Lecture 19: Data Storage and Indexes

Monday, May 14, 2007

Final Exam

- Will reschedule to Tuesday, June 5th, in the morning (exact time TBD)
- If you have serious conflicts, send me email

Outline

- Representing data elements (12)
- Index structures (13.1, 13.2)
- B-trees (13.3)

Files and Tables

- A disk = a sequence of blocks
- A file = a subsequence of blocks, usually contiguous
- Need to store tables/records/indexes in files/block

Representing Data Elements

- Relational database elements:

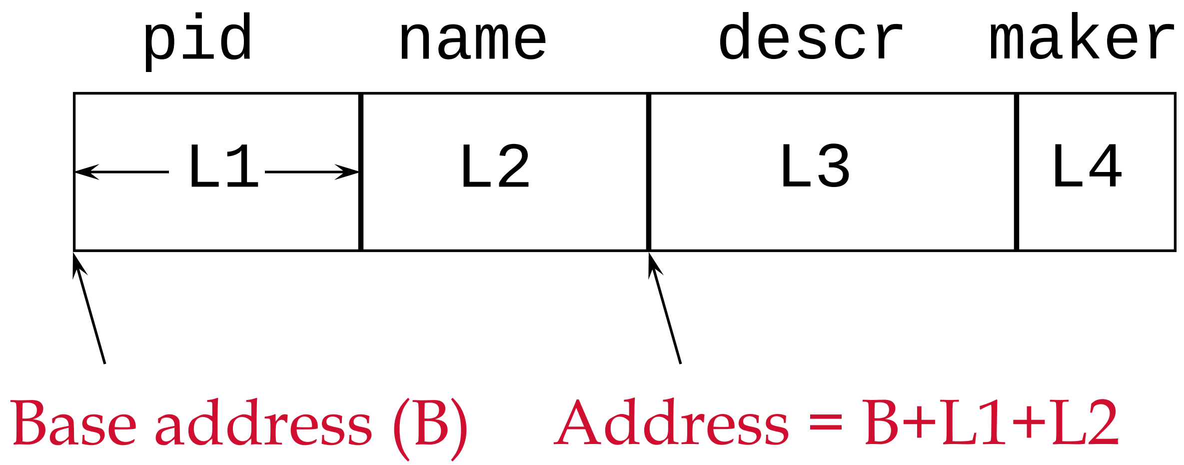
```
CREATE TABLE Product (

    pid INT PRIMARY KEY,
    name CHAR(20),
    description VARCHAR(200),
    maker CHAR(10) REFERENCES Company(name)
)
```

- A tuple is represented as a record
- The table is a sequence of records

Issues

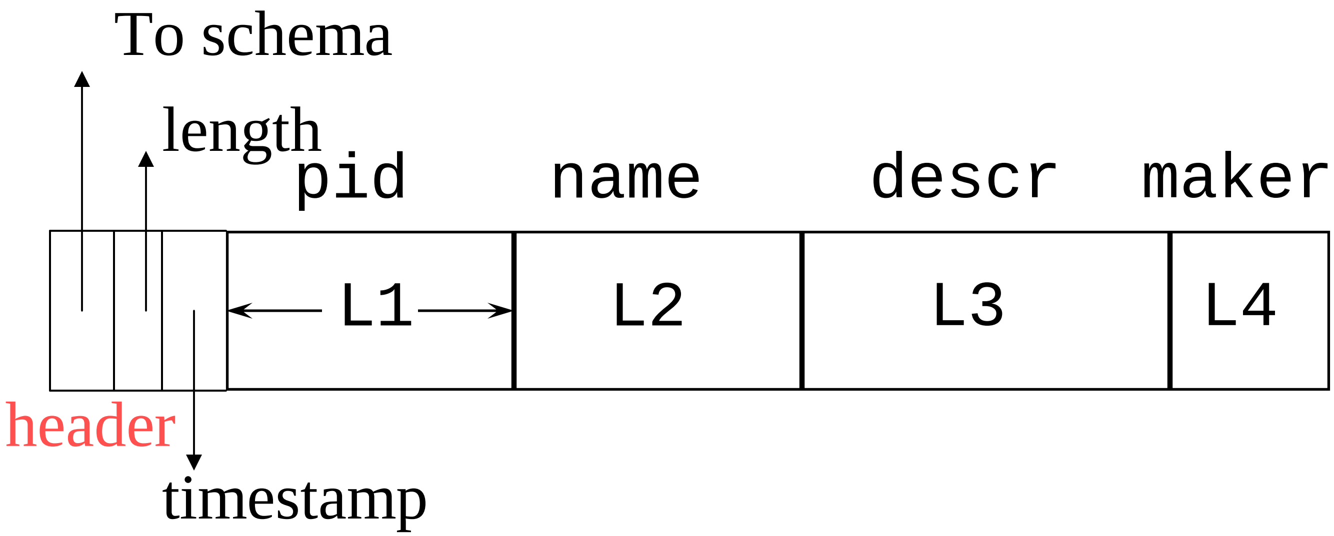
- Represent attributes inside the records
- Represent the records inside the blocs

Record Formats: Fixed Length



- Information about field types same for all records in a file; stored in *system catalogs.*
- Finding *i'th* field requires scan of record.
- **Note the importance of schema information!**

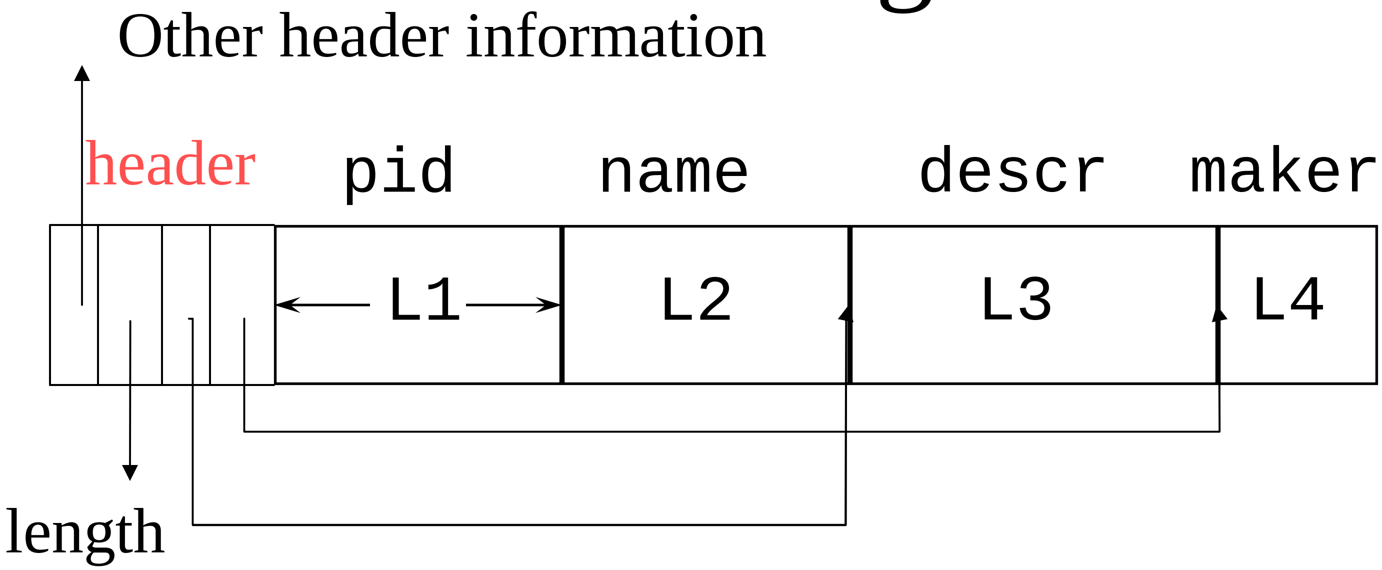
Record Header



Need the header because:

- The schema may change
  - for a while new+old may coexist
- Records from different relations may coexist

Variable Length Records



Place the fixed fields first: F1

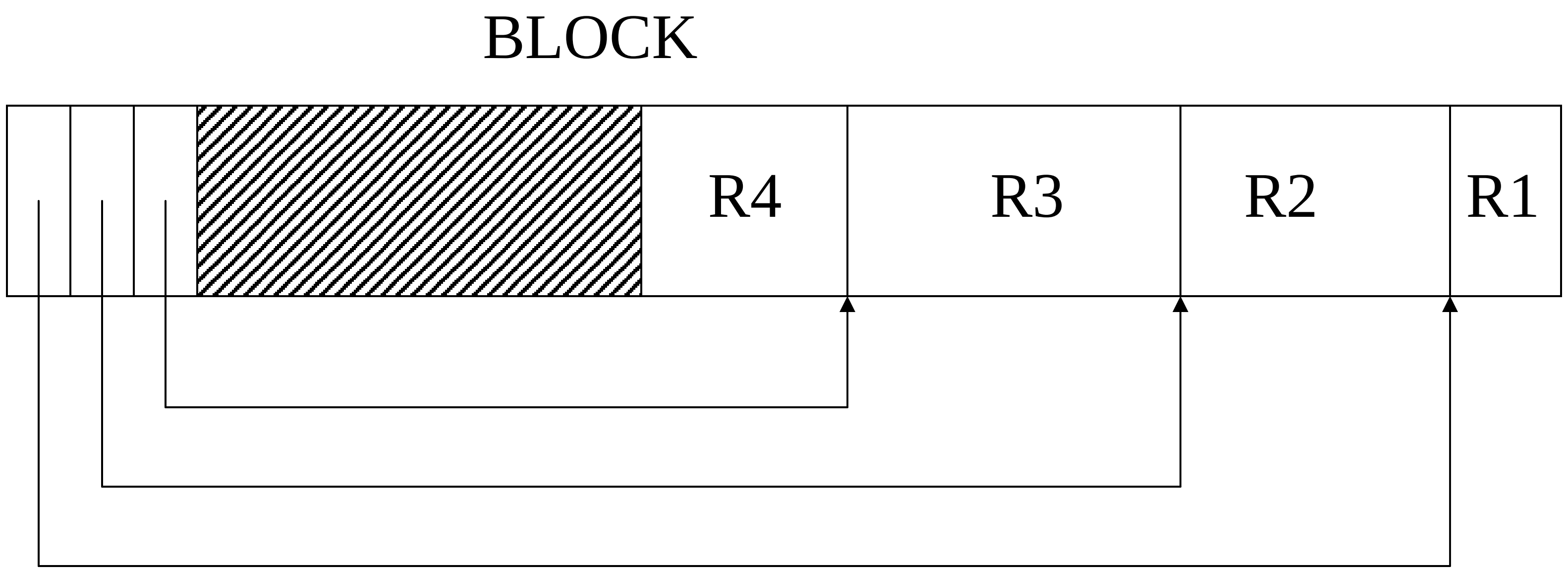
Then the variable length fields: F2, F3, F4

Null values take 2 bytes only

Sometimes they take 0 bytes (when at the end)

Storing Records in Blocks

- Blocks have fixed size (typically 4k – 8k)



BLOB

- Binary large objects
- Supported by modern database systems
- E.g. images, sounds, etc.
- Storage: attempt to cluster blocks together

CLOB = character large objec

- Supports only restricted operations

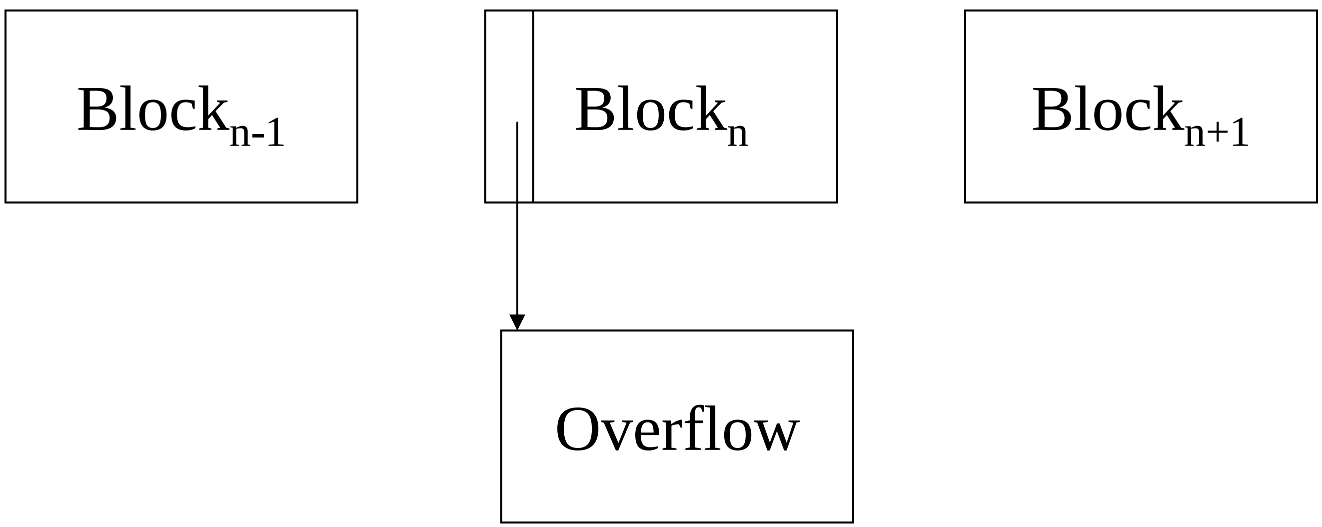
File Types

- Unsorted (heap)
- Sorted (e.g. by pid)

Modifications: Insertion

- File is unsorted: add it to the end (easy ☺)
- File is sorted:
  - Is there space in the right block ?
    - Yes: we are lucky, store it there
  - Is there space in a neighboring block ?
    - Look 1-2 blocks to the left/right, shift records
  - If anything else fails, create *overflow block*

Overflow Blocks



- After a while the file starts being dominated by overflow blocks: time to reorganize

Modifications: Deletions

- Free space in block, shift records
- May be able to eliminate an overflow block
- Can never really eliminate the record, because others may *point* to it
  - Place a tombstone instead (a NULL record)

How can we *point* to a record in an RDBMS ?

Modifications: Updates

- If new record is shorter than previous, easy ☺
- If it is longer, need to shift records, create overflow blocks

Pointers

Logical pointer to a record consists of:

- Logical block number
- An offset in the block’s header

We use pointers in Indexes and in Log entries

Note: review what a pointer in C is

Indexes

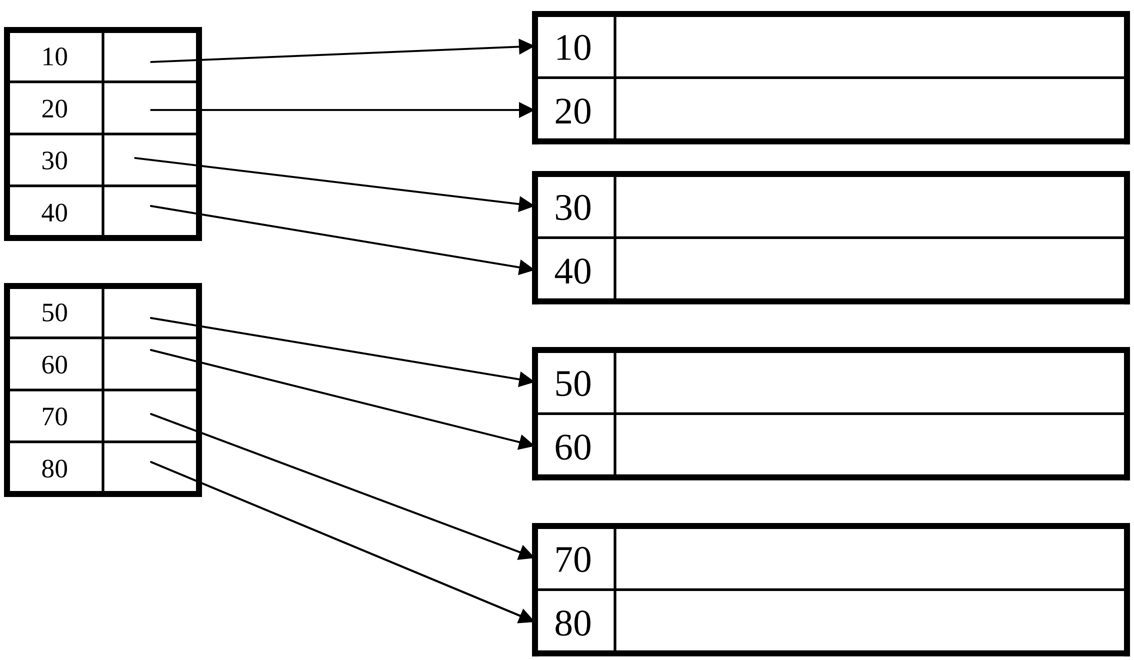
- An *index* on a file speeds up selections on the *search key fields* for the index.
  - Any subset of the fields of a relation can be the search key for an index on the relation.
  - *Search key* is not the same as *key* (minimal set of fields that uniquely identify a record in a relation).
- An index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

Index Classification

- Clustered/unclustered
  - Clustered = records close in the index are close in the data; same as saying that the table is ordered by the index key
  - Unclustered = records close in the index may be far in the data
- Primary/secondary:
  - Interpretation 1:
    - Primary = is over attributes part of the primary
    - Secondary = cannot reorder data
  - Interpretation 2: means the same as clustered/unclustured
- B+ tree or Hash table

