Database Applications (15-415)

DBMS Internals- Part IV
Lecture 12, February 24, 2014

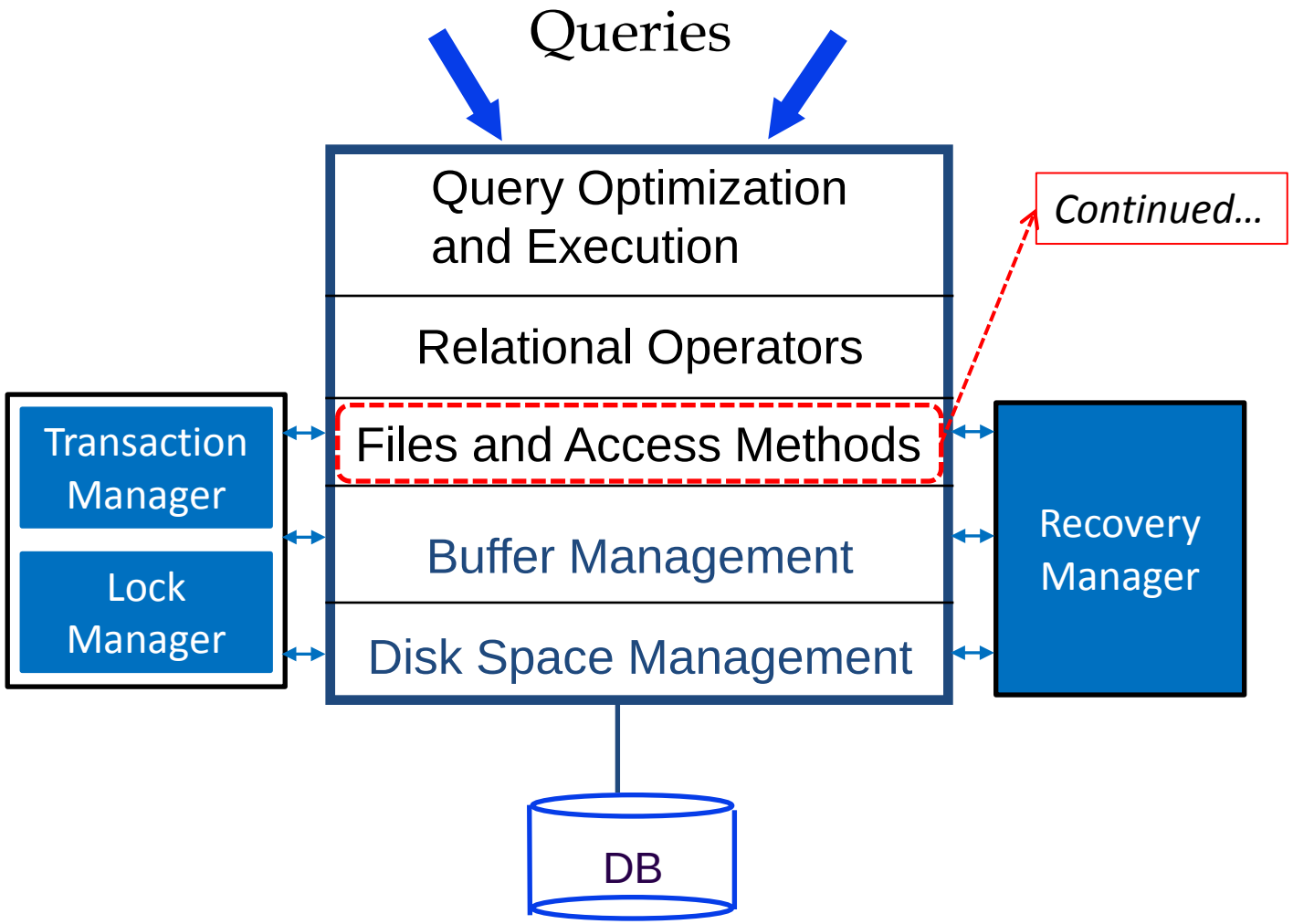
Mohammad Hammoud

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

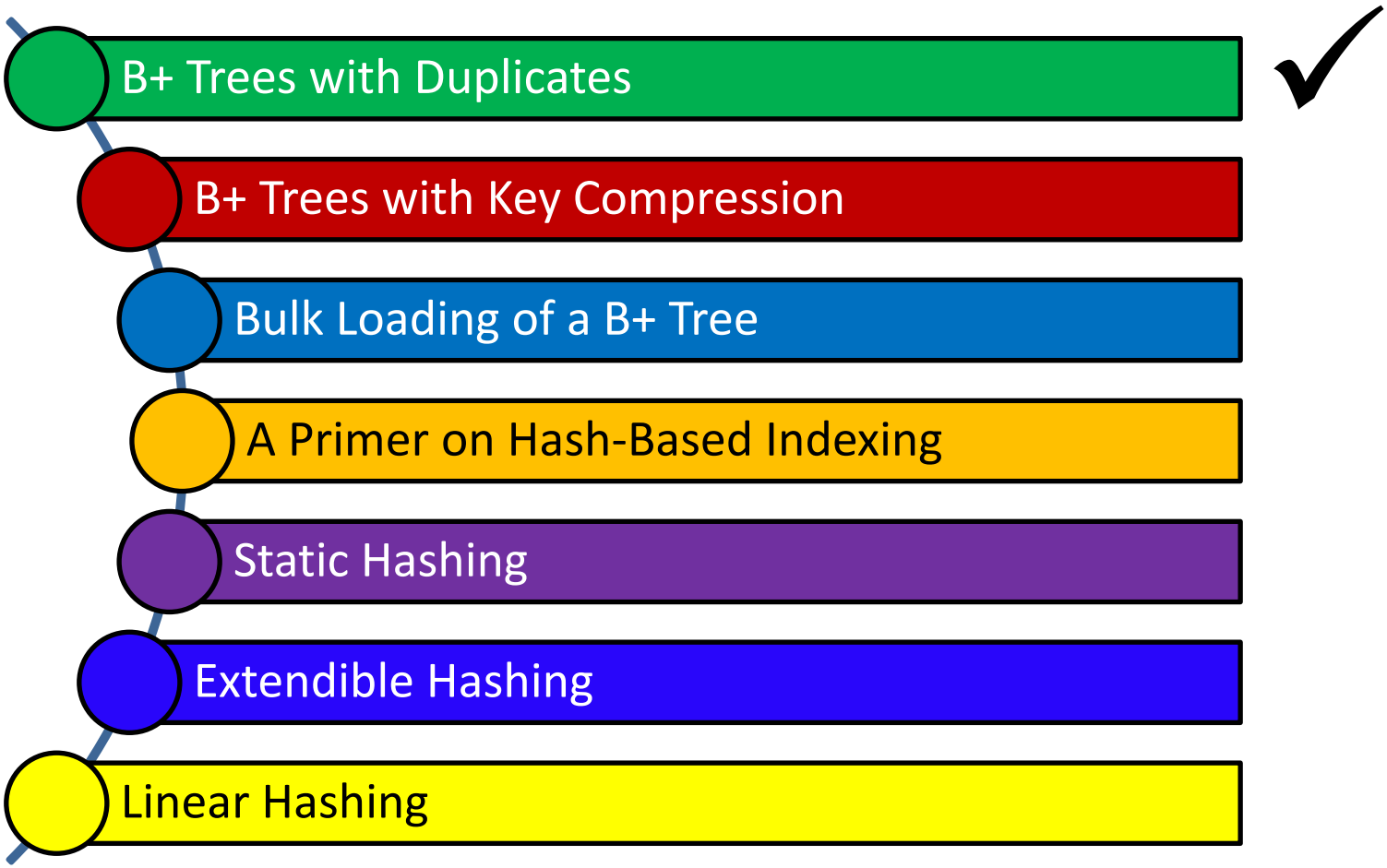
Today…

- Last Session:
  - DBMS Internals- Part III
    - Tree-based indexes: ISAM, B and B+ trees
- Today's Session:
  - DBMS Internals- Part IV
    - Tree-based (B+ tree- cont'd) and Hash-based indexes
- Announcements:
  - PS2 grades are out
  - PS3 is now posted. It is due on March 03, 2014
  - The midterm exam is on Wednesday Feb 26 (*all* material are included)

DBMS Layers


Queries
Query Optimization and Execution
Relational Operators
Files and Access Methods
Buffer Management
Disk Space Management
Transaction Manager
Lock Manager
Recovery Manager
Continued…
DB

Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

B+ Trees With Duplicates

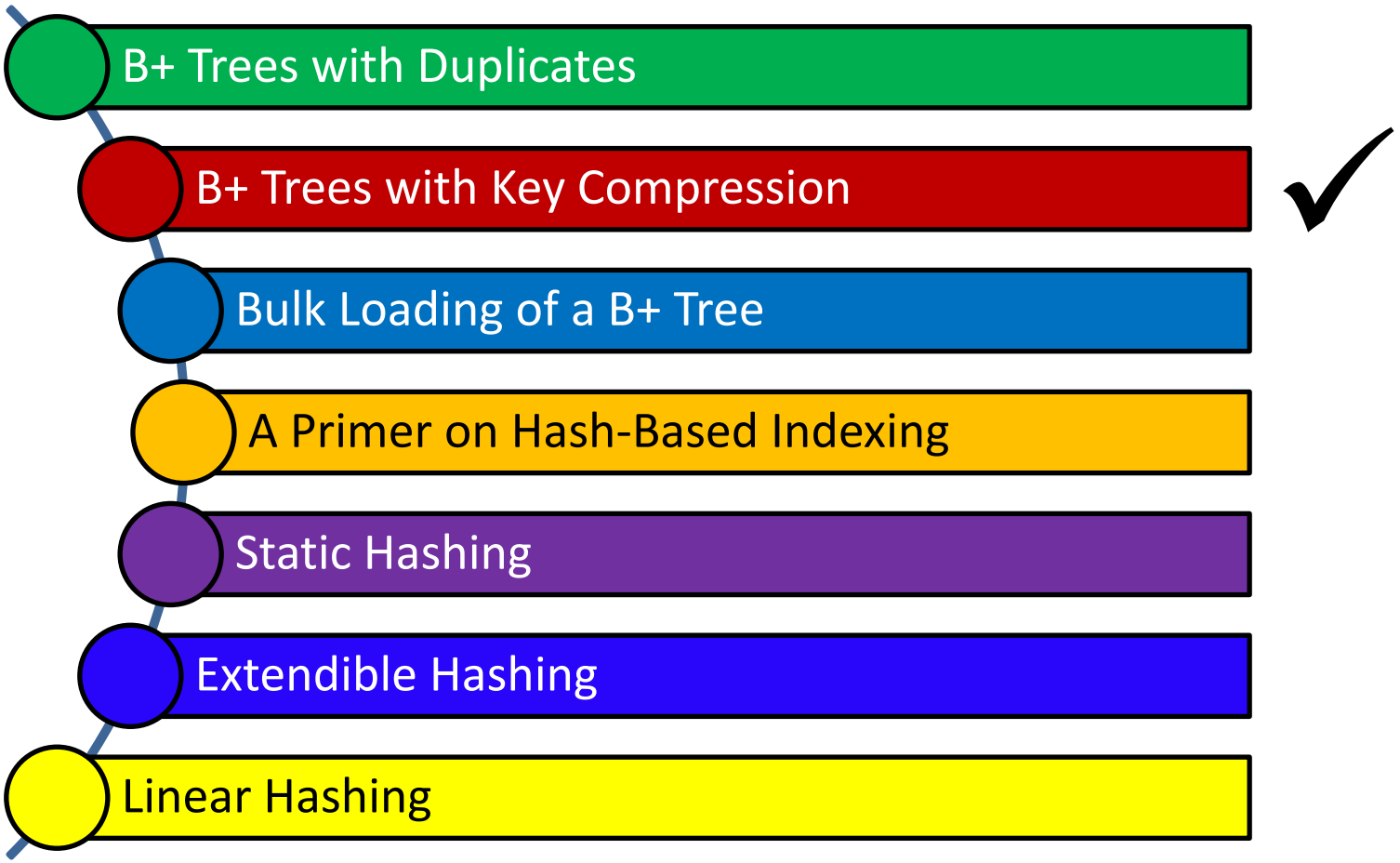
- Thus far, we only discussed unique indices (no *duplicate keys- i.e., several data entries with the same key value*)

- How can we handle duplicate keys?
  1. Use overflow pages to keep all entries of a given key value on a single leaf page (natural for ISAM)

  2. Treat duplicates like any other entries
     - Several leaf pages will contain entries of a given key value
     - How to search/delete?

  3. Make the *rid* value part of the search key

Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

The Height of a B+ Tree

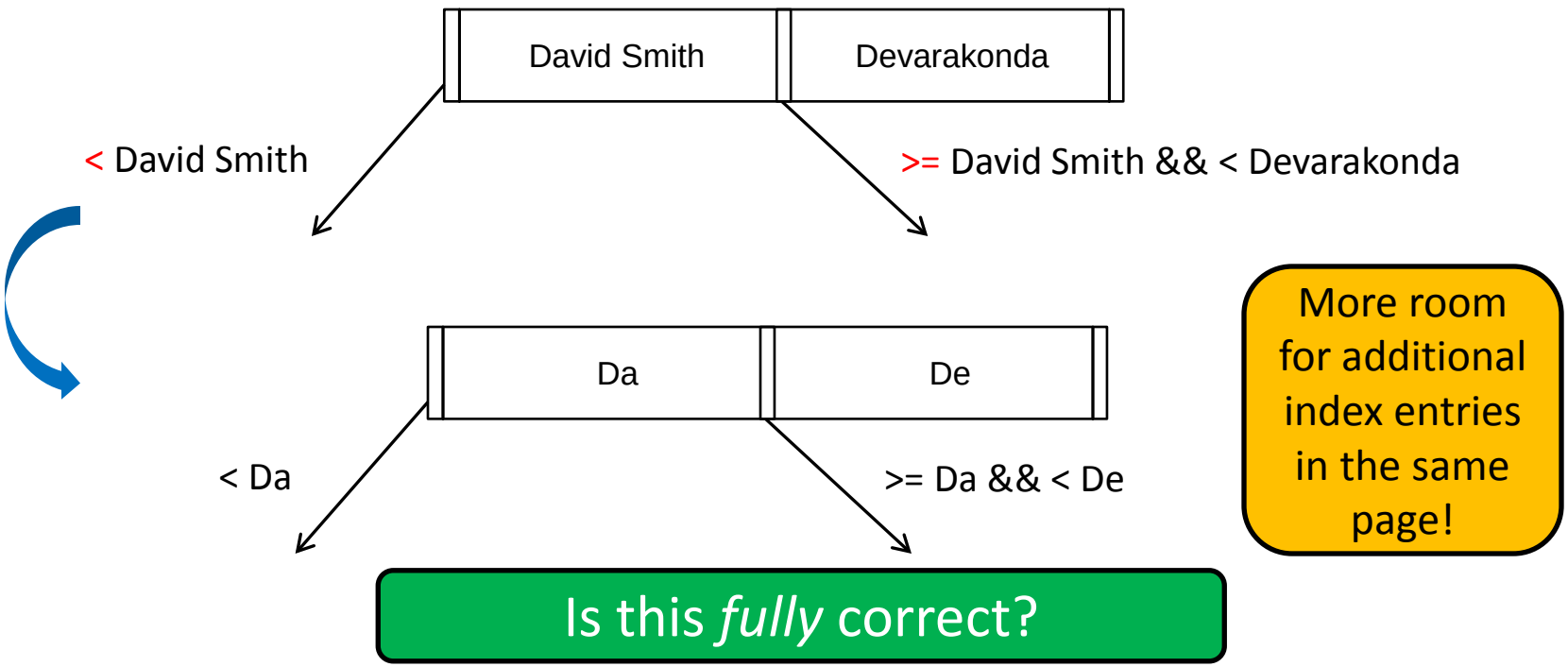
- What are the factors that define the height of a B+ tree?
  - Number of data entries
  - The order of occupancy
- The order of occupancy dictates the *fan-out* of the tree
- The height of the tree is proportional to $log_{fan\text{-}out}$ *(# of DEs)*
- What is the number of disk I/Os to retrieve a data entry?
  - $log_{fan\text{-}out}$ *(# of DEs)*
- How to minimize the height?
  - Maximize the fan-out

Towards Maximizing the Fan-Out?

- What does an index entry contain?
  - A search key
  - A page pointer
- Hence, the size of an index entry depends primarily on the size of the search key value!
- What if the search key values are very long?
  - Not many index entries will fit on a page
  - Fan-out will be low
  - The height of the tree will be large

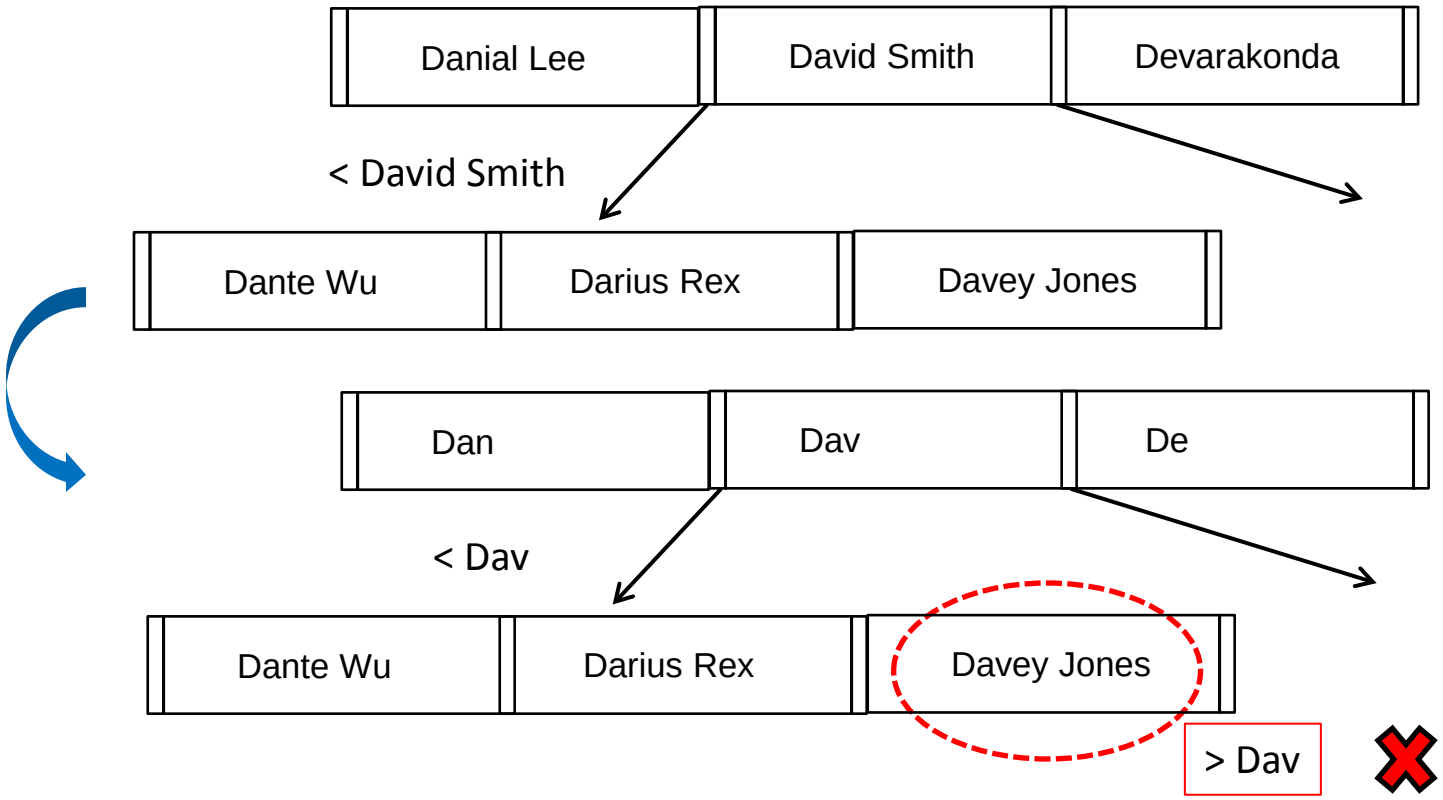
Key Compression: A Way to Maximize the Fan-Out

- How can we reduce the size of search key values?
  - Apply key compression, especially that keys are only used to direct traffic to the appropriate leaves



