CS 225

Data Structures

October 19 – Intro Kd-trees and Btrees

G Carl Evans
Range-based Searches

Balanced BSTs are useful structures for range-based and nearest-neighbor searches.

Q: Consider points in 1D: \( p = \{p_1, p_2, \ldots, p_n\} \).
   ...what points fall in \([11, 42]\)?

Ex:

3 6 11 33 41 44 55
Range-based Searches

Q: Consider points in 1D: $p = \{p_1, p_2, \ldots, p_n\}$.
   ...what points fall in [11, 42]?

Tree construction:
Range-based Searches

Q: Consider points in 1D: $p = \{p_1, p_2, \ldots, p_n\}$. ...what points fall in $[11, 42]$?
Range-based Searches

Consider points in 2D: \( p = \{ p_1, p_2, \ldots, p_n \} \).

Q: What points are in the rectangle: \( [ (x_1, y_1), (x_2, y_2) ] \)?

Q: What is the nearest point to \( (x_1, y_1) \)?
Range-based Searches

Consider points in 2D: $p = \{p_1, p_2, ..., p_n\}$.

Tree construction:
Range-based Searches
kD-Trees
B-Trees

Q: Can we always fit our data in main memory?

Q: Where else can we keep our data?

However, Our big-O has assumed uniform time for all operations.
Vast Differences in Time

A 3GHz CPU performs 3m operations in _______.

Old Argument: “Disk Storage is Slow”
- Bleeding-edge storage is pretty fast:
  SSD
- Large Disks (25 TB+) still have slow throughout:

New Argument: “The Cloud is Slow!”
AVLs on Disk

![AVL tree diagram]

The image shows an AVL tree with nodes labeled from 1 to 12.
Real Application

Imagine storing TicTok profiles for everyone in the US:

How many records?

How much data in total?

How deep is the AVL tree?
BTree Motivations

Knowing that we have large seek times for data, we want to:
BTree (of order m)

**Goal:** Minimize the number of reads!

Build a tree that uses _________________ / node

- [1 network packet]
- [1 disk block]
BTree Insertion

A **BTree** of order \( m \) is an m-way tree:
- All keys within a node are ordered
- All leaves contain hold no more than \( m-1 \) keys.

\( m=5 \)
BTree Insertion

When a BTree node reaches $m$ keys:
BTree Recursive Insert

```
<table>
<thead>
<tr>
<th>-3</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>42</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>43</td>
<td>55</td>
</tr>
</tbody>
</table>
```

\( m=3 \)
BTree Recursive Insert

$\begin{array}{c}
\text{-3} & 8 & 25 & 31 & 43 & 55
\end{array}$

$m=3$
BTree Visualization/Tool

https://www.cs.usfca.edu/~galles/visualization/BTree.html
Btree Properties

A BTrees of order m is an m-way tree:
- All keys within a node are ordered
- All leaves contain hold no more than \( m-1 \) keys.

- All internal nodes have exactly one more child than keys
- Root nodes can be a leaf or have \([2, m]\) children.
- All non-root, internal nodes have \([\lceil m/2 \rceil, m]\) children.

- All leaves are on the same level
BTree
BTree Search

```
  23
 /   \
-3    42 55
 /     /  \
-11   8   25 31 43 60
```
BTree Search

```cpp
bool Btree::_exists(BTreeNode & node, const K & key) {
    unsigned i;
    for (i = 0; i < node.keys_ct_ && key < node.keys_[i]; i++) { }
    if (i < node.keys_ct_ && key == node.keys_[i]) {
        return true;
    }
    if (node.isLeaf()) {
        return false;
    } else {
        BTreeNode nextChild = node._fetchChild(i);
        return _exists(nextChild, key);
    }
}
```
BTree Analysis

The height of the BTree determines maximum number of ______________ possible in search data.

...and the height of the structure is: ______________.

Therefore: The number of seeks is no more than ____________.

...suppose we want to prove this!
BTree Analysis

In our AVL Analysis, we saw finding an upper bound on the height (given \( n \)) is the same as finding a lower bound on the nodes (given \( h \)).

We want to find a relationship for BTrees between the number of keys (\( n \)) and the height (\( h \)).
BTree Analysis

Strategy:
We will first count the number of nodes, level by level.

Then, we will add the minimum number of keys per node (n).

The minimum number of nodes will tell us the largest possible height (h), allowing us to find an upper-bound on height.
BTree Analysis

The minimum number of nodes for a BTree of order m at each level:

root:

level 1:

level 2:

level 3:

... level h:
BTree Analysis

The **total number of nodes** is the sum of all of the levels:
BTREE Analysis

The total number of keys:
BTree Analysis

The **smallest total number of keys** is:

So an inequality about \( n \), the total number of keys:

Solving for \( h \), since \( h \) is the number of seek operations:
BTree Analysis

Given $m=101$, a tree of height $h=4$ has:

Minimum Keys:

Maximum Keys: