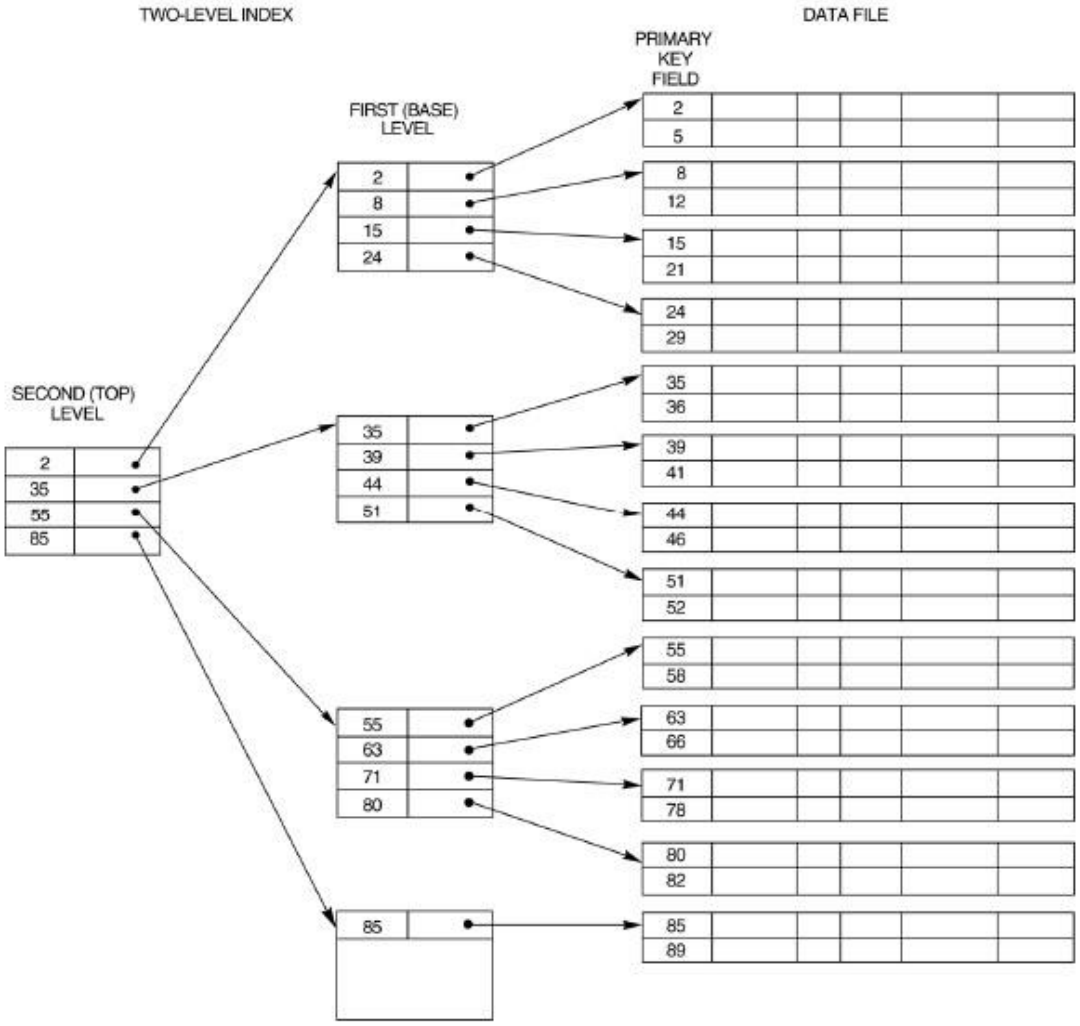
[a]Yes if every distinct value of the ordering field starts a new block; no otherwise.

[b]For option 1.

[c]For options 2 and 3.

