CS 6530: Advanced Database Systems Fall 2022

# Lecture 03
# In-memory indexing
# (Trees, Tries, Skip Lists)

Prashant Pandey

prashant.pandey@utah.edu

# Some reminders…

- We have a course website:
  - https://www.cs.utah.edu/~pandey/courses/cs6530/fall22/index.html
- Paper report deadlines are posted
- Project #1 logistics should be available the end this week
- Join online discussion through Piazza

  https://piazza.com/utah/fall2022/cs6530001fall2022

# Announcement

- The School of Computing would like to send a few student representatives to the 2022 Rocky Mountain Celebration of Women in Computing.
- The conference is September 29-30, 2022 in Boulder, CO. For more information, see
  - https://toilers.mines.edu/RMCWiC/2022/home.html
- Students who attend will present a research poster.
- Travel expenses and conference registration will be paid by the School.
- If you would like to be considered, send a current resume, and a title and abstract for your poster.
- Please send text or PDF documents only. The deadline is 5p Thursday, September 8.
- If you have any questions, don't hesitate to ask Prof. Tucker Hermans.

# In-memory indexing

| Bandwidth | Latency | Size | |
|---|---|---|---|
| | | | CPU |
| 1 TB/sec | 1-4 n sec | 256 KB | on-chip cache |
| ~400 GB/sec | ~30 n sec | 8 - 12 MB | on-board cache |
| ~50 GB/sec | 80-100 n sec | 4 GB - 1 TB | main memory |



Acknowledgement: Prof. Manos Athanassoulis, BU

# In-memory indexing

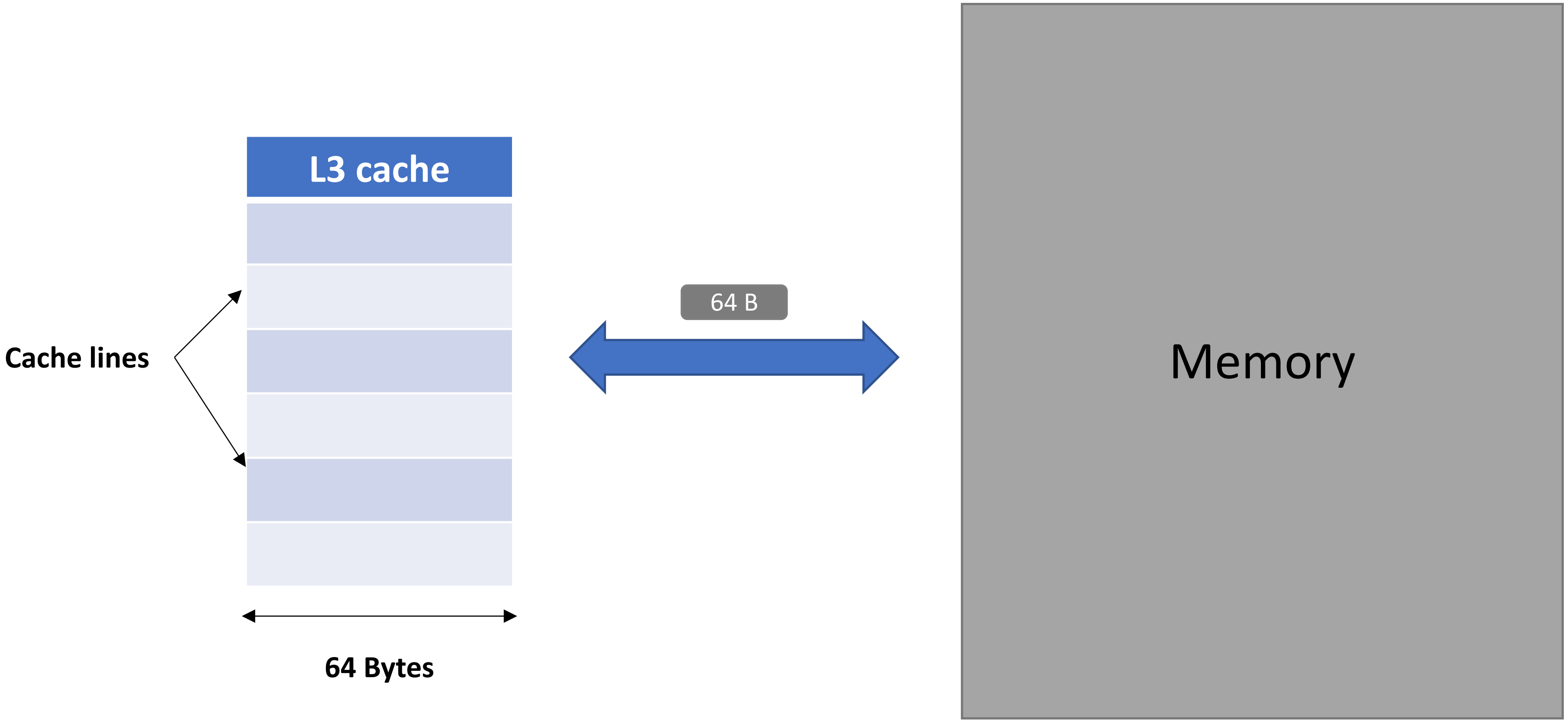
| Bandwidth | Latency | Size | |
|---|---|---|---|
| | | | CPU |
| 1 TB/sec | 1-4 n sec | 256 KB | on-chip cache |
| ~400 GB/sec | ~30 n sec | 8 - 12 MB | on-board cache |
| ~50 GB/sec | 80-100 n sec | 4 GB - 1 TB | main memory |

**Data transfer between L3 and DRAM is the dominating factor.**



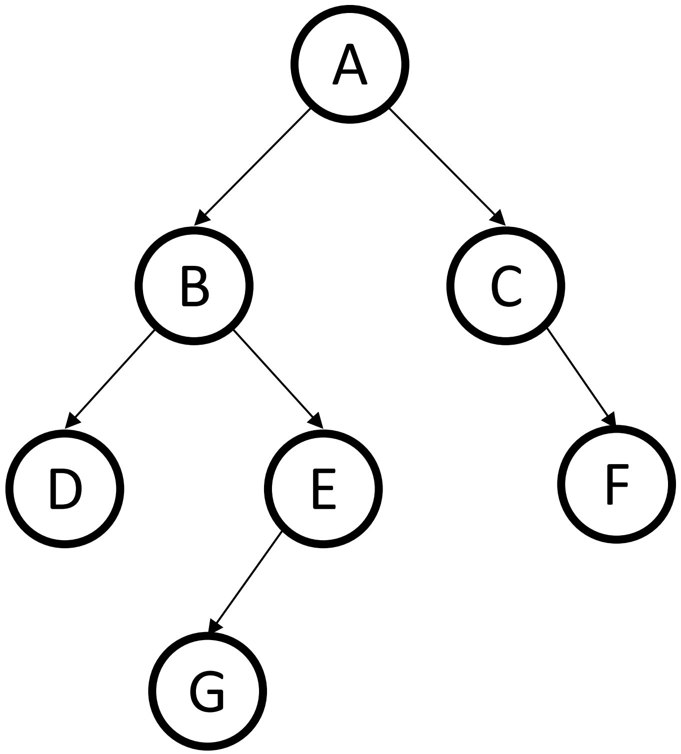
Acknowledgement: Prof. Manos Athanassoulis, BU

# Data movement b/w L3 cache and memory



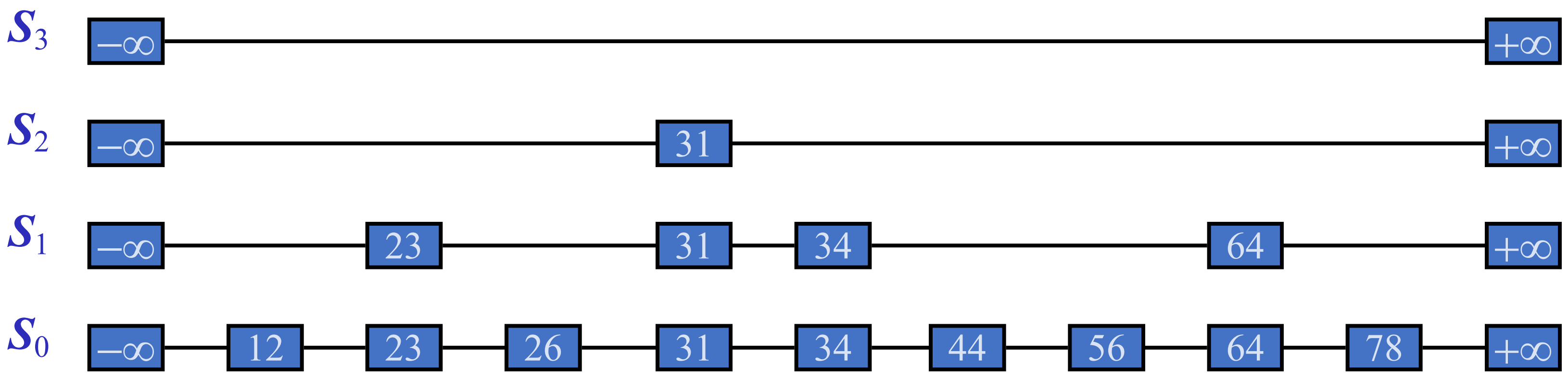
# Binary tree

- Binary tree: Each node has at most 2 children (branching factor 2)
- Binary tree is
  - A root *(with data)*
  - A left subtree *(may be empty)*
  - A right subtree *(may be empty)*
- Search/Insert/Delete O($\log(N)$)



# What is a Skip List

- A skip list for a set $S$ of distinct (key, element) items is a series of lists $S_0, S_1, \ldots, S_h$ such that
  - Each list $S_i$ contains the special keys $+\infty$ and $-\infty$
  - List $S_0$ contains the keys of $S$ in non-decreasing order
  - Each list is a subsequence of the previous one, i.e.,
    $$S_0 \supseteq S_1 \supseteq \ldots \supseteq S_h$$
  - List $S_h$ contains only the two special keys
- Skip lists are one way to implement the dictionary ADT



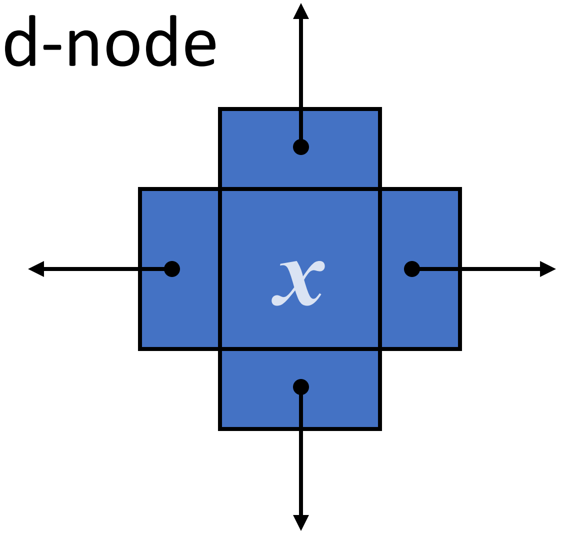



Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Implementation

- We can implement a skip list with quad-nodes
- A quad-node stores:
  - item
  - link to the node before
  - link to the node after
  - link to the node below
- Also, we define special keys PLUS_INF and MINUS_INF, and we modify the key comparator to handle them

quad-node






Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

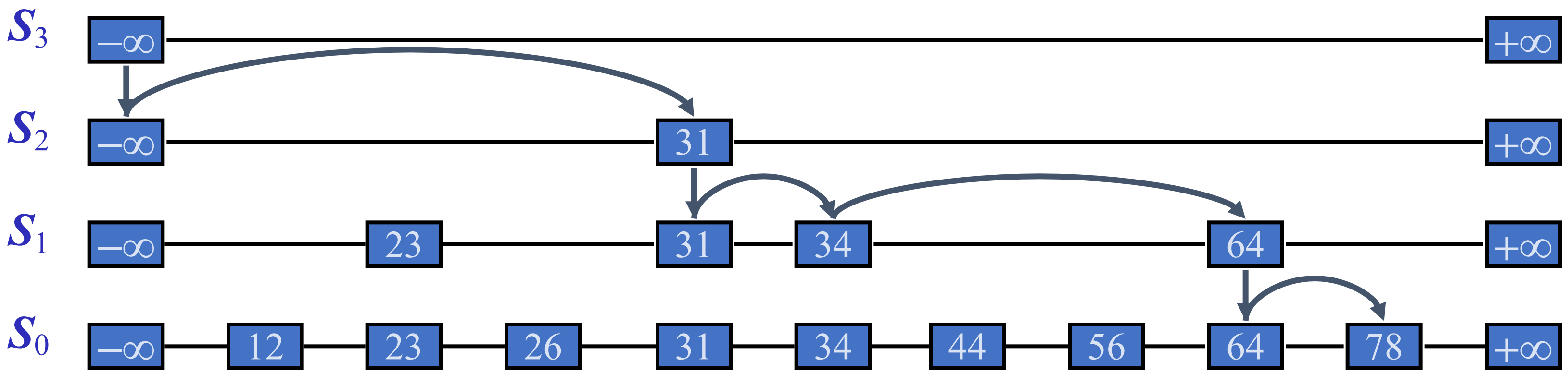
# Search

- We search for a key $x$ in a a skip list as follows:
  - We start at the first position of the top list
  - At the current position $p$, we compare $x$ with $y \leftarrow key(after(p))$