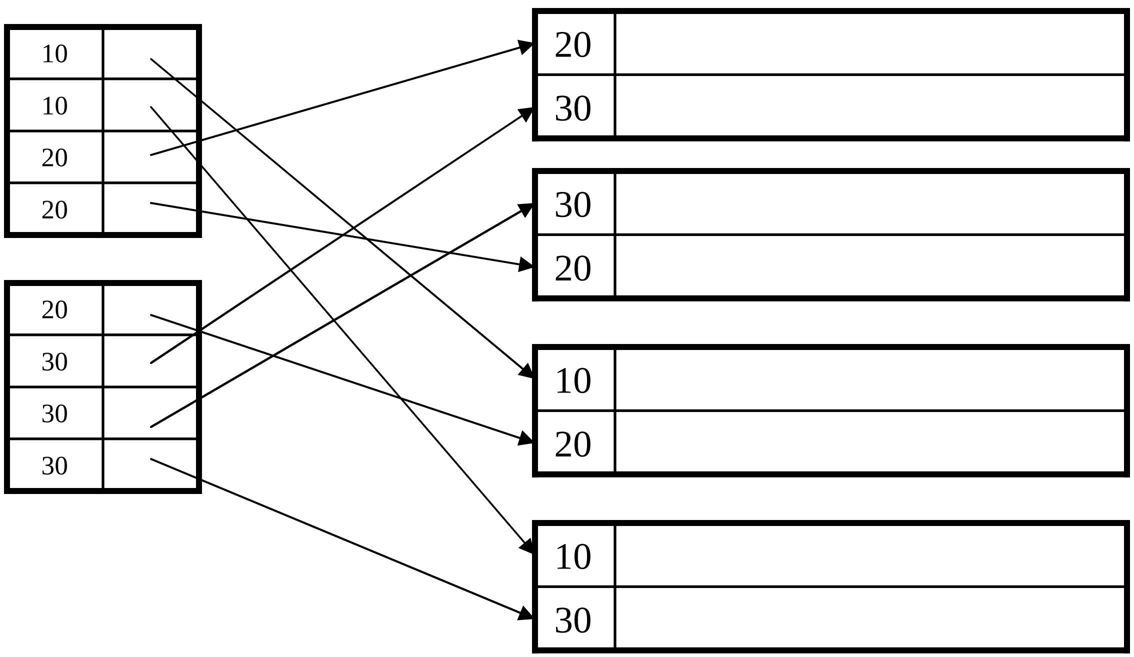
Clustered Index

- File is sorted on the index attribute
- Only one per table

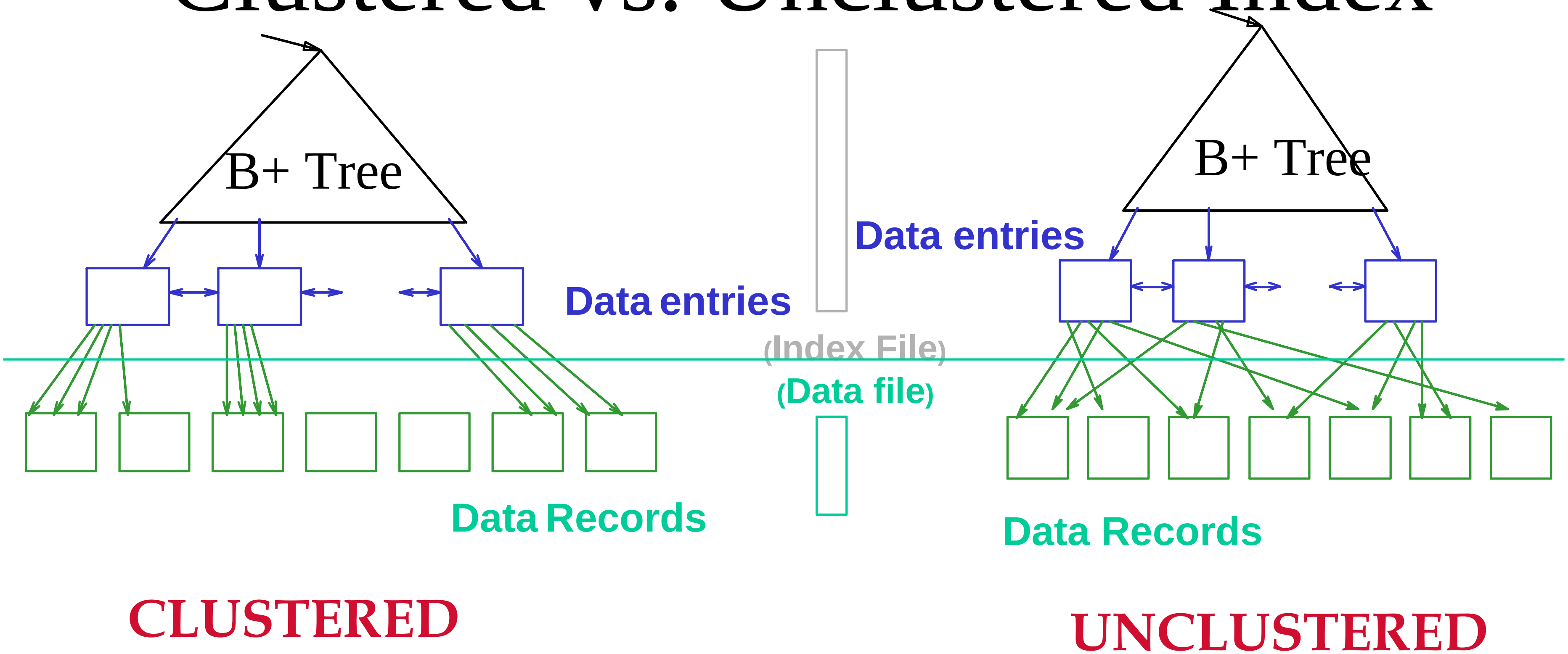


Unclustered Index

- Several per table



Clustered vs. Unclustered Index

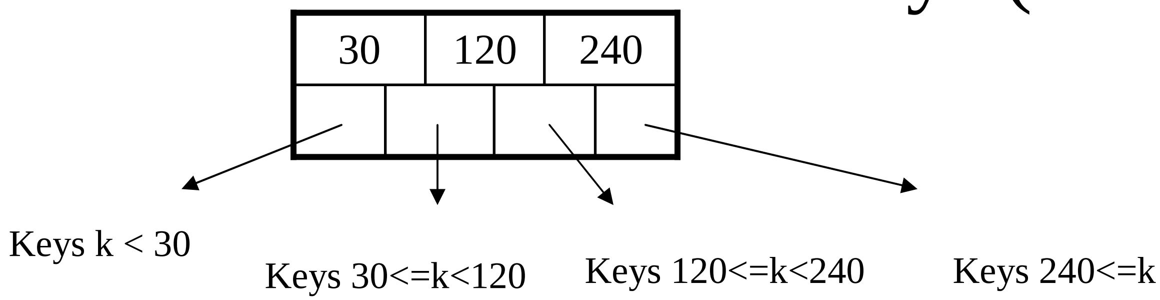


B+ Trees

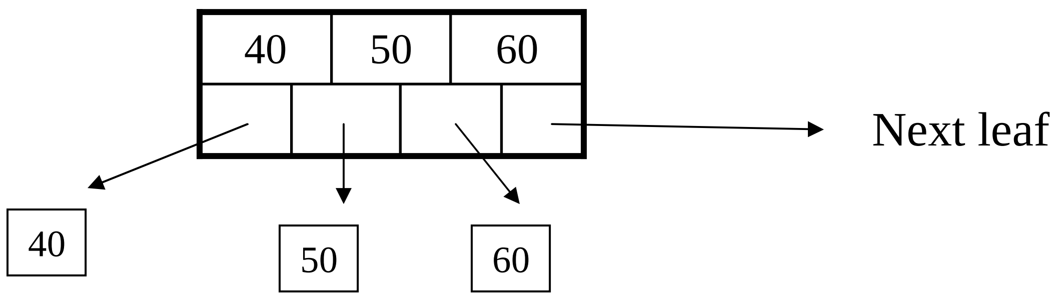
- Search trees
- Idea in B Trees:
  - make 1 node = 1 block
- Idea in B+ Trees:
  - Make leaves into a linked list (range queries are easier)

B+ Trees Basics

- Parameter d = the *degree*
- Each node has >= d and <= 2d keys (except root)

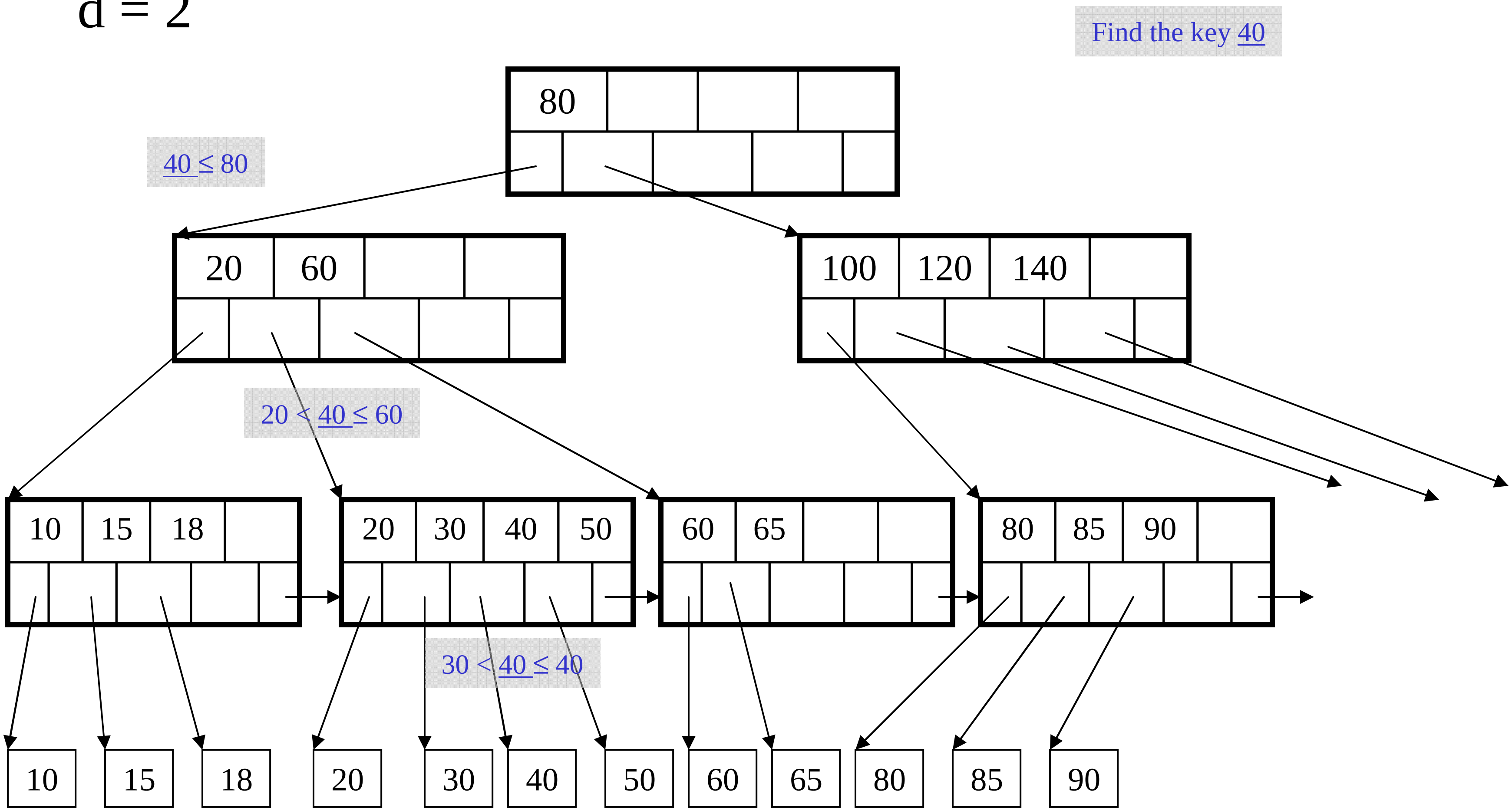


- Each leaf has >=d and <= 2d keys:



B+ Tree Example

d = 2



B+ Tree Design

- How large d ?
- Example:
  - Key size = 4 bytes
  - Pointer size = 8 bytes
  - Block size = 4096 byes
- 2d x 4 + (2d+1) x 8 <= 4096
- d = 170

Searching a B+ Tree

- Exact key values:
  - Start at the root
  - Proceed down, to the leaf

```
Select name
From people
Where age = 25
```

- Range queries:
  - As above
  - Then sequential traversal

```
Select name
From people
Where 20 <= age
  and  age <= 30
```