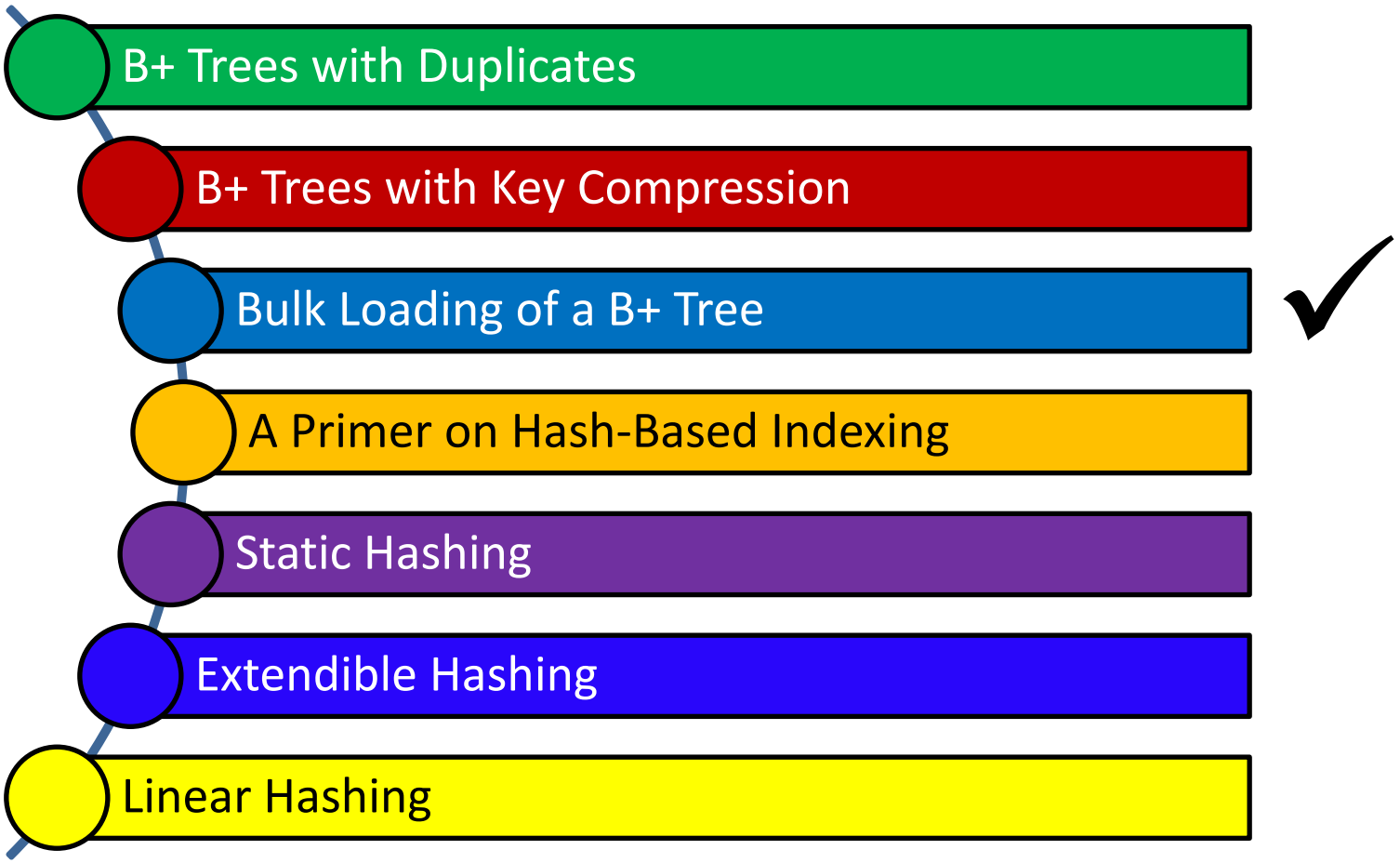
Key Compression: A Way to Maximize the Fan-Out (Cont'd)

- What about the following example?



To ensure correct semantics, the **largest key value in the left sub-tree** and the **smallest key value in the right sub-tree** must be examined!

Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

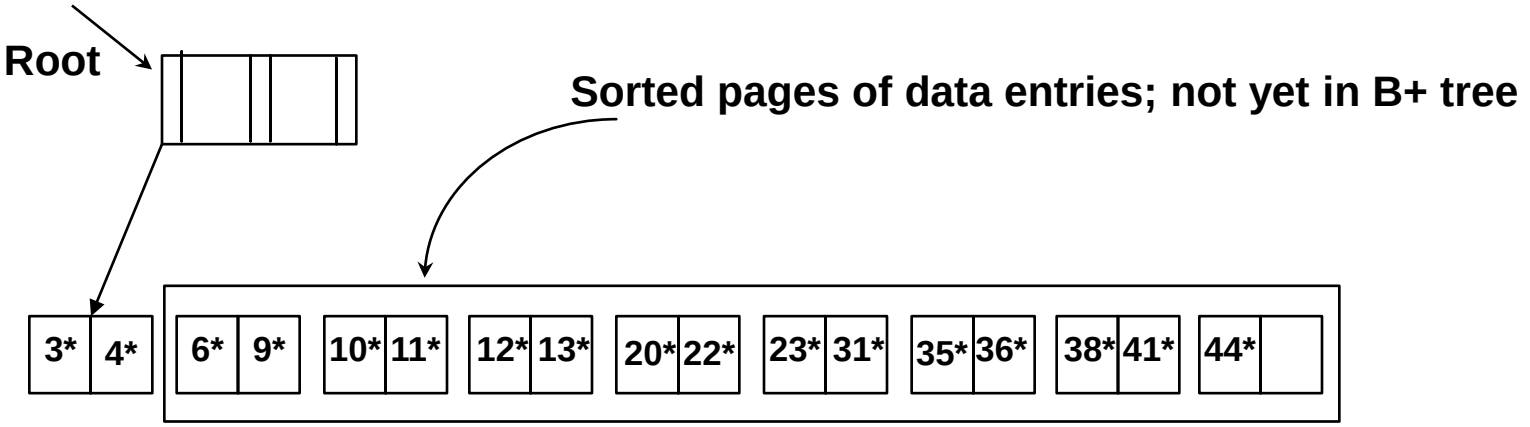
B+ Tree: Bulk Loading

- Assume a collection of data records with an *existing* B+ tree index on it
  - How to add a new record to it?
    - Use the B+ tree insert() function

- What if we have a collection of data records for which we want to create a B+ tree index? (i.e., we want to *bulk load* the B+ tree)
  - Starting with an empty tree and using the insert() function for each data record, *one at a time*, is expensive!
    - This is because for each entry we would require starting again from the root and going down to the appropriate leaf page

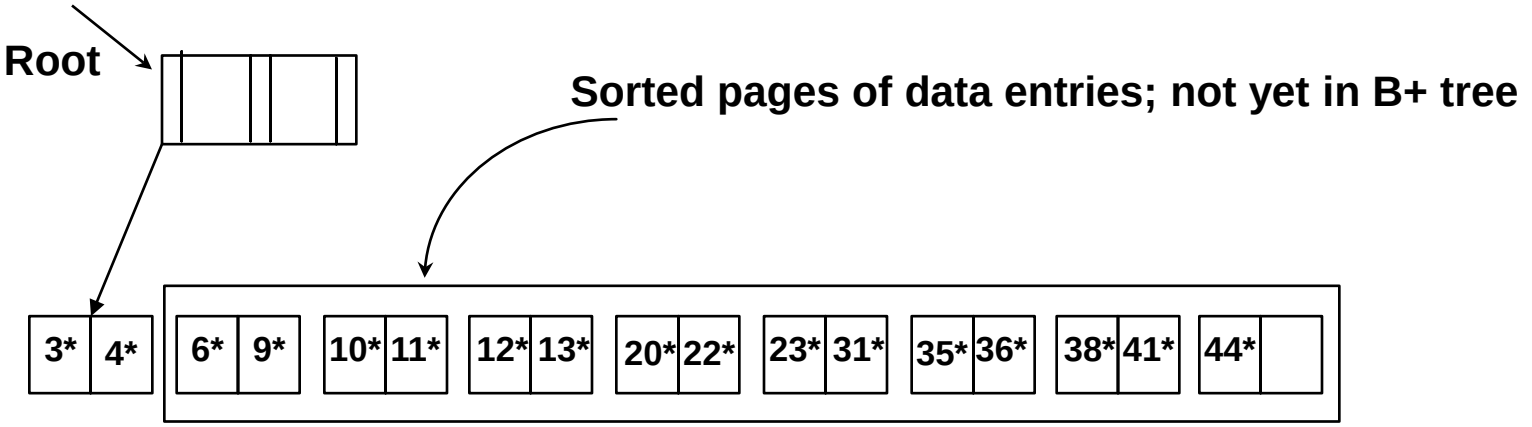
B+ Tree: Bulk Loading

- What to do?
  - *Initialization*: Sort all data entries, insert pointer to first (leaf) page in a new (root) page



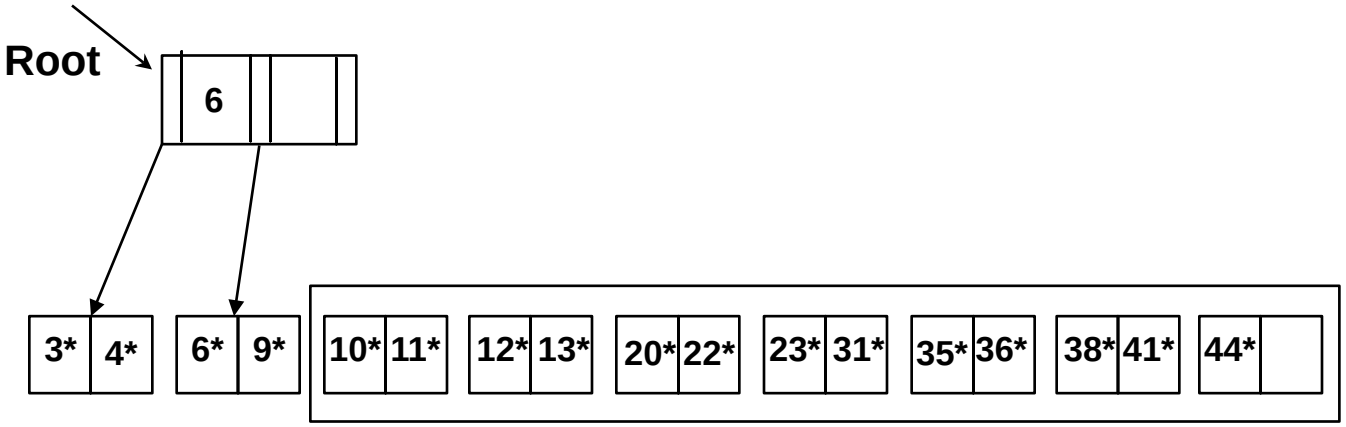
B+ Tree: Bulk Loading

- What to do?
  - Add one entry to the root page for each subsequent page of the sorted data entries *(i.e., <lowest key value on page, pointer to the page>)*



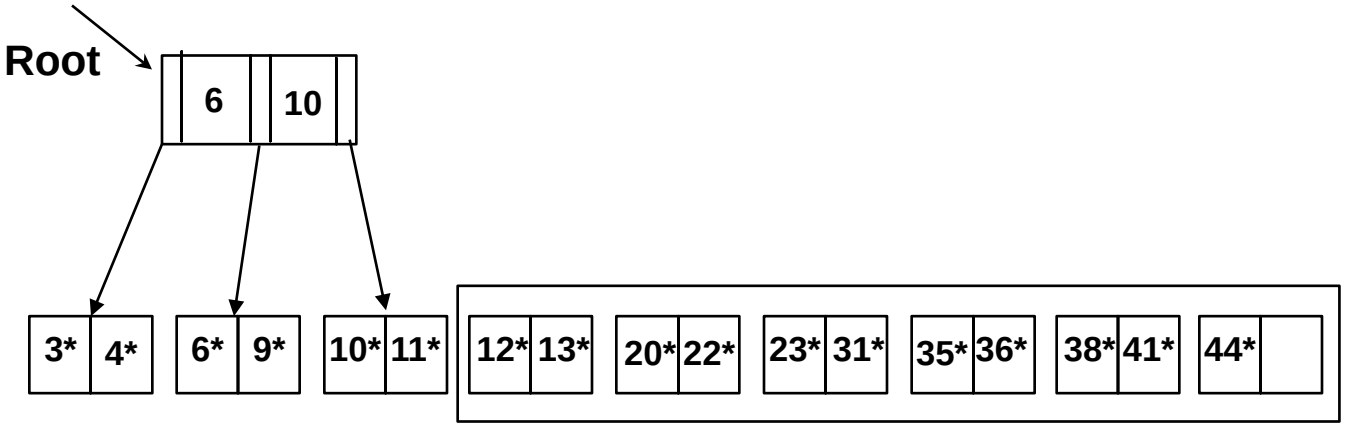
B+ Tree: Bulk Loading

- What to do?
  - Add one entry to the root page for each subsequent page of the sorted data entries *(i.e., <lowest key value on page, pointer to the page>)*



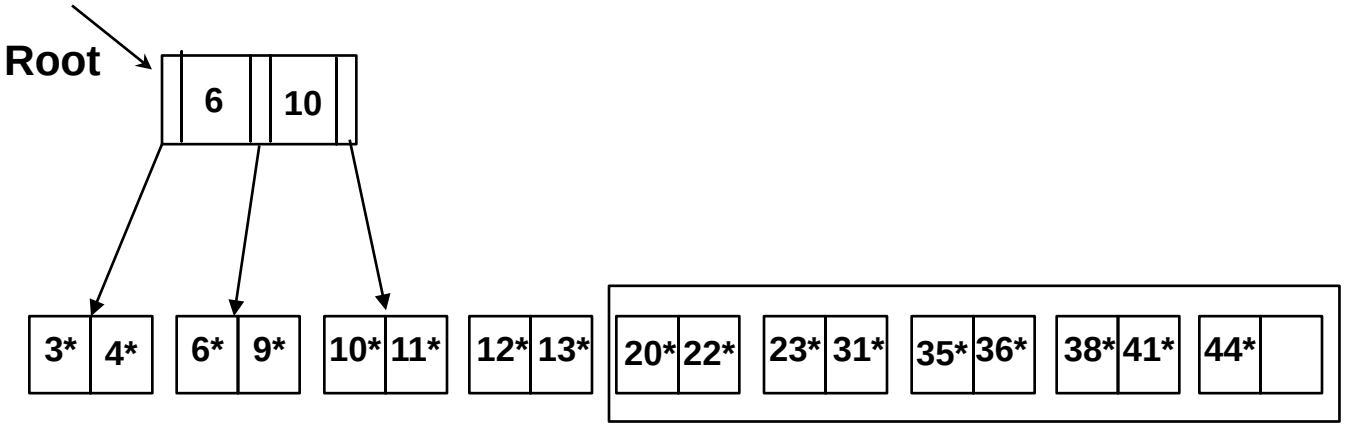
B+ Tree: Bulk Loading

- What to do?
  - Add one entry to the root page for each subsequent page of the sorted data entries *(i.e., <lowest key value on page, pointer to the page>)*



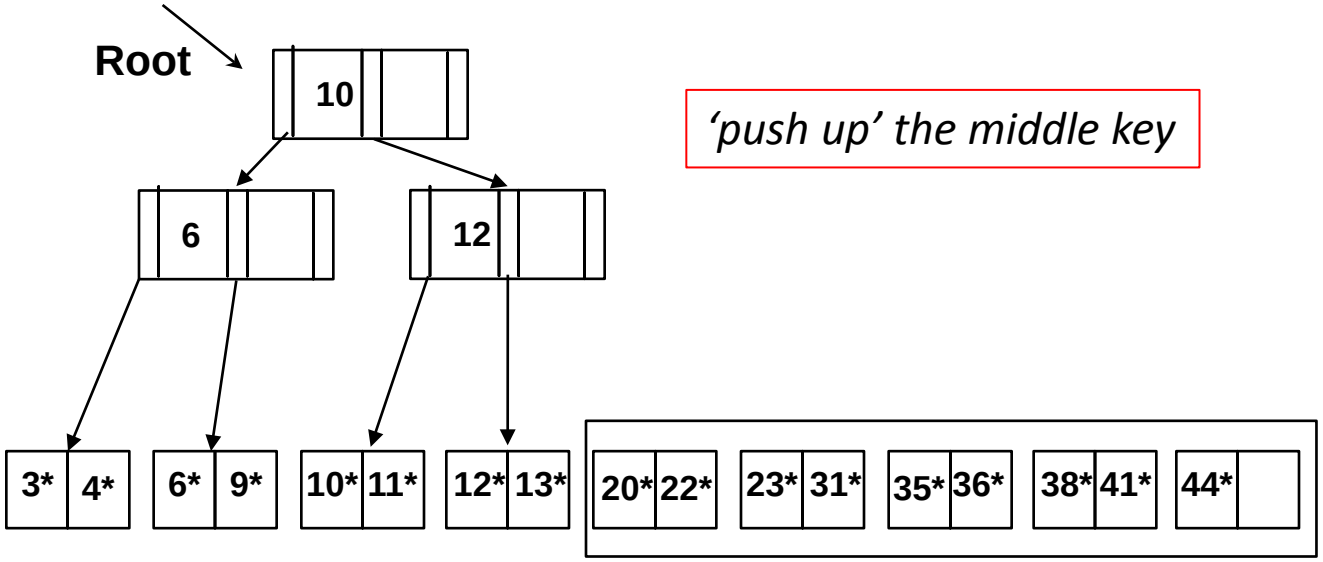
B+ Tree: Bulk Loading

- What to do?
  - Split the root and create a new root page



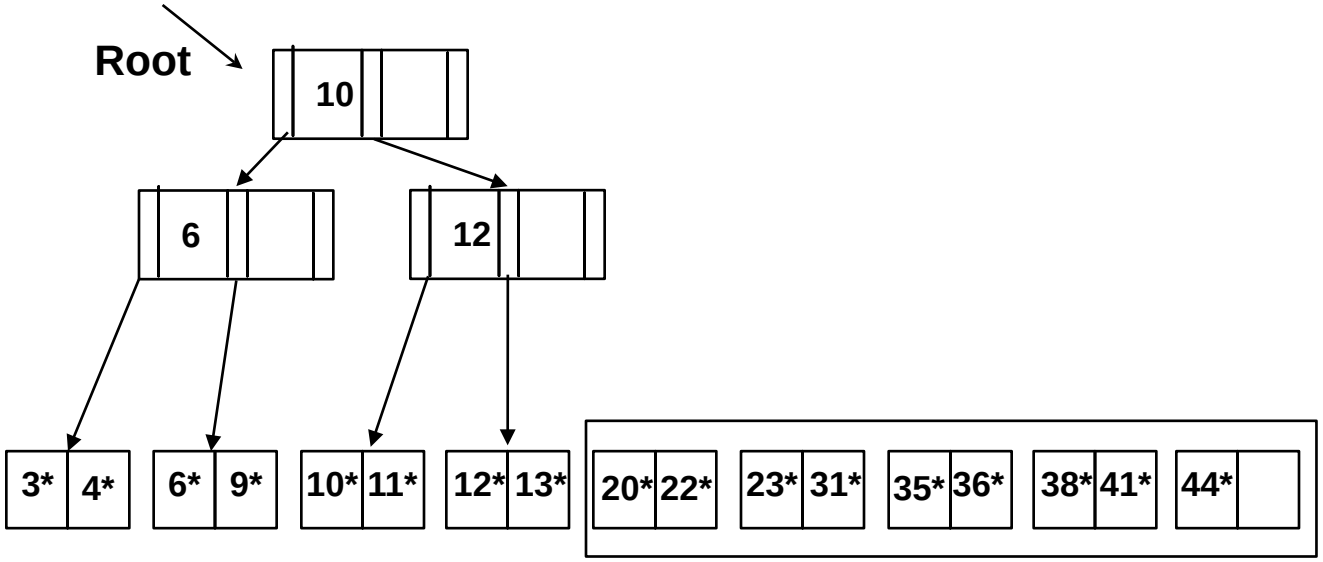
B+ Tree: Bulk Loading

- What to do?
  - Split the root and create a new root page



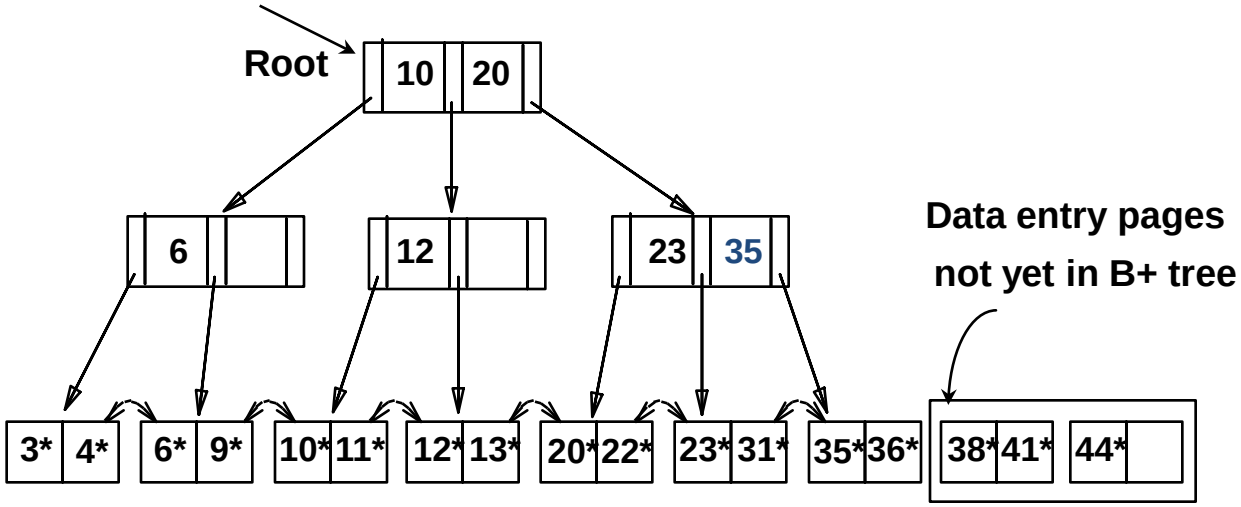
B+ Tree: Bulk Loading

- What to do?
  - Continue by inserting entries into the right-most index page just above the leaf page; split when fills up