    $x = y$: we return $element(after(p))$

    $x > y$: we "scan forward"

    $x < y$: we "drop down"
  - If we try to drop down past the bottom list, we return $NO\_SUCH\_KEY$
- Example: search for 78

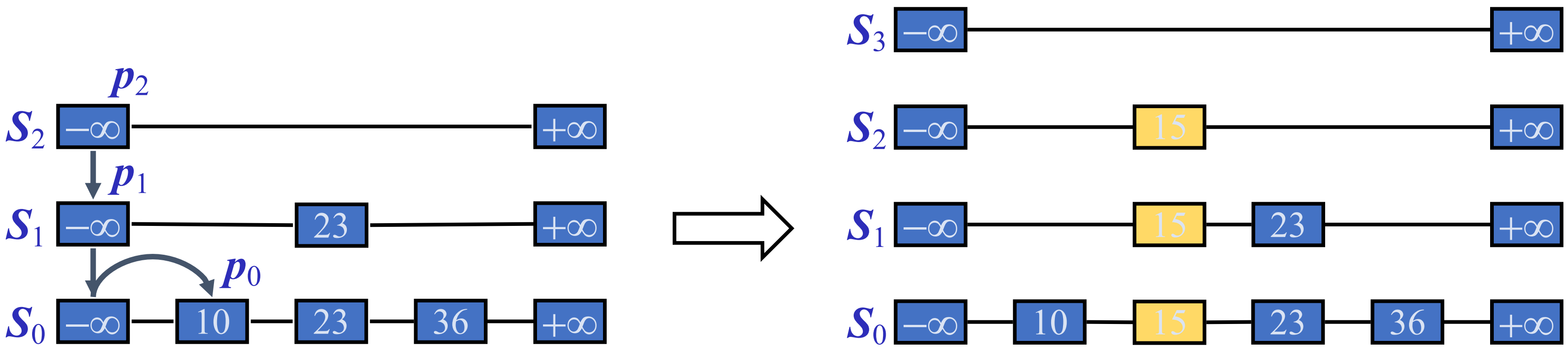





Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Insertion

- To insert an item $(x, o)$ into a skip list, we use a randomized algorithm:
  - We repeatedly toss a coin until we get tails, and we denote with $i$ the number of times the coin came up heads
  - If $i \geq h$, we add to the skip list new lists $S_{h+1}, \ldots, S_{i+1}$, each containing only the two special keys
  - We search for $x$ in the skip list and find the positions $p_0, p_1, \ldots, p_i$ of the items with largest key less than $x$ in each list $S_0, S_1, \ldots, S_i$
  - For $j \leftarrow 0, \ldots, i$, we insert item $(x, o)$ into list $S_j$ after position $p_j$
- Example: insert key $15$, with $i = 2$

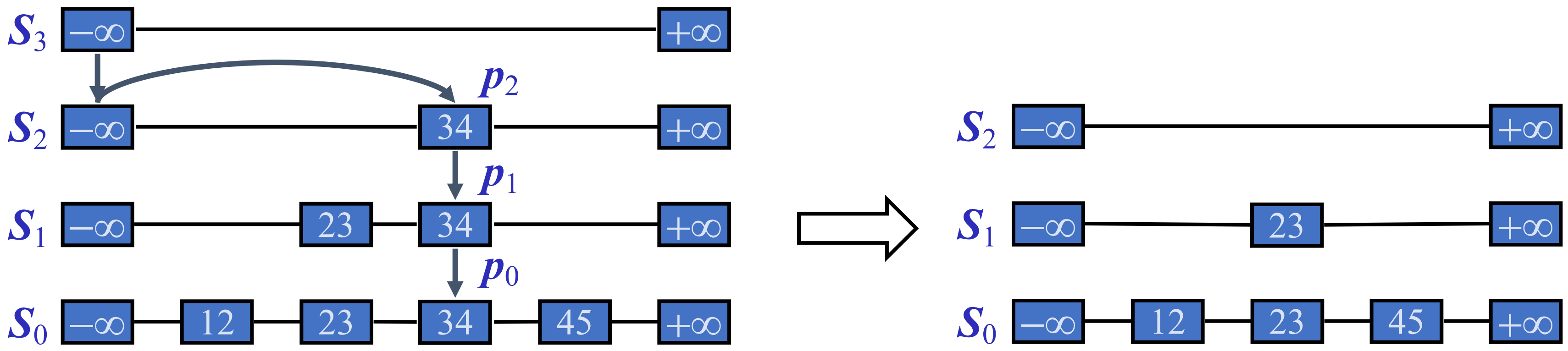





Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Deletion

- To remove an item with key $\boldsymbol{x}$ from a skip list, we proceed as follows:
  - We search for $\boldsymbol{x}$ in the skip list and find the positions $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_i$ of the items with key $\boldsymbol{x}$, where position $\boldsymbol{p}_j$ is in list $\boldsymbol{S}_j$
  - We remove positions $\boldsymbol{p}_0, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_i$ from the lists $\boldsymbol{S}_0, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_i$
  - We remove all but one list containing only the two special keys
- Example: remove key 34






Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Randomized Algorithms

- A randomized algorithm controls its execution through random selection (e.g., coin tosses)
- It contains statements like:

$$
\begin{aligned}
&\boldsymbol{b} \leftarrow \boldsymbol{randomBit}()\\
&\textbf{if }\ \boldsymbol{b} = 0\\
&\quad \text{do A} \ldots\\
&\textbf{else } \{ \boldsymbol{b} = 1\}\\
&\quad \text{do B} \ldots
\end{aligned}
$$

- Its running time depends on the outcomes of the coin tosses

- Through probabilistic analysis we can derive the expected running time of a randomized algorithm
- We make the following assumptions in the analysis:
  - the coins are unbiased
  - the coin tosses are independent
- The worst-case running time of a randomized algorithm is often large but has very low probability (e.g., it occurs when all the coin tosses give "heads")
- We use a randomized algorithm to insert items into a skip list to insert in expected O(log n)-time
- When randomization is used in data structures they are referred to as probabilistic data structures



Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Space Usage

- The space used by a skip list depends on the random bits used by each invocation of the insertion algorithm
- We use the following two basic probabilistic facts:

  Fact 1: The probability of getting $\boldsymbol{i}$ consecutive heads when flipping a coin is $1/2^{\boldsymbol{i}}$

  Fact 2: If each of $\boldsymbol{n}$ items is present in a set with probability $\boldsymbol{p}$, the expected size of the set is $\boldsymbol{np}$

- Consider a skip list with $\boldsymbol{n}$ items
  - By Fact 1, we insert an item in list $\boldsymbol{S_i}$ with probability $1/2^{\boldsymbol{i}}$
  - By Fact 2, the expected size of list $\boldsymbol{S_i}$ is $\boldsymbol{n}/2^{\boldsymbol{i}}$
- The expected number of nodes used by the skip list is

$$\sum_{i=0}^{h} \frac{n}{2^i} = n \sum_{i=0}^{h} \frac{1}{2^i} < 2n$$

- Thus, the expected space usage of a skip list with $\boldsymbol{n}$ items is $\boldsymbol{O}(\boldsymbol{n})$



Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Height

- The running time of the search and insertion algorithms is affected by the height $\boldsymbol{h}$ of the skip list
- We show that with high probability, a skip list with $\boldsymbol{n}$ items has height $\boldsymbol{O}(\log \boldsymbol{n})$
- We use the following additional probabilistic fact:

  Fact 3: If each of $\boldsymbol{n}$ events has probability $\boldsymbol{p}$, the probability that at least one event occurs is at most $\boldsymbol{np}$

- Consider a skip list with $\boldsymbol{n}$ items
  - By Fact 1, we insert an item in list $\boldsymbol{S_i}$ with probability $1/2^{\boldsymbol{i}}$
  - By Fact 3, the probability that list $\boldsymbol{S_i}$ has at least one item is at most $\boldsymbol{n}/2^{\boldsymbol{i}}$
- By picking $\boldsymbol{i} = 3\log \boldsymbol{n}$, we have that the probability that $\boldsymbol{S}_{3\log \boldsymbol{n}}$ has at least one item is at most

$$\boldsymbol{n}/2^{3\log \boldsymbol{n}} = \boldsymbol{n}/\boldsymbol{n}^3 = 1/\boldsymbol{n}^2$$

- Thus, a skip list with $\boldsymbol{n}$ items has height at most $3\log \boldsymbol{n}$ with probability at least $1 - 1/\boldsymbol{n}^2$



Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Height

- The running time of the search and insertion algorithms is affected by the height $\boldsymbol{h}$ of the skip list
- We show that with high probability, a skip list with $\boldsymbol{n}$ items has height $\boldsymbol{O}(\log \boldsymbol{n})$
- We use the following additional probabilistic fact:

  Fact 3: If each of $\boldsymbol{n}$ events has probability $\boldsymbol{p}$, the probability that at least one event occurs is at most $\boldsymbol{np}$

- Consider a skip list with $\boldsymbol{n}$ items
  - By Fact 1, we insert an item in list $\boldsymbol{S}_i$ with probability $1/2^{\boldsymbol{i}}$
  - By Fact 3, the probability that list $\boldsymbol{S}_i$ has at least one item is at most $\boldsymbol{n}/2^{\boldsymbol{i}}$
- By picking $\boldsymbol{i} = 3\log \boldsymbol{n}$, we have that the probability that $\boldsymbol{S}_{3\log \boldsymbol{n}}$ has at least one item is at most

$$\boldsymbol{n}/2^{3\log \boldsymbol{n}} = \boldsymbol{n}/\boldsymbol{n}^3 = 1/\boldsymbol{n}^2$$

- Thus, a skip list with $\boldsymbol{n}$ items has height at most $3\log \boldsymbol{n}$ with probability at least $1 - 1/\boldsymbol{n}^2$

**With High Probability (WHP)**



Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Height

- The running time of the search and insertion algorithms is affected by the height $\boldsymbol{h}$ of the skip list

  Fact 3: If each of $\boldsymbol{n}$ events has probability $\boldsymbol{p}$, the probability that at least one event occurs is at most $\boldsymbol{np}$

- Consider a skip list with $\boldsymbol{n}$ items
  - By Fact 1, we insert an item in list $\boldsymbol{S_i}$ with probability $1/2^{\boldsymbol{i}}$
  - By Fact 3, the probability that list $\boldsymbol{S_i}$ has at least one item is at most

at most

$$\boldsymbol{n}/2^{3\log \boldsymbol{n}} = \boldsymbol{n}/\boldsymbol{n}^3 = 1/\boldsymbol{n}^2$$