Multi-Level Indexes

- Because a single-level index is an ordered file, we can create a primary index *to the index itself*; in this case, the original index file is called the *first-level index* and the index to the index is called the *second-level index.*
- We can repeat the process, creating a third, fourth, ..., top level until all entries of the *top level* fit in one disk block
- A multi-level index can be created for any type of first-level index (primary, secondary, clustering) as long as the first-level index consists of *more than one* disk block

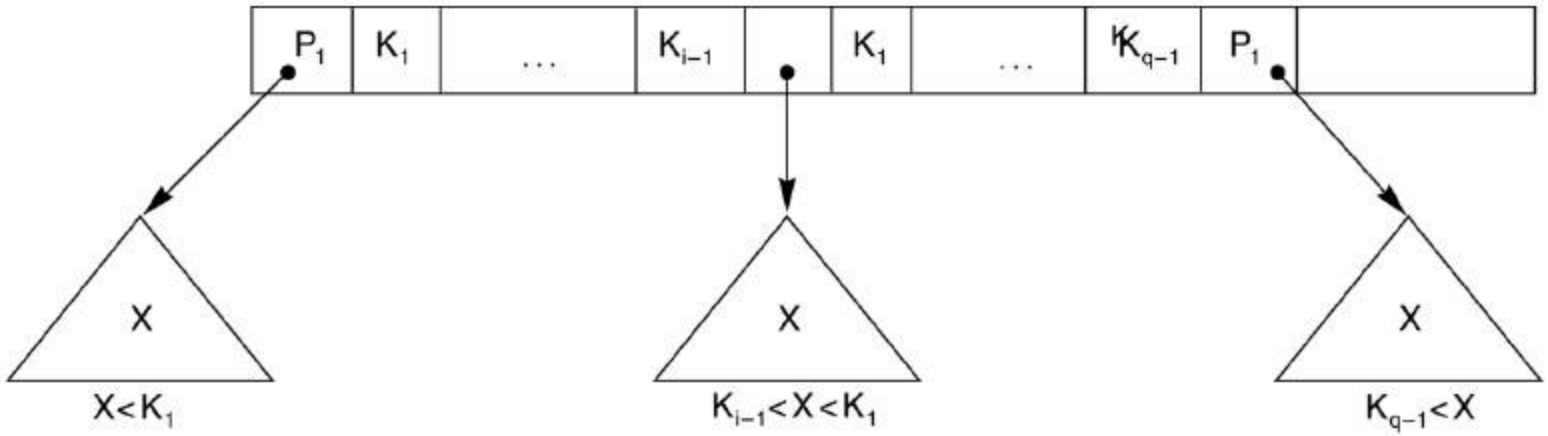


A two-level primary index resembling ISAM (Indexed Sequential Access Method) organization.

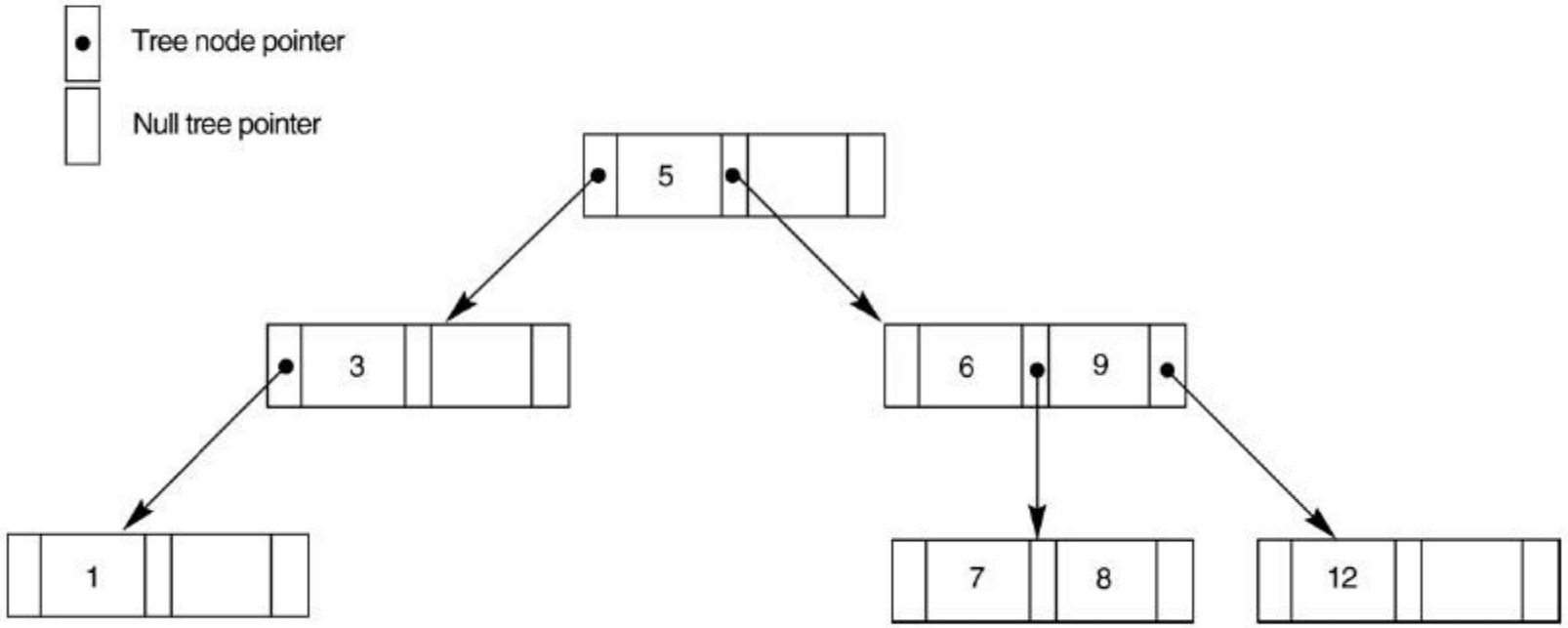
Multi-Level Indexes

- Such a multi-level index is a form of *search tree*; however, insertion and deletion of new index entries is a severe problem because every level of the index is an *ordered file.*

A node in a search tree with pointers to subtrees below it.



A search tree of order p = 3.



Dynamic Multilevel Indexes Using B-Trees and B+-Trees

- Because of the insertion and deletion problem, most multi-level indexes use B-tree or B+-tree data structures, which leave space in each tree node (disk block) to allow for new index entries
- These data structures are variations of search trees that allow efficient insertion and deletion of new search values.
- In B-Tree and B+-Tree data structures, each node corresponds to a disk block.
- Each node is kept between half-full and completely full.

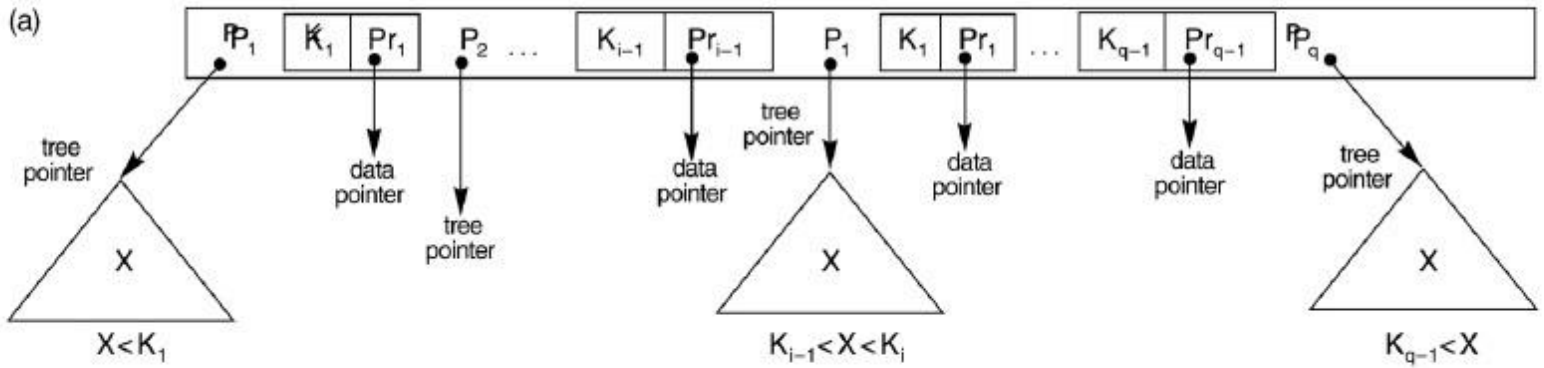
Dynamic Multilevel Indexes Using B-Trees and B+-Trees (cont.)