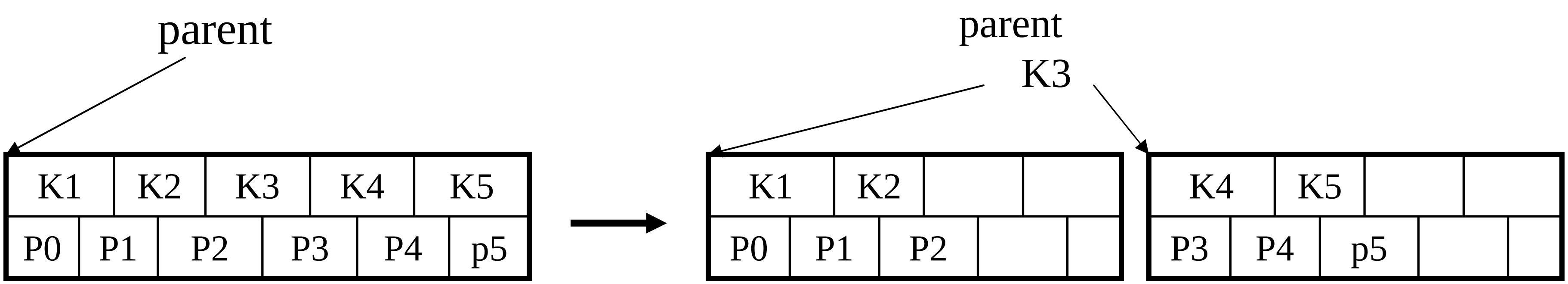
B+ Trees in Practice

- Typical order: 100. Typical fill-factor: 67%.
  - average fanout = 133
- Typical capacities:
  - Height 4: 133^4 = 312,900,700 records
  - Height 3: 133^3 = 2,352,637 records
- 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

Insertion in a B+ Tree

Insert (K, P)

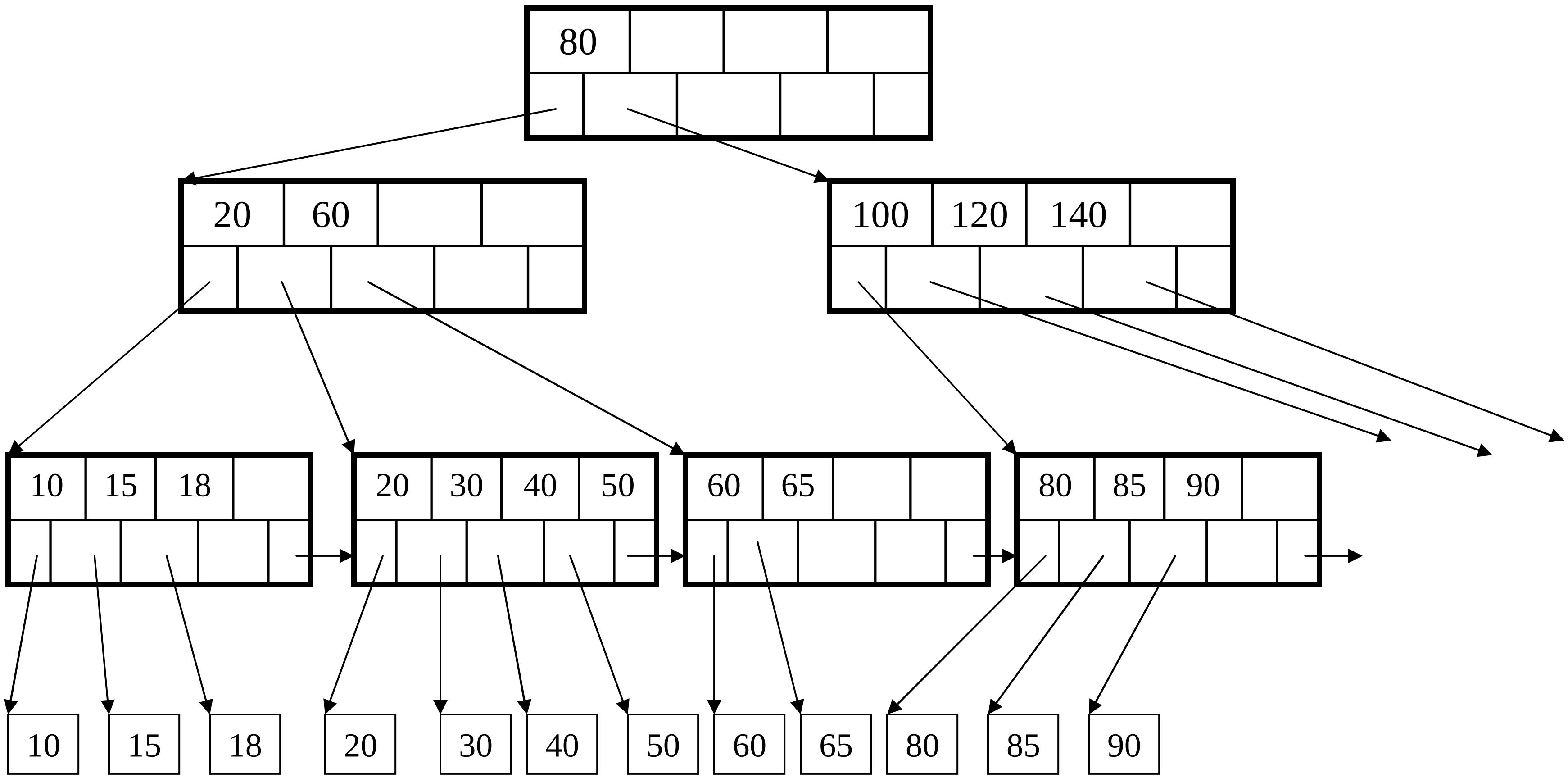
- Find leaf where K belongs, insert
- If no overflow (2d keys or less), halt
- If overflow (2d+1 keys), split node, insert in parent:



- If leaf, keep K3 too in right node
- When root splits, new root has 1 key only

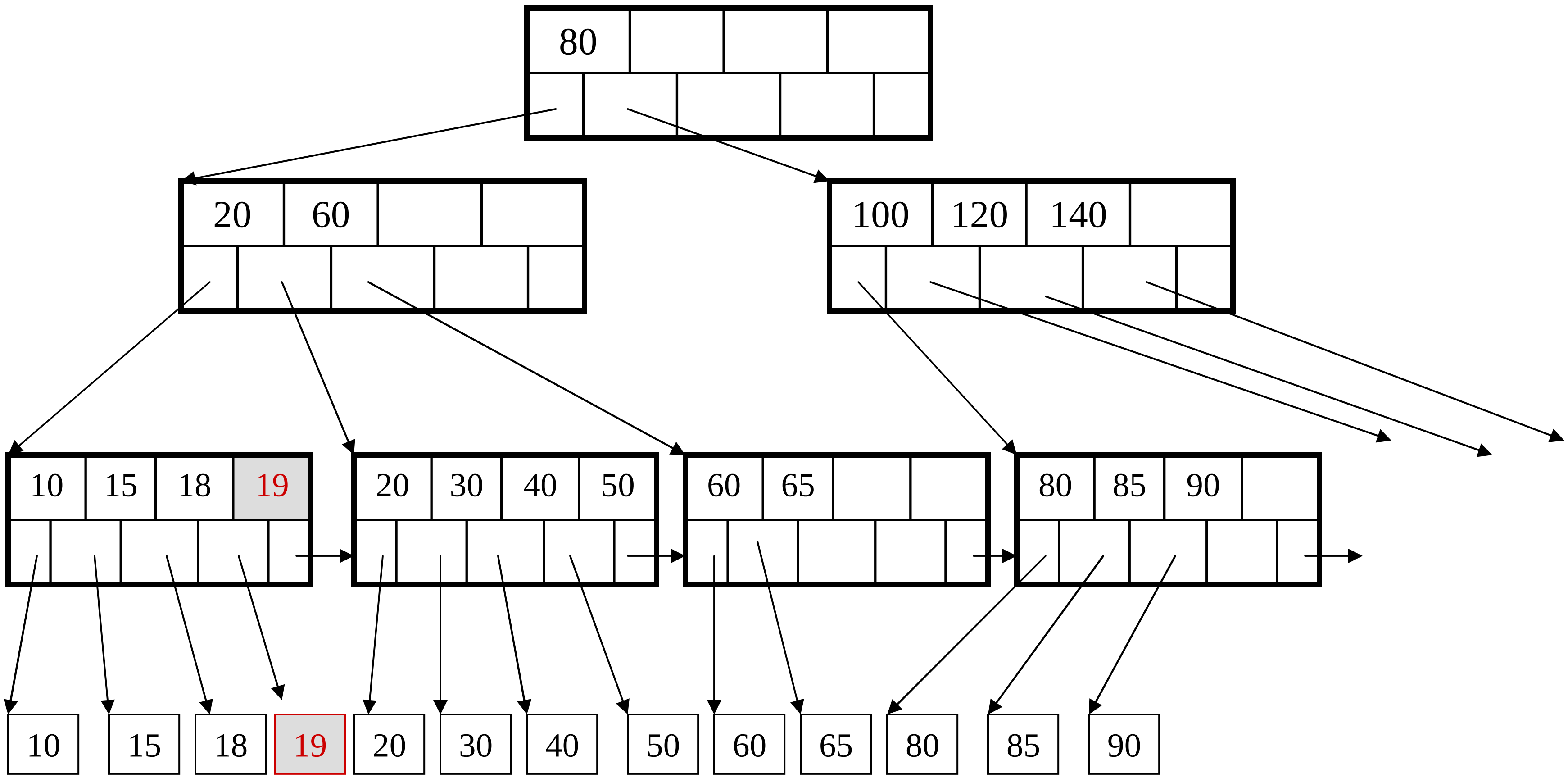
Insertion in a B+ Tree

Insert K=19



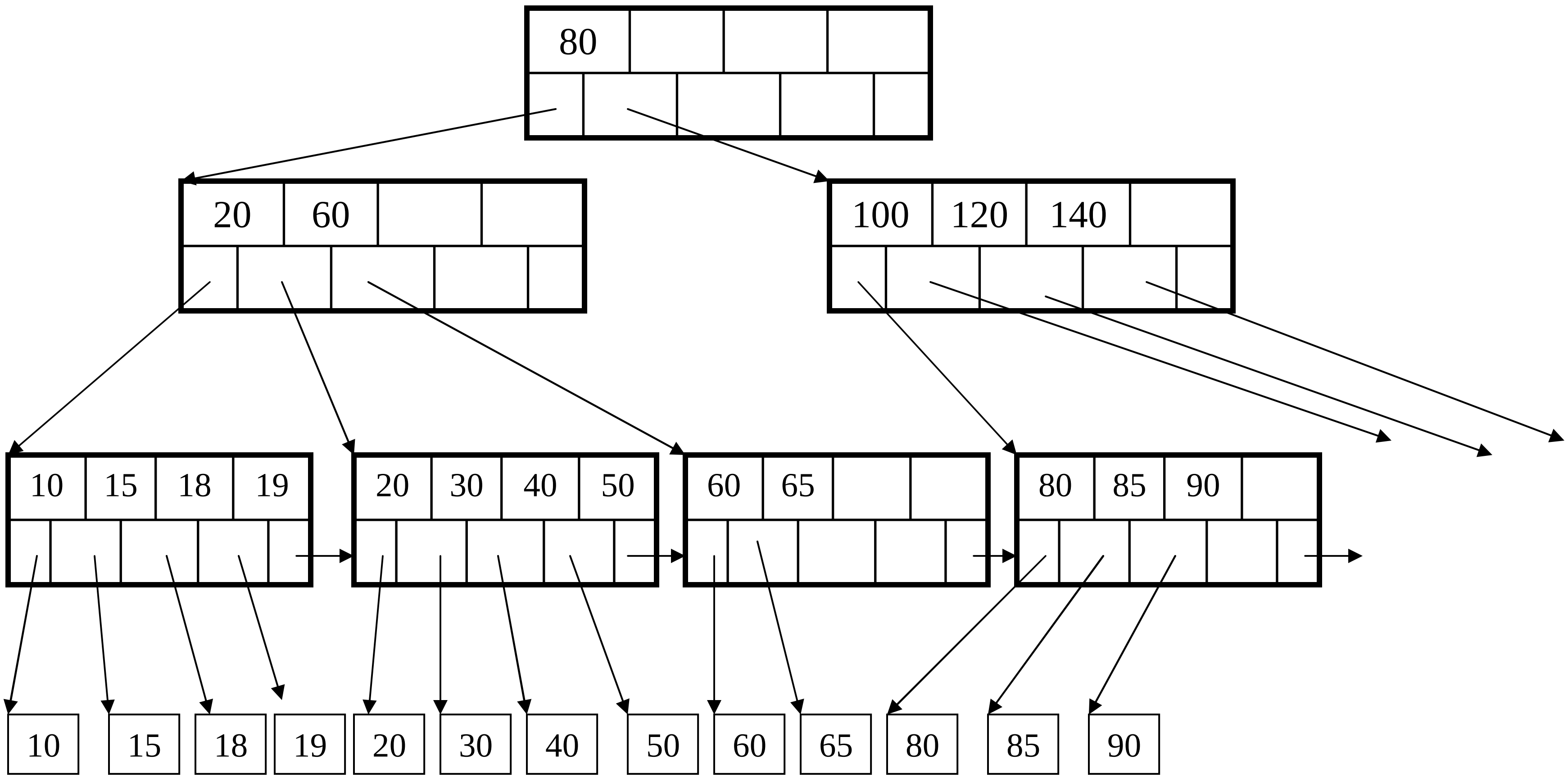
Insertion in a B+ Tree

After insertion