- Thus, a skip list with $\boldsymbol{n}$ items has height at most $3\log \boldsymbol{n}$ with probability at least $1 - 1/\boldsymbol{n}^2$

An event that occurs ***with high probability*** (WHP) is one whose probability depends on a certain number *n* and goes to 1 as *n* goes to infinity. [Wikipedia]

**With High Probability (WHP)**



Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Search and Update Times

- The search time in a skip list is proportional to
  - the number of drop-down steps, plus
  - the number of scan-forward steps
- The drop-down steps are bounded by the height of the skip list and thus are $O(\log n)$ with high probability
- To analyze the scan-forward steps, we use yet another probabilistic fact:

  Fact 4: The expected number of coin tosses required in order to get tails is 2

- When we scan forward in a list, the destination key does not belong to a higher list
  - A scan-forward step is associated with a former coin toss that gave tails
- By Fact 4, in each list the expected number of scan-forward steps is 2
- Thus, the expected number of scan-forward steps is $O(\log n)$
- We conclude that a search in a skip list takes $O(\log n)$ expected time
- The analysis of insertion and deletion gives similar results



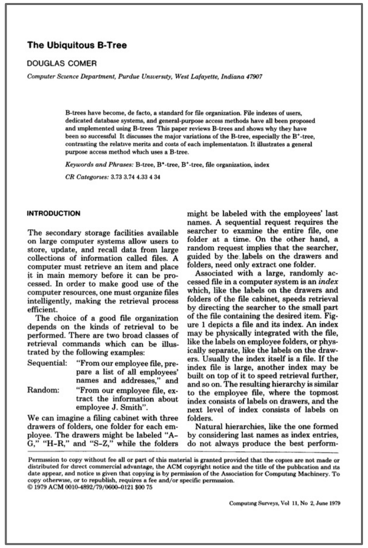
Acknowledgement: https://people.engr.tamu.edu/schaefer/teaching/221_Fall2018/Lectures/SkipLists.ppt

# Question?

Are Binary trees and skip lists optimal for in-memory indexing?

# B+ Trees

- A **B+Tree** is a self-balancing tree data structure that keeps data sorted and allows searches, sequential access, insertions, and deletions in $\mathbf{O}(\boldsymbol{log_B(N)})$.
  - The fanout of the tree is *B*
  - Generalization of a binary search tree in that a node can have more than two children.
  - Optimized for systems that read and write large blocks of data.



**The Ubiquitous B-Tree**

DOUGLAS COMER

*Computer Science Department, Purdue University, West Lafayette, Indiana 47907*

B-trees have become, de facto, a standard for file organization. File indexes of users, dedicated database systems, and general-purpose access methods have all been proposed and implemented using B-trees This paper reviews B-trees and shows why they have been so successful It discusses the major variations of the B-tree, especially the $B^+$-tree, contrasting the relative merits and costs of each implementation. It illustrates a general purpose access method which uses a B-tree.

*Keywords and Phrases:* B-tree, $B^*$-tree, $B^+$-tree, file organization, index

*CR Categories:* 3.73 3.74 4.33 4 34

**INTRODUCTION**

The secondary storage facilities available on large computer systems allow users to store, update, and recall data from large collections of information called files. A computer must retrieve an item and place it in main memory before it can be processed. In order to make good use of the computer resources, one must organize files intelligently, making the retrieval process efficient.

The choice of a good file organization depends on the kinds of retrieval to be performed. There are two broad classes of retrieval commands which can be illustrated by the following examples:

Sequential: "From our employee file, prepare a list of all employees' names and addresses," and

Random: "From our employee file, extract the information about employee J. Smith".

We can imagine a filing cabinet with three drawers of folders, one folder for each employee. The drawers might be labeled "A–G," "H–R," and "S–Z," while the folders might be labeled with the employees' last names. A sequential request requires the searcher to examine the entire file, one folder at a time. On the other hand, a random request implies that the searcher, guided by the labels on the drawers and folders, need only extract one folder.

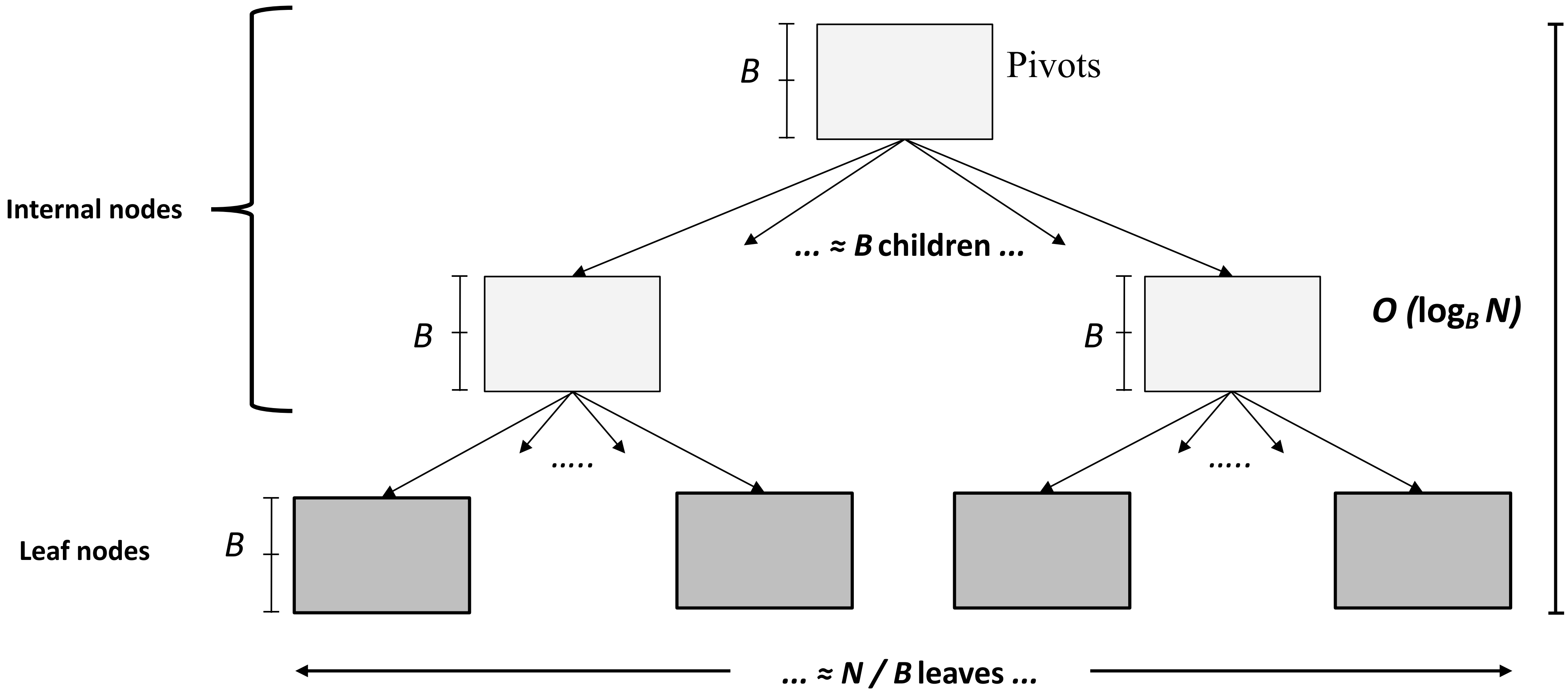
Associated with a large, randomly accessed file in a computer system is an *index* which, like the labels on the drawers and folders of the file cabinet, speeds retrieval by directing the searcher to the small part of the file containing the desired item. Figure 1 depicts a file and its index. An index may be physically integrated with the file, like the labels on employee folders, or physically separate, like the labels on the drawers. Usually the index itself is a file. If the index file is large, another index may be built on top of it to speed retrieval further, and so on. The resulting hierarchy is similar to the employee file, where the topmost index consists of labels on drawers, and the next level of index consists of labels on folders.

Natural hierarchies, like the one formed by considering last names as index entries, do not always produce the best perform-

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1979 ACM 0010-4892/79/0600-0121 $00 75

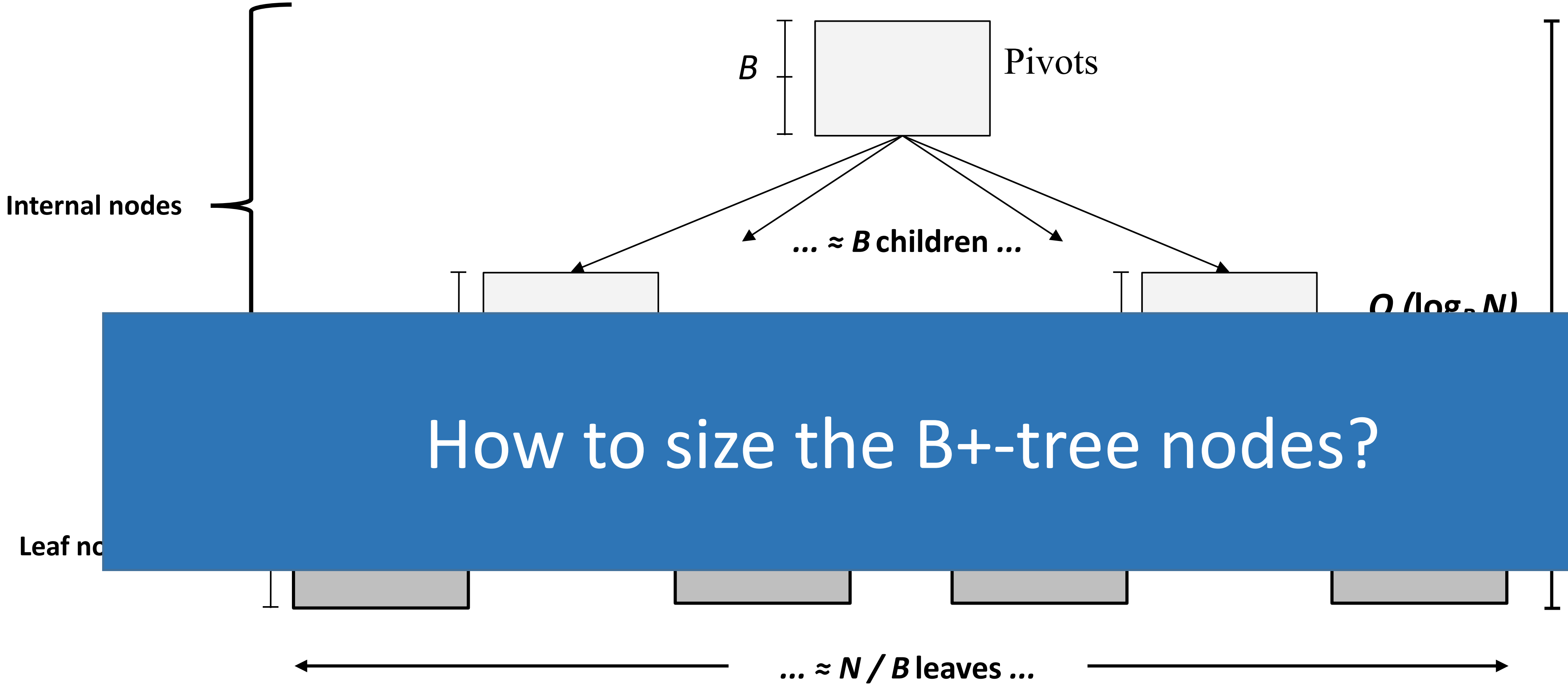
Computing Surveys, Vol 11, No 2, June 1979

# B+ Trees



Internal nodes
B
Pivots
... ≈ B children ...
B
B
O (log$_B$ N)
.....
.....
Leaf nodes
B
... ≈ N / B leaves ...