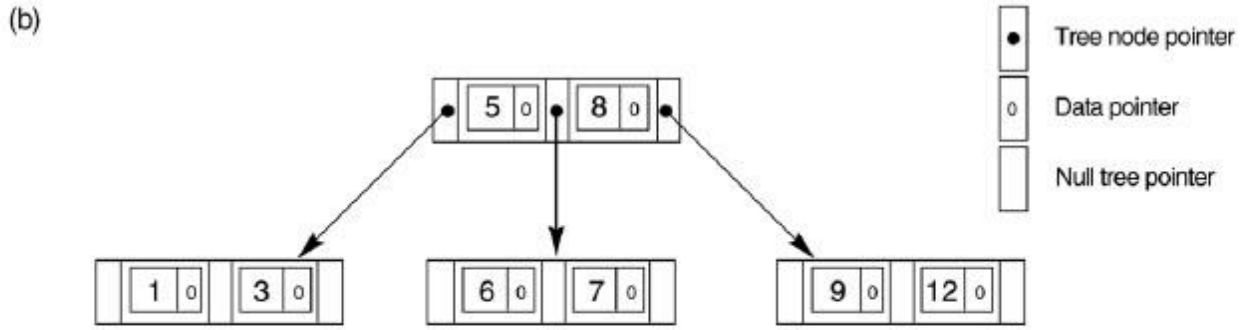
- An insertion into a node that is not full is quite efficient; if a node is full, the insertion causes a split into two nodes
- Splitting may propagate to other tree levels
- A deletion is quite efficient if a node does not become less than half full
- If a deletion causes a node to become less than half full, it must be merged with neighboring nodes

Difference between B-tree and B+-tree

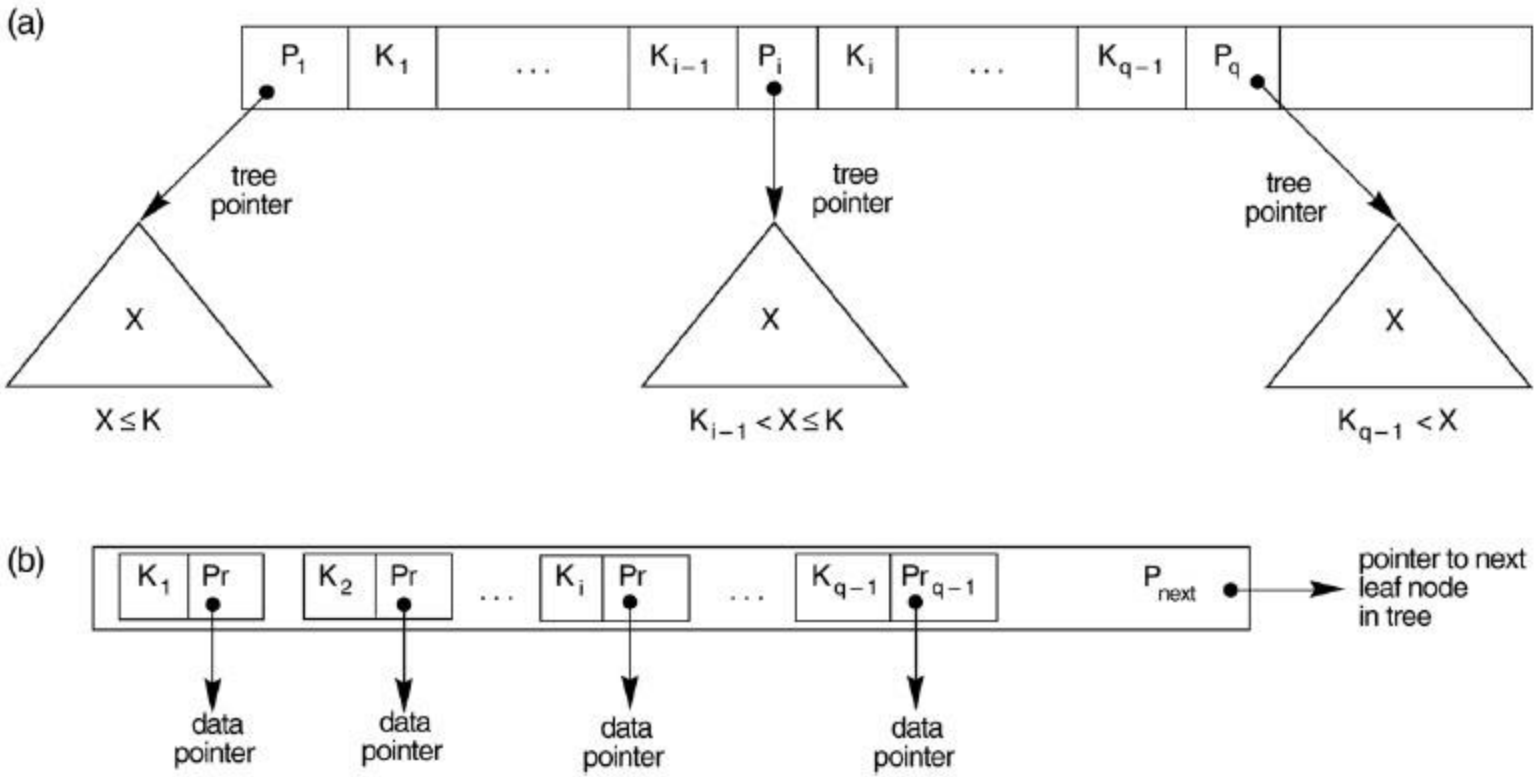
- In a B-tree, pointers to data records exist at all levels of the tree
- In a B+-tree, all pointers to data records exist at the leaf-level nodes
- A B+-tree can have less levels (or higher capacity of search values) than the corresponding B-tree