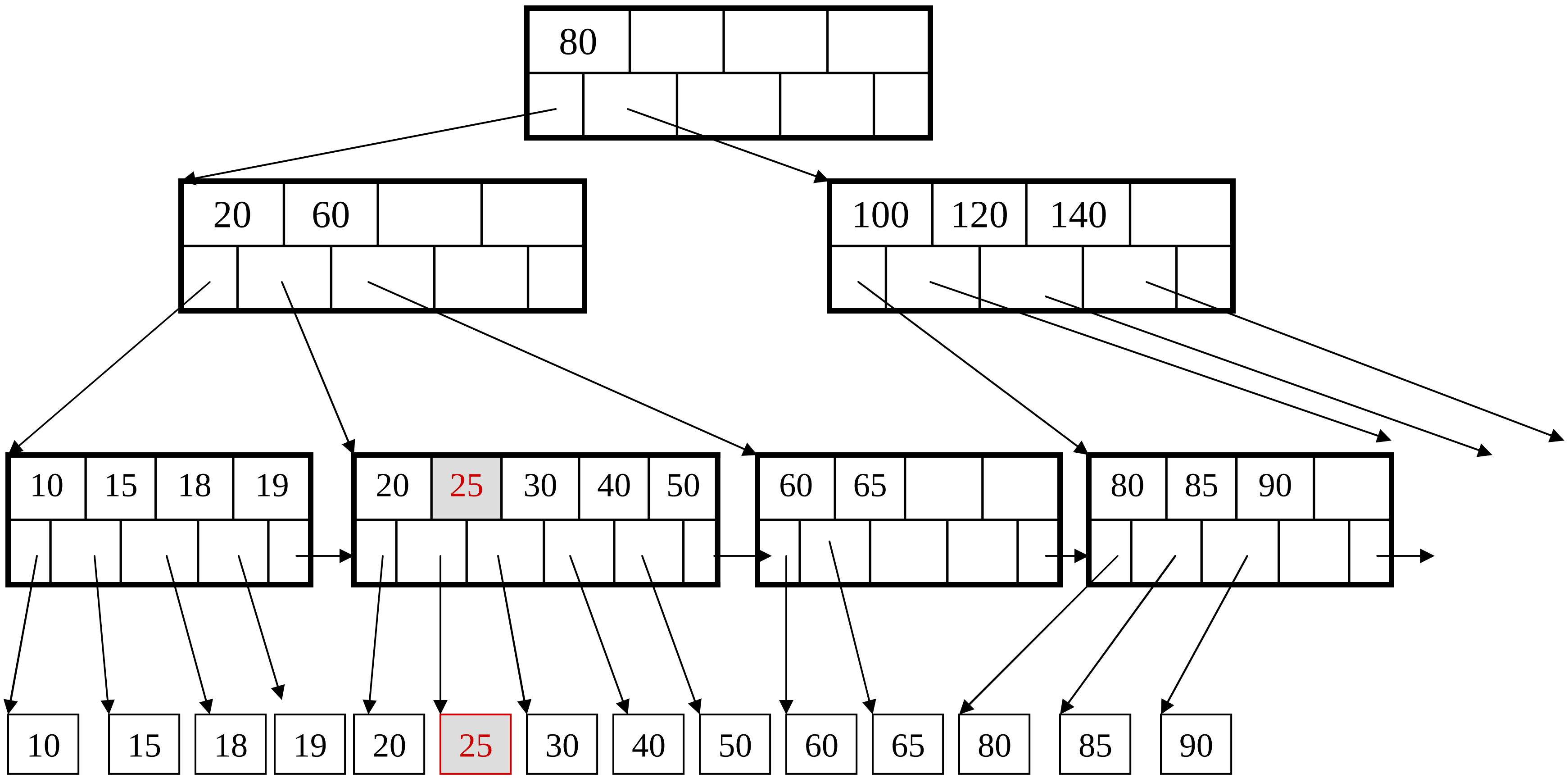
Insertion in a B+ Tree

Now insert 25



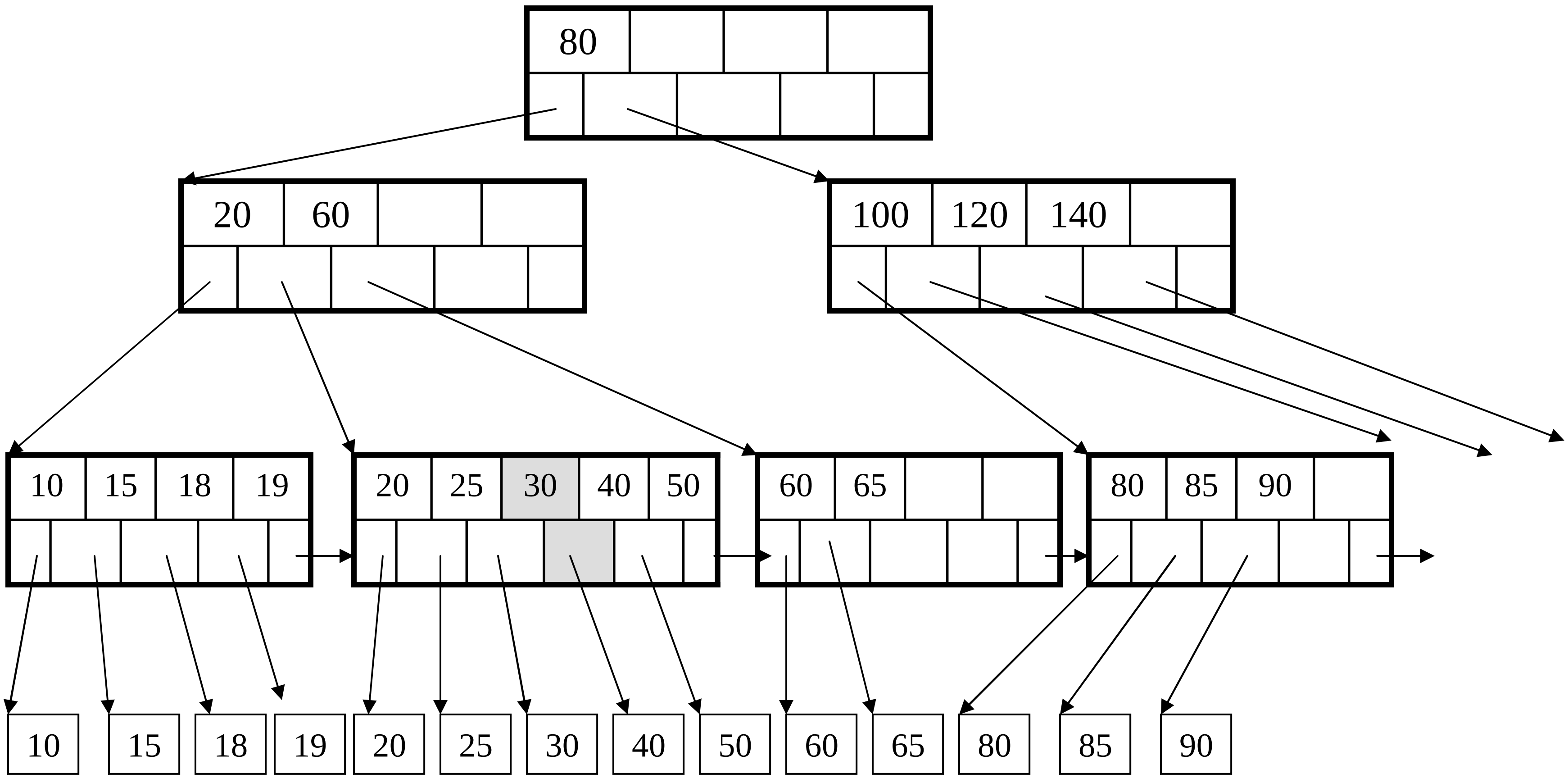
Insertion in a B+ Tree

After insertion



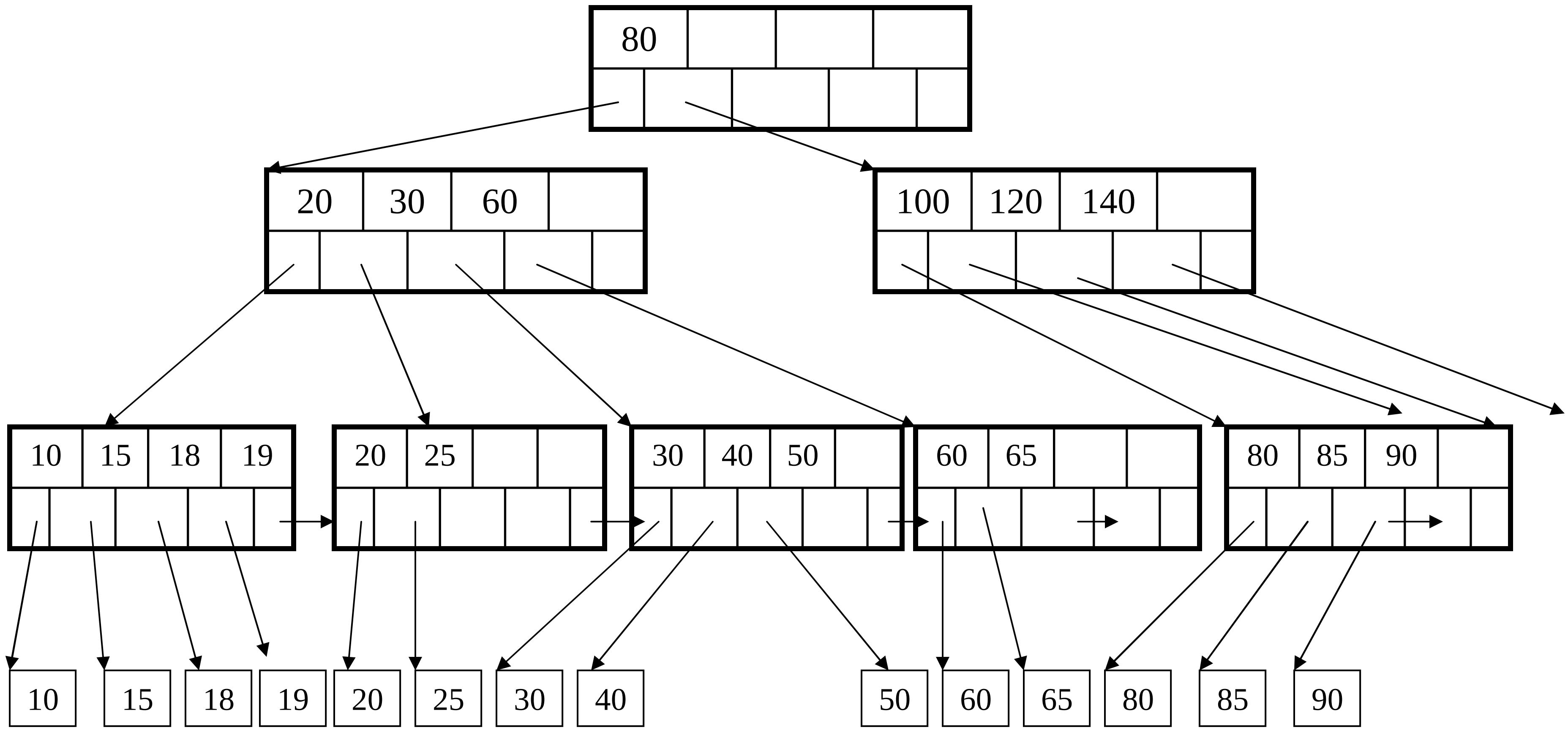
Insertion in a B+ Tree

But now have to split !