# B+ Trees

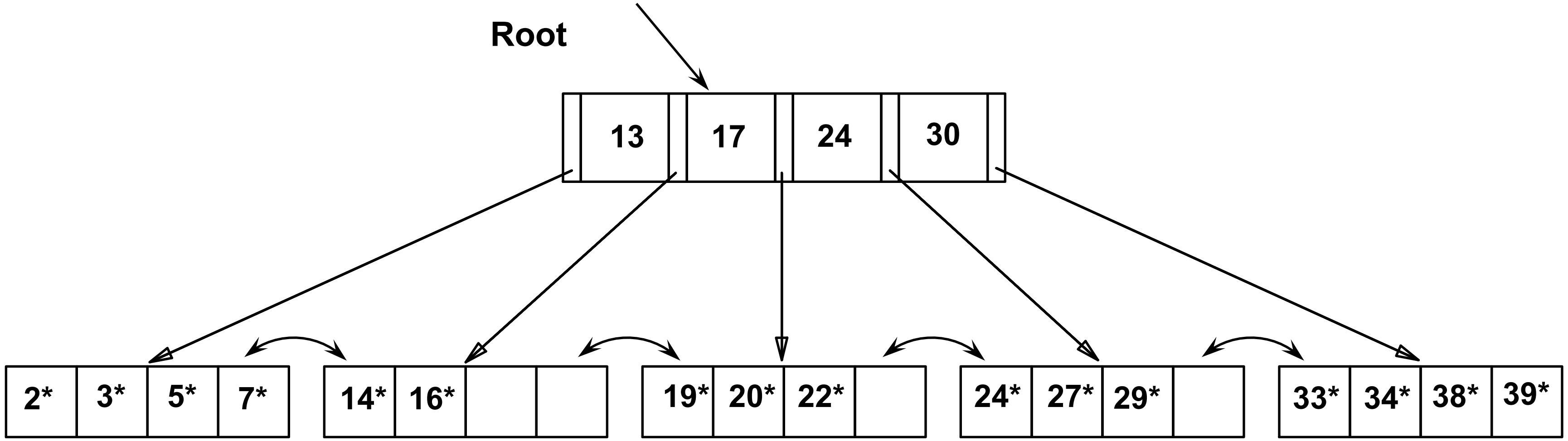



How to size the B+-tree nodes?

# B+ Trees

Search begins at root, and key comparisons direct it to a leaf.

Search for 5*, 15*, all data entries >= 24* ...