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.





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.



EXAMPLE 4: Suppose the search field is V = 9 bytes long, the disk block size is B = 512 bytes, a record (data) pointer is P_r = 7 bytes, and a block pointer is P = 6 bytes. Each B-tree node can have *at most* p tree pointers, p − 1 data pointers, and p − 1 search key field values (see Figure 14.10a). These must fit into a single disk block if each B-tree node is to correspond to a disk block. Hence, we must have:

$$(p * P) + ((p - 1) * (P_r + V)) \leq B$$

$$(p * 6) + ((p - 1) * (7 + 9)) \leq 512$$

$$(22 * p) \leq 528$$

We can choose p to be a large value that satisfies the above inequality, which gives p = 23 (p = 24 is not chosen because of the reasons given next).

EXAMPLE 5: Suppose that the search field of Example 4 is a nonordering key field, and we construct a B-tree on this field. Assume that each node of the B-tree is 69 percent full. Each node, on the average, will have p * 0.69 = 23 * 0.69 or approximately 16 pointers and, hence, 15 search key field values. The **average fan-out** fo =16. We can start at the root and see how many values and pointers can exist, on the average, at each subsequent level:

Root:	1 node	15 entries	16 pointers
Level 1:	16 nodes	240 entries	256 pointers
Level 2:	256 nodes	3840 entries	4096 pointers
Level 3:	4096 nodes	61,440 entries	

At each level, we calculated the number of entries by multiplying the total number of pointers at the previous level by 15, the average number of entries in each node. Hence, for the given block size, pointer size, and search key field size, a two-level B-tree holds 3840 + 240 + 15 = 4095 entries on the average; a three-level B-tree holds 65,535 entries on the average.