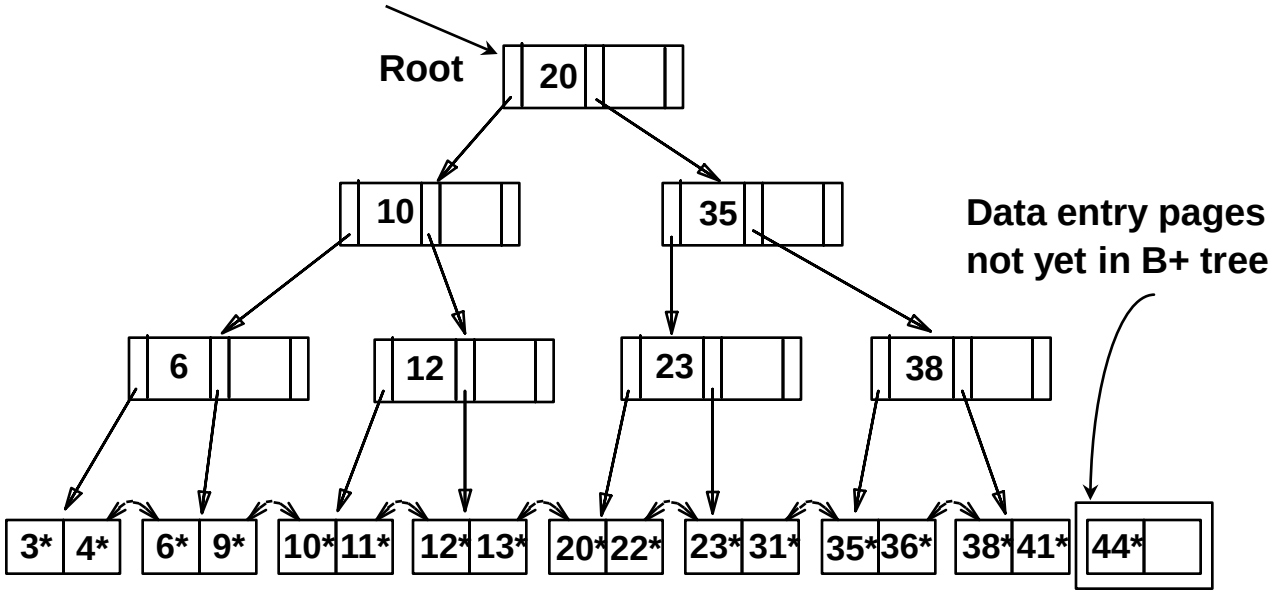
B+ Tree: Bulk Loading

- What to do?
  - Continue by inserting entries into the right-most index page just above the leaf page; split when fills up



B+ Tree: Bulk Loading

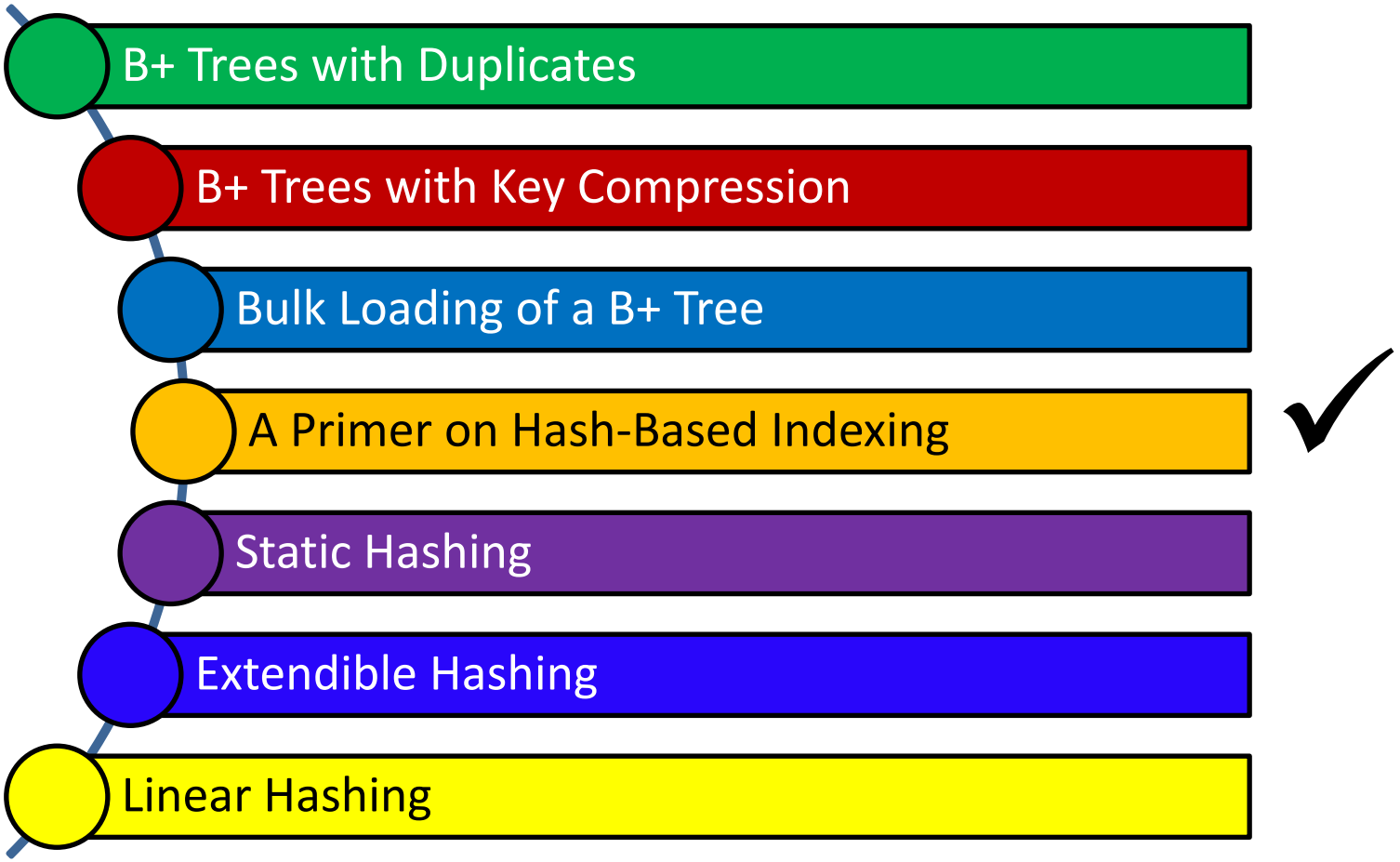
- What to do?
  - Continue by inserting entries into the right-most index page just above the leaf page; split when fills up



B+ Tree: Bulk Loading

- What is the cost of bulk loading?
  1. Creating the leaf-level entries
     - Scanning the data entries and writing out all the leaf-level entries (i.e., ***K****)
     - Hence, (***R***+***E***) I/Os, where ***R*** is the number of pages containing data entries and ***E*** is the number of pages containing ***K****entries
  2. Sorting leaf-level entries
     - 3***E*** I/Os (*when discussing sorting, we will see how*)
  3. Building the index from the sorted leaf-level entries
     - The cost of writing out all index-level pages (*will be an exercise in the recitation*)!

Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

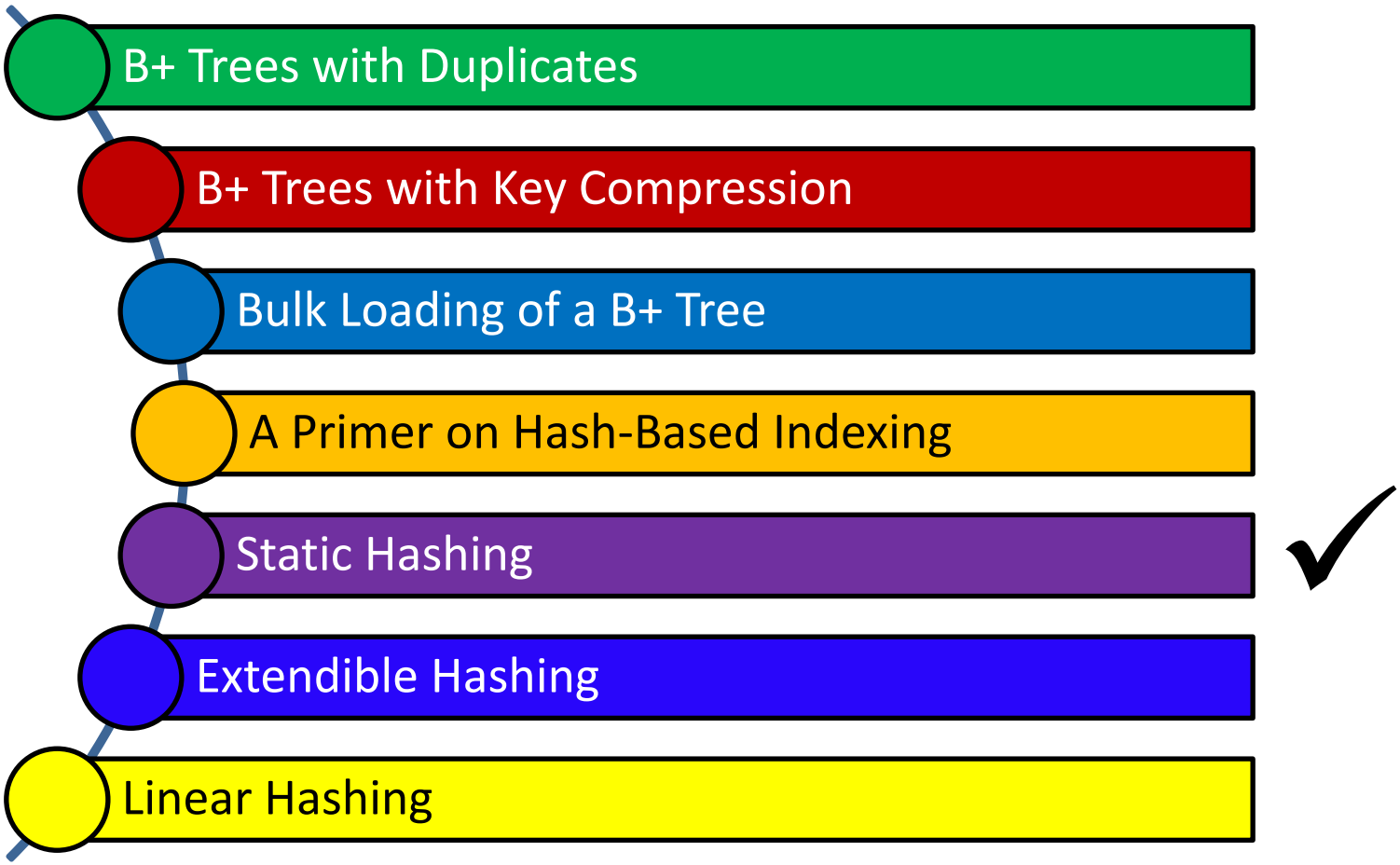
Hash-Based Indexing

- What indexing technique can we use to support *range searches* (e.g., “Find s_name where gpa >= 3.0)?
  - Tree-Based Indexing

- What about *equality selections* (e.g., “Find s_name where sid = 102”?
  - Tree-Based Indexing
  - Hash-Based Indexing (*cannot support range searches!*)

- Hash-based indexing, however, proves to be very useful in implementing relational operators (e.g., joins)

Outline

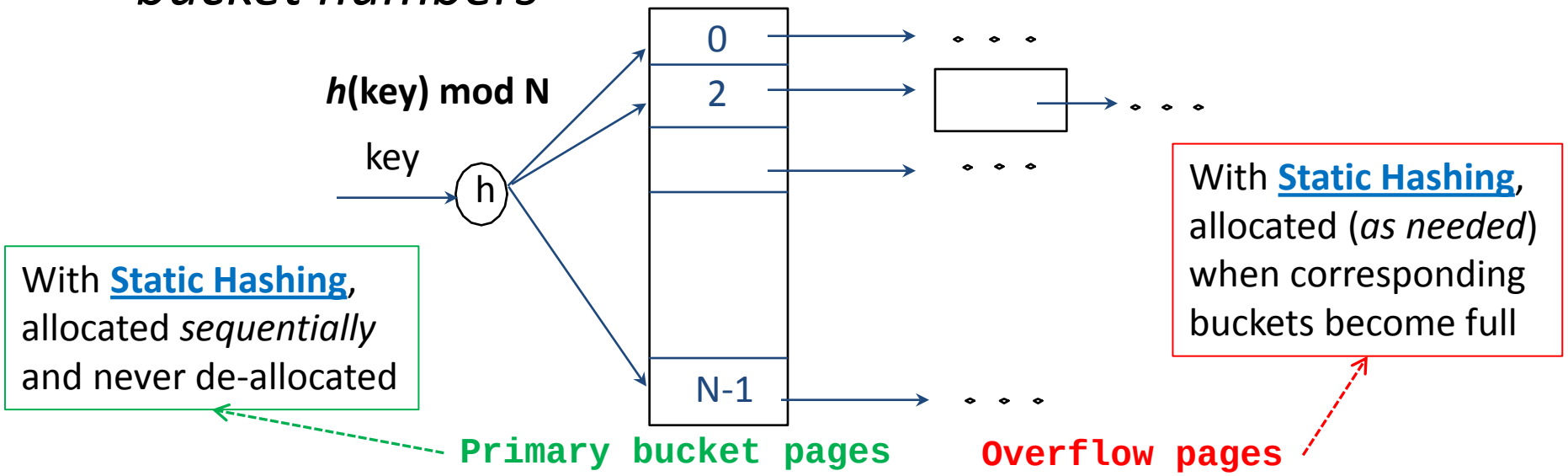


جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Static Hashing

- A hash structure (or table or file) is a *generalization* of the simpler notion of an ordinary array
  - In an array, an arbitrary position can be examined in O(1)

- A hash function ***h*** is used to map keys into a range of *bucket numbers*



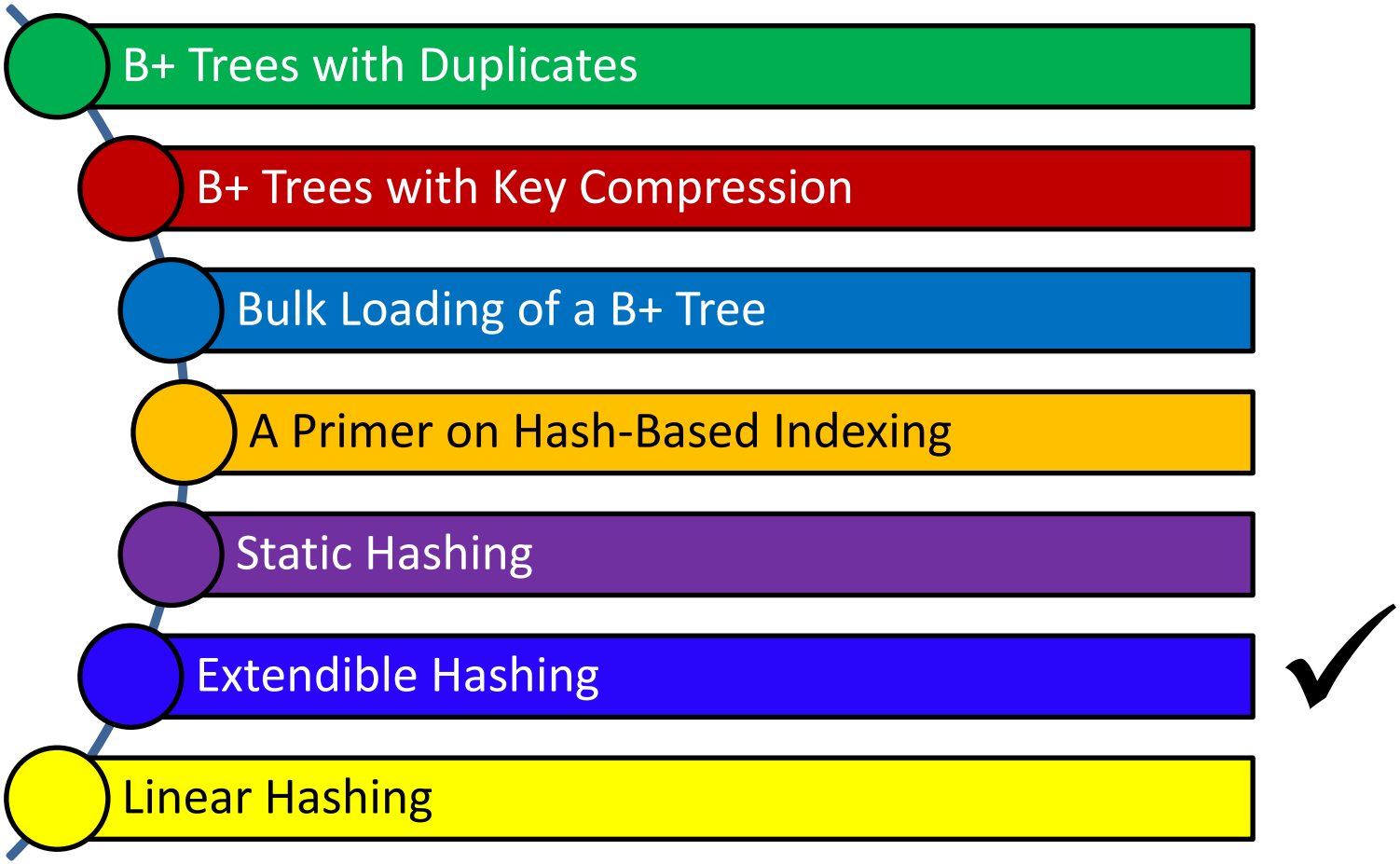
Static Hashing

- Data entries can be any of the three alternatives (**A (1)**, **A (2)** or **A (3)**- see previous lecture)
- Data entries can be *sorted* in buckets to speed up searches
- The hash function ***h*** is used to identify the bucket to which a given key belongs and subsequently *insert*, *delete* or *locate* a respective data record
  - A hash function of the form ***h***(key) = (**a** * key + **b**) works well in practice
- A search *ideally* requires 1 disk I/O, while an insertion or a deletion necessitates 2 disk I/Os

Static Hashing: Some Issues

- Similar to ISAM, the number of buckets is fixed!
  - Cannot deal with insertions and deletions gracefully
- Long overflow chains can develop easily and degrade performance!
  - Pages can be initially kept *only* 80% full
- *Dynamic* hashing techniques can be used to fix the problem
  - Extendible Hashing (EH)
  - Liner Hashing (LH)

Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Directory of Pointers

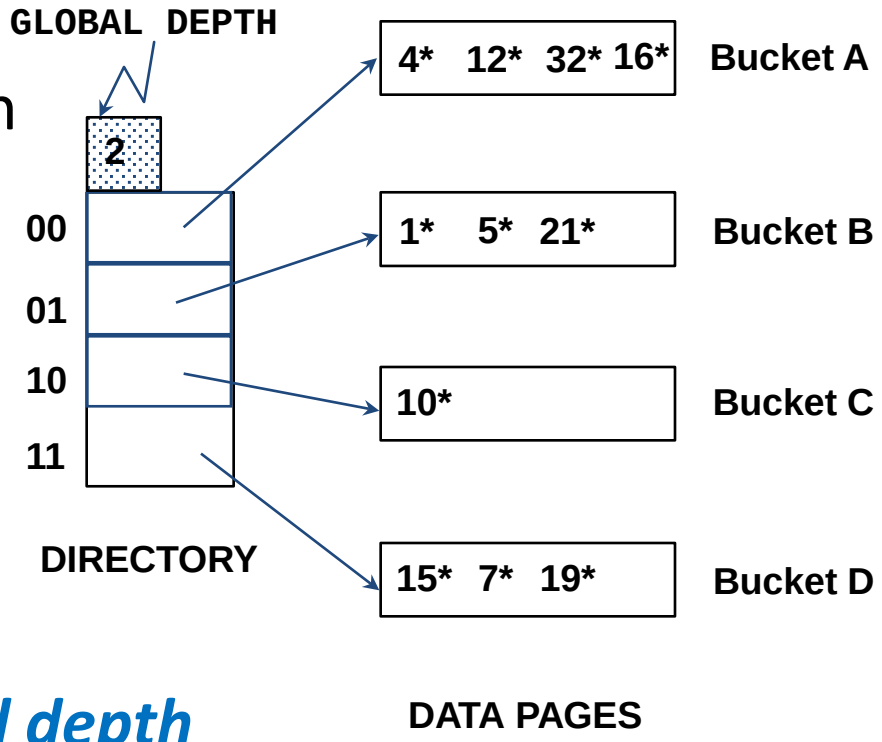
- How else (*as opposed to overflow pages*) can we add a data record to a full bucket in a *static* hash file?
  - Reorganize the table (e.g., by doubling the number of buckets and redistributing the entries across the new set of buckets)
  - But, reading and writing all pages is expensive!