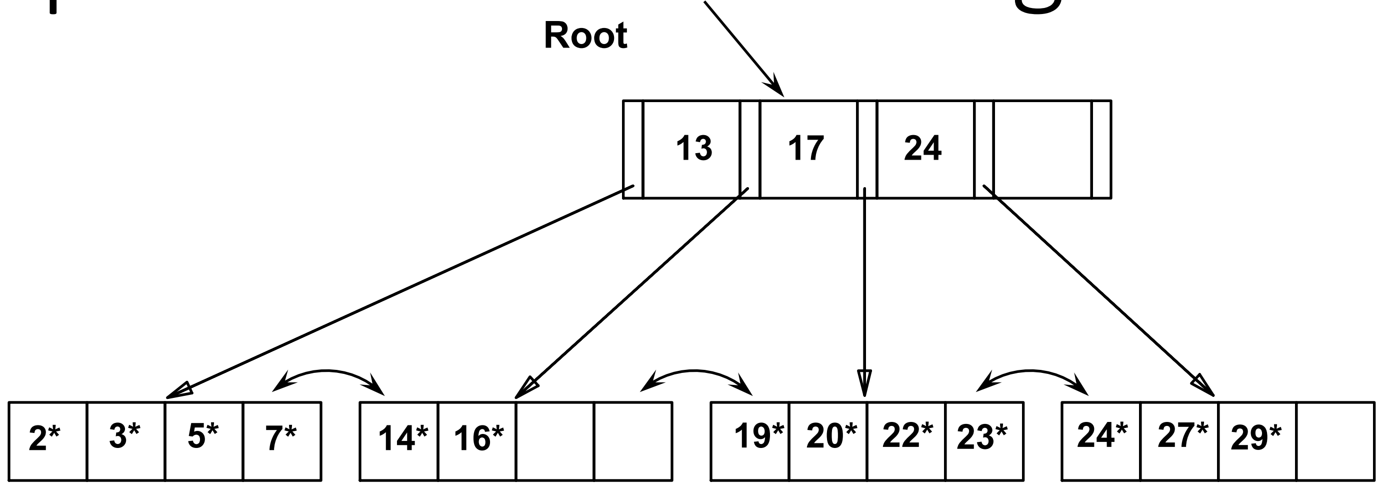



*Based on the search for 15*, we know it is not in the tree!*



Acknowledgement: Prof. Manos Athanassoulis, BU

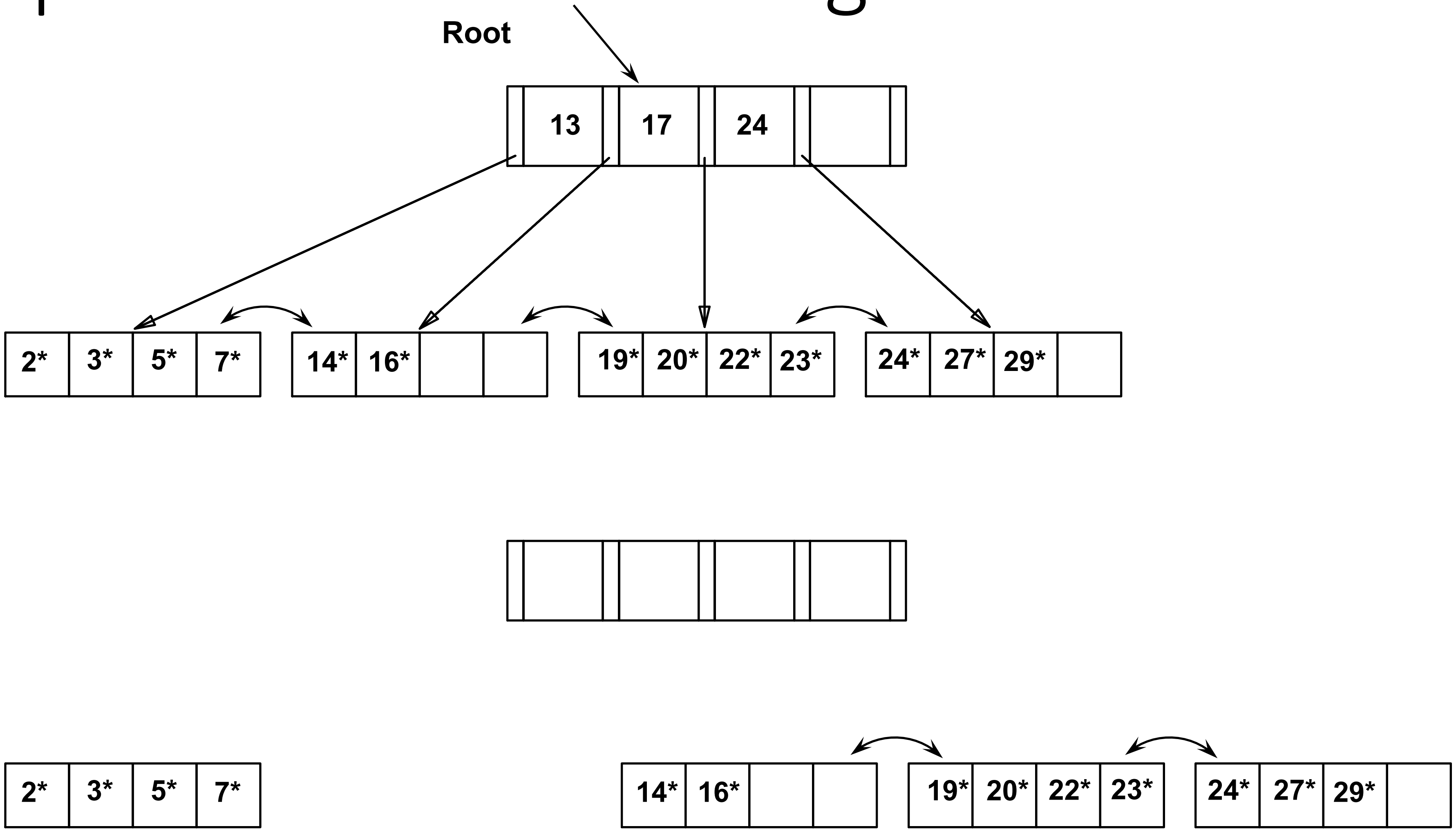
# Example B+ Tree - Inserting 8*






Acknowledgement: Prof. Manos Athanassoulis, BU

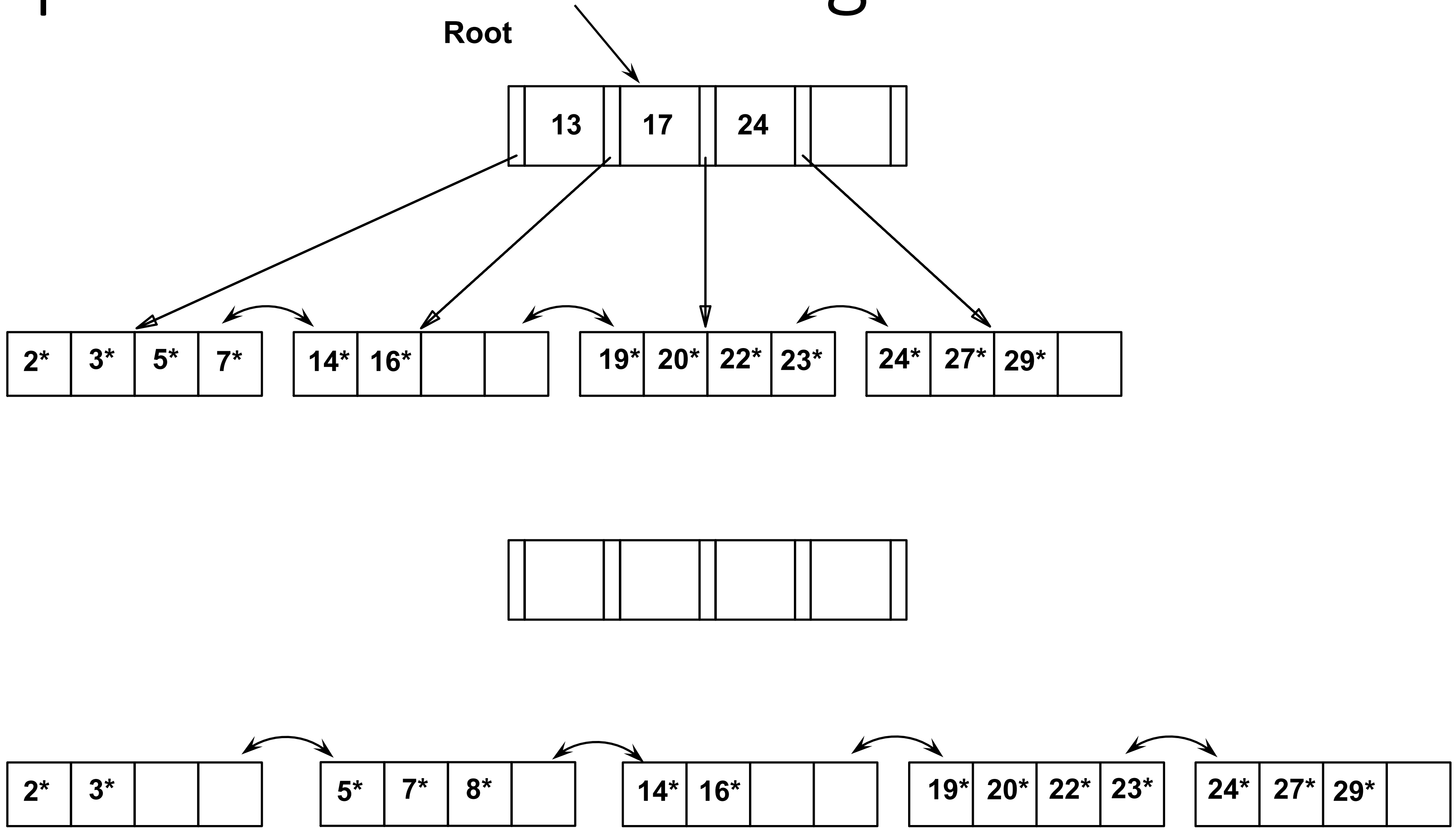
# Example B+ Tree - Inserting 8*






Acknowledgement: Prof. Manos Athanassoulis, BU

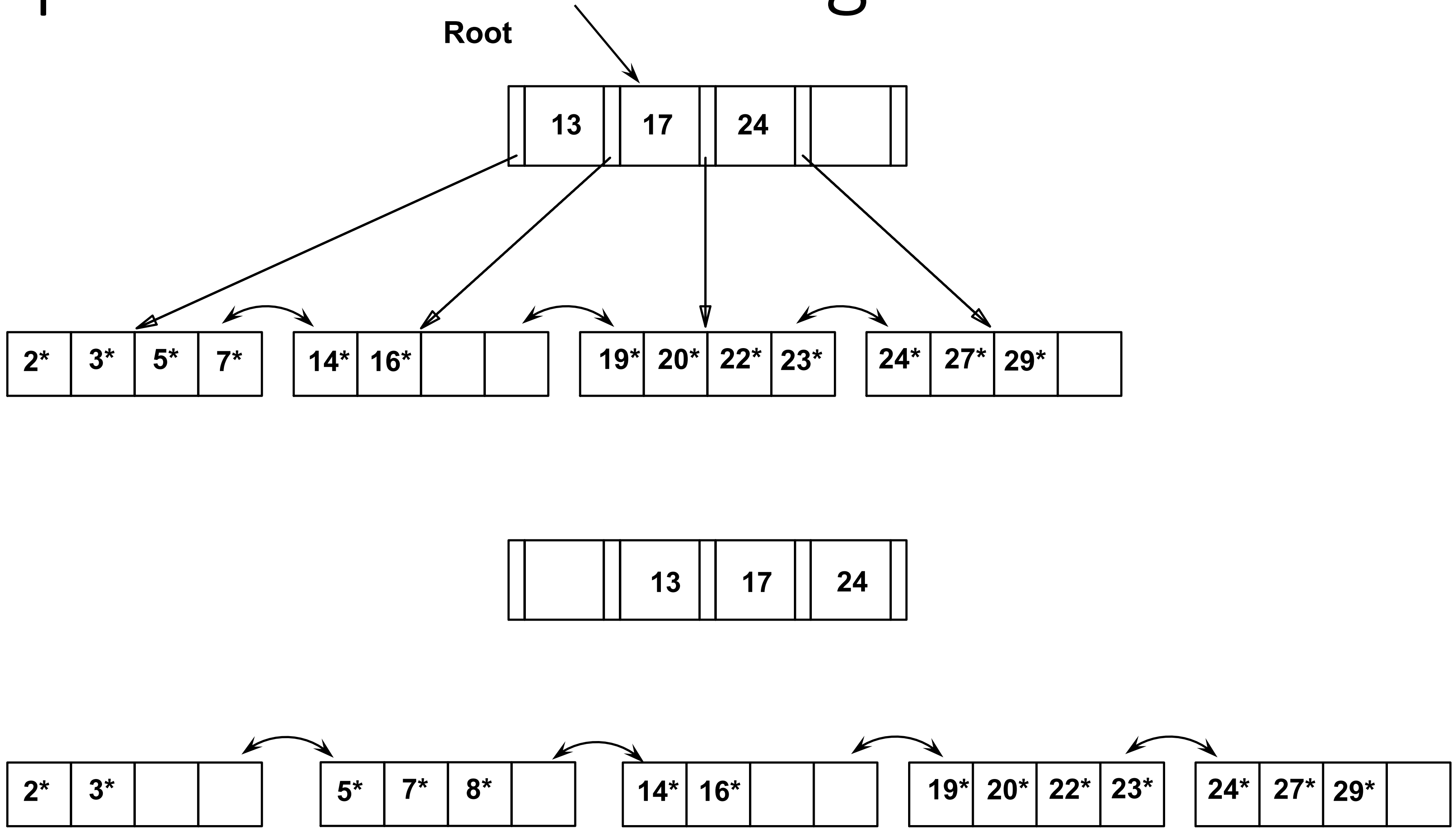
# Example B+ Tree - Inserting 8*






Acknowledgement: Prof. Manos Athanassoulis, BU

# Example B+ Tree - Inserting 8*

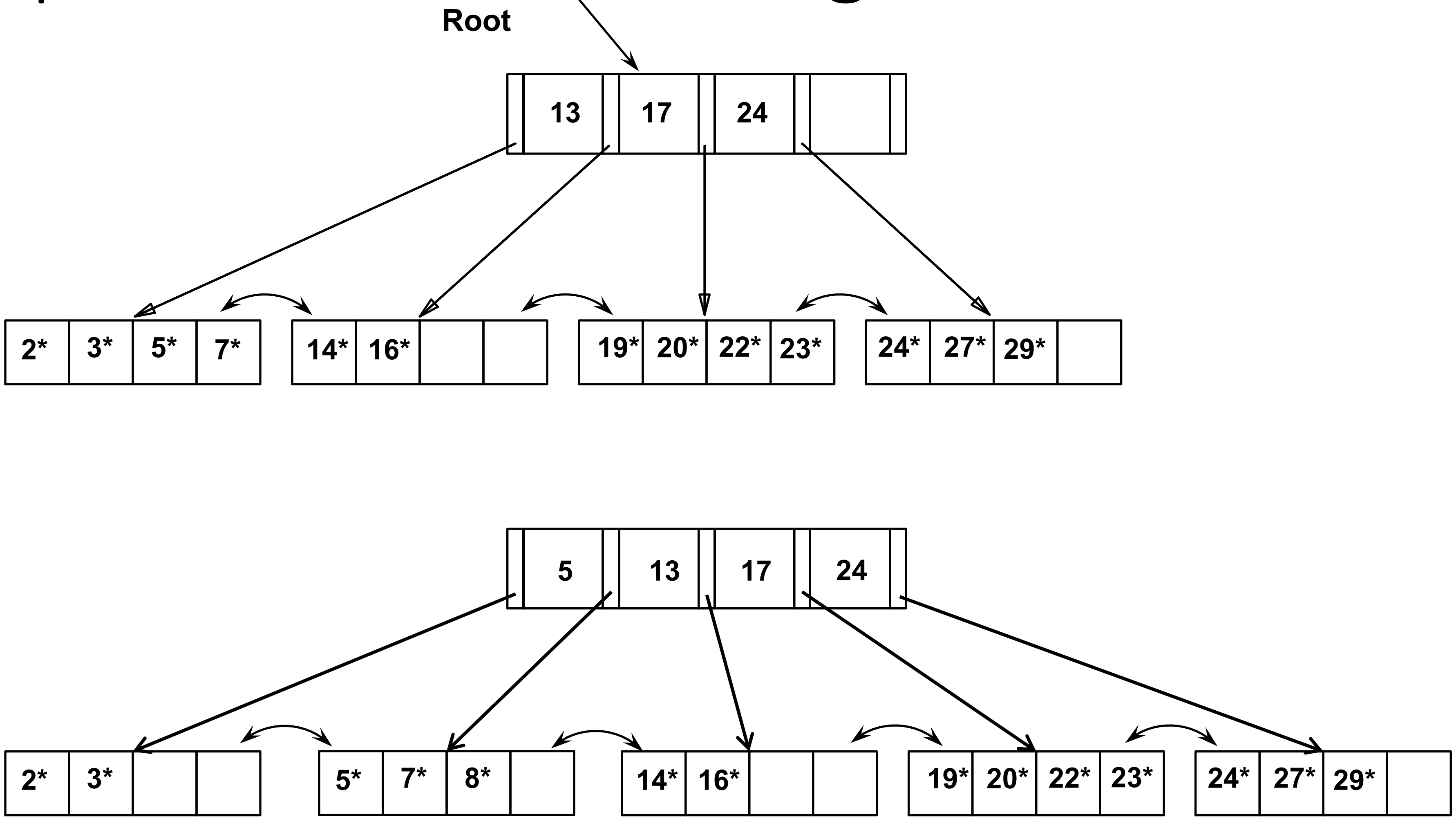







Acknowledgement: Prof. Manos Athanassoulis, BU

# Example B+ Tree - Inserting 8*

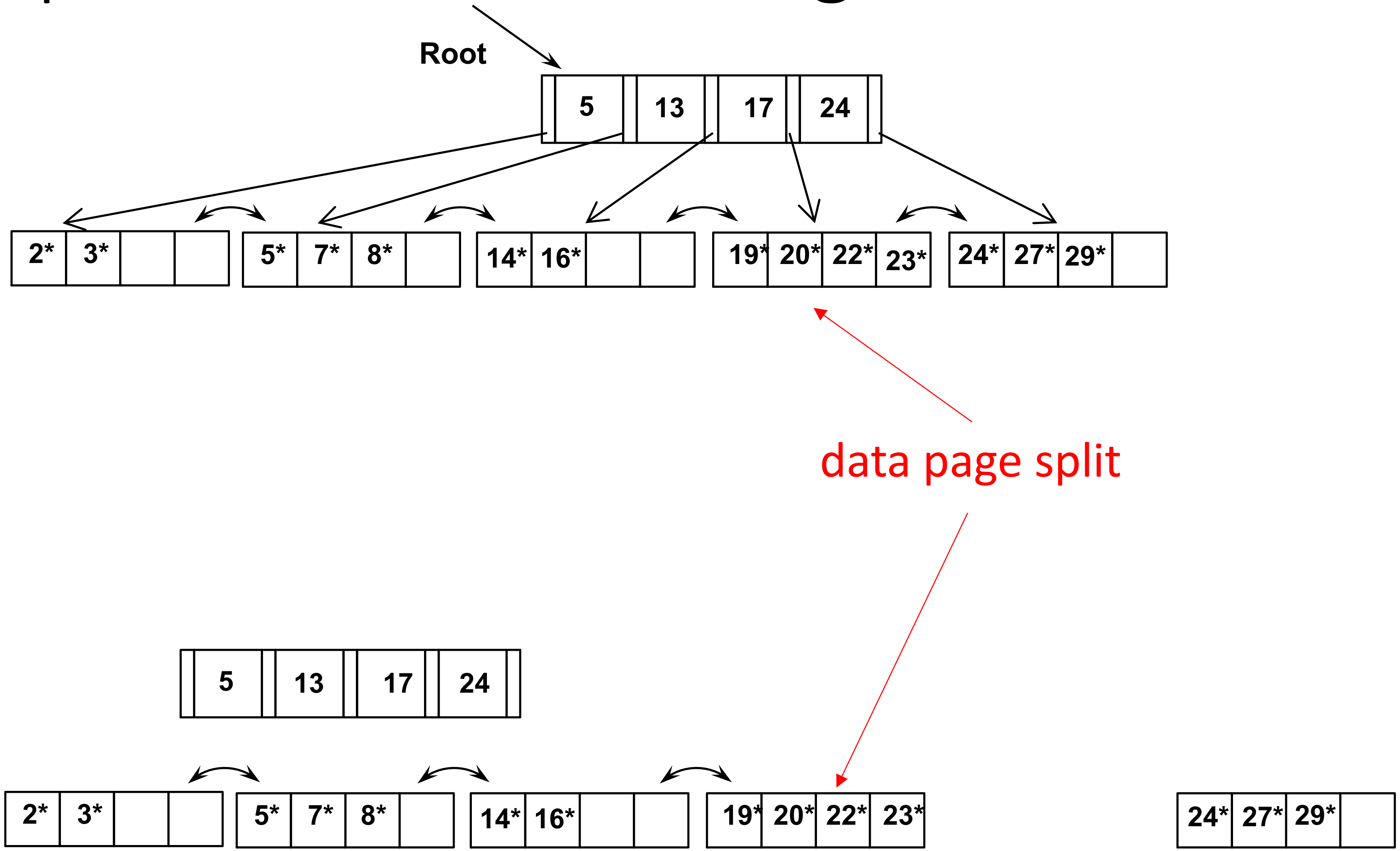







