B-Trees

Antonio Carzaniga

Faculty of Informatics
Università della Svizzera italiana

April 30, 2024

Outline

- Search in secondary storage
- B-Trees
  - properties
  - search
  - insertion

Complexity Model

- Basic assumption so far: *data structures fit completely in main memory (RAM)*
  - all basic operations have the same cost
    - even this is a rough approximation, since the main-memory system is not at all “flat”

Complexity Model

- Basic assumption so far: *data structures fit completely in main memory (RAM)*
  - all basic operations have the same cost
    - even this is a rough approximation, since the main-memory system is not at all “flat”
- However, some applications require more storage than what fits in main memory
  - we must use data structures that reside in *secondary storage* (i.e., disk)

- Basic assumption so far: *data structures fit completely in main memory (RAM)*
  - all basic operations have the same cost
    - even this is a rough approximation, since the main-memory system is not at all “flat”
- However, some applications require more storage than what fits in main memory
  - we must use data structures that reside in *secondary storage* (i.e., disk)

Disk is 10,000–100,000 times slower than RAM

Memory access/transfer	CPU cycles (≈ 1ns)
Register	*1*

Memory access/transfer	CPU cycles (≈ 1ns)
Register	***1***
L1 cache	4
L2 cache	10
Local L3 cache	40–75
Remote L3 cache	100–300
Local DRAM	60
Remote DRAM (main memory)	***100***

Memory access/transfer	**CPU cycles (≈ 1ns)**
Register	***1***
L1 cache	4
L2 cache	10
Local L3 cache	40–75
Remote L3 cache	100–300
Local DRAM	60
Remote DRAM (main memory)	***100***
SSD seek	***20,000***

Memory access/transfer	**CPU cycles (≈ 1ns)**
Register	***1***
L1 cache	4
L2 cache	10
Local L3 cache	40–75
Remote L3 cache	100–300
Local DRAM	60
Remote DRAM (main memory)	***100***
SSD seek	***20,000***
Send 2K bytes over 1 Gbps network	20,000
Read 1 MB sequentially from memory	250,000
Round trip within a datacenter	500,000

Memory access/transfer	**CPU cycles (≈ 1ns)**
Register	***1***
L1 cache	4
L2 cache	10
Local L3 cache	40–75
Remote L3 cache	100–300
Local DRAM	60
Remote DRAM (main memory)	***100***
SSD seek	***20,000***
Send 2K bytes over 1 Gbps network	20,000
Read 1 MB sequentially from memory	250,000
Round trip within a datacenter	500,000
HDD seek	***10,000,000***
Read 1 MB sequentially from network	10,000,000
Read 1 MB sequentially from disk	30,000,000
Round-trip time USA–Europe	150,000,000

Modeling Disk Access

- Let *x* be a pointer to some (possibly complex) object

- Let x be a pointer to some (possibly complex) object
- When the object is in memory, x can be used directly as a reference to the object
  - e.g., $\ell = x.size$ or $x.root = y$

- Let x be a pointer to some (possibly complex) object

- When the object is in memory, x can be used directly as a reference to the object
  - e.g., $\ell = x.size$ or $x.root = y$

- When the object is on disk, we must first perform a disk-read operation

  Disk-Read(x) reads the object into memory, allowing us to refer to it (and modify it) through x

- Let x be a pointer to some (possibly complex) object
- When the object is in memory, x can be used directly as a reference to the object
  - e.g., $\ell = x.size$ or $x.root = y$
- When the object is on disk, we must first perform a disk-read operation

  DISK-READ(x) reads the object into memory, allowing us to refer to it (and modify it) through x
- Any changes to the object in memory must be eventually saved onto the disk

  DISK-WRITE(x) writes the object onto the disk (if the object was modified)

Binary Trees on Disk

- Assume each node *x* is stored on disk

Binary Trees on Disk

- Assume each node x is stored on disk

```
ITERATIVE-TREE-SEARCH(T, k)
x = T.root
while x ≠ NIL
        DISK-READ(x)
        if k == x.key
                return x
        elseif k < x.key
                x = x.left
        else x = x.right
return x
```

- Assume each node x is stored on disk

```
ITERATIVE-TREE-SEARCH(T, k)
  x = T.root
  while x ≠ NIL
        DISK-READ(x)
        if k == x.key
              return x
        elseif k < x.key
              x = x.left
        else x = x.right
  return x
```

cost

Binary Trees on Disk

- Assume each node x is stored on disk

ITERATIVE-TREE-SEARCH(T, k)	*cost*
$x = T.root$	c
while $x \neq$ NIL	c
DISK-READ(x)	$100000c$
if $k == x.key$	c
return x	c
elseif $k < x.key$	c
$x = x.left$	c
else $x = x.right$	c
return x	c

