5.2 Tries
Review: summary of the performance of symbol-table implementations

Frequency of operations.

<table>
<thead>
<tr>
<th>implementation</th>
<th>typical case</th>
<th>ordered operations</th>
<th>operations on keys</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>search</td>
<td>insert</td>
<td>delete</td>
</tr>
<tr>
<td>red-black BST</td>
<td>1.00 lg N</td>
<td>1.00 lg N</td>
<td>1.00 lg N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hashing</td>
<td>1 †</td>
<td>1 †</td>
<td>1 †</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† under uniform hashing assumption

Q. Can we do better?
A. Yes, if we can avoid examining the entire key, as with string sorting.
String symbol table basic API

**String symbol table.** Symbol table specialized to string keys.

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>public class StringST&lt;Value&gt;</code></td>
<td>string symbol table type</td>
</tr>
<tr>
<td><code>StringST()</code></td>
<td>create an empty symbol table</td>
</tr>
<tr>
<td><code>void put(String key, Value val)</code></td>
<td>put key-value pair into the symbol table</td>
</tr>
<tr>
<td><code>Value get(String key)</code></td>
<td>return value paired with given key</td>
</tr>
<tr>
<td><code>boolean contains(String key)</code></td>
<td>is there a value paired with the given key?</td>
</tr>
</tbody>
</table>

**Goal.** As fast as hashing, more flexible than binary search trees.
String symbol table implementations cost summary

<table>
<thead>
<tr>
<th>implementation</th>
<th>search (hit)</th>
<th>search (miss)</th>
<th>insert</th>
<th>space (links)</th>
<th>moby.txt</th>
<th>actors.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>red-black BST</td>
<td>L + c ( \lg^2 N )</td>
<td>c ( \lg^2 N )</td>
<td>c ( \lg^2 N )</td>
<td>4 N</td>
<td>1.40</td>
<td>97.4</td>
</tr>
<tr>
<td>hashing</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>4 N to 16 N</td>
<td>0.76</td>
<td>40.6</td>
</tr>
</tbody>
</table>

Parameters
- \( N \) = number of strings
- \( L \) = length of string
- \( R \) = radix

<table>
<thead>
<tr>
<th>file</th>
<th>size</th>
<th>words</th>
<th>distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>moby.txt</td>
<td>1.2 MB</td>
<td>210 K</td>
<td>32 K</td>
</tr>
<tr>
<td>actors.txt</td>
<td>82 MB</td>
<td>11.4 M</td>
<td>900 K</td>
</tr>
</tbody>
</table>

**Challenge.** Efficient performance for string keys.
tries
TSTs
string symbol table API
Tries

Tries. [from retrieval, but pronounced "try"]
- Store characters and values in nodes (not keys).
- Each node has $R$ children, one for each possible character.

Ex. she sells sea shells by the
Search in a trie

Follow links corresponding to each character in the key.
- **Search hit**: node where search ends has a non-null value.
- **Search miss**: reach a null link or node where search ends has null value.
Search in a trie

Follow links corresponding to each character in the key.
- **Search hit**: node where search ends has a non-null value.
- **Search miss**: reach a null link or node where search ends has null value.

get("shell")

get("shore")

no link for the o, so return null

no value in the node corresponding to the last key character, so return null
Insertion into a trie

Follow links corresponding to each character in the key.

• Encounter a null link: create new node.
• Encounter the last character of the key: set value in that node.
Trie construction example

- **she**: 0
  - Value is in node corresponding to last character
  - Nodes corresponding to characters at the end of the key do not exist, so create them and set the value of the last one

- **sells**: 1
  - One node for each key character

- **sea**: 2
  - Key is sequence of characters from root to value

- **shells**: 3
  - Value is in node corresponding to last character

- **by**: 4
  - Node corresponding to the last key character exists, so reset its value

- **shore**: 7
Node. A value, plus references to R nodes.

private static class Node
{
    private Object value;
    private Node[] next = new Node[R];
}
Trie representation: Java implementation

**Node.** A value, plus references to R nodes.

```java
private static class Node {
    private Object value;
    private Node[] next = new Node[R];
}
```