Acknowledgement: Prof. Manos Athanassoulis, BU

# Example B+ Tree - Inserting 21*

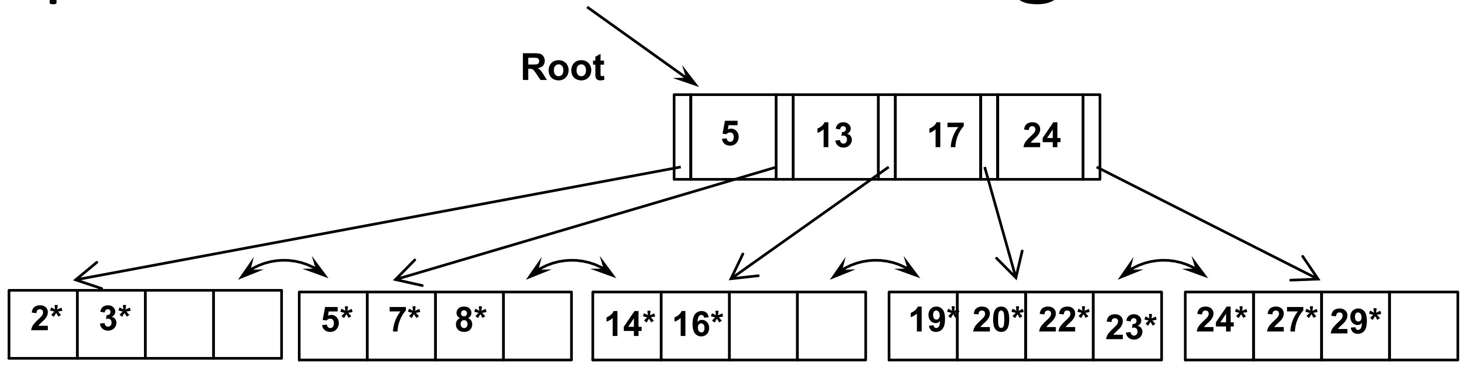


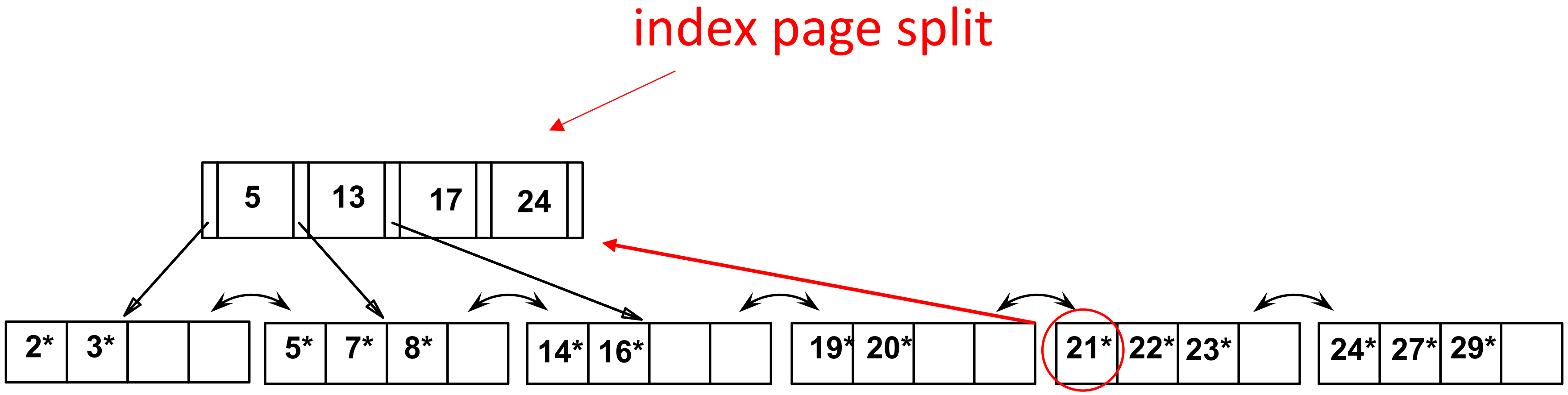



SCHOOL OF COMPUTING
UNIVERSITY OF UTAH

Acknowledgement: Prof. Manos Athanassoulis, BU

# Example B+ Tree - Inserting 21*



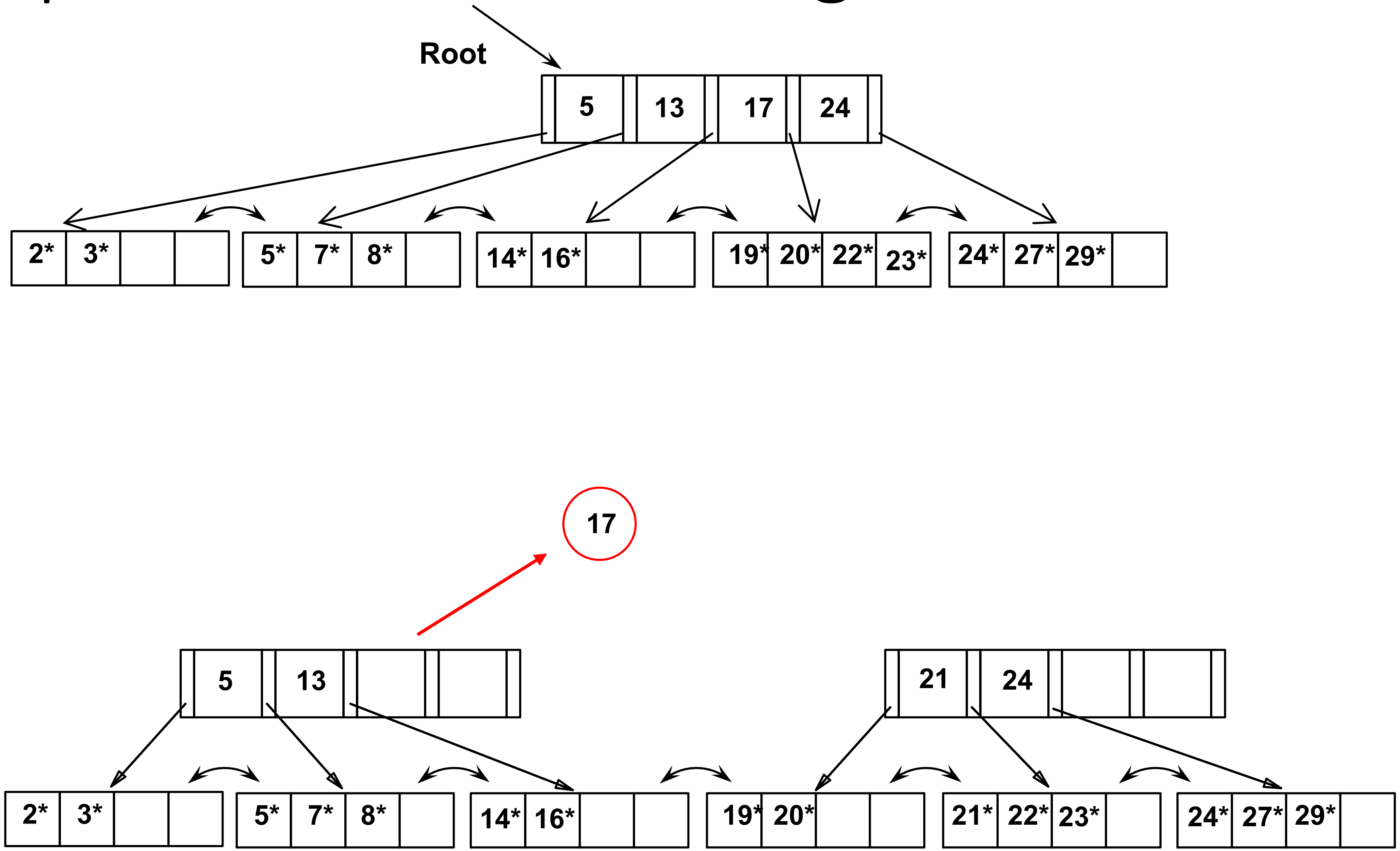






SCHOOL OF COMPUTING
UNIVERSITY OF UTAH

Acknowledgement: Prof. Manos Athanassoulis, BU

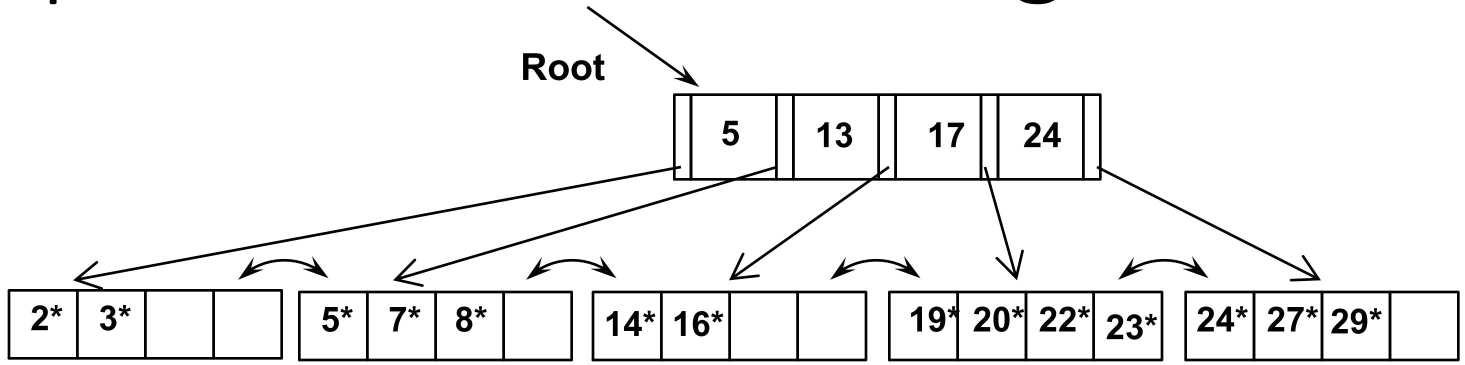
# Example B+ Tree - Inserting 21*

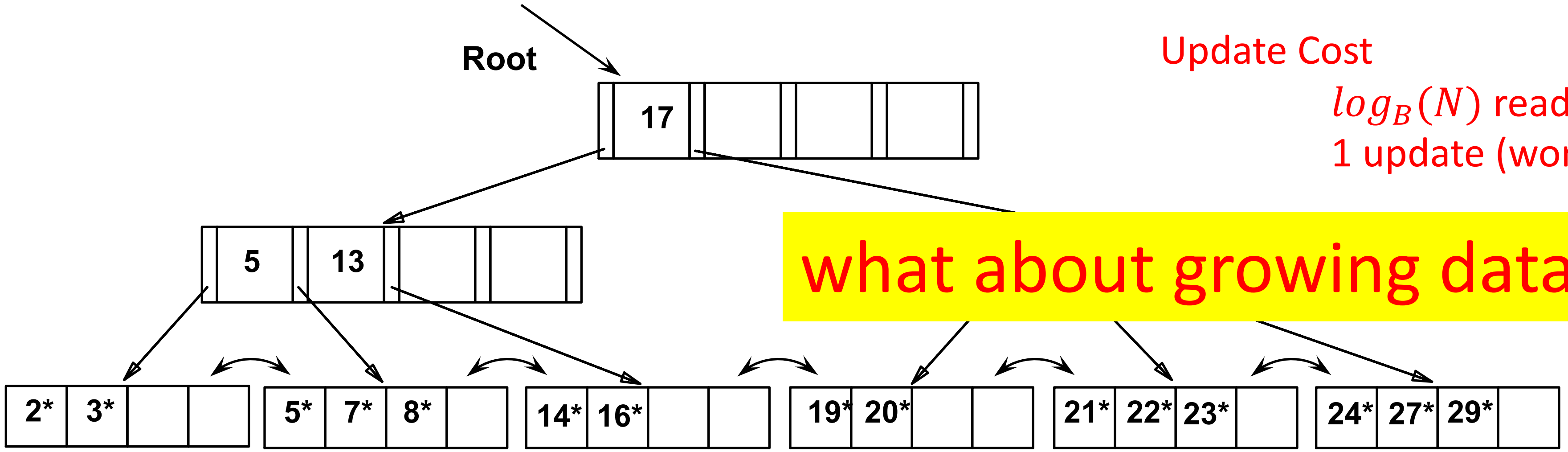





Acknowledgement: Prof. Manos Athanassoulis, BU

# Example B+ Tree - Inserting 21*




Read Cost: $log_B(N)$

Update Cost

$log_B(N)$ reads

1 update (worse case $log_B(N)$)

what about growing dataset size?



Acknowledgement: Prof. Manos Athanassoulis, BU

# Observation

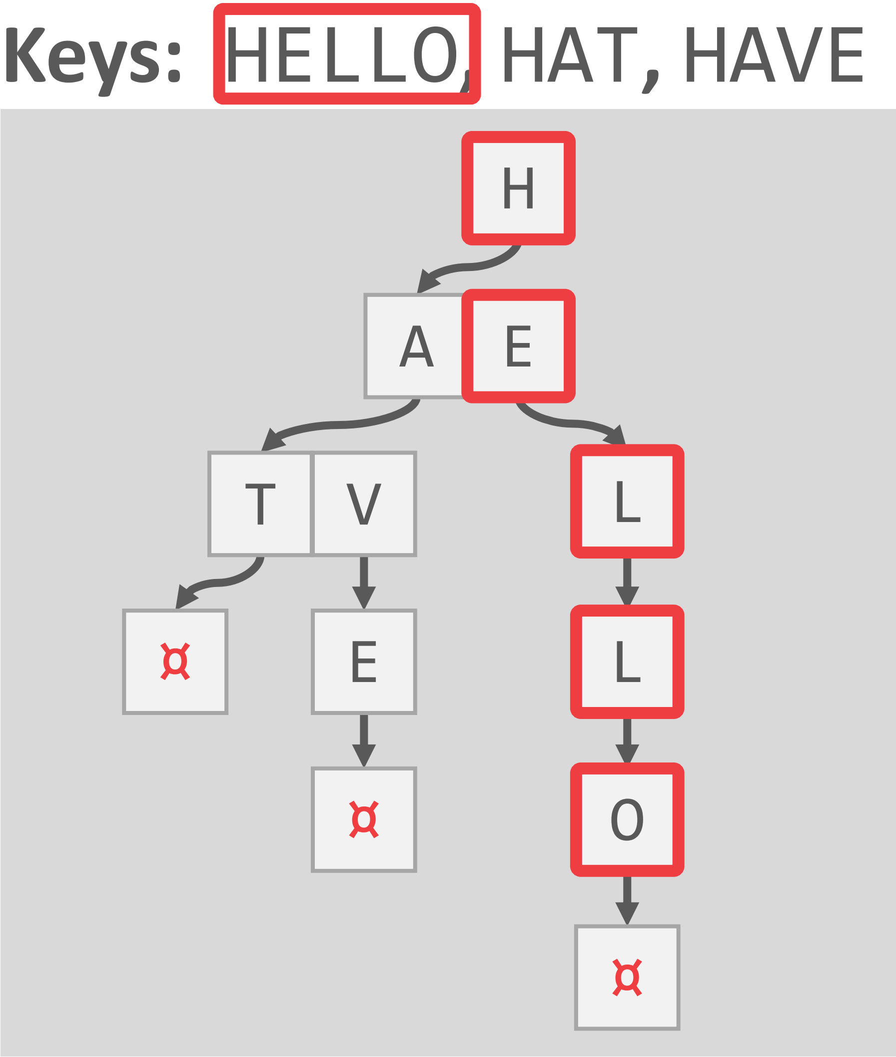
- The inner node keys in a B+tree cannot tell you whether a key exists in the index. You always must traverse to the leaf node.

- This means that you could have (at least) one cache miss per level in the tree.



Acknowledgement: Prof. Manos Athanassoulis, BU

# Trie index




- Use a digital representation of keys to examine prefixes one-by-one instead of comparing entire key.
    - Also known as ***Digital Search Tree***, ***Prefix Tree***.


Acknowledgement: Prof. Andy Pavlo, CMU

# Trie index properties

- Shape only depends on key space and lengths.
  - Does not depend on existing keys or insertion order.
  - Does not require rebalancing operations.

- All operations have **O(*k*)** complexity where ***k*** is the length of the key.
  - The path to a leaf node represents the key of the leaf
  - Keys are stored implicitly and can be reconstructed from paths.


Acknowledgement: Prof. Andy Pavlo, CMU

# Trie index properties

- Shape only depends on key space and lengths.
  - Does not depend on existing keys or insertion order.
  - Does not require rebalancing operations.

**History independent**

- All operations have **O(*k*)** complexity where ***k*** is the length of the key.
  - The path to a leaf node represents the key of the leaf
  - Keys are stored implicitly and can be reconstructed from paths.


Acknowledgement: Prof. Andy Pavlo, CMU

# Trie key span

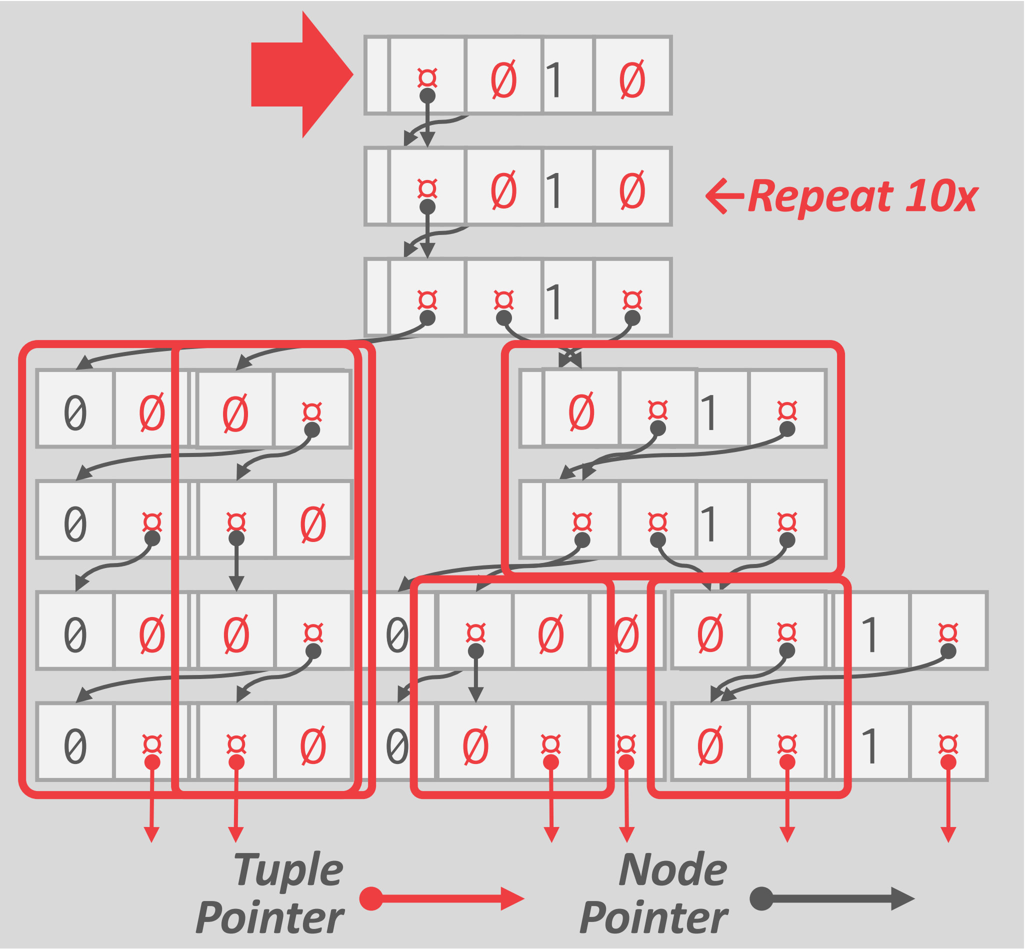
- The **span** of a trie level is the number of bits that each partial key / digit represents.
  - If the digit exists in the corpus, then store a pointer to the next level in the trie branch. Otherwise, store null.

- This determines the **fan-out** of each node and the physical **height** of the tree.
  - ***n***-way Trie = Fan-Out of ***n***


Acknowledgement: Prof. Andy Pavlo, CMU

# Trie key span

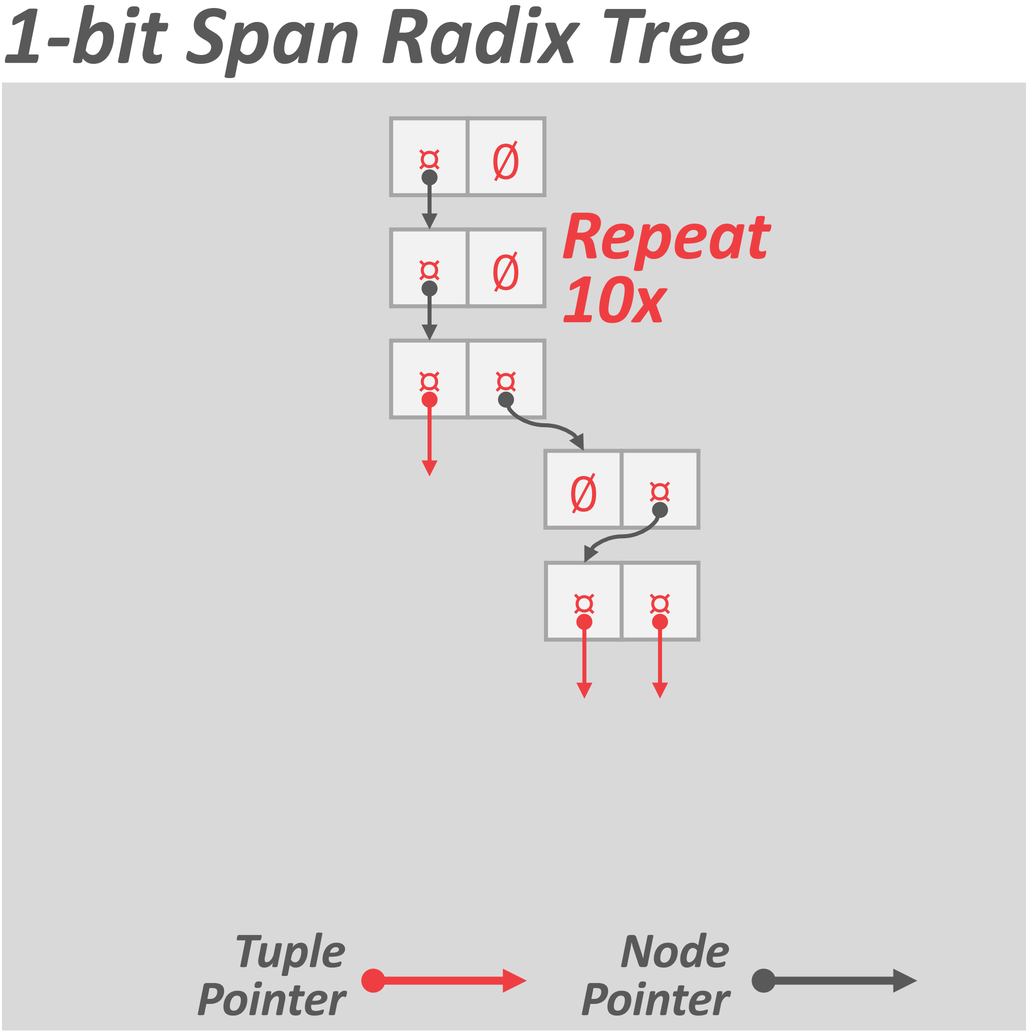
## *1-bit Span Trie*




```
K10→ 00000000 00001010
K25→ 00000000 00011001
K31→ 00000000 00011111
```


Acknowledgement: Prof. Andy Pavlo, CMU

# Radix tree




- Omit all nodes with only a single child.
  - Also known as ***Patricia Tree***.
- Can produce false positives, so the DBMS always checks the original tuple to see whether a key matches.