Basic Intuition

- Assume we store the nodes of a search tree on disk
  1. node accesses should be reduced to a minimum
  2. spending more than a few basic operations for each node is not a problem

Basic Intuition

- Assume we store the nodes of a search tree on disk
  1. node accesses should be reduced to a minimum
  2. spending more than a few basic operations for each node is not a problem

- Rationale
  - basic in-memory operations are much cheaper
  - the bottleneck is with node accesses, which involve **Disk-Read** and **Disk-Write** operations

Idea

Idea

- In a balanced *binary* tree, n keys require a tree of height $h = \lfloor \log_2 n \rfloor$
  - all the important operations require access to $O(h)$ nodes
  - each one accounting for *one or very few* basic operations

Idea

- In a balanced *binary* tree, n keys require a tree of height $h = \lfloor \log_2 n \rfloor$
  - all the important operations require access to $O(h)$ nodes
  - each one accounting for *one or very few* basic operations

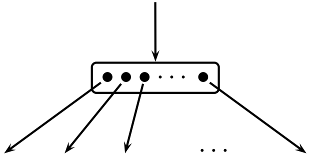
- **Idea:** store several keys and pointers to children nodes in a single node

- In a balanced *binary* tree, n keys require a tree of height $h = \lfloor \log_2 n \rfloor$
  - all the important operations require access to $O(h)$ nodes
  - each one accounting for *one or very few* basic operations

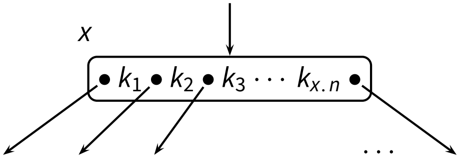
- **Idea:** store several keys and pointers to children nodes in a single node
  - in practice we ***increase the degree*** (or *branching factor*) of each node up to $d > 2$, so $h = \lfloor \log_d n \rfloor$
    - in practice d can be as high as a few thousands

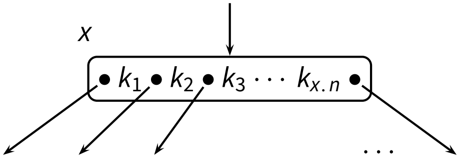
- In a balanced *binary* tree, n keys require a tree of height $h = \lfloor \log_2 n \rfloor$
  - all the important operations require access to $O(h)$ nodes
  - each one accounting for *one or very few* basic operations

- **Idea:** store several keys and pointers to children nodes in a single node
  - in practice we ***increase the degree*** (or *branching factor*) of each node up to $d > 2$, so $h = \lfloor \log_d n \rfloor$
    - in practice d can be as high as a few thousands

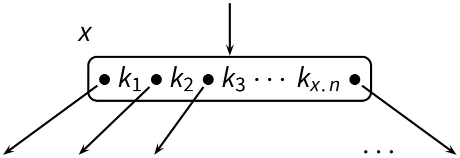


E.g., if $d = 1000$, then ***only three accesses*** ($h = 2$) cover ***up to one billion keys***

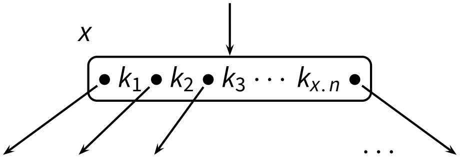

x
k1 k2 k3 · · · kx.n
. . .



- Every node x has the following fields
  - $x.n$ is the number of keys stored at each node

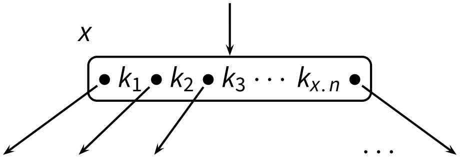


- Every node x has the following fields
  - $x.n$ is the number of keys stored at each node
  - $x.key[1] \leq x.key[2] \leq \ldots x.key[x.n]$ are the $x.n$ keys stored in nondecreasing order