- Trie representation (R = 26)
- Each node has an array of links and a value
- Characters are implicitly defined by link index
- Use `Object` instead of `Value` since no generic array creation in Java
public class TrieST<Value> {
    private static final int R = 256;
    private Node root;

    private static class Node {
        /* see previous slide */
    }

    public void put(String key, Value val) {
        root = put(root, key, val, 0);
    }

    private Node put(Node x, String key, Value val, int d) {
        if (x == null) x = new Node();
        if (d == key.length()) { x.val = val; return x; }
        char c = key.charAt(d);
        x.next[c] = put(x.next[c], key, val, d+1);
        return x;
    }
}
R-way trie: Java implementation (continued)

```java
public boolean contains(String key)
{  return get(key) != null;  }

public Value get(String key)
{  
    Node x = get(root, key, 0);
    if (x == null) return null;
    return (Value) x.val;
}

private Node get(Node x, String key, int d)
{  
    if (x == null) return null;
    if (d == key.length()) return x;
    char c = key.charAt(d);
    return get(x.next[c], key, d+1);
}
```
Trie performance

Search miss.
• Could have mismatch on first character.
• Typical case: examine only a few characters.

Search hit. Need to examine all L characters for equality.

Space. R null links at each leaf.
(but sublinear space possible if many short strings share common prefixes)

Bottom line. Fast search hit, sublinear-time search miss, wasted space.
String symbol table implementations cost summary

<table>
<thead>
<tr>
<th>implementation</th>
<th>character accesses (typical case)</th>
<th>dedup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>search hit</td>
<td>search miss</td>
</tr>
<tr>
<td>red-black BST</td>
<td>$L + c \lg^2 N$</td>
<td>$c \lg^2 N$</td>
</tr>
<tr>
<td>hashing</td>
<td>$L$</td>
<td>$L$</td>
</tr>
<tr>
<td>R-way trie</td>
<td>$L$</td>
<td>$\log_R N$</td>
</tr>
</tbody>
</table>

**R-way trie.**
- Method of choice for small $R$.
- Too much memory for large $R$.

**Challenge.** Use less memory, e.g., 65,536-way trie for Unicode!
Digression: out of memory?

“640 K ought to be enough for anybody.”
— attributed to Bill Gates, 1981
(commenting on the amount of RAM in personal computers)

“64 MB of RAM may limit performance of some Windows XP features; therefore, 128 MB or higher is recommended for best performance.”
— Windows XP manual, 2002

“64 bit is coming to desktops, there is no doubt about that. But apart from Photoshop, I can't think of desktop applications where you would need more than 4GB of physical memory, which is what you have to have in order to benefit from this technology. Right now, it is costly.”
— Bill Gates, 2003
Digression: out of memory?

A short (approximate) history.

<table>
<thead>
<tr>
<th>machine</th>
<th>year</th>
<th>address bits</th>
<th>addressable memory</th>
<th>typical actual memory</th>
<th>cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDP-8</td>
<td>1960s</td>
<td>12</td>
<td>6 KB</td>
<td>6 KB</td>
<td>$16K</td>
</tr>
<tr>
<td>PDP-10</td>
<td>1970s</td>
<td>18</td>
<td>256 KB</td>
<td>256 KB</td>
<td>$1M</td>
</tr>
<tr>
<td>IBM S/360</td>
<td>1970s</td>
<td>24</td>
<td>4 MB</td>
<td>512 KB</td>
<td>$1M</td>
</tr>
<tr>
<td>VAX</td>
<td>1980s</td>
<td>32</td>
<td>4 GB</td>
<td>1 MB</td>
<td>$1M</td>
</tr>
<tr>
<td>Pentium</td>
<td>1990s</td>
<td>32</td>
<td>4 GB</td>
<td>1 GB</td>
<td>$1K</td>
</tr>
<tr>
<td>Xeon</td>
<td>2000s</td>
<td>64</td>
<td>enough</td>
<td>4 GB</td>
<td>$100</td>
</tr>
<tr>
<td>??</td>
<td>future</td>
<td>128+</td>
<td>enough</td>
<td>enough</td>
<td>$1</td>
</tr>
</tbody>
</table>

“512-bit words ought to be enough for anybody.”
— RS, 1995
A modest proposal

Number of atoms in the universe (estimated). \( \leq 2^{266} \).
Age of universe (estimated). 14 billion years \( \sim 2^{59} \) seconds \( \leq 2^{89} \) nanoseconds.

Q. How many bits address every atom that ever existed?
A. Use a unique 512-bit address for every atom at every time quantum.

Ex. Use 256-way trie to map atom to location.
• Represent atom as 64 8-bit chars (512 bits).
• 256-way trie wastes 255/256 actual memory.
• Need better use of memory.
- tries
- TSTs
- string symbol table API
Ternary search tries

**TST.** [Bentley-Sedgewick, 1997]

- Store characters and values in nodes (not keys).
- Each node has three children: smaller (left), equal (middle), larger (right).
**Ternary search tries**

**TST.** [Bentley-Sedgewick, 1997]

- Store characters and values in nodes (not keys).
- Each node has three children: smaller (left), equal (middle), larger (right).

![Ternary search tries diagram](image-url)
Follow links corresponding to each character in the key.
• If less, take left link; if greater, take right link.
• If equal, take the middle link and move to the next key character.

Search hit. Node where search ends has a non-null value.
Search miss. Reach a null link or node where search ends has null value.
26-way trie vs. TST

26-way trie. 26 null links in each leaf.

TST. 3 null links in each leaf.
A TST node is five fields:

- A value.
- A character c.
- A reference to a left TST.
- A reference to a middle TST.
- A reference to a right TST.

```java
private class Node {
    private Value val;
    private char c;
    private Node left, mid, right;
}
```
public class TST<Value> {
    private Node root;

    private class Node {
        /* see previous slide */
    }

    public void put(String key, Value val) {
        root = put(root, key, val, 0);
    }

    private Node put(Node x, String key, Value val, int d) {
        char c = s.charAt(d);
        if (x == null) {
            x = new Node();
            x.c = c;
        }
        if (c < x.c) x.left = put(x.left, key, val, d);
        else if (c > x.c) x.right = put(x.right, key, val, d);
        else if (d < s.length() - 1) x.mid = put(x.mid, key, val, d+1);
        else x.val = val;
        return x;
    }
}
public boolean contains(String key)
{  return get(key) != null;  }

public Value get(String key)
{
    Node x = get(root, key, 0);
    if (x == null) return null;
    return x.val;
}

private Node get(Node x, String key, int d)
{
    if (x == null) return null;
    char c = s.charAt(d);
    if      (c < x.c)              return get(x.left,  key, d);
    else if (c > x.c)              return get(x.right, key, d);
    else if (d < key.length() - 1) return get(x.mid,   key, d+1);
    else                           return x;
}
String symbol table implementation cost summary

<table>
<thead>
<tr>
<th>Implementation</th>
<th>Character accesses (typical case)</th>
<th>Dedup</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>search hit</td>
<td>search miss</td>
</tr>
<tr>
<td>red-black BST</td>
<td>L + c lg^2 N</td>
<td>c lg^2 N</td>
</tr>
<tr>
<td>hashing</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>R-way trie</td>
<td>L</td>
<td>log_R N</td>
</tr>
<tr>
<td>TST</td>
<td>L + ln N</td>
<td>ln N</td>
</tr>
</tbody>
</table>

**Remark.** Can build balanced TSTs via rotations to achieve $L + \log N$ worst-case guarantees.

**Bottom line.** TST is as fast as hashing (for string keys), space efficient.
TST with $R^2$ branching at root

Hybrid of R-way trie and TST.
- Do $R^2$-way branching at root.
- Each of $R^2$ root nodes points to a TST.

Q. What about one- and two-letter words?
## String symbol table implementation cost summary

<table>
<thead>
<tr>
<th>implementation</th>
<th>search hit</th>
<th>search miss</th>
<th>insert</th>
<th>space (links)</th>
<th>moby.txt</th>
<th>actors.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>red-black BST</td>
<td>$L + c \log^2 N$</td>
<td>$c \log^2 N$</td>
<td>$c \log^2 N$</td>
<td>$4N$</td>
<td>1.40</td>
<td>97.4</td>
</tr>
<tr>
<td>hashing</td>
<td>$L$</td>
<td>$L$</td>
<td>$L$</td>
<td>$4N$ to $16N$</td>
<td>0.76</td>
<td>40.6</td>
</tr>
<tr>
<td>R-way trie</td>
<td>$L$</td>
<td>$\log_R N$</td>
<td>$L$</td>
<td>$(R + 1)N$</td>
<td>1.12</td>
<td>out of memory</td>
</tr>
<tr>
<td>TST</td>
<td>$L + \ln N$</td>
<td>$\ln N$</td>
<td>$L + \ln N$</td>
<td>$4N$</td>
<td>0.72</td>
<td>38.7</td>
</tr>
<tr>
<td>TST with $R^2$</td>
<td>$L + \ln N$</td>
<td>$\ln N$</td>
<td>$L + \ln N$</td>
<td>$4N + R^2$</td>
<td>0.51</td>
<td>32.7</td>
</tr>
</tbody>
</table>
TST vs. hashing

Hashing.
- Need to examine entire key.
- Search hits and misses cost about the same.
- Need good hash function for every key type.
- No help for ordered symbol table operations.

TSTs.
- Works only for strings (or digital keys).
- Only examines just enough key characters.
- Search miss may only involve a few characters.
- Can handle ordered symbol table operations (plus others!).