- In contrast, we can use a directory of pointers to buckets
  - Buckets number can be doubled by doubling just the directory and *splitting only the bucket that overflowed*
  - The *trick* lies on how the hash function can be adjusted!

Extendible Hashing

- Extendible Hashing uses a directory of pointers to buckets

- The result of applying a hash function ***h*** is treated as a *binary number* and the last ***d*** bits are interpreted as an offset into the directory

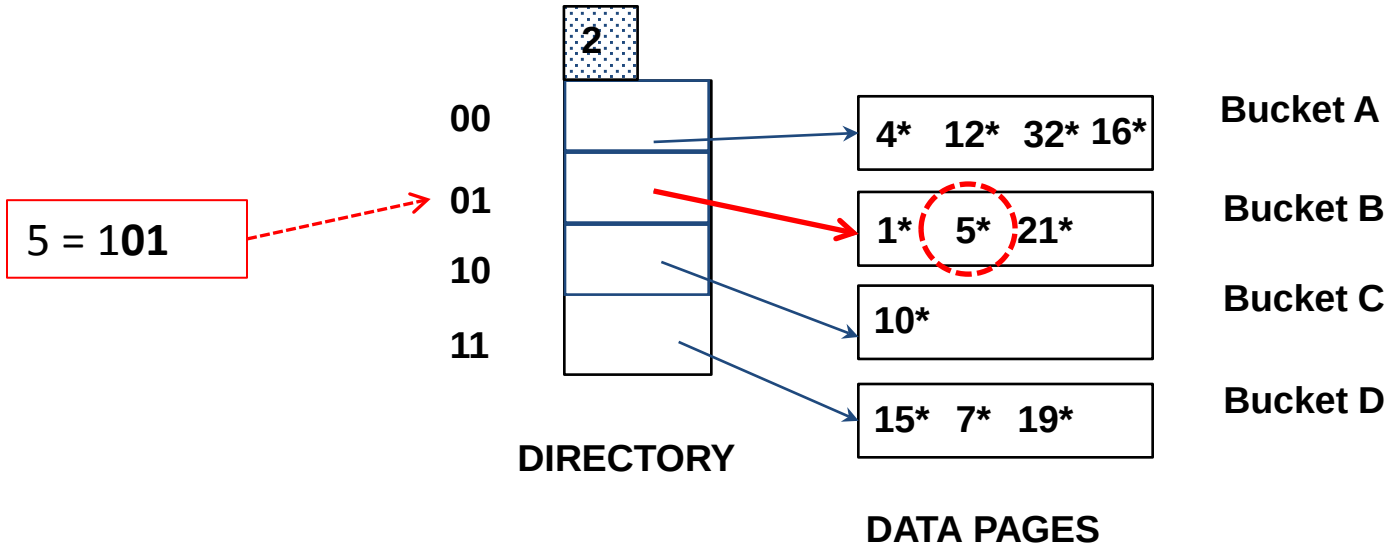


- ***d*** is referred to as the ***global depth*** of the hash file and is kept as part of the header of the file

Extendible Hashing: Searching for Entries

- To search for a data entry, apply a hash function ***h*** to the key and take the last ***d*** bits of its binary representation to get the bucket number

- Example: search for 5*



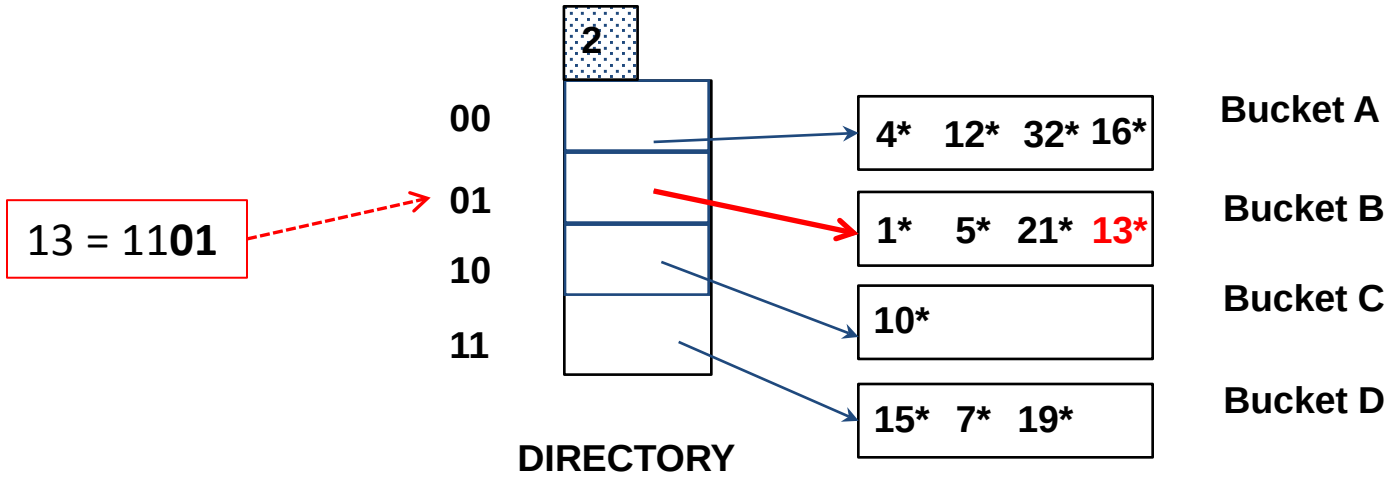
Extendible Hashing: Inserting Entries

- An entry can be inserted as follows:
  - Find the appropriate bucket (*as in search*)
  - Split the bucket *if full* and *redistribute* contents (including the new entry to be inserted) across the old bucket and its ***"split image"***
  - Double the directory *if necessary*
  - Insert the given entry

Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

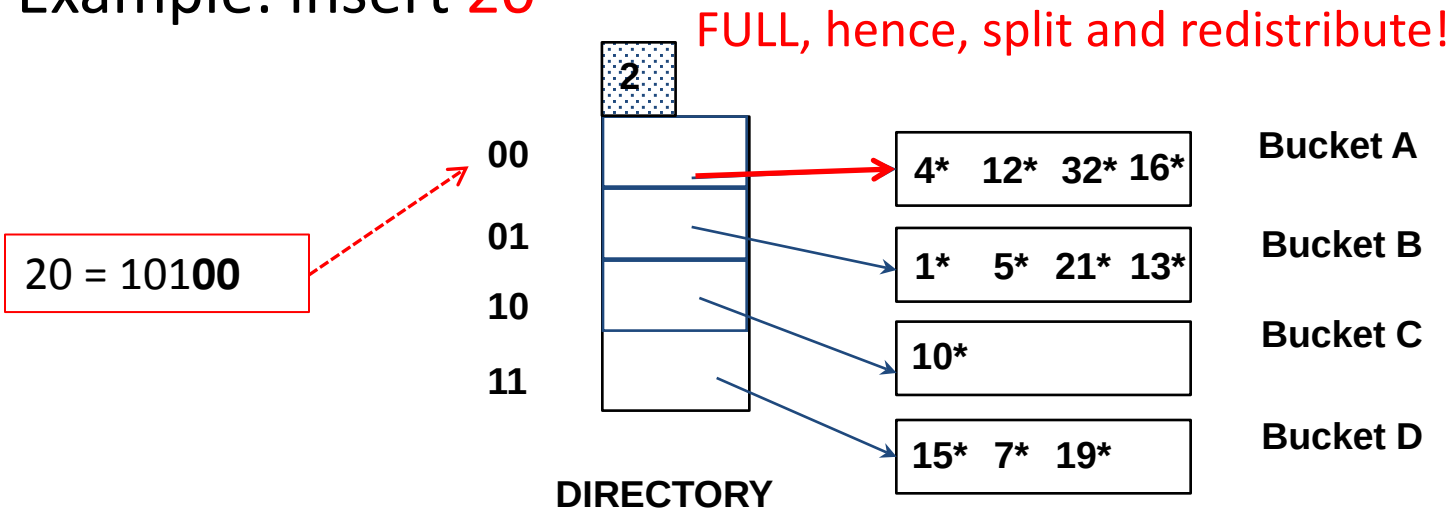
- Example: insert 13*



Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

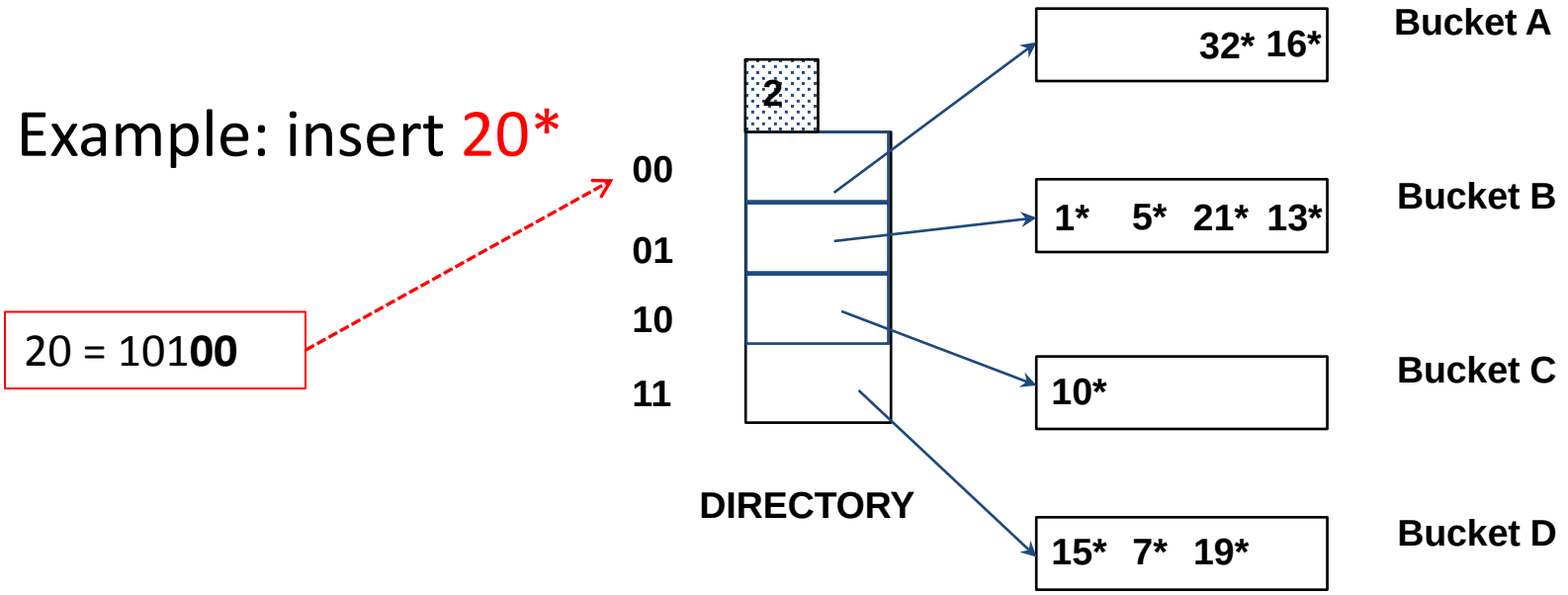
- Example: insert 20*



Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

- Example: insert 20*

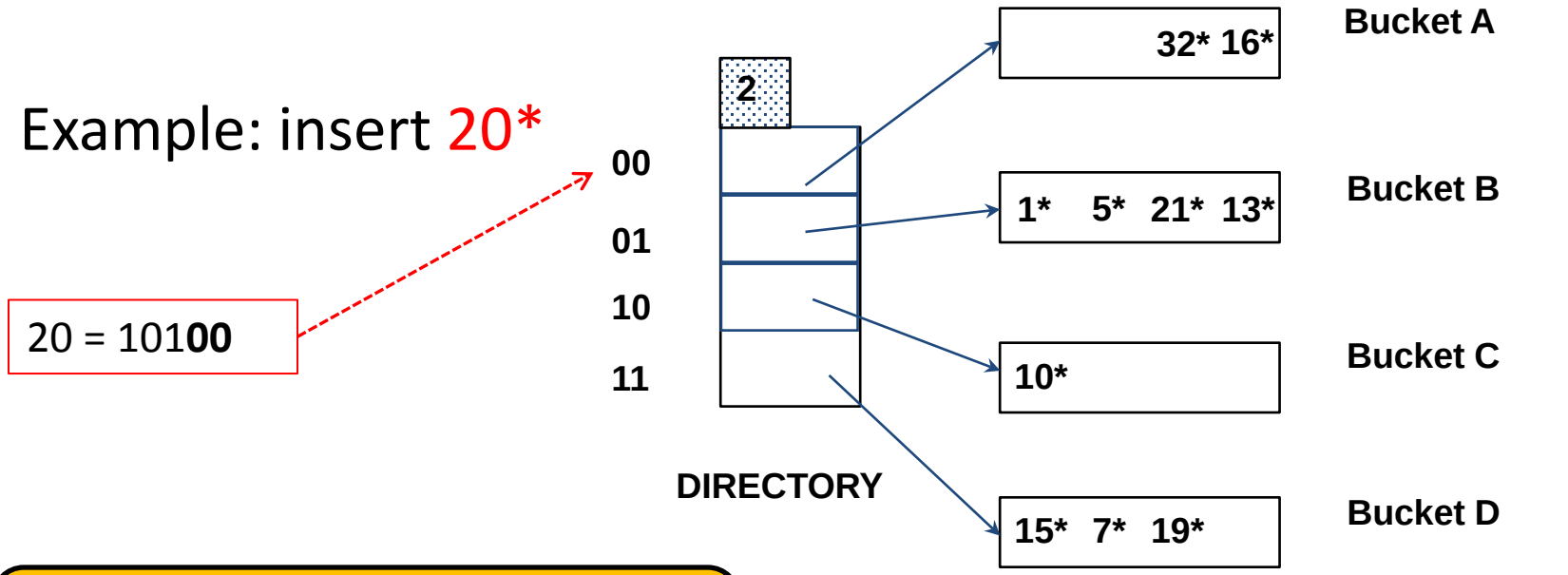


Is this enough?

Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

- Example: insert 20*

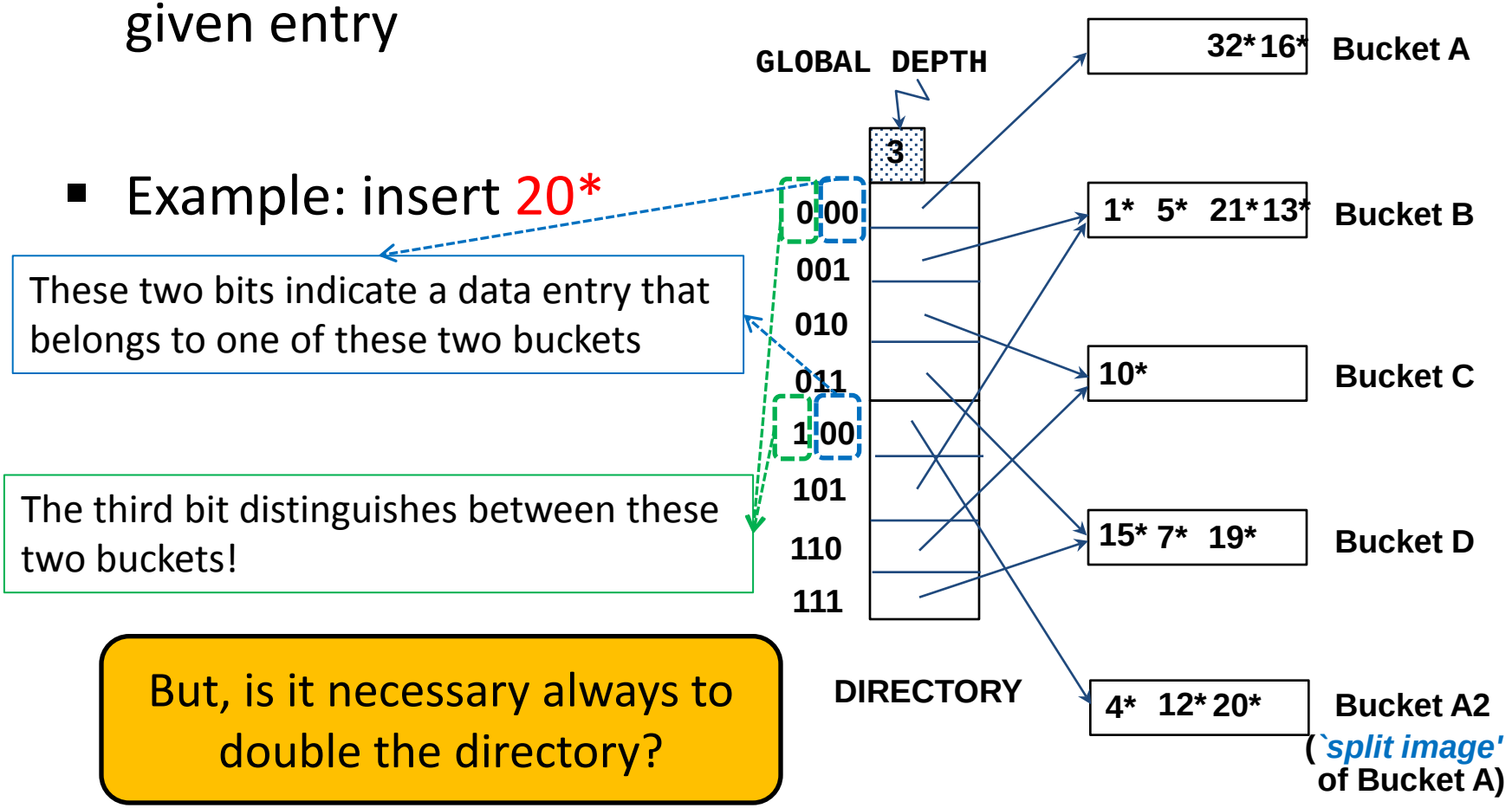


Double the directory and increase the *global depth*

Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

- Example: insert 20*

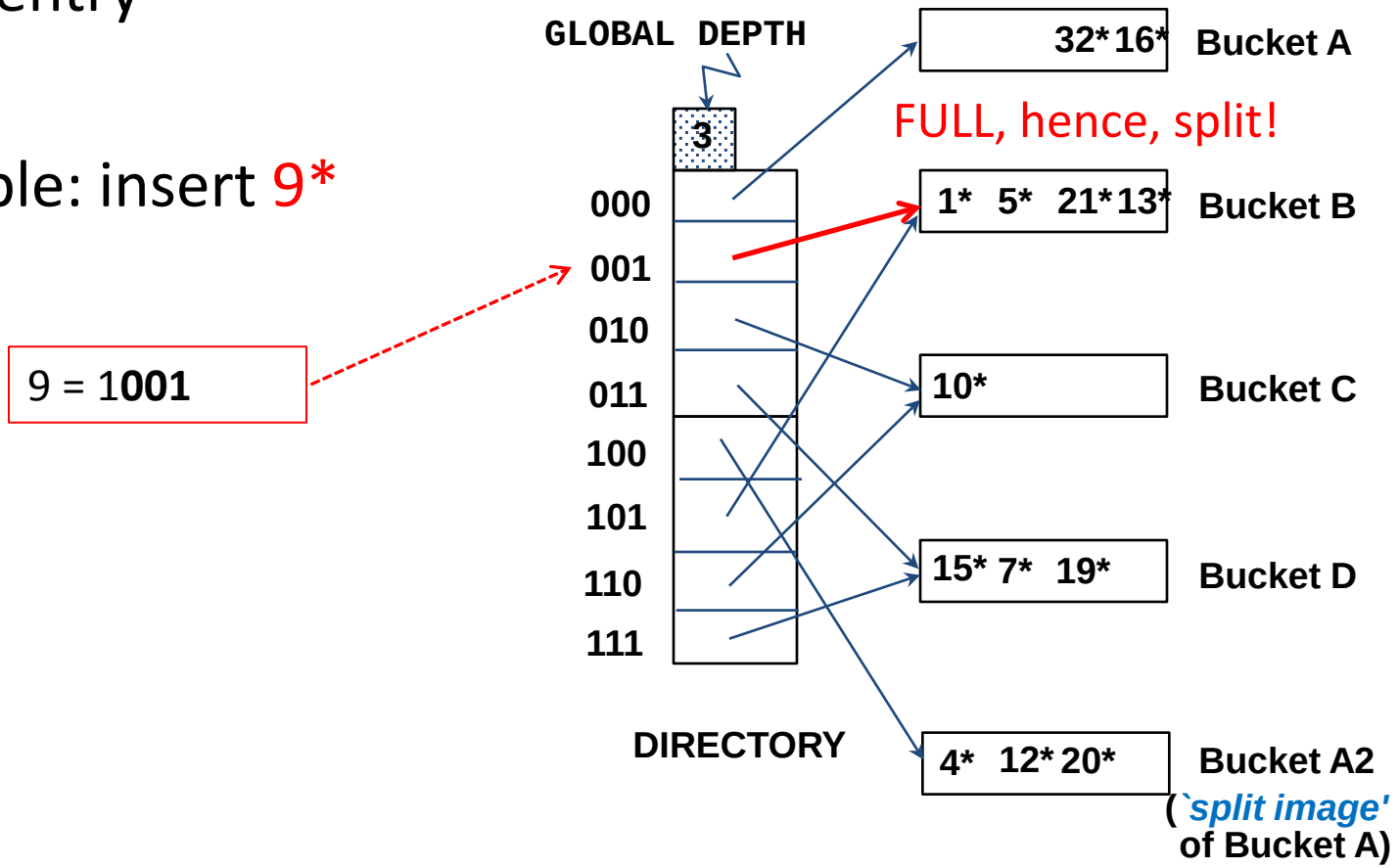


But, is it necessary always to double the directory?

Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

- Example: insert 9*



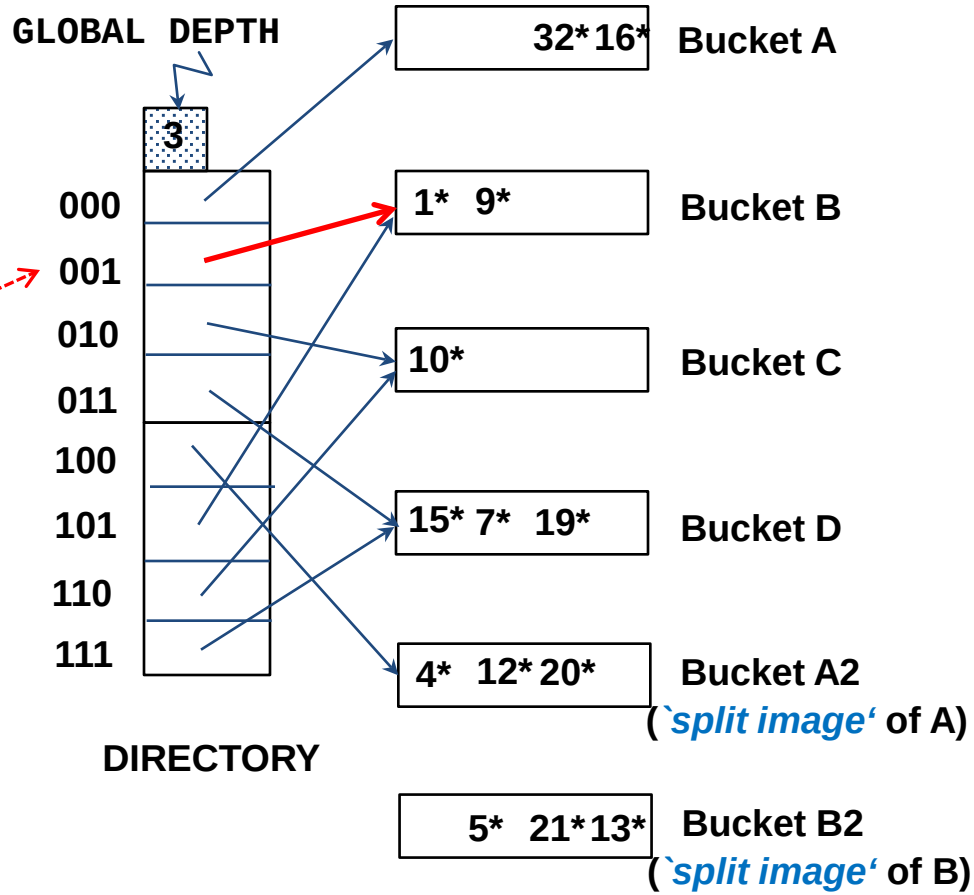
Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

- Example: insert 9*

9 = 1**001**

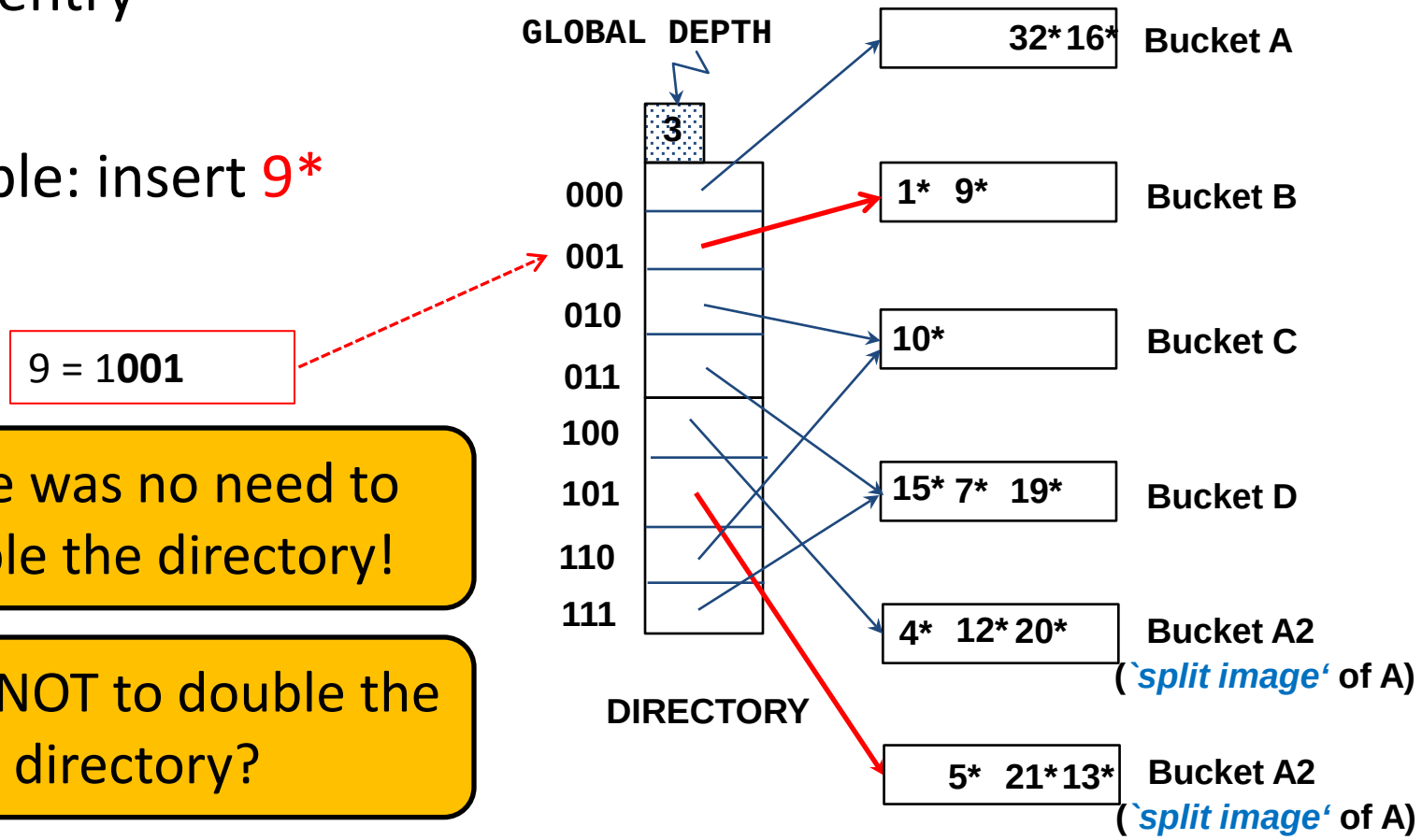
Almost there…



Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

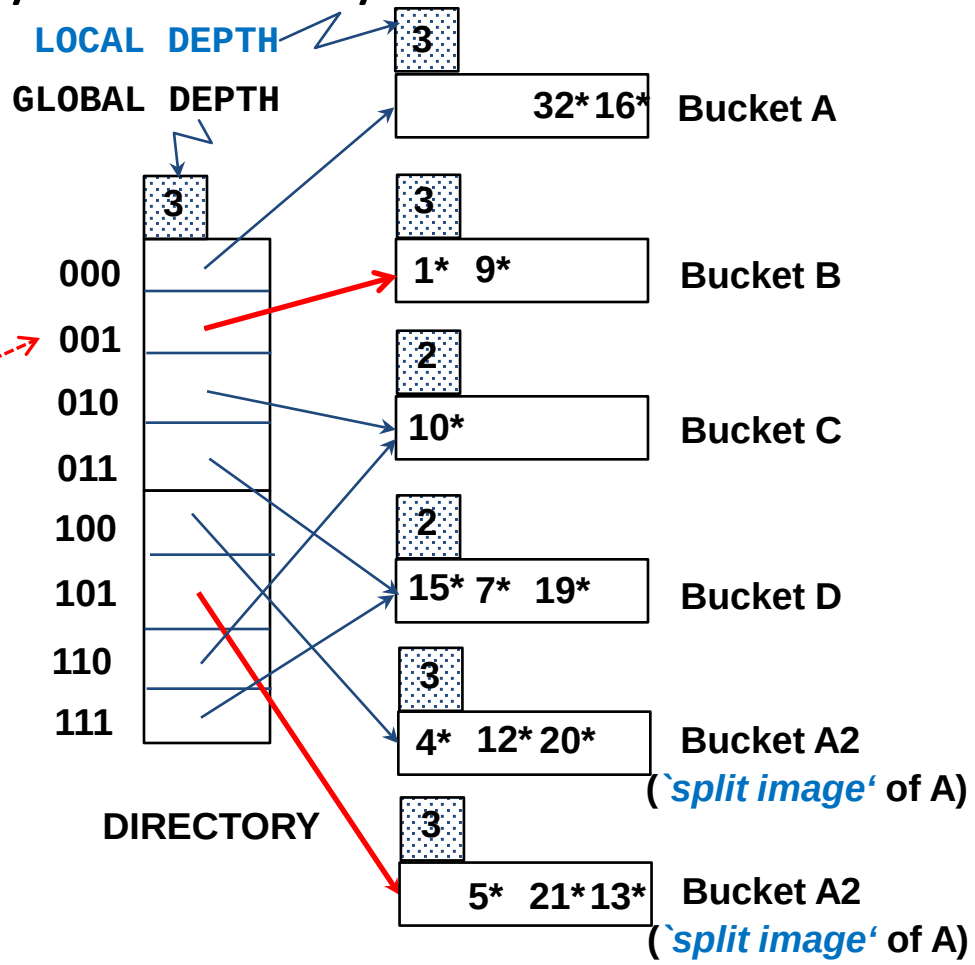
- Example: insert 9*



Extendible Hashing: Inserting Entries

- Find the appropriate bucket (as in search), split the bucket if full, double the directory if necessary and insert the given entry

- Example: insert 9*

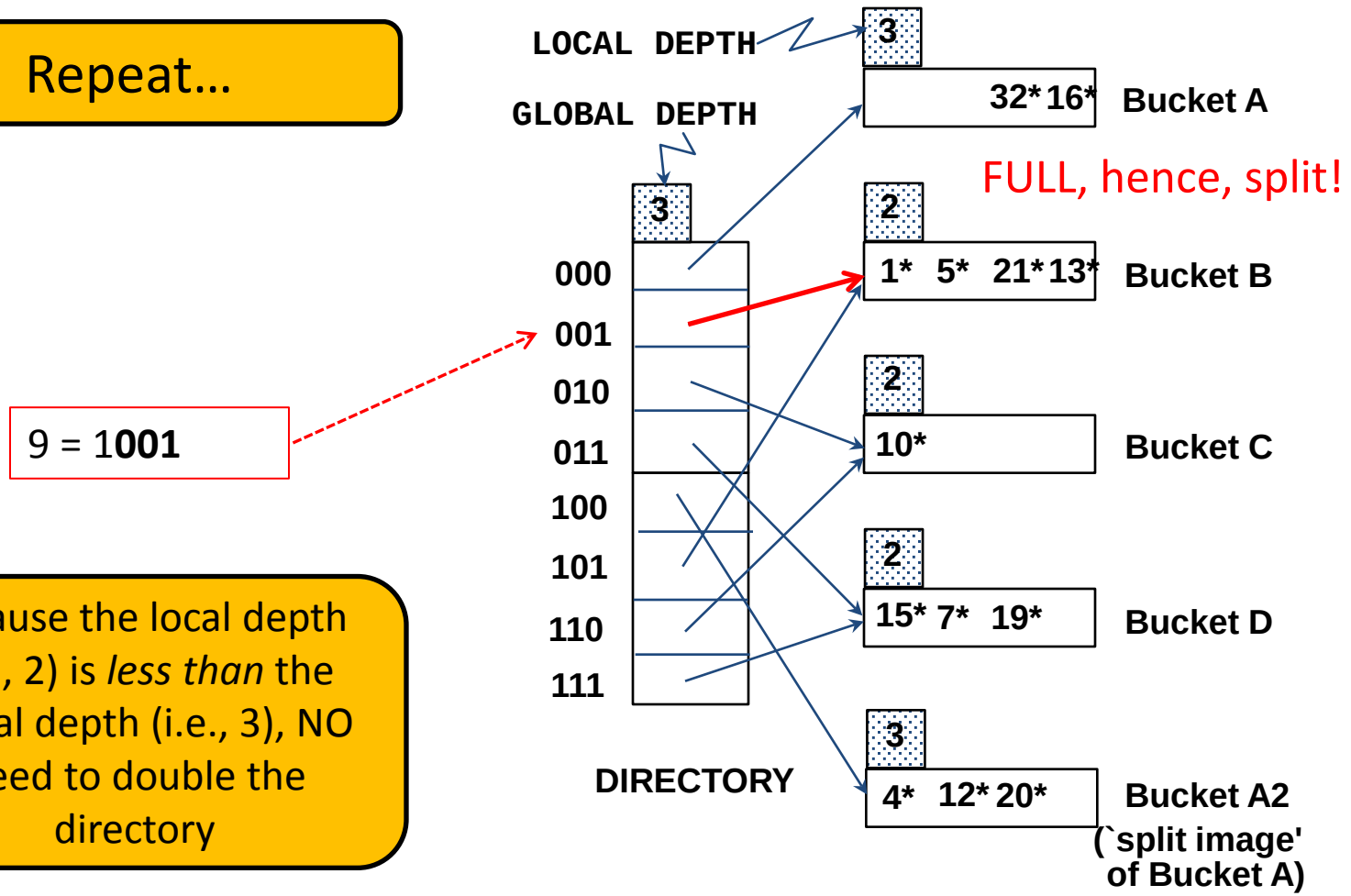


If a bucket whose **local depth equals to the global depth** is split, the directory *must* be doubled