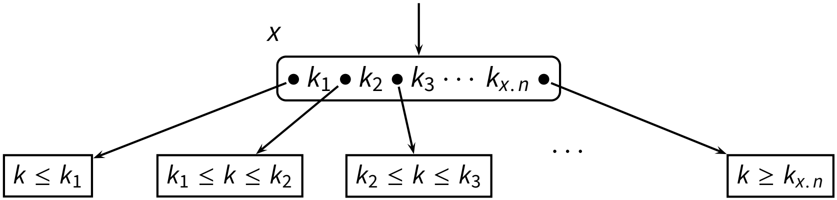
- Every node x has the following fields
  - $x.n$ is the number of keys stored at each node
  - $x.key[1] \leq x.key[2] \leq \ldots x.key[x.n]$ are the $x.n$ keys stored in nondecreasing order
  - $x.leaf$ is a Boolean flag that is TRUE if x is a *leaf node* or FALSE if x is an *internal node*

Definition of a B-Tree

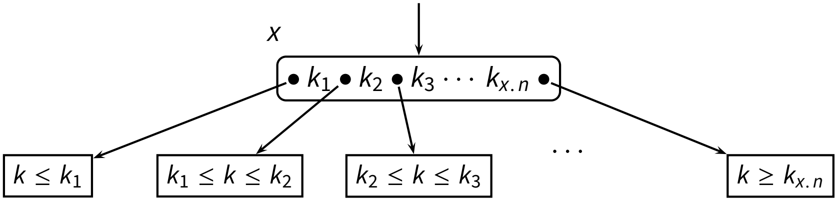


- Every node x has the following fields
  - $x.n$ is the number of keys stored at each node
  - $x.key[1] \leq x.key[2] \leq \ldots x.key[x.n]$ are the $x.n$ keys stored in nondecreasing order
  - $x.leaf$ is a Boolean flag that is TRUE if x is a *leaf node* or FALSE if x is an *internal node*
  - $x.c[1], x.c[2], \ldots, x.c[x.n+1]$ are the $x.n+1$ pointers to its children, if x is an *internal node*

Definition of a B-Tree (2)

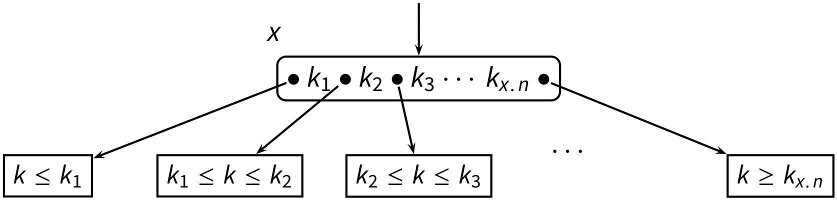


Definition of a B-Tree (2)



- The keys $x.key[i]$ delimit the ranges of keys stored in each subtree

Definition of a B-Tree (2)



- The keys $x.key[i]$ delimit the ranges of keys stored in each subtree

  $x.c[1] \longrightarrow$ subtree containing keys $k \leq x.key[1]$

  $x.c[2] \longrightarrow$ subtree containing keys $k, x.key[1] \leq k \leq x.key[2]$

  $x.c[3] \longrightarrow$ subtree containing keys $k, x.key[2] \leq k \leq x.key[3]$

  $\ldots$

  $x.c[x.n+1] \longrightarrow$ subtree containing keys $k, k \geq x.key[x.n]$

Definition of a B-Tree (3)

Definition of a B-Tree (3)

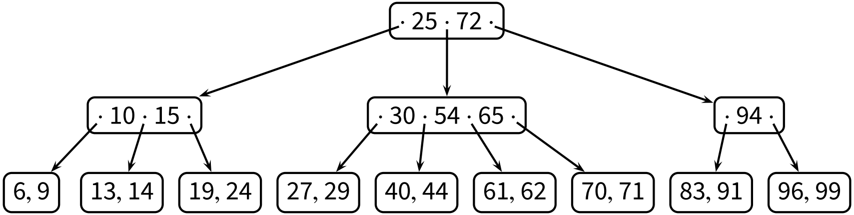
- ***All leaves have the same depth***

Definition of a B-Tree (3)

- ***All leaves have the same depth***

- Let $t \geq 2$ be the ***minimum degree*** of the B-tree
  - every node other than the root must have ***at least*** $t - 1$ ***keys***
  - every node must contain ***at most*** $2t - 1$ ***keys***
    - a node is *full* when it contains exactly $2t - 1$ keys
    - a full node has $2t$ children

Example



· 25 · 72 ·
· 10 · 15 ·
· 30 · 54 · 65 ·
· 94 ·
6, 9
13, 14
19, 24
27, 29
40, 44
61, 62
70, 71
83, 91
96, 99