Acknowledgement: Prof. Andy Pavlo, CMU

# Trie variants

- Judy Arrays (HP)
- ART Index (HyPer)
- Masstree (Silo)


Acknowledgement: Prof. Andy Pavlo, CMU

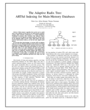
# Judy arrays

- Variant of a 256-way radix tree. First known radix tree that supports adaptive node representation.
- Three array types
  - **Judy1**: Bit array that maps integer keys to true/false.
  - **JudyL**: Map integer keys to integer values.
  - **JudySL**: Map variable-length keys to integer values.
- Open-Source Implementation (LGPL). Patented by HP in 2000. Expires in 2022.
  - Not an issue according to authors.
  - http://judy.sourceforge.net/


Acknowledgement: Prof. Andy Pavlo, CMU

# Adaptive radix tree (ART)

- Developed for TUM HyPer DBMS in 2013.
- 256-way radix tree that supports different node types based on its population.
  - Stores meta-data about each node in its header.
- Concurrency support was added in 2015.



**THE ADAPTIVE RADIX TREE: ARTFUL INDEXING FOR MAIN-MEMORY DATABASES**
*ICDE 2013*



Acknowledgement: Prof. Andy Pavlo, CMU

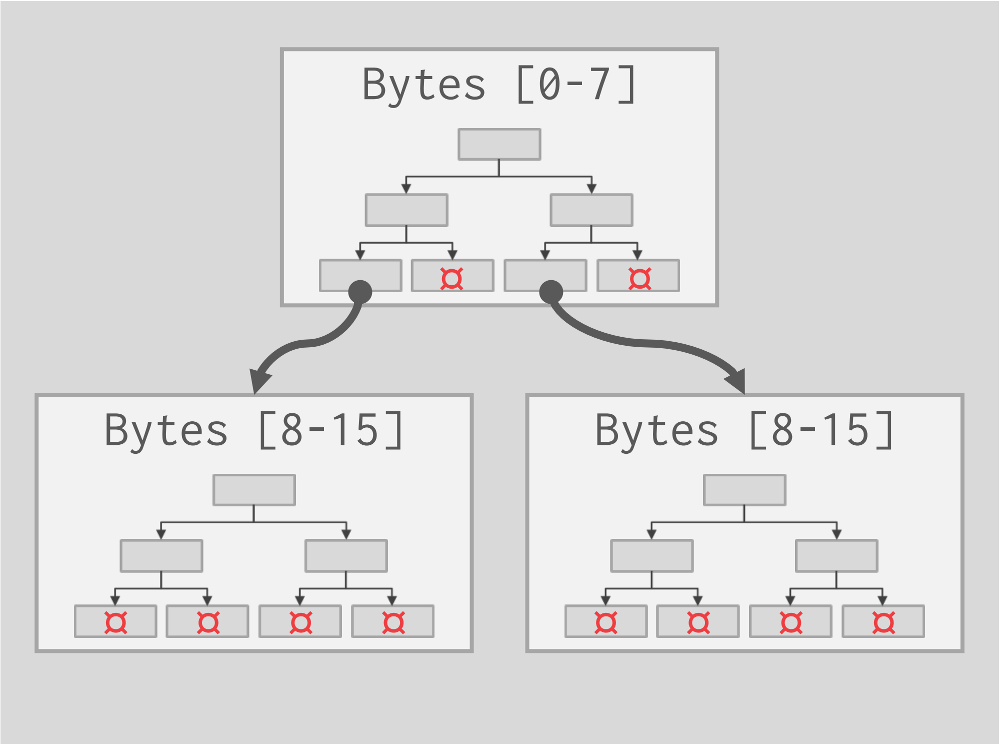
# ART vs. JUDY

- **Difference #1: Node Types**
  - Judy has three node types with different organizations.
  - ART has four nodes types that (mostly) vary in the maximum number of children.
- **Difference #2: Purpose**
  - Judy is a general-purpose associative array. It "owns" the keys and values.
  - ART is a table index and does not need to cover the full keys. Values are pointers to tuples.


Acknowledgement: Prof. Andy Pavlo, CMU

# MASSTREE

## *Masstree*




- Instead of using different layouts for each trie node based on its size, use an entire B+Tree.
  - Each B+tree represents 8-byte span.
  - Optimized for long keys.
  - Uses a latching protocol that is similar to versioned latches.

- Part of the Harvard Silo project.

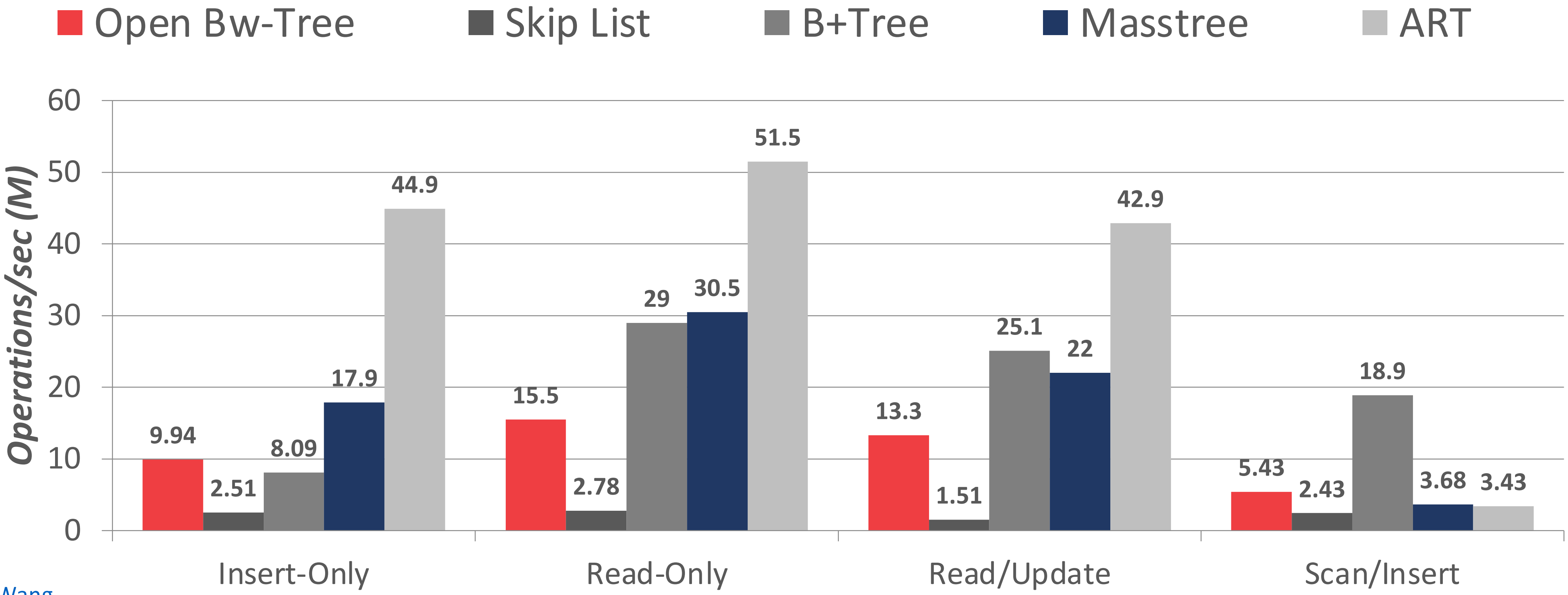
**CACHE CRAFTINESS FOR FAST MULTICORE KEY-VALUE STORAGE**
***EUROSYS 2012***



Acknowledgement: Prof. Andy Pavlo, CMU

# IN-MEMORY INDEXES

***Processor: 1 socket, 10 cores w/ 2×HT***
***Workload: 50m Random Integer Keys (64-bit)***

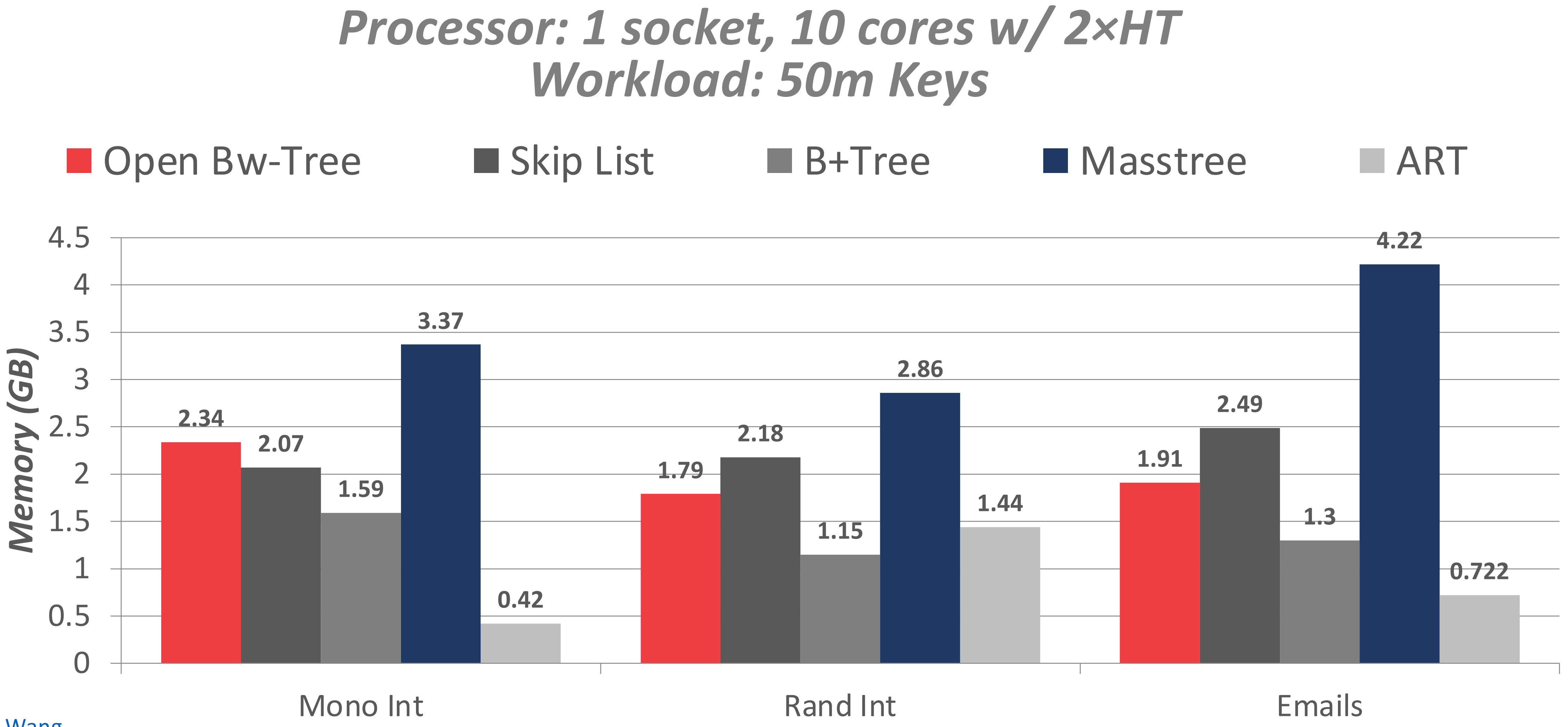



Source: [Ziqi Wang]



Acknowledgement: Prof. Andy Pavlo, CMU

# IN-MEMORY INDEXES




Source: [Ziqi Wang]




Acknowledgement: Prof. Andy Pavlo, CMU

# PARTING THOUGHTS

- B+ trees are the go to in-memory indexing data structures.
- Radix trees have interesting properties, but a well-written B+tree is still a solid design choice.
- Skip lists are amazing if you don’t want to implement self balancing binary trees


Acknowledgement: Prof. Andy Pavlo, CMU

# Next class

- In-memory indexing (hash tables/filters)

***Make sure to read the related papers from the reading list***