Extendible Hashing: Inserting Entries

- Example: insert 9*

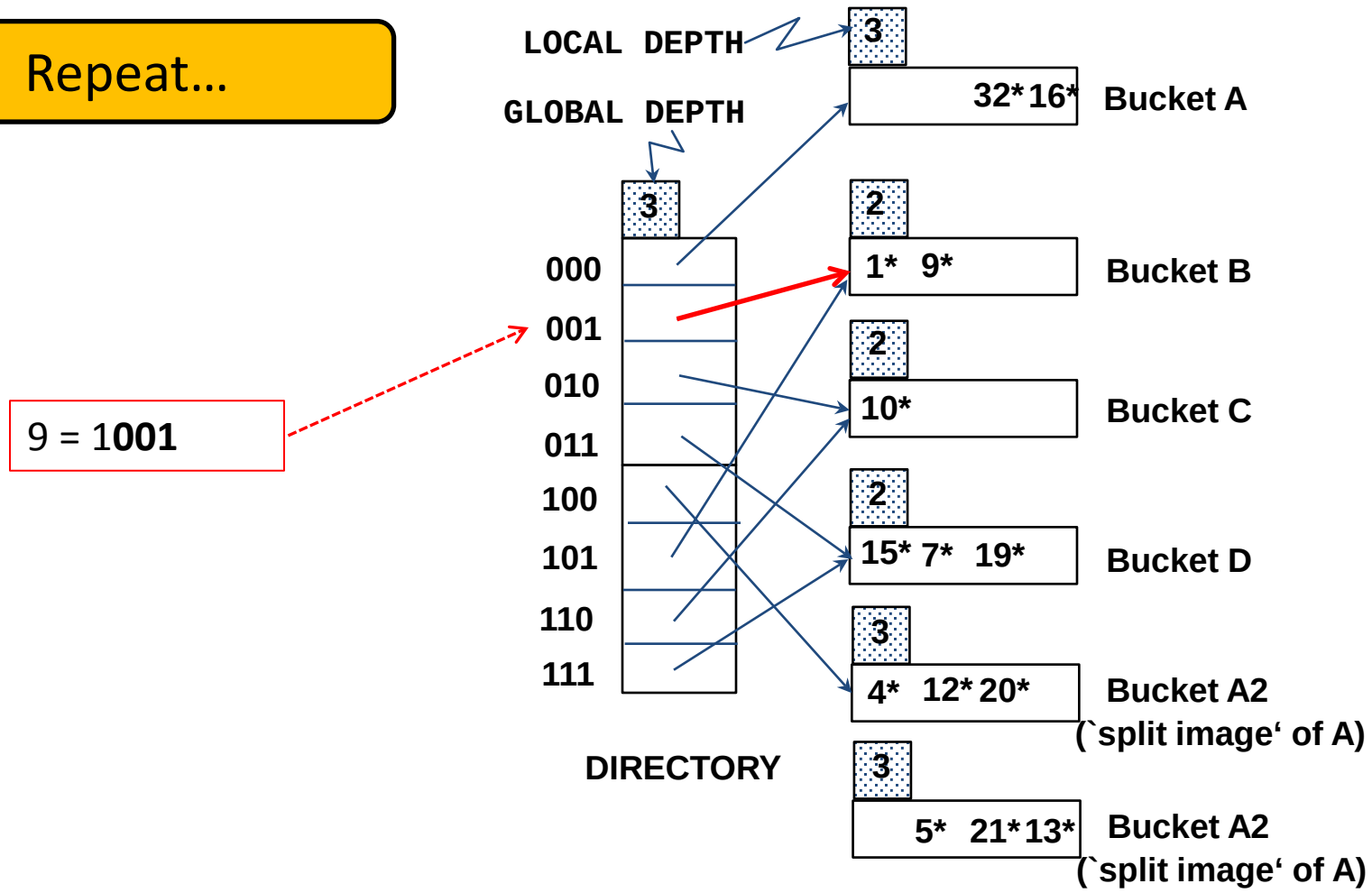
Repeat…



Extendible Hashing: Inserting Entries

- Example: insert 9*

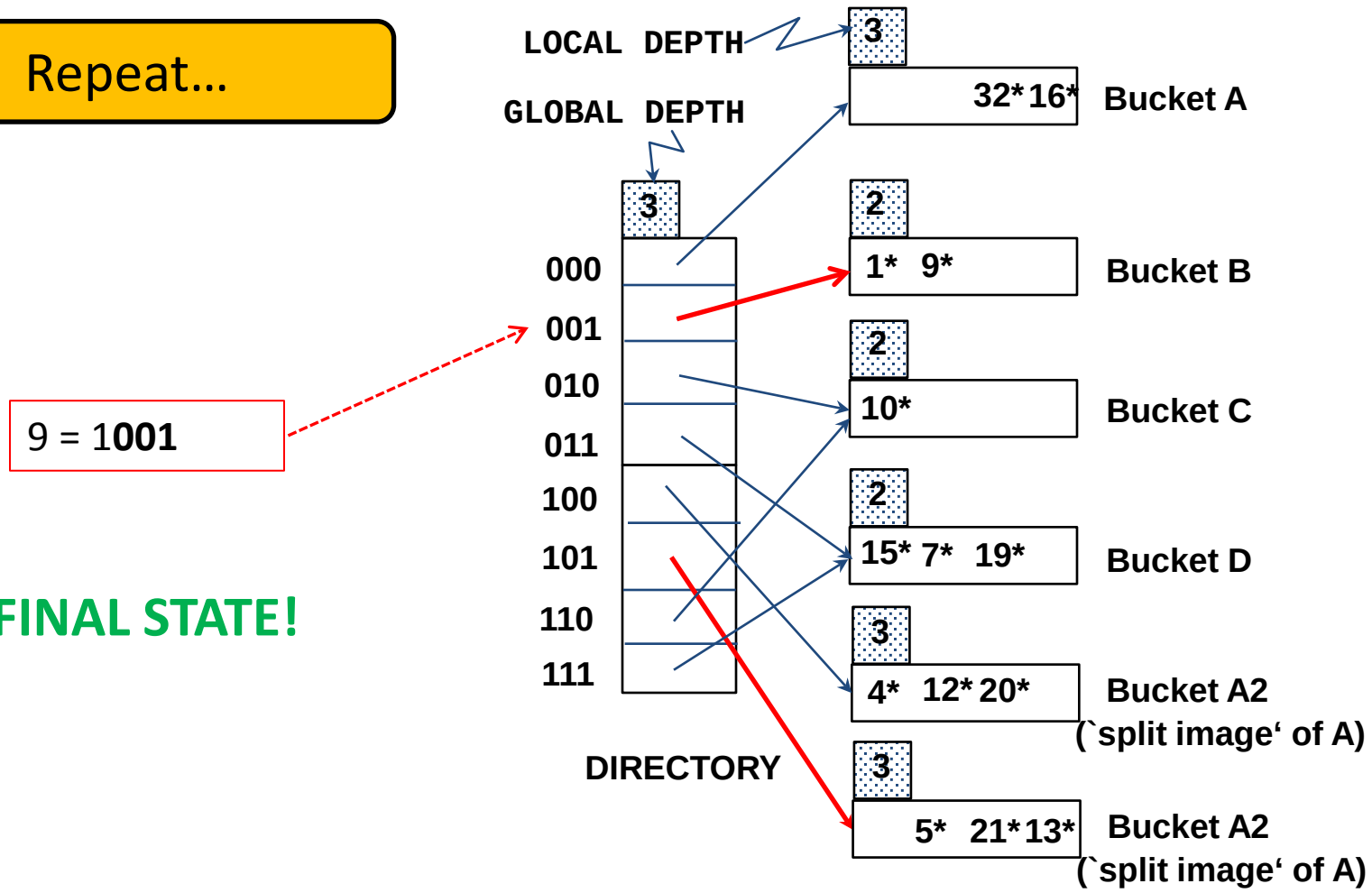
Repeat…



Extendible Hashing: Inserting Entries

- Example: insert 9*

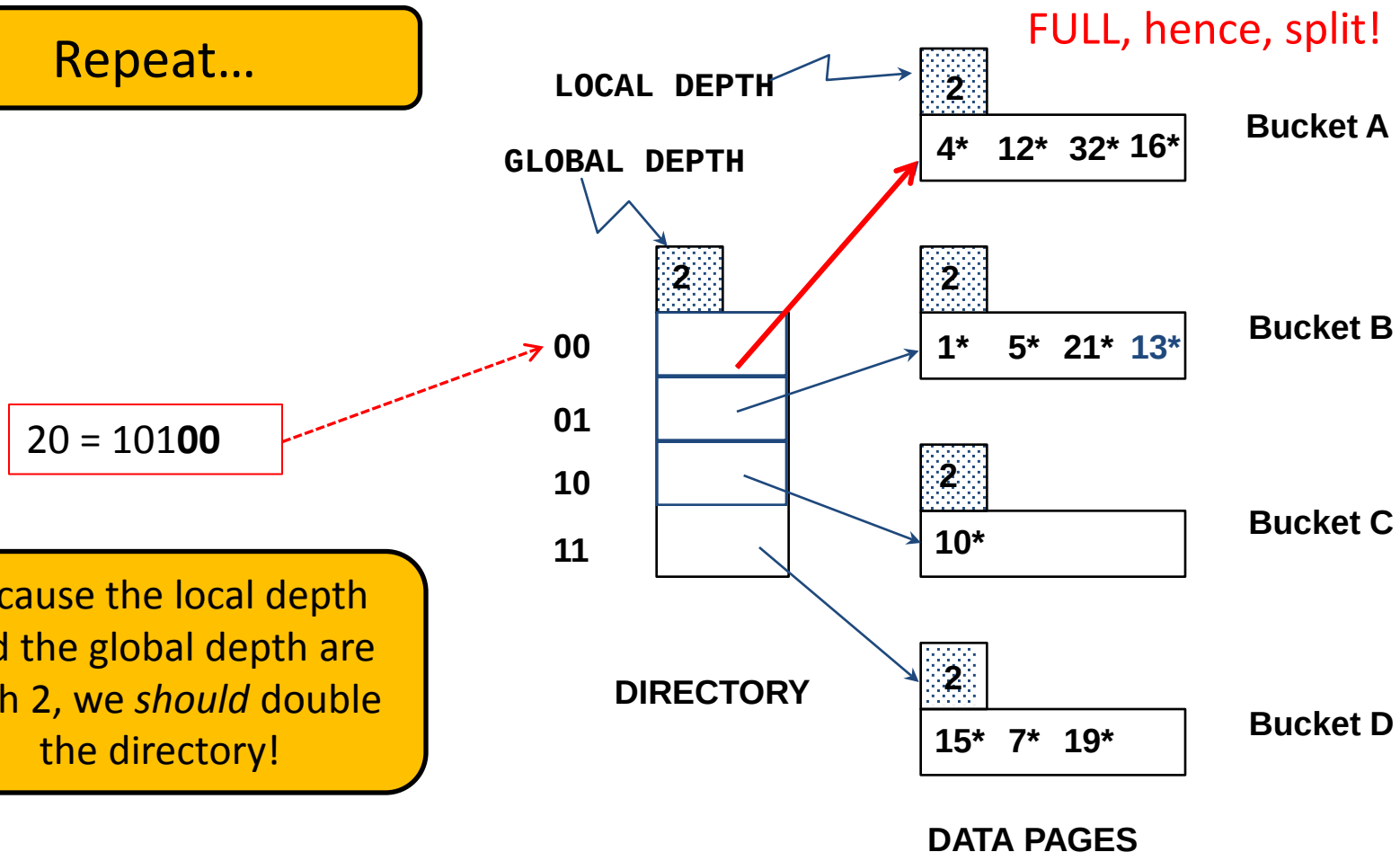
Repeat…



Extendible Hashing: Inserting Entries

- Example: insert 20*

Repeat…

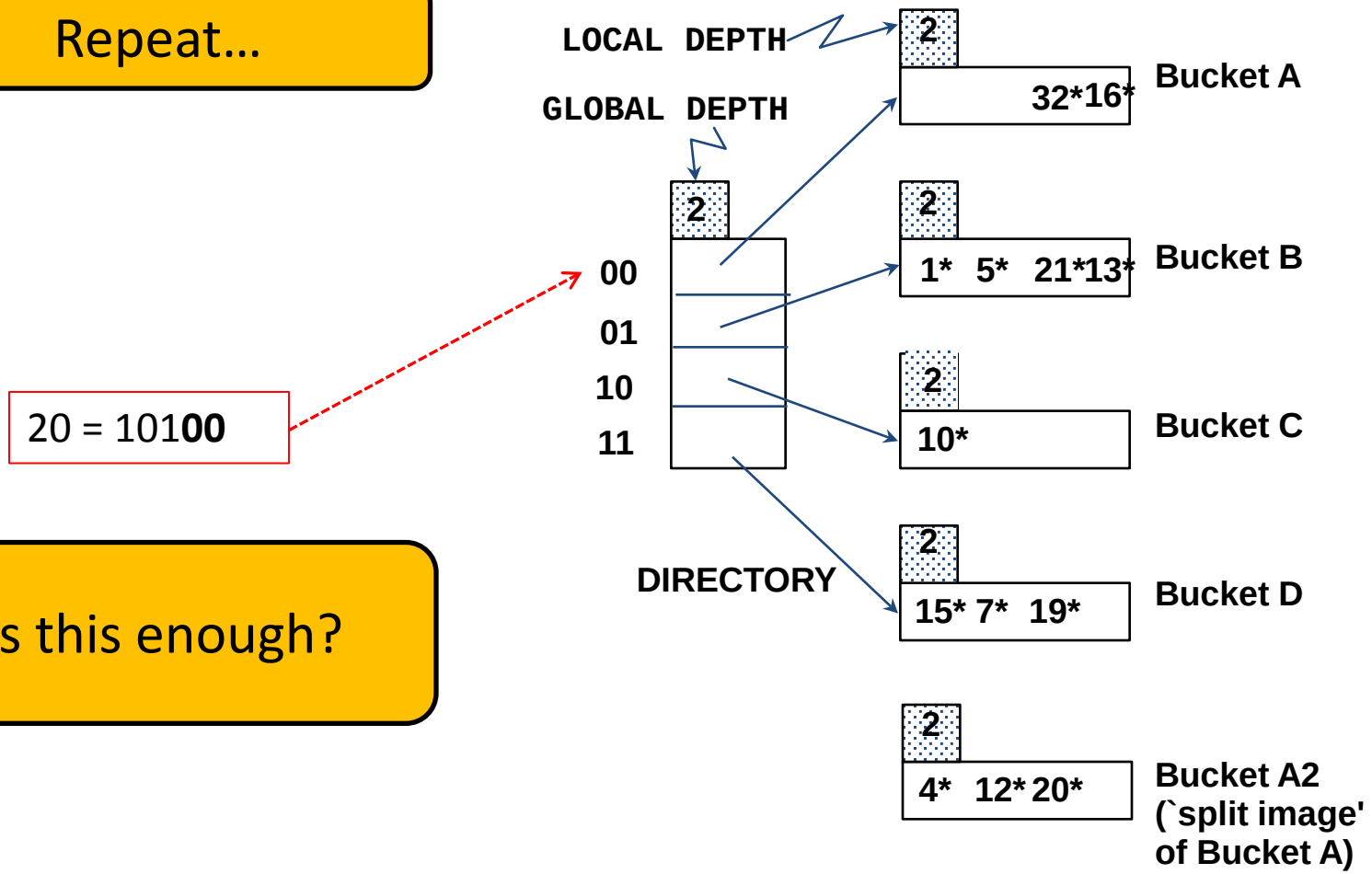


Because the local depth and the global depth are both 2, we *should* double the directory!

Extendible Hashing: Inserting Entries

- Example: insert 20*

Repeat…



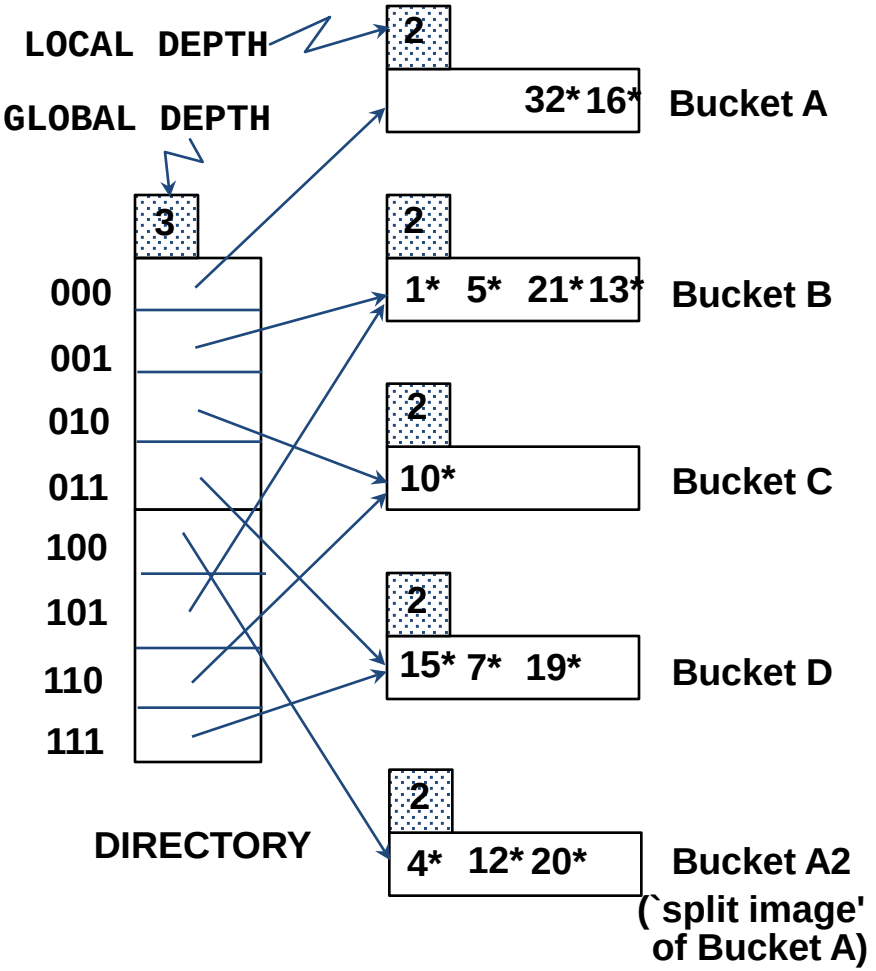
Is this enough?

Extendible Hashing: Inserting Entries

- Example: insert 20*

Repeat…

Is this enough?

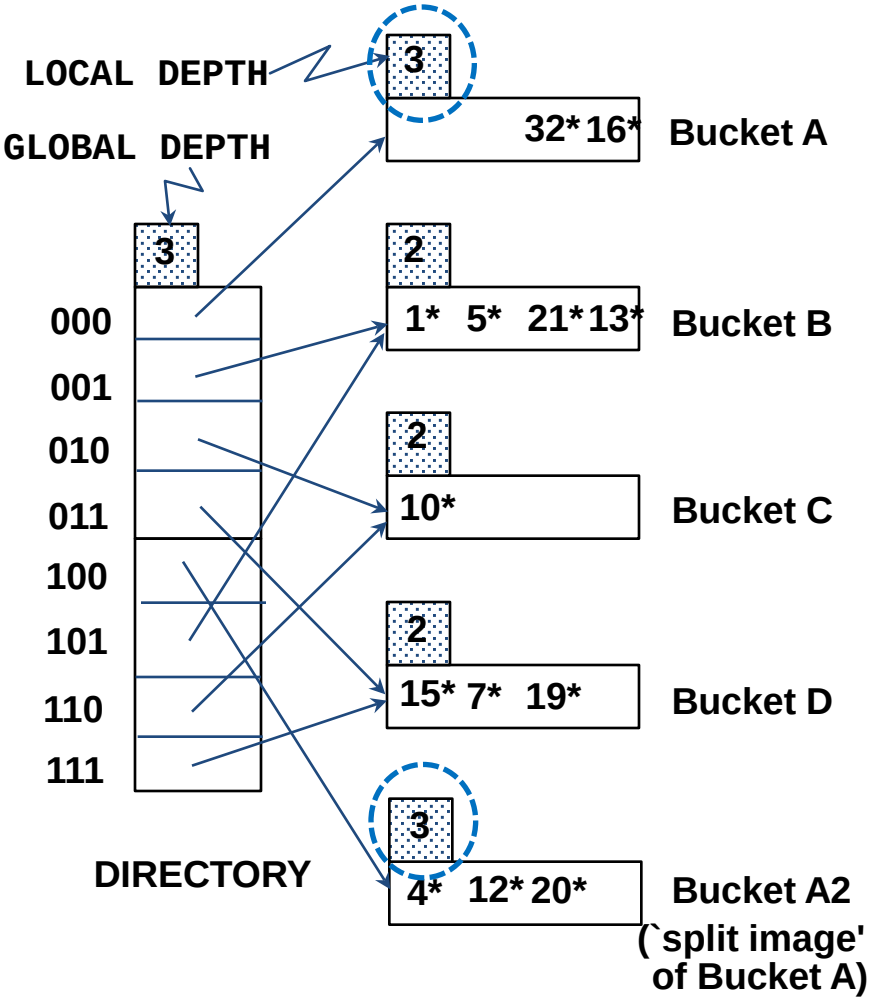


Extendible Hashing: Inserting Entries

- Example: insert 20*

Repeat…

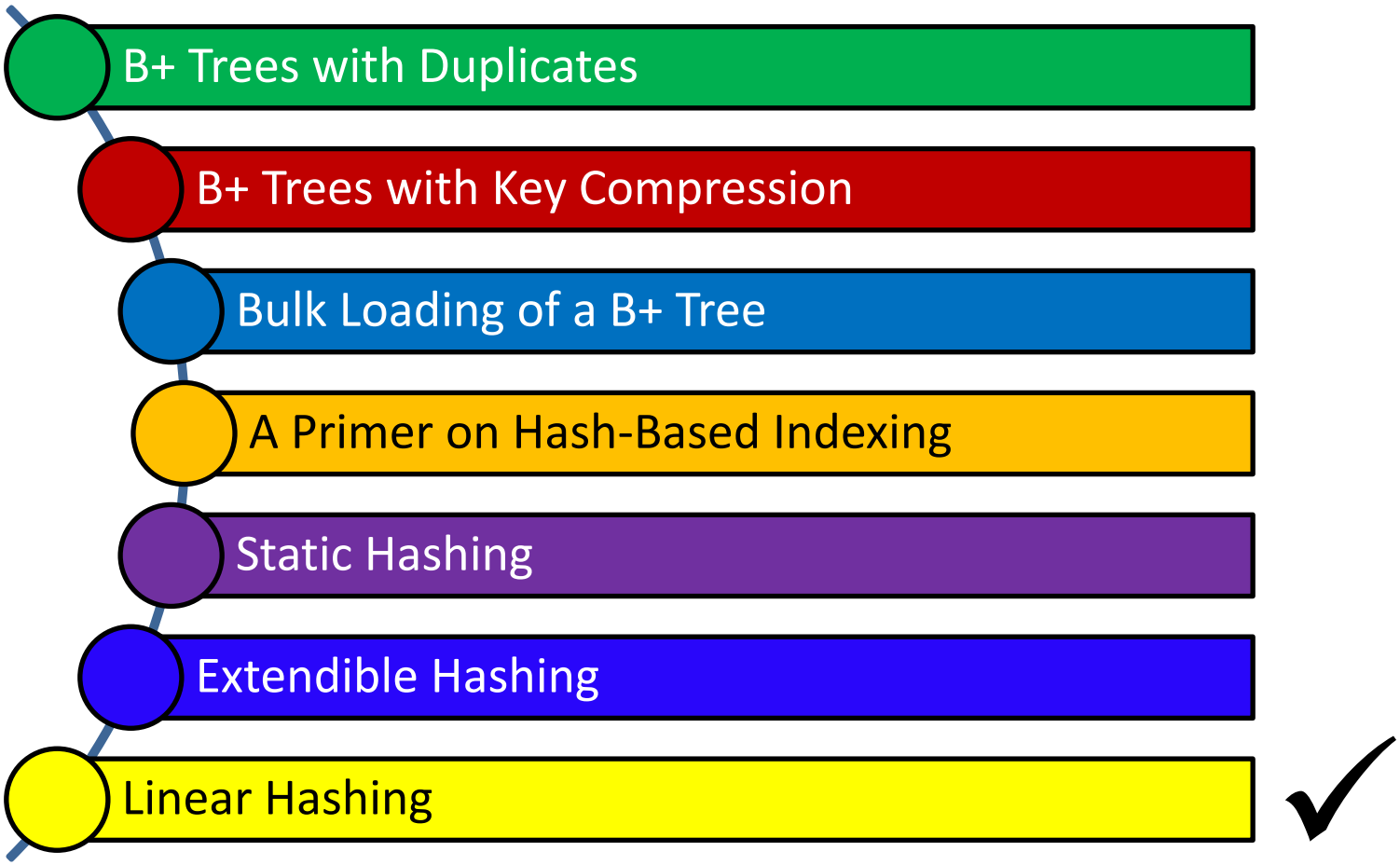
FINAL STATE!



Extendible Hashing: Deleting Entries

- For a deletion, the data entry is located and removed
- If the deletion leaves the bucket empty, it can be *merged* with its split image
  - Merging buckets decreases the local depth
- If each directory element points to the same bucket as its split image, the directory can be *halved* and the global depth decremented
- The insertion examples can be worked out backwards as examples of deletions!