Search in B-Trees

```
B-TREE-SEARCH(x, k)
i = 1
while i ≤ x.n and k > x.key[i]
        i = i + 1
if i ≤ x.n and k == x.key[i]
        return (x, i)
if x.leaf
        return NIL
else DISK-READ(x.c[i])
        return B-TREE-SEARCH(x.c[i], k)
```

Height of a B-Tree

- **Theorem:** the height of a B-tree containing $n \geq 1$ keys and with a minimum degree $t \geq 2$ is

$$h \leq \log_t \frac{n+1}{2}$$

- **Theorem:** the height of a B-tree containing $n \geq 1$ keys and with a minimum degree $t \geq 2$ is

$$h \leq \log_t \frac{n+1}{2}$$

Proof:

  - $n \geq 1$, so the root has at least one key (and therefore two children)

- **Theorem:** the height of a B-tree containing $n \geq 1$ keys and with a minimum degree $t \geq 2$ is

$$h \leq \log_t \frac{n+1}{2}$$

Proof:

  - $n \geq 1$, so the root has at least one key (and therefore two children)
  - every other node has at least t children

- **Theorem:** the height of a B-tree containing $n \geq 1$ keys and with a minimum degree $t \geq 2$ is

$$h \leq \log_t \frac{n+1}{2}$$

Proof:

  - $n \geq 1$, so the root has at least one key (and therefore two children)
  - every other node has at least t children
  - in the worst case, there are two subtrees (of the root) each one containing a total of $(n-1)/2$ keys, and each one consisting of t-degree nodes, with each node containing $t-1$ keys

- **Theorem:** the height of a B-tree containing $n \geq 1$ keys and with a minimum degree $t \geq 2$ is

$$h \leq \log_t \frac{n+1}{2}$$

Proof:

  - $n \geq 1$, so the root has at least one key (and therefore two children)
  - every other node has at least t children
  - in the worst case, there are two subtrees (of the root) each one containing a total of $(n-1)/2$ keys, and each one consisting of t-degree nodes, with each node containing $t-1$ keys
  - each subtree contains $1 + t + t^2 \cdots + t^{h-1}$ nodes, each one containing $t-1$ keys

- **Theorem:** the height of a B-tree containing $n \geq 1$ keys and with a minimum degree $t \geq 2$ is

$$h \leq \log_t \frac{n+1}{2}$$

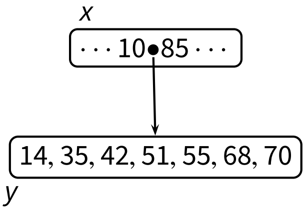
Proof:

- $n \geq 1$, so the root has at least one key (and therefore two children)
- every other node has at least t children
- in the worst case, there are two subtrees (of the root) each one containing a total of $(n-1)/2$ keys, and each one consisting of t-degree nodes, with each node containing $t-1$ keys
- each subtree contains $1 + t + t^2 \cdots + t^{h-1}$ nodes, each one containing $t-1$ keys, so

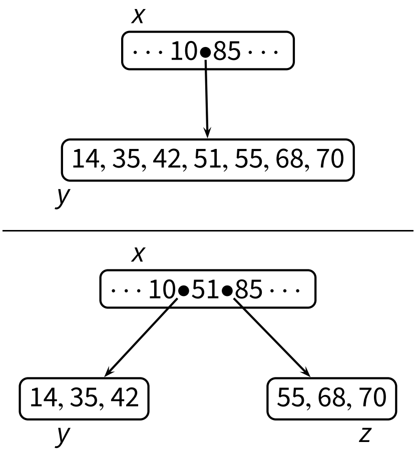
$$n \geq 1 + 2(t^h - 1)$$

Splitting

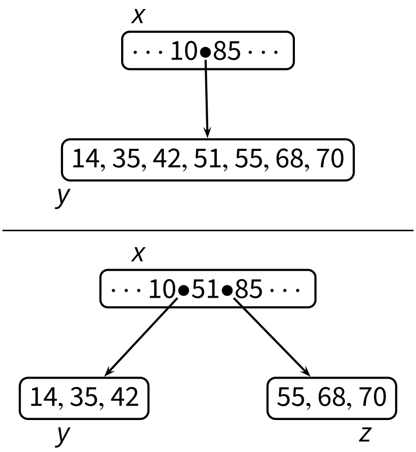
Splitting



Splitting



Splitting




```
B-TREE-SPLIT-CHILD(x, i, y)
z = ALLOCATE-NODE()
z.leaf = y.leaf
z.n = t - 1
for j = 1 to t - 1
    z.key[j] = y.key[j + t]
if not y.leaf
    for j = 1 to t
        z.c[j] = y.c[j + t]
y.n = t - 1
for j = x.n + 1 downto i + 1
    x.c[j + 1] = x.c[j]
for j = x.n downto i
    x.key[j + 1] = x.key[j]
x.key[i] = y.key[t]
x.n = x.n + 1
DISK-WRITE(y)
DISK-WRITE(z)
DISK-WRITE(x)
```

Complexity of B-TREE-SPLIT-CHILD

- What is the complexity of **B-TREE-SPLIT-CHILD**?