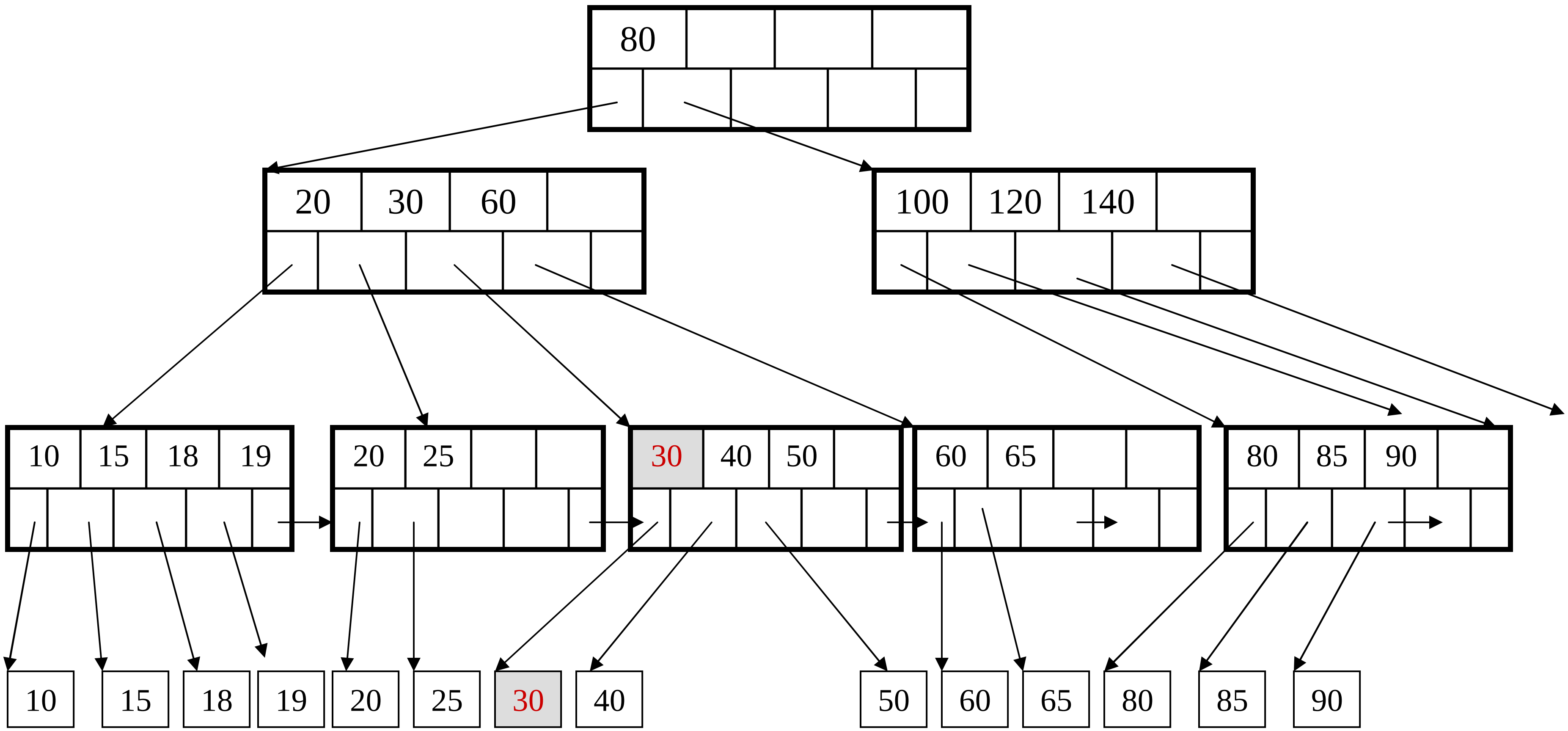
Insertion in a B+ Tree

After the split



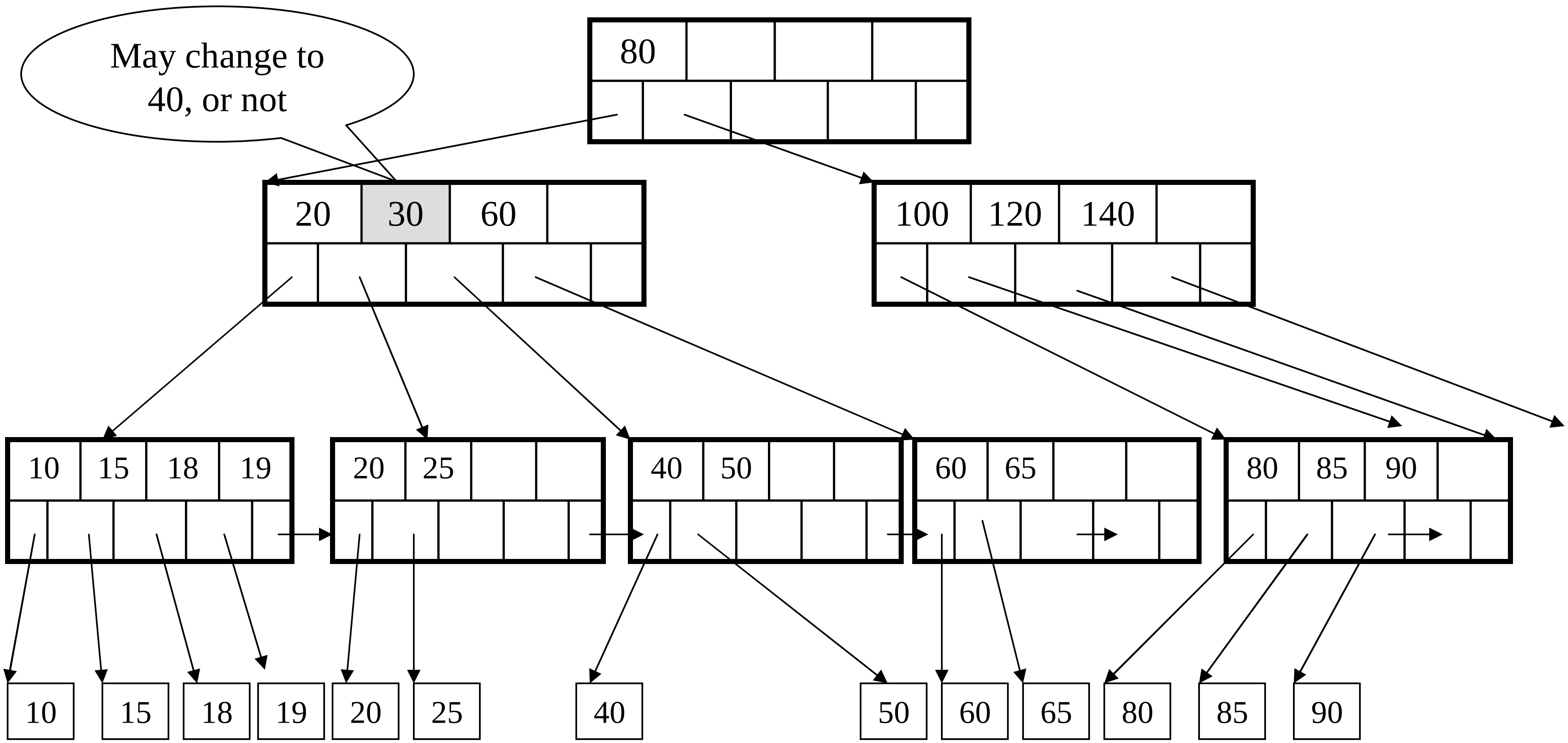
Deletion from a B+ Tree

Delete 30



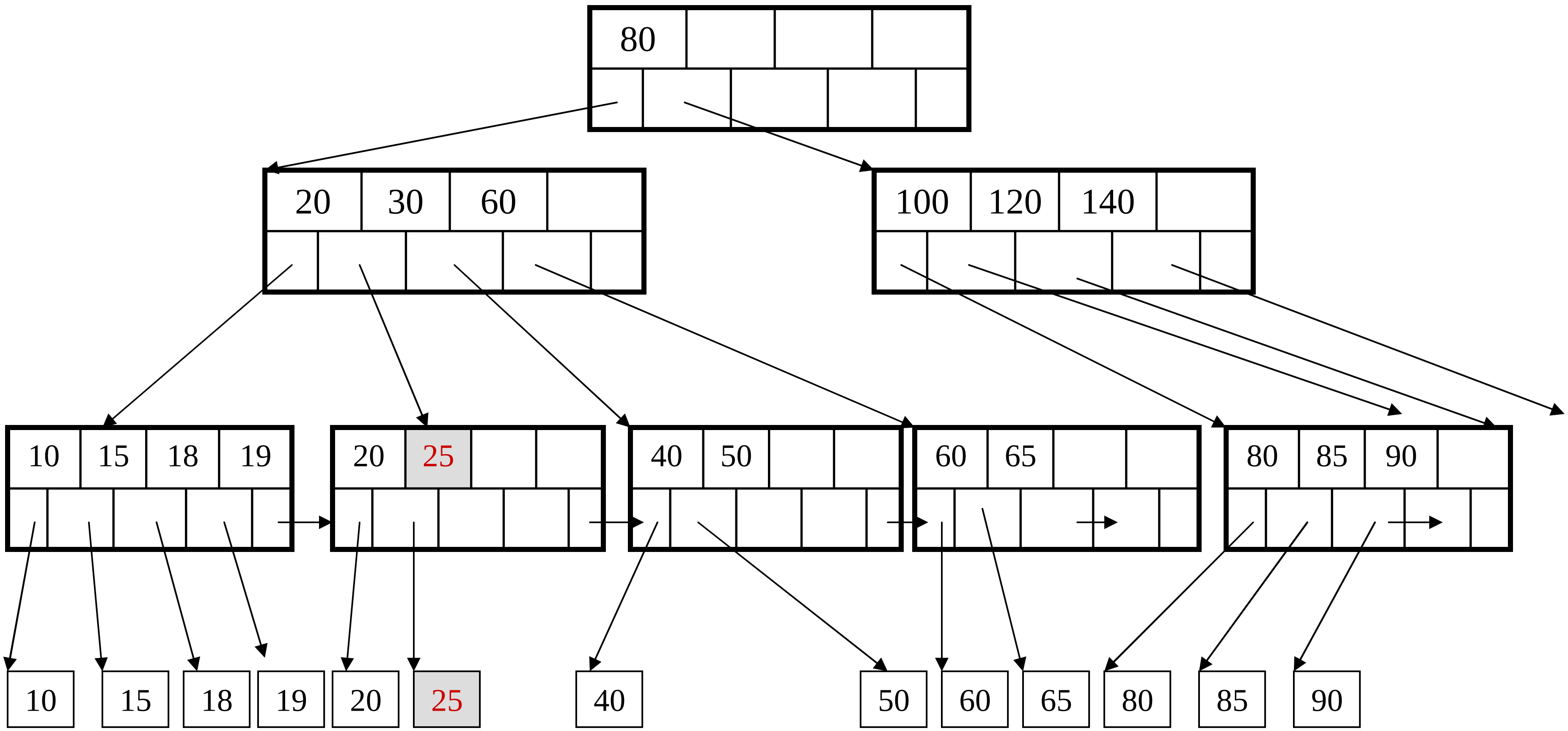
Deletion from a B+ Tree

After deleting 30



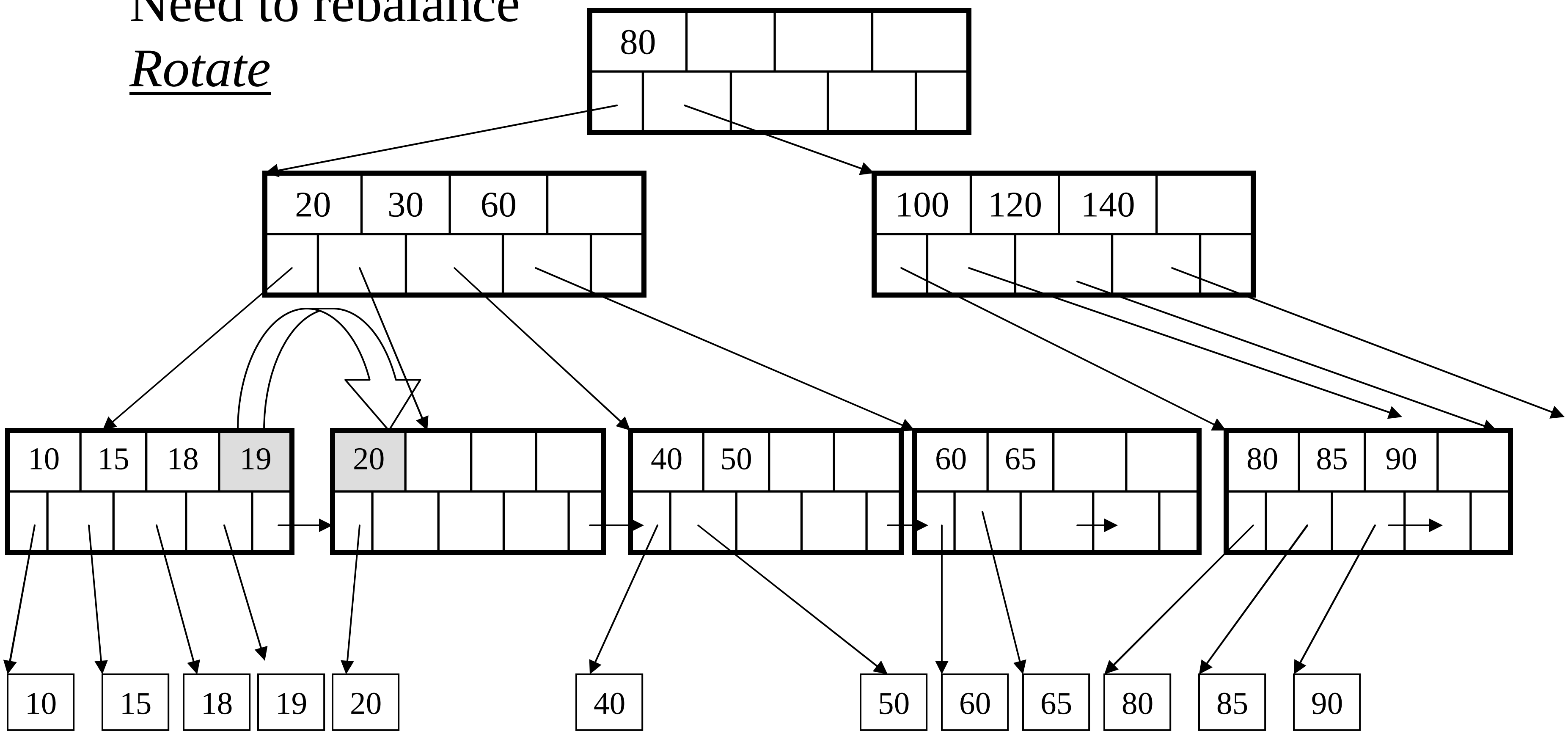
Deletion from a B+ Tree

Now delete 25



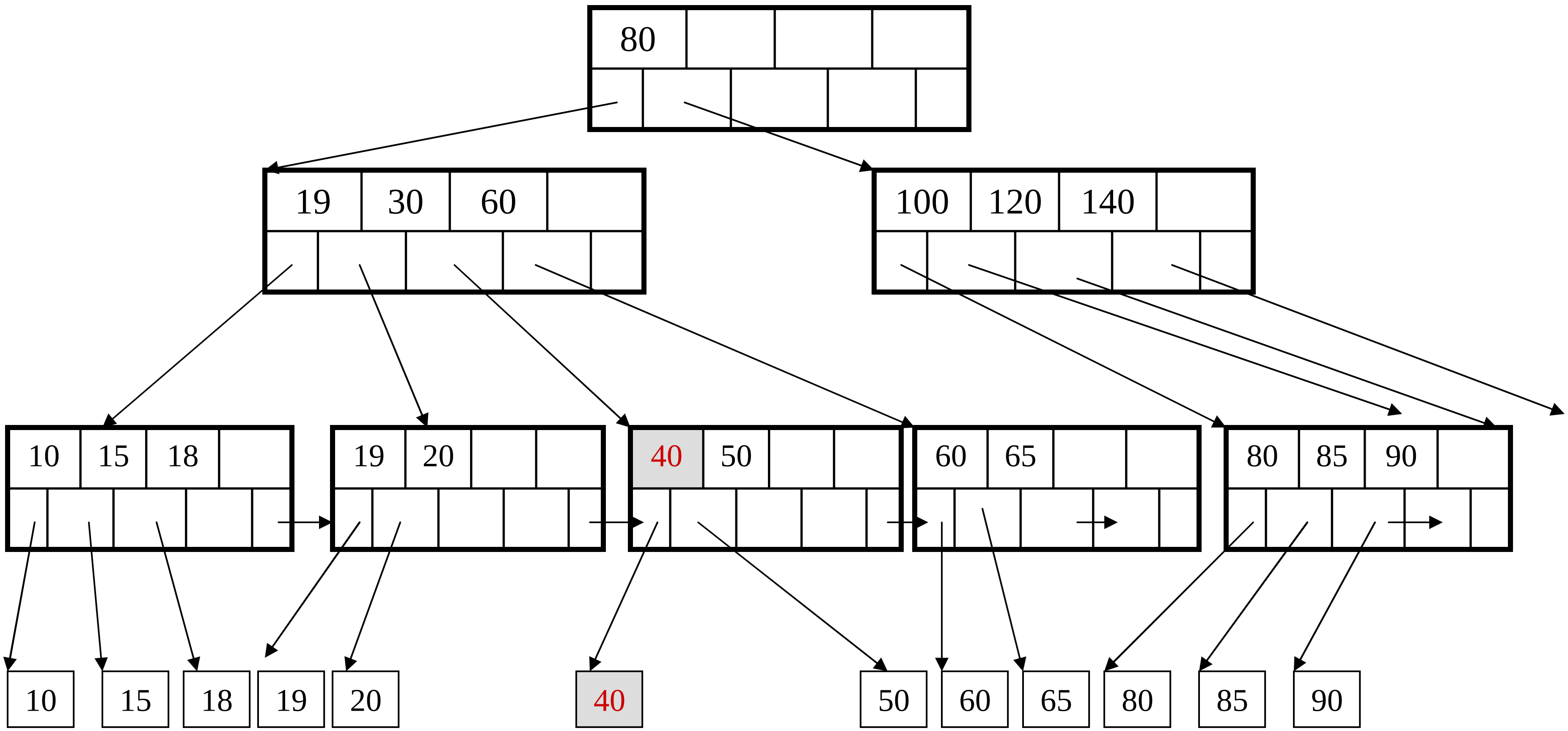
Deletion from a B+ Tree

After deleting 25
Need to rebalance
Rotate



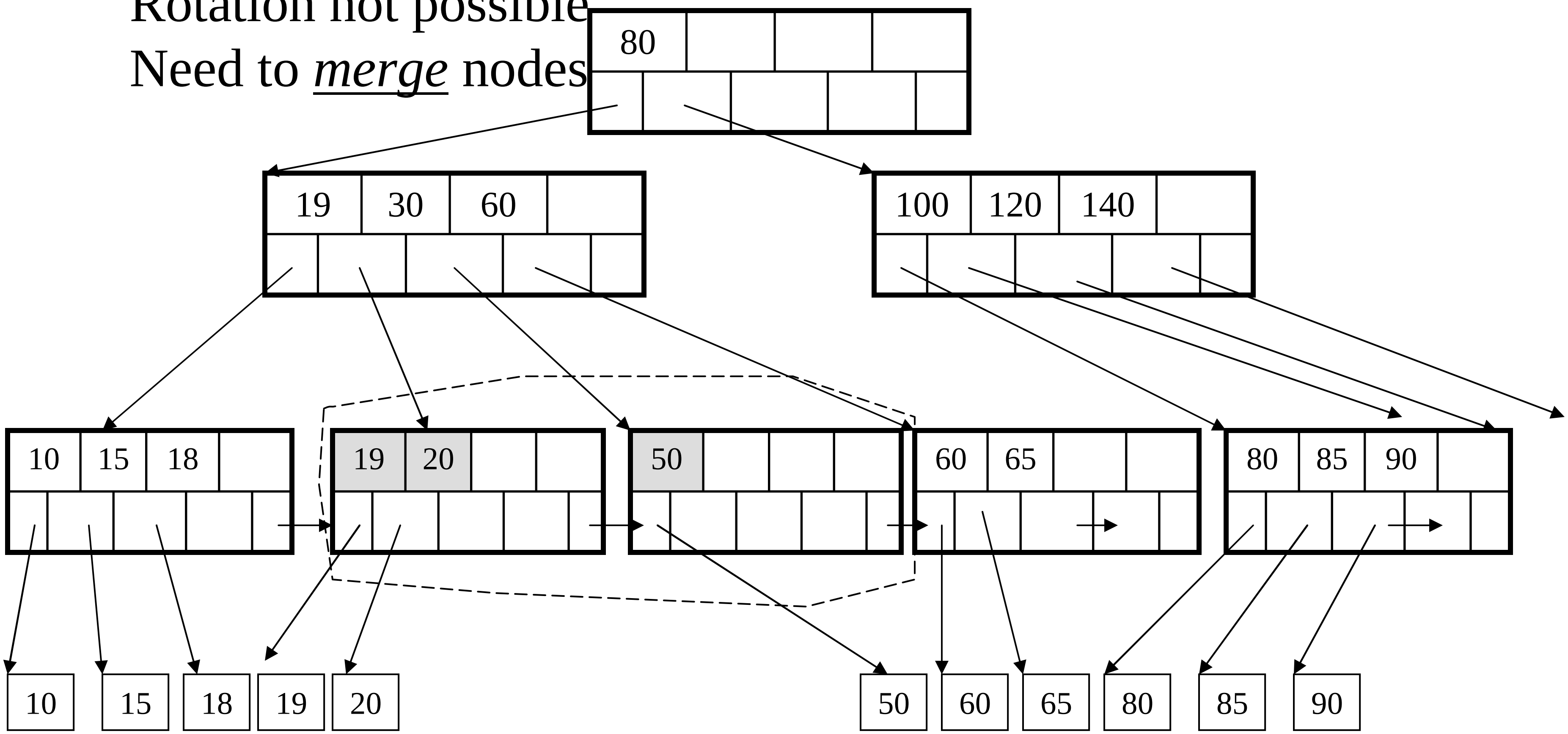