23*0.69 = 15.87 ≈ 16

EXAMPLE 6: To calculate the order p of a B^+-tree, suppose that the search key field is V = 9 bytes long, the block size is B = 512 bytes, a record pointer is P_r = 7 bytes, and a block pointer is P = 6 bytes, as in Example 4. An internal node of the B^+-tree can have up to p tree pointers and p − 1 search field values; these must fit into a single block. Hence, we have:

$$(p * P) + ((p - 1) * V) \leq B$$

$$(p * 6) + ((p - 1) * 9) \leq 512$$

$$(15 * p) \leq 521$$

We can choose p to be the largest value satisfying the above inequality, which gives p = 34. This is larger than the value of 23 for the B-tree, resulting in a larger fan-out and more entries in each internal node of a B^+-tree than in the corresponding B-tree. The leaf nodes of the B^+-tree will have the same number of values and pointers, except that the pointers are data pointers and a next pointer. Hence, the order p_{leaf} for the leaf nodes can be calculated as follows:

$$(p_{leaf} * (P_r + V)) + P \leq B$$

$$(p_{leaf} * (7 + 9)) + 6 \leq 512$$

$$(16 * p_{leaf}) \leq 506$$

It follows that each leaf node can hold up to p_{leaf} = 31 key value/data pointer combinations, assuming that the data pointers are record pointers.

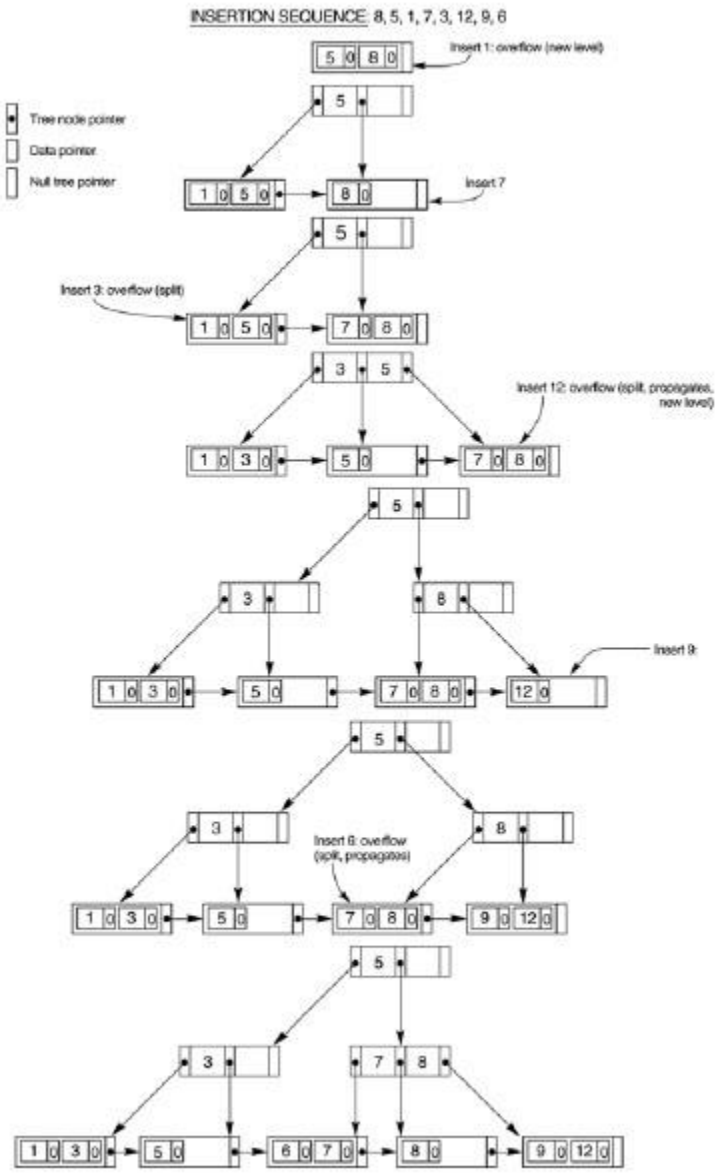
EXAMPLE 7: Suppose that we construct a B^+-tree on the field of Example 6. To calculate the approximate number of entries of the B^+-tree, we assume that each node is 69 percent full. On the average, each internal node will have 34 * 0.69 or approximately 23 pointers, and hence 22 values. Each leaf node, on the average, will hold 0.69 * p_{leaf} = 0.69 * 31 or approximately 21 data record pointers. A B^+-tree will have the following average number of entries at each level:

Root:	1 node	22 entries	23 pointers
Level 1:	23 nodes	506 entries	529 pointers
Level 2:	529 nodes	11,638 entries	12,167 pointers
Leaf level:	12,167 nodes	255,507 record pointers	

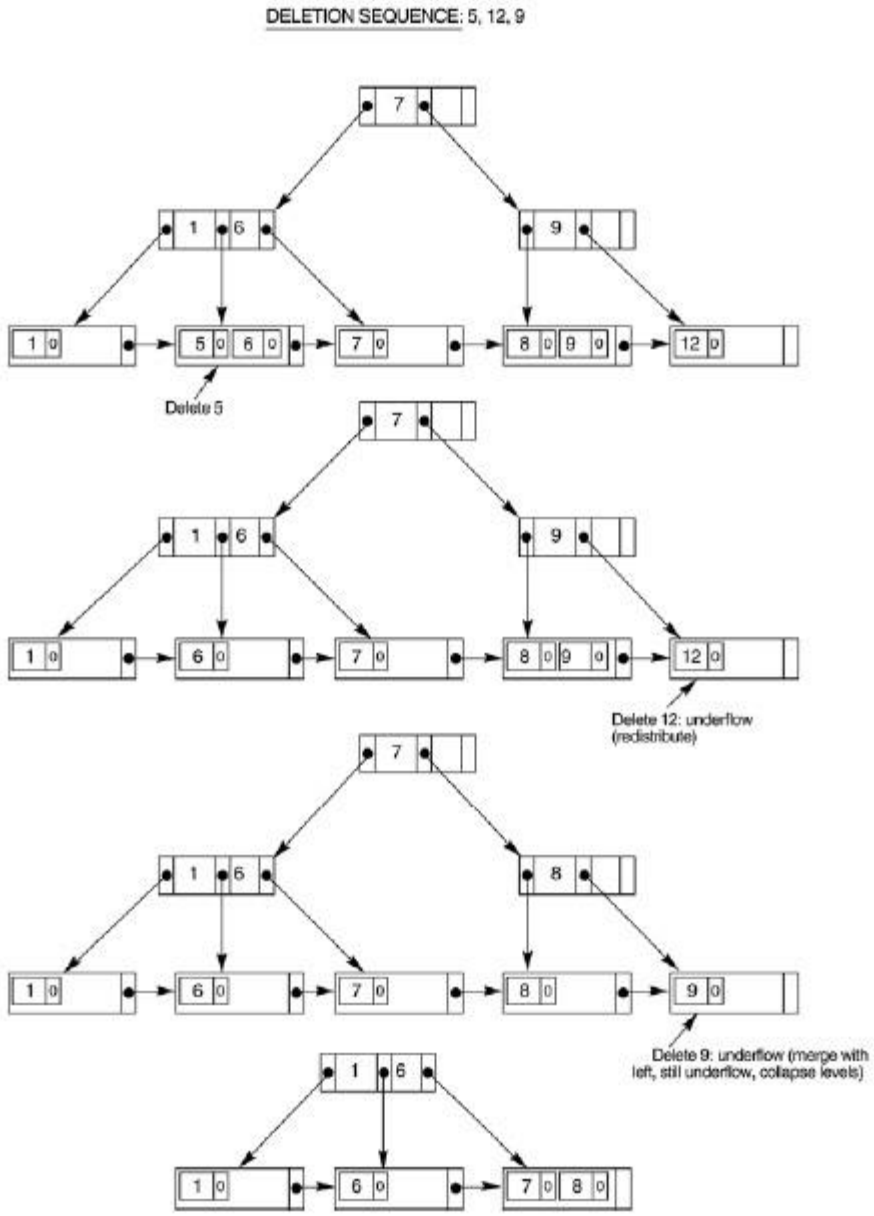
For the block size, pointer size, and search field size given above, a three-level B^+-tree holds up to 255,507 record pointers, on the average. Compare this to the 65,535 entries for the corresponding B-tree in Example 5.

34*0.69 = 23.46 ≈ 23

31*0.69 = 21.39 ≈ 21



An example of insertion in a B+-tree with q = 3 and p_{leaf} = 2.



An example of deletion from a B+-tree.