Complexity of B-TREE-SPLIT-CHILD

- What is the complexity of **B-TREE-SPLIT-CHILD**?

- $\Theta(t)$ basic CPU operations

Complexity of B-Tree-Split-Child

- What is the complexity of **B-Tree-Split-Child**?
- $\Theta(t)$ basic CPU operations
- 3 **Disk-Write** operations

```
B-TREE-SPLIT-CHILD(x, i, y)
    z = ALLOCATE-NODE()
    z.leaf = y.leaf
    z.n = t − 1
    for j = 1 to t − 1
        x.key[j] = x.key[j + t]
    if not x.leaf
        for j = 1 to t
            z.c[j] = y.c[j + t]
    y.n = t − 1
    for j = x.n + 1 downto i + 1
        x.c[j + 1] = x.c[j]
    for j = x.n downto i
        x.key[j + 1] = x.key[j]
    x.key[i] = y.key[t]
    x.n = x.n + 1
    DISK-WRITE(y)
    DISK-WRITE(z)
    DISK-WRITE(x)
```

Insertion Under Non-Full Node

Insertion Under Non-Full Node

```
B-TREE-INSERT-NONFULL(x, k)
i = x.n                                  // assume x is not full
if x.leaf
    while i ≥ 1 and k < x.key[i]
        x.key[i + 1] = x.key[i]
        i = i − 1
    x.key[i + 1] = k
    x.n = x.n + 1
    DISK-WRITE(x)
else while i ≥ 1 and k < x.key[i]
        i = i − 1
    i = i + 1
    DISK-READ(x.c[i])
    if x.c[i].n == 2t − 1                // child x.c[i] is full
        B-TREE-SPLIT-CHILD(x, i, x.c[i])
        if k > x.key[i]
            i = i + 1
    B-TREE-INSERT-NONFULL(x.c[i], k)
```

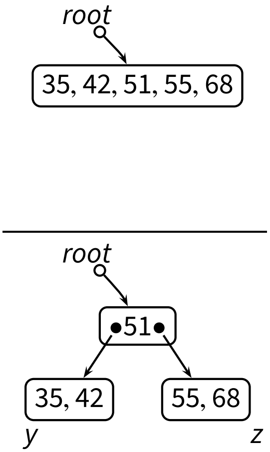
Insertion Procedure

Insertion Procedure

```
B-TREE-INSERT(T, k)
r = T.root
if r.n == 2t − 1
    s = ALLOCATE-NODE()
    T.root = s
    s.leaf = FALSE
    s.n = 0
    s.c[1] = r
    B-TREE-SPLIT-CHILD(s, 1, r)
    B-TREE-INSERT-NONFULL(s, k)
else B-TREE-INSERT-NONFULL(r, k)
```

Insertion Procedure

```
B-TREE-INSERT(T, k)
r = T.root
if r.n == 2t − 1
        s = ALLOCATE-NODE()
        T.root = s
        s.leaf = FALSE
        s.n = 0
        s.c[1] = r
        B-TREE-SPLIT-CHILD(s, 1, r)
        B-TREE-INSERT-NONFULL(s, k)
else B-TREE-INSERT-NONFULL(r, k)
```



Complexity of Insertion

- What is the complexity of **B-TREE-INSERT**?

Complexity of Insertion

- What is the complexity of **B-Tree-Insert**?
- $O(th) = O(t \log_t n)$ basic CPU steps operations

Complexity of Insertion

- What is the complexity of **B-TREE-INSERT**?
- $O(th) = O(t \log_t n)$ basic CPU steps operations
- $O(h) = O(\log_t n)$ disk-access operations

- What is the complexity of **B-Tree-Insert**?
- $O(th) = O(t \log_t n)$ basic CPU steps operations
- $O(h) = O(\log_t n)$ disk-access operations
- The best value for t can be determined according to
  - the ratio between CPU (RAM) speed and disk-access time
  - the *block-size* of the disk, which determines the maximum size of an object that can be accessed (read/write) in one shot