Deletion from a B+ Tree

Now delete 40



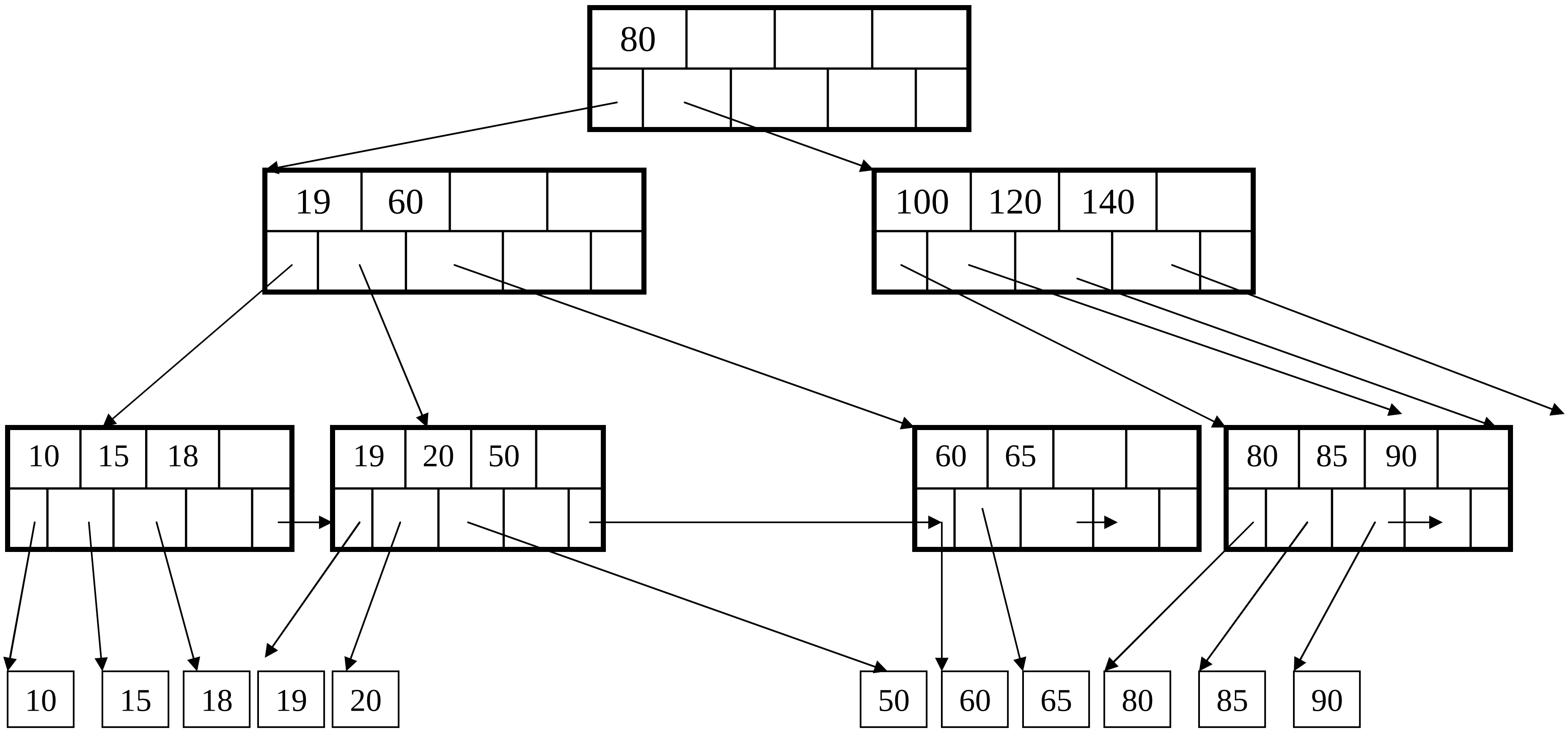
Deletion from a B+ Tree

After deleting 40
Rotation not possible
Need to *merge* nodes



Deletion from a B+ Tree

Final tree



Summary on B+ Trees

- Default index structure on most DBMS
- Very effective at answering ‘point’ queries:
  productName = ‘gizmo’
- Effective for range queries:
  50 < price AND price < 100
- Less effective for multirange:
  50 < price < 100 AND 2 < quant < 20