Outline



جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Linear Hashing

- Another way of adapting gracefully to insertions and deletions (i.e., pursuing dynamic hashing) is to use Linear Hashing (LH)

- In contrast to Extendible Hashing, LH
  - Does not require a directory
  - Deals naturally with collisions
  - Offers a lot of flexibility w.r.t the timing of bucket split (allowing trading off greater overflow chains for higher average space utilization)

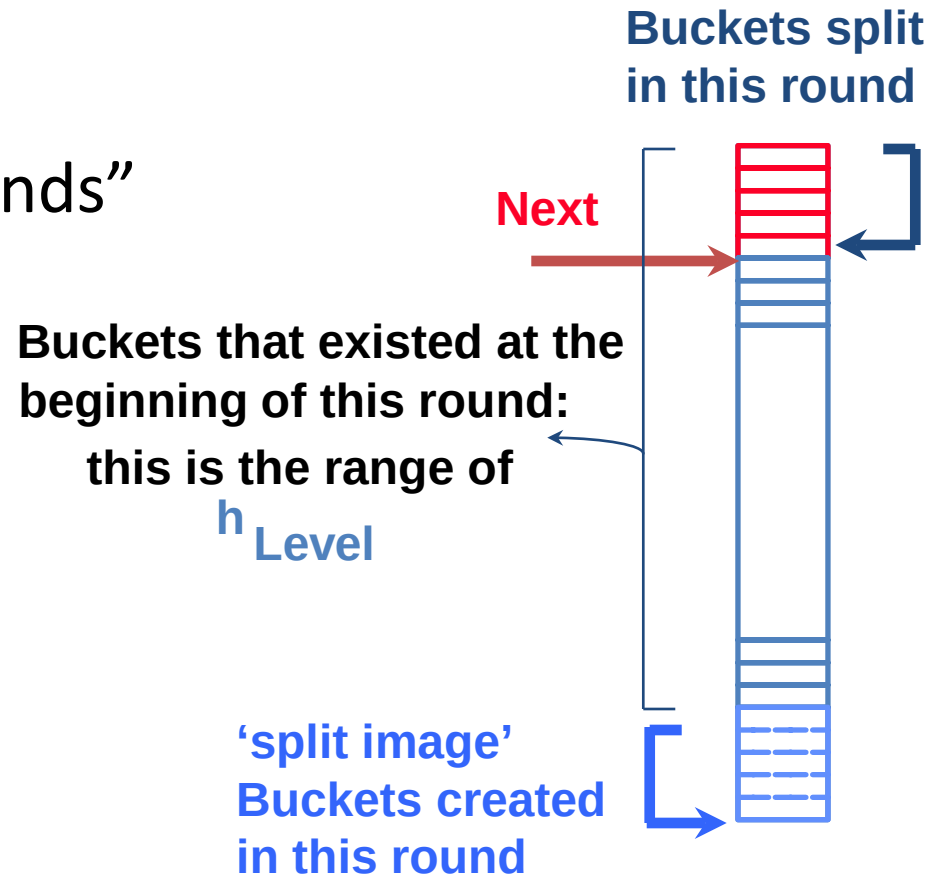
How Linear Hashing Works?

- LH uses a family of hash functions $\boldsymbol{h_0}$, $\boldsymbol{h_1}$, $\boldsymbol{h_2}$, ...
  - $\boldsymbol{h_i}(key) = \boldsymbol{h}(key) \bmod (2^i N)$; N = initial # buckets
  - $\boldsymbol{h}$ is some hash function (range is *not* 0 to N-1)
  - If $N = 2^{d0}$, for some *d0*, $\boldsymbol{h_i}$ consists of applying $\boldsymbol{h}$ and looking at the last *di* bits, where *di* = *d0* + *I*
  - $\boldsymbol{h_{i+1}}$ doubles the range of $\boldsymbol{h_i}$ (*similar to directory doubling*)

How Linear Hashing Works? (Cont'd)

- LH uses overflow pages, and chooses buckets to split in a *round-robin* fashion

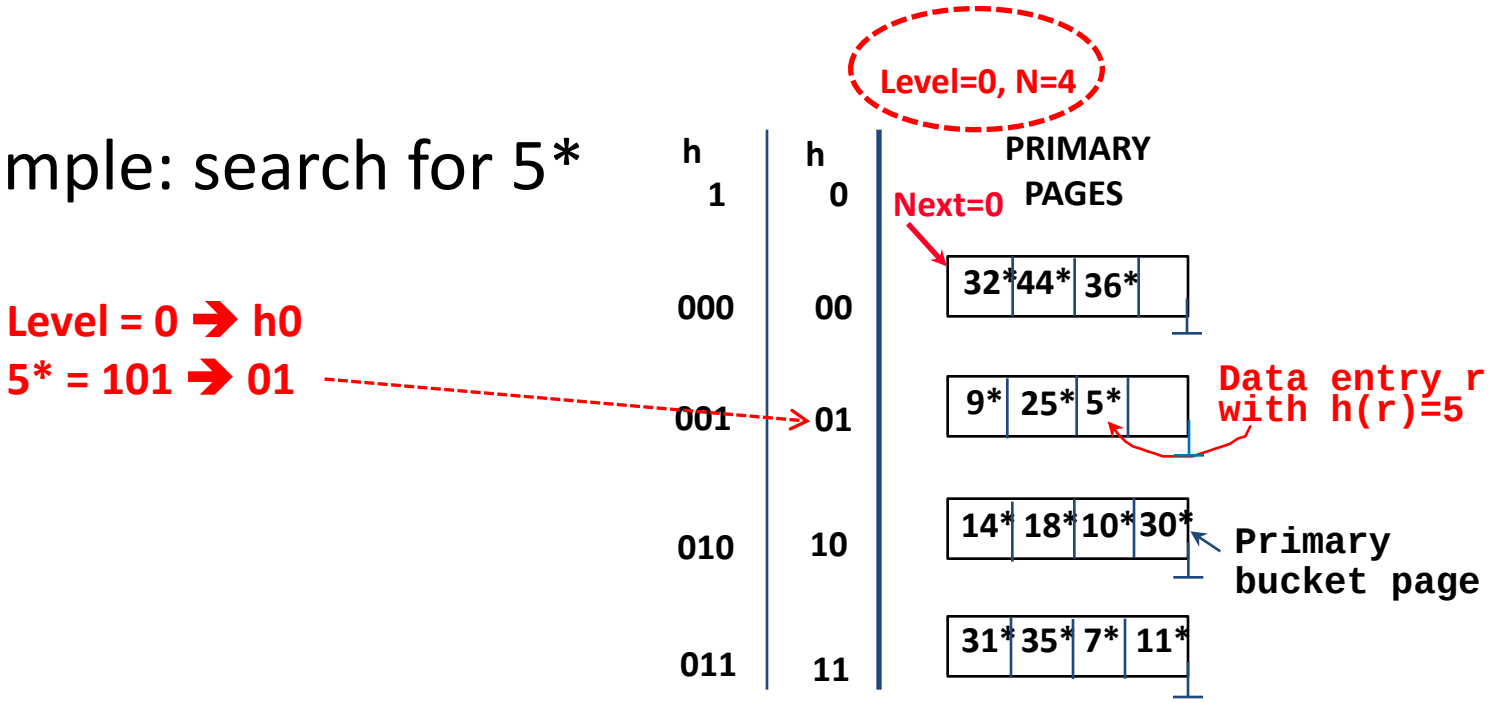
- Splitting proceeds in "rounds"
  - A round ends when all N_R (for round R) initial buckets are split
  - Buckets 0 to *Next-1* have been split; *Next* to N_R yet to be split
  - Current round number is referred to as *Level*



Linear Hashing: Searching For Entries

- To find bucket for data entry *r,* find $\boldsymbol{h}_{Level}(r)$*:*
  - If $\boldsymbol{h}_{Level}(r)$ in range `*Next* to N_R' , *r* belongs there
  - Else, *r* could belong to bucket $\boldsymbol{h}_{Level}(r)$ or bucket $\boldsymbol{h}_{Level}(r) + N_R$; must apply $\boldsymbol{h}_{Level+1}(r)$ to find out

- Example: search for 5*



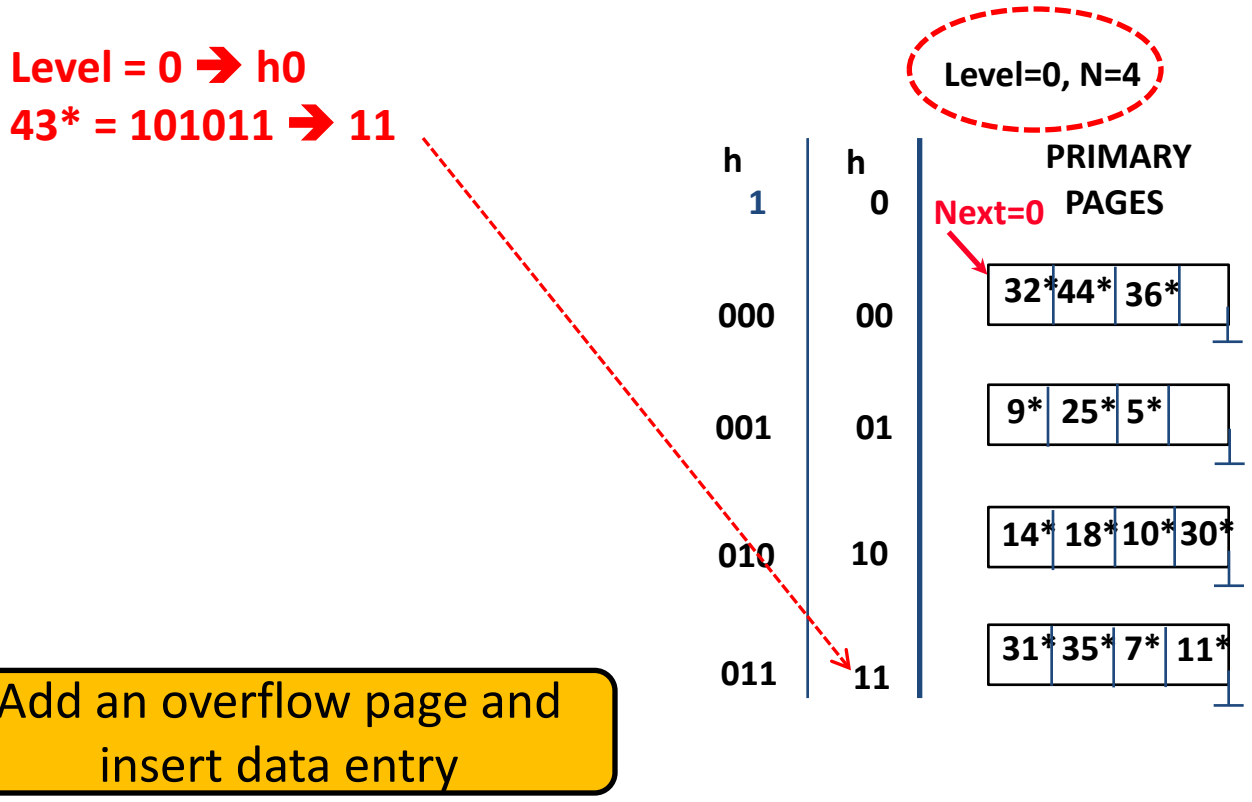
Linear Hashing: Inserting Entries

- Find bucket as in search
  - If the bucket to insert the data entry into is full:
    - Add an overflow page and insert data entry
    - (*Maybe*) Split *Next* bucket and increment *Next*

- Some points to Keep in mind:
  - Unlike Extendible Hashing, when an insert triggers a split, the bucket into which the data entry is inserted is not necessarily the bucket that is split
  - As in Static Hashing, an overflow page is added to store the newly inserted data entry
  - However, since the bucket to split is chosen in a round-robin fashion, eventually *all* buckets will be split

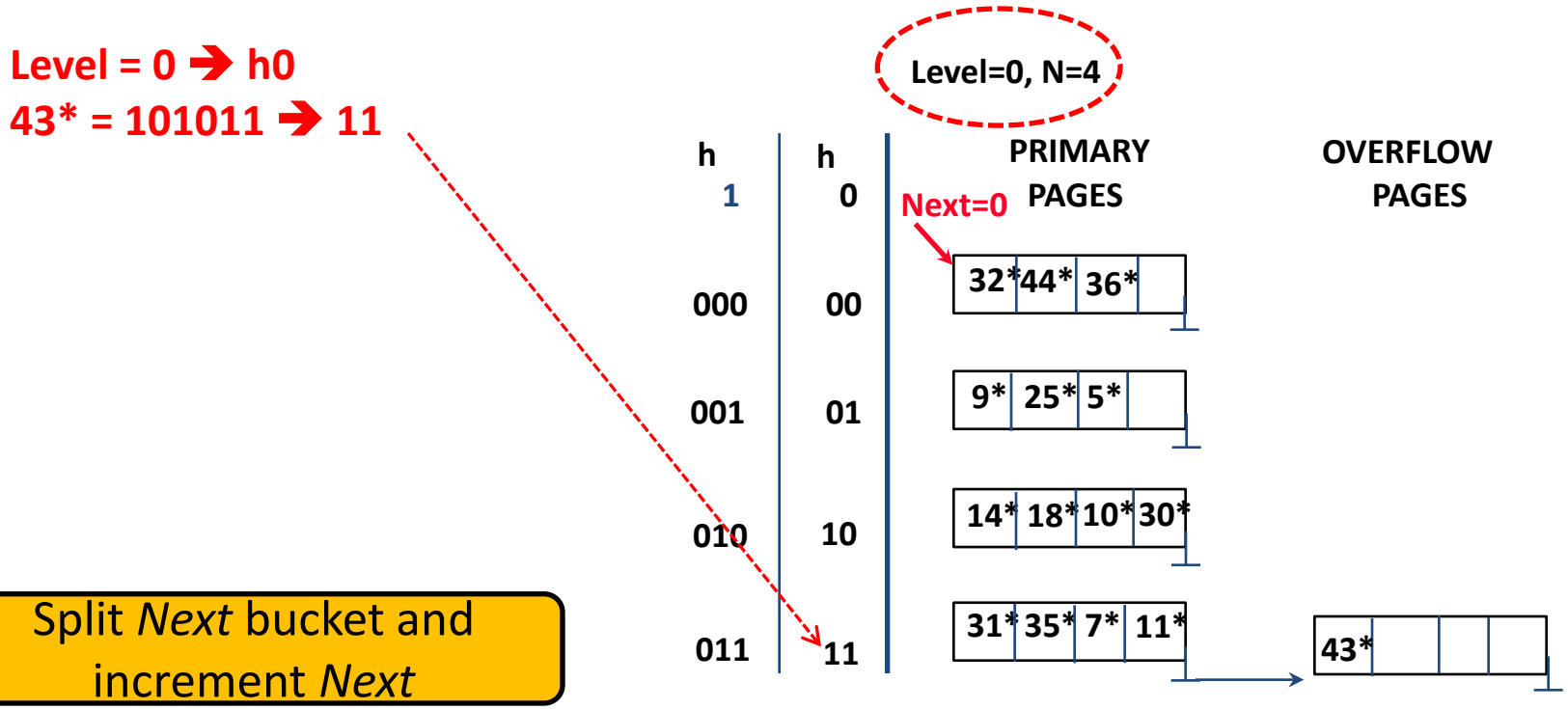
Linear Hashing: Inserting Entries

- Example: insert 43*



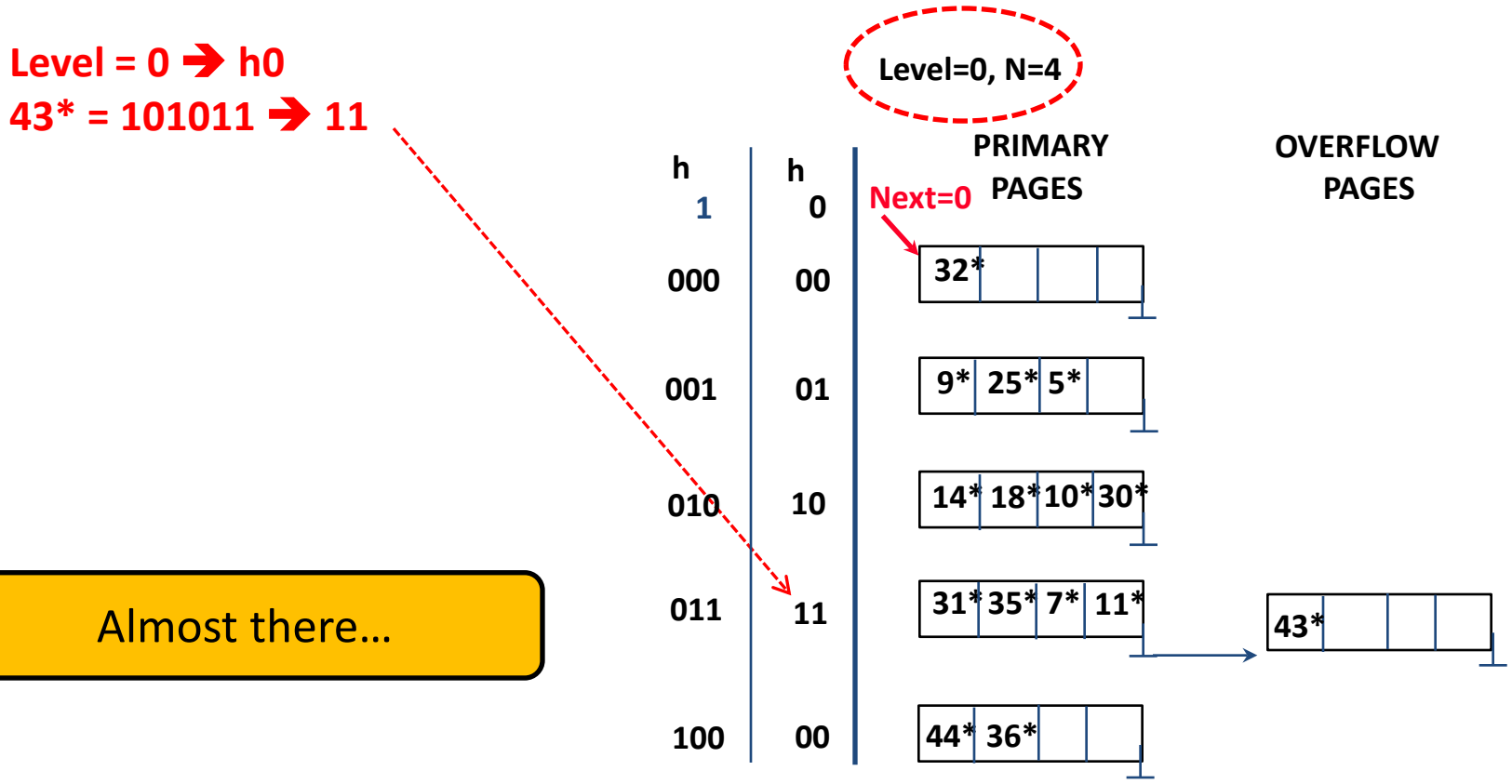
Linear Hashing: Inserting Entries

- Example: insert 43*



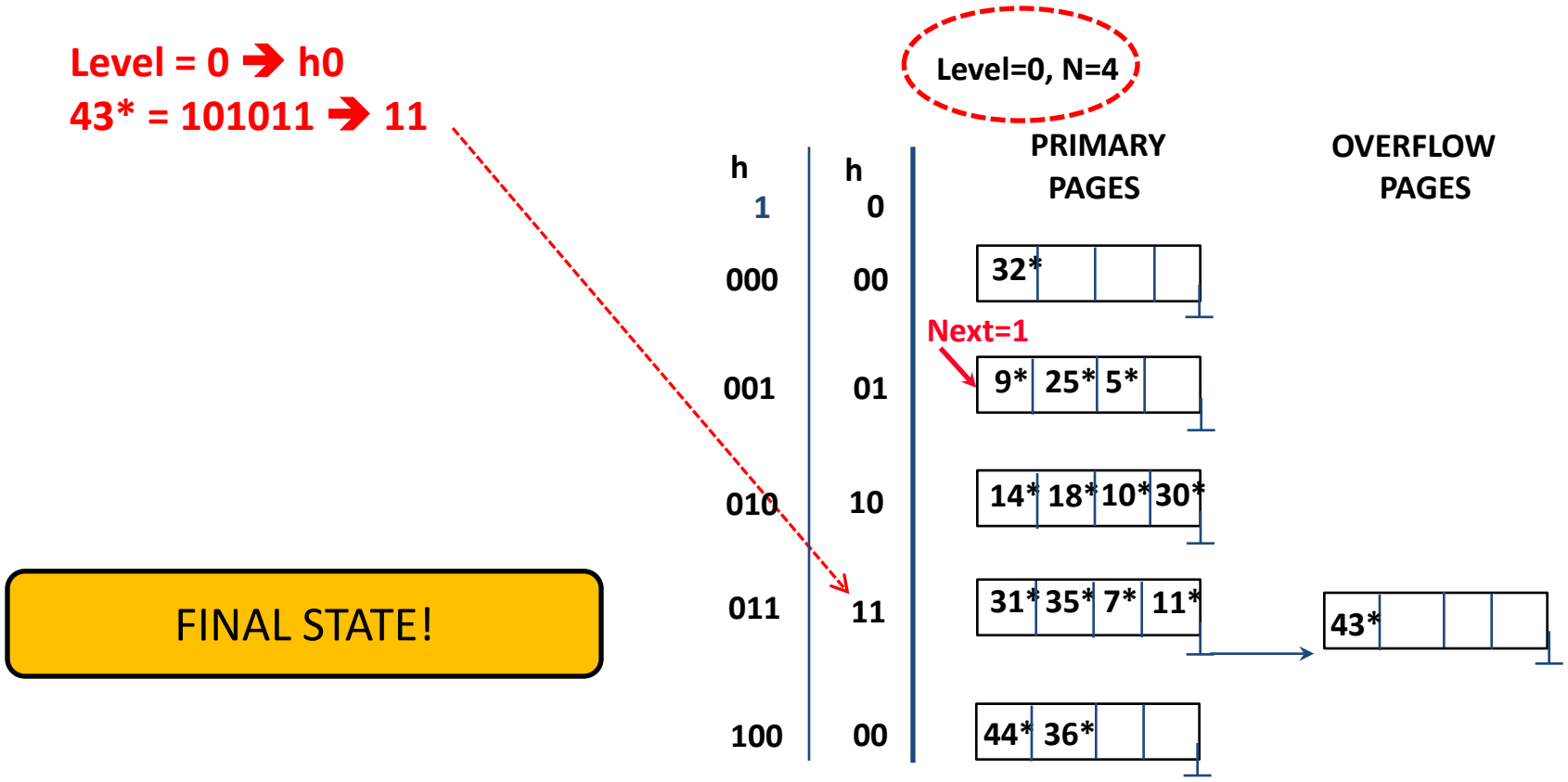
Linear Hashing: Inserting Entries

- Example: insert 43*



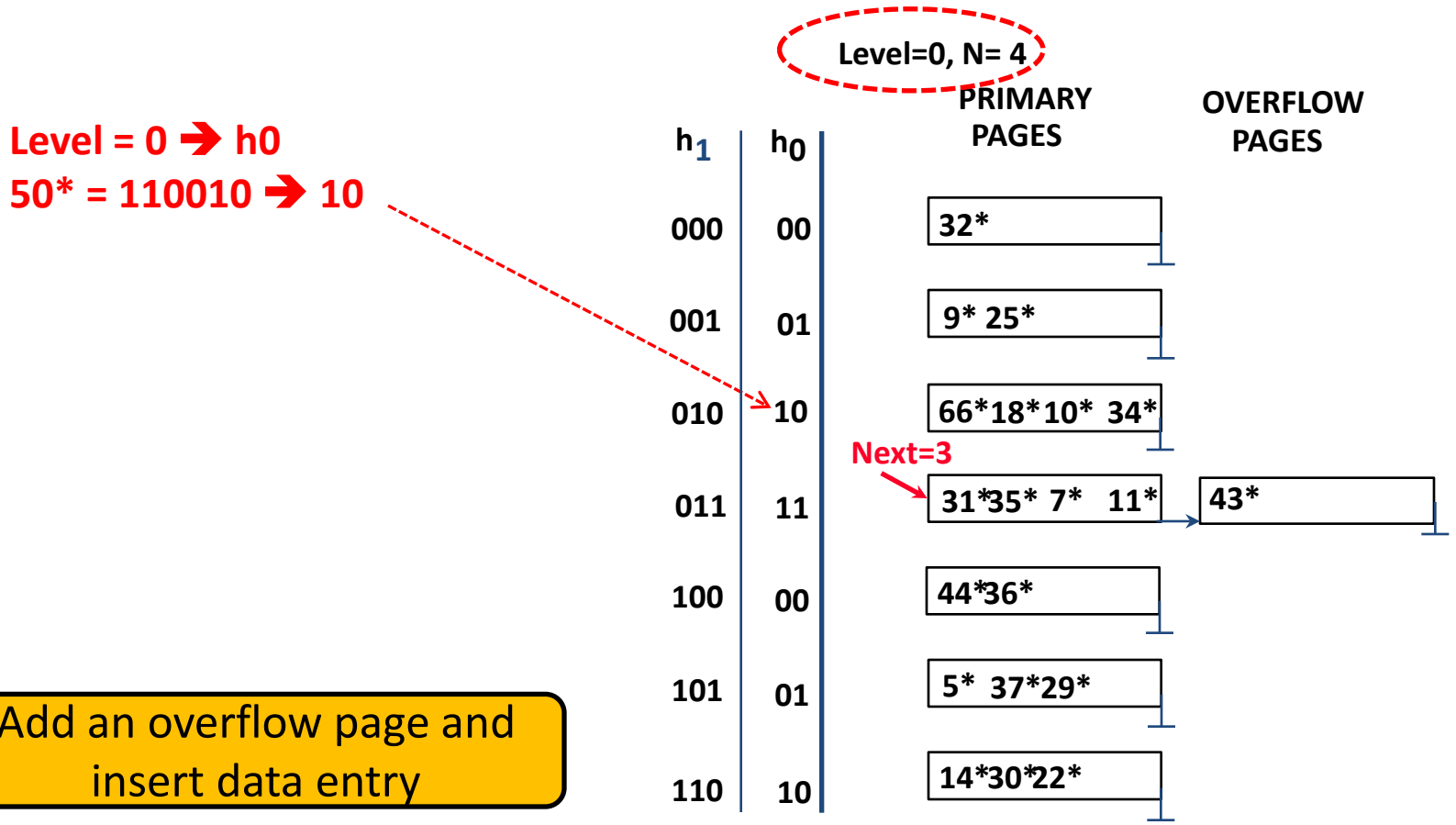
Linear Hashing: Inserting Entries

- Example: insert 43*



Linear Hashing: Inserting Entries

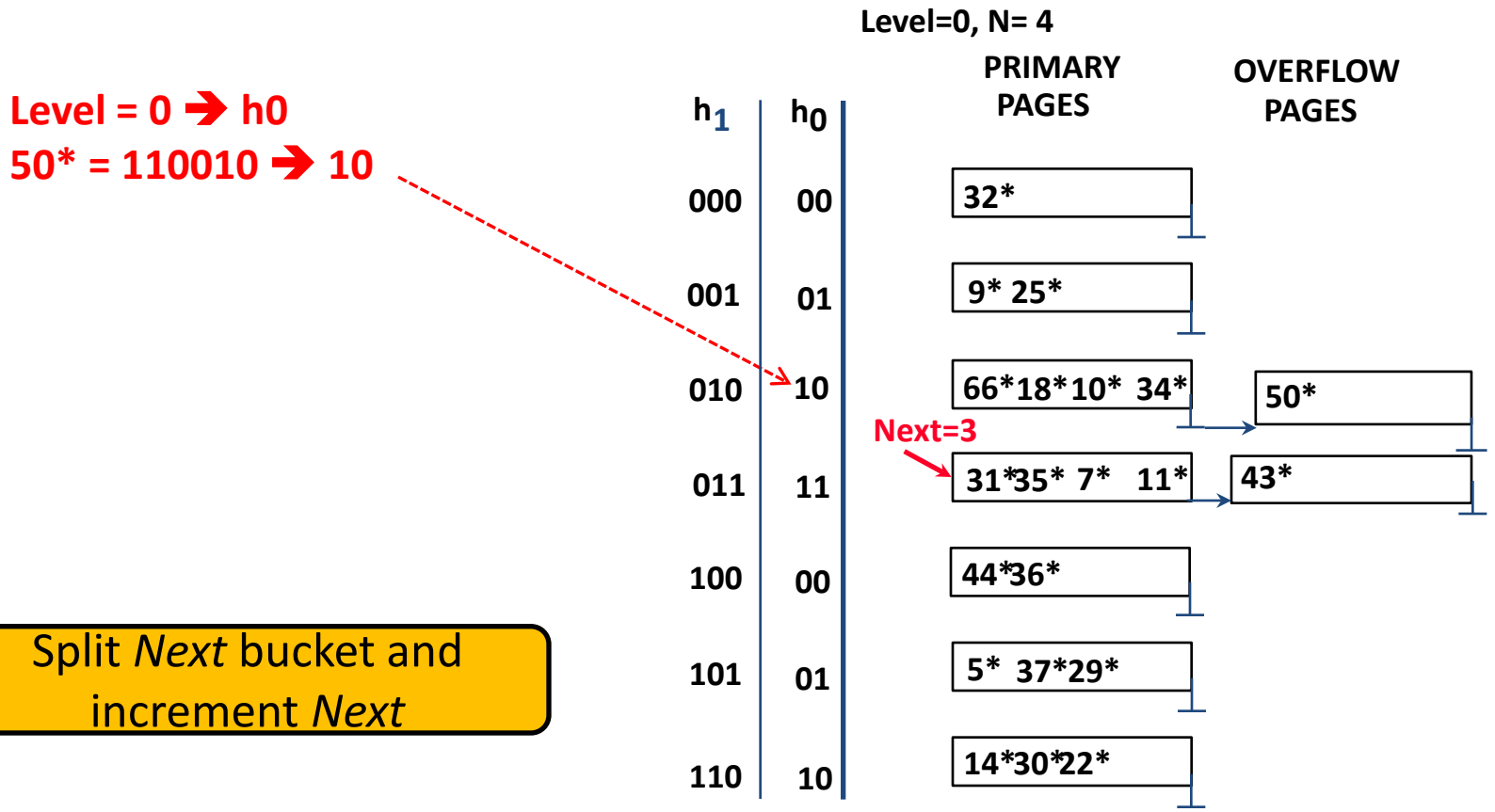
- Another Example: insert 50*



Add an overflow page and insert data entry

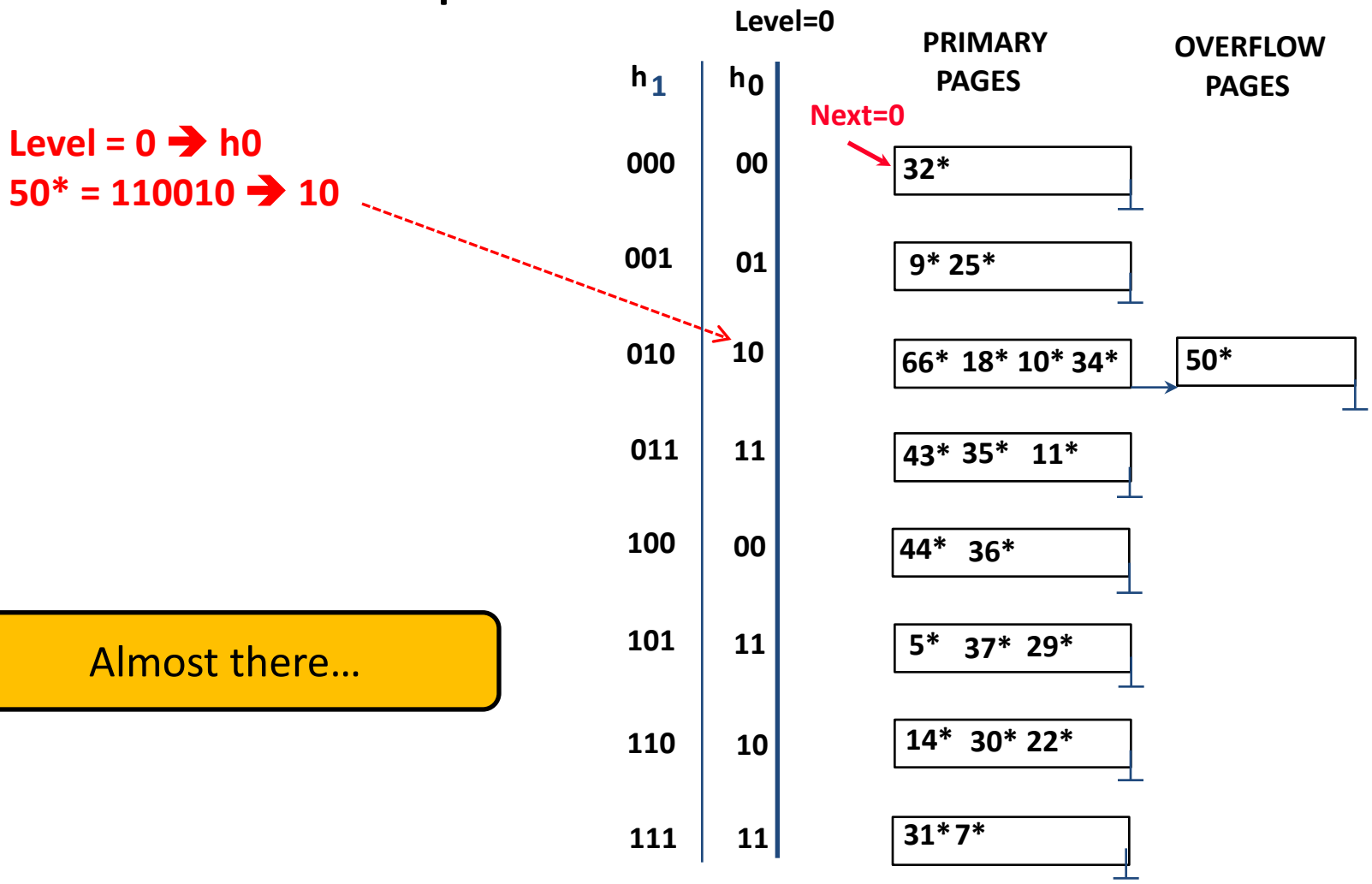
Linear Hashing: Inserting Entries

- Another Example: insert 50*



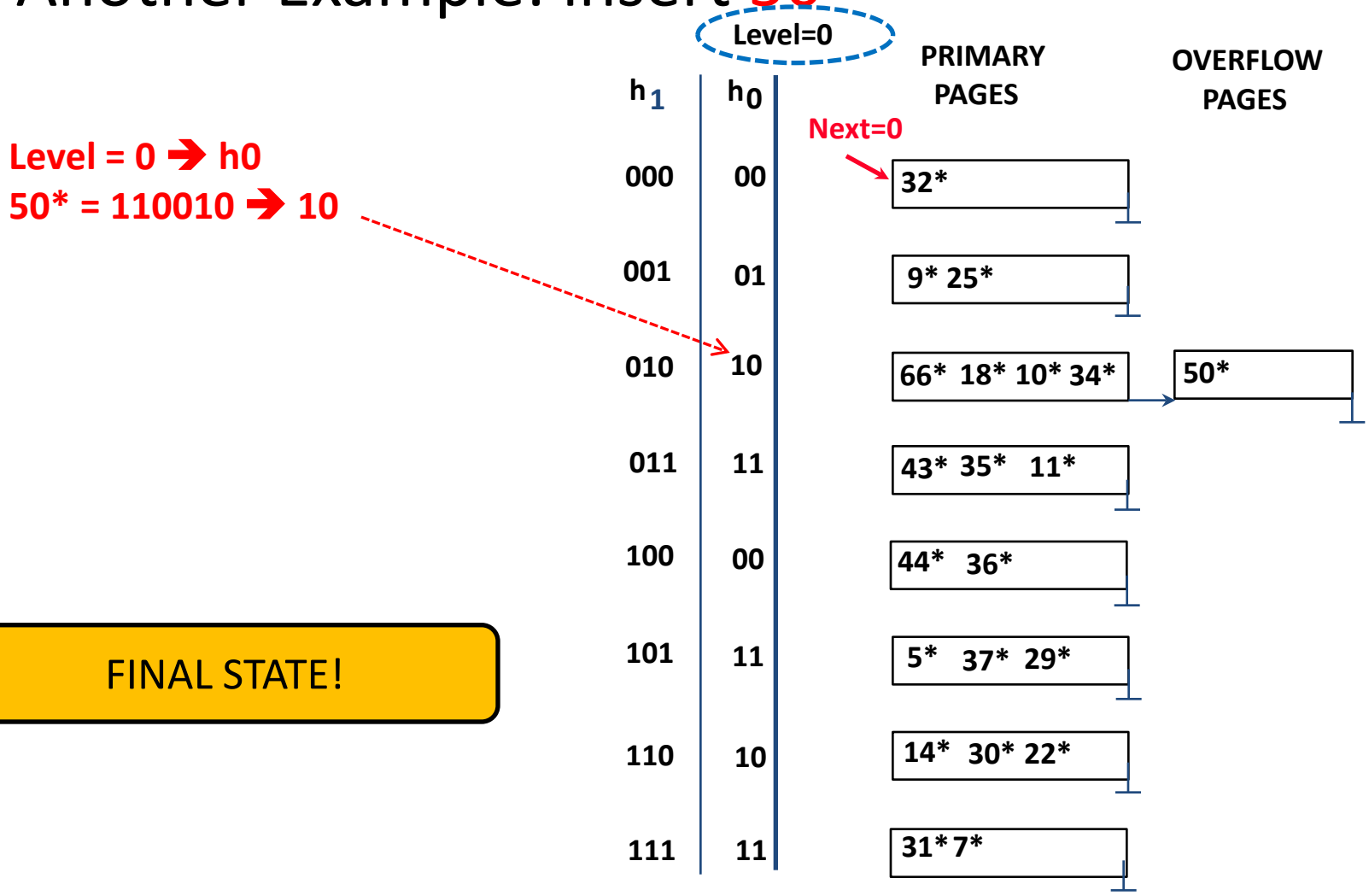
Linear Hashing: Inserting Entries

- Another Example: insert 50*



Linear Hashing: Inserting Entries

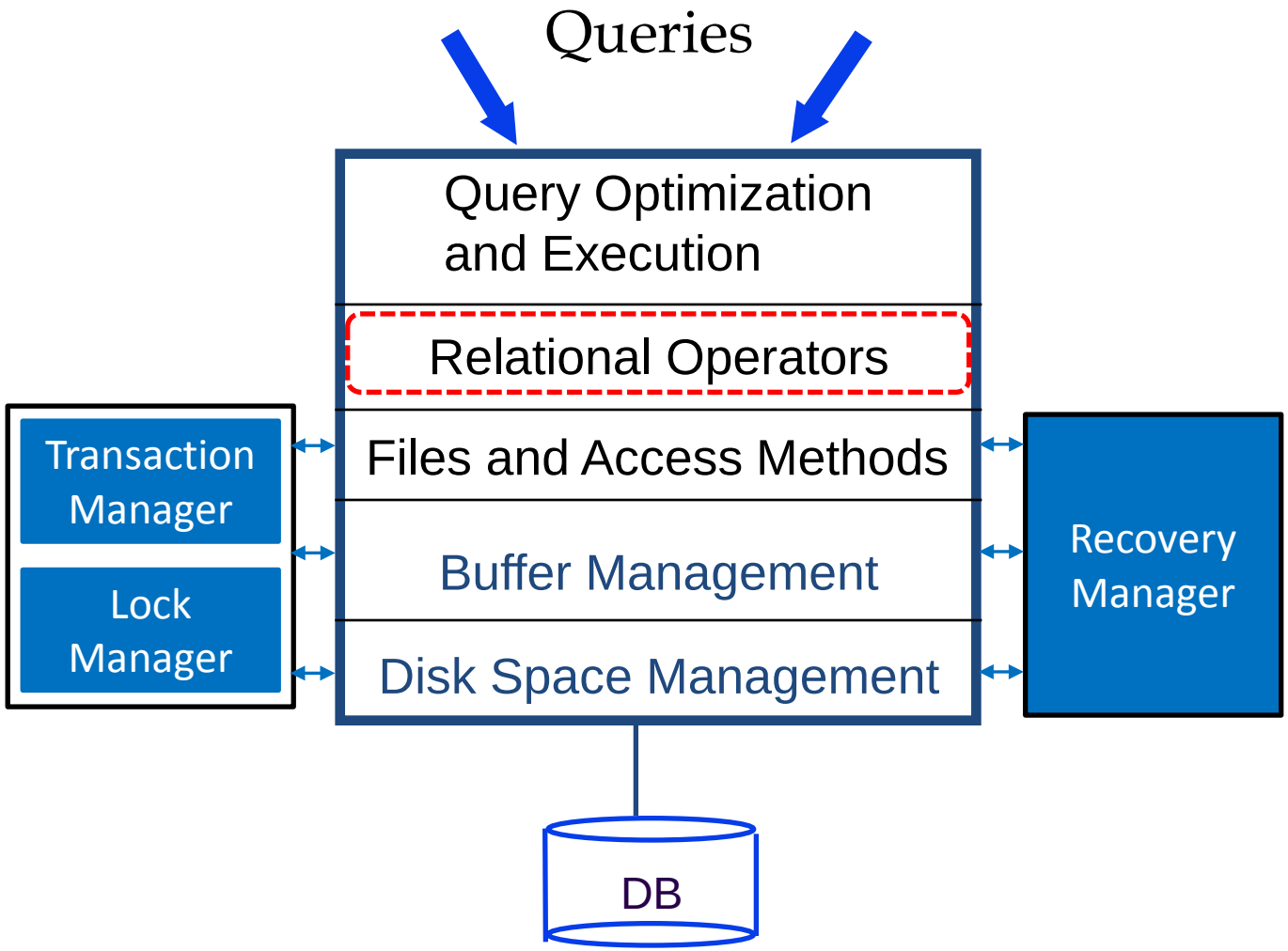
- Another Example: insert 50*



Linear Hashing: Deleting Entries

- Deletion is essentially the inverse of insertion
- If the last bucket in the file is empty, it can be removed and *Next* can be decremented
- If *Next* is zero and the last bucket becomes empty
  - *Next* is made to point to bucket ***M***/2 -1 (where ***M*** is the current number of buckets)
  - *Level* is decremented
  - The empty bucket is removed
- The insertion examples can be worked out backwards as examples of deletions!

Next Class


Queries
Query Optimization and Execution
Relational Operators
Files and Access Methods
Buffer Management
Disk Space Management
Transaction Manager
Lock Manager
Recovery Manager
DB