Bottom line. TSTs are:
- Faster than hashing (especially for search misses).
  More flexible than red-black trees (next).
- tries
- TSTs
- string symbol table API
String symbol table API

Character-based operations. The string symbol table API supports several useful character-based operations.

Prefix match. The keys with prefix "sh" are "she", "shells", and "shore".

Longest prefix. The key that is the longest prefix of "shellsort" is "shells".

Wildcard match. The key that match ".he" are "she" and "the".
### String symbol table API

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>StringST()</code></td>
<td>create a symbol table with string keys</td>
</tr>
<tr>
<td><code>StringST(Alphabet alpha)</code></td>
<td>create a symbol table with string keys whose characters are taken from alpha.</td>
</tr>
<tr>
<td><code>void put(String key, Value val)</code></td>
<td>put key-value pair into the symbol table (remove key from table if value is null)</td>
</tr>
<tr>
<td><code>Value get(String key)</code></td>
<td>value paired with key (null if key is absent)</td>
</tr>
<tr>
<td><code>void delete(String key)</code></td>
<td>remove key (and its value) from table</td>
</tr>
<tr>
<td><code>boolean contains(String key)</code></td>
<td>is there a value paired with key?</td>
</tr>
<tr>
<td><code>boolean isEmpty()</code></td>
<td>is the table empty?</td>
</tr>
<tr>
<td><code>String longestPrefixOf(String s)</code></td>
<td>return the longest key that is a prefix of s</td>
</tr>
<tr>
<td><code>Iterable&lt;String&gt; keysWithPrefix(String s)</code></td>
<td>all the keys having s as a prefix.</td>
</tr>
<tr>
<td><code>Iterable&lt;String&gt; keysThatMatch(String s)</code></td>
<td>all the keys that match s (where . matches any character).</td>
</tr>
<tr>
<td><code>int size()</code></td>
<td>number of key-value pairs in the table</td>
</tr>
<tr>
<td><code>Iterable&lt;String&gt; keys()</code></td>
<td>all the keys in the symbol table</td>
</tr>
</tbody>
</table>

**Remark.** Can also add other ordered ST methods, e.g., `floor()` and `rank()`. 

---

**API for a symbol table with string keys**
Deletion in an R-way trie

To delete a key-value pair:
- Find the node corresponding to key and set value to null.
- If that node has all null links, remove that node (and recur).

```plaintext
delete("shells");
```

Deleting a key (and its associated value) from a trie
Ordered iteration

To iterate through all keys in sorted order:
• Do inorder traversal of trie; add keys encountered to a queue.
• Maintain sequence of characters on path from root to node.

```
key      q
  b
 by
  s
 se
 sea
 sel
 sell
 sells
 sh
 she
 shell
 shells
 sho
 shor
 shore
 t
 th
 the

keysWithPrefix("");
```

Collecting the keys in a trie (trace)
To iterate through all keys in sorted order:
• Do inorder traversal of trie; add keys encountered to a queue.
• Maintain sequence of characters on path from root to node.

```java
public Iterable<String> keys()
{
    Queue<String> queue = new Queue<String>();
    collect(root, "", queue);
    return queue;
}

private void collect(Node x, String prefix, Queue<String> q)
{
    if (x == null) return;
    if (x.val != null) q.enqueue(prefix);
    for (char c = 0; c < R; c++)
        collect(x.next[c], prefix + c, q);
}
```
Prefix matches

Find all keys in symbol table starting with a given prefix.

Ex. Autocomplete in a cell phone, search bar, text editor, or shell.
- User types characters one at a time.
- System reports all matching strings.
Find all keys in symbol table starting with a given prefix.

```java
public Iterable<String> keysWithPrefix(String prefix) {
    Queue<String> queue = new Queue<String>();
    Node x = get(root, prefix, 0);
    collect(x, prefix, queue);
    return queue;
}
```

Prefix match in a trie
Longest prefix

Find longest key in symbol table that is a prefix of query string.

Ex. Search IP database for longest prefix matching destination IP, and route packets accordingly.

```
"128"
"128.112"
"128.112.055"
"128.112.055.15"
"128.112.136"
"128.112.155.11"
"128.112.155.13"
"128.222"
"128.222.136"
```

```
prefix("128.112.136.11") = "128.112.136"
prefix("128.166.123.45") = "128"
```

Q. Why isn't longest prefix match the same as floor or ceiling?
Longest prefix

Find longest key in symbol table that is a prefix of query string.
• Search for query string.
• Keep track of longest key encountered.
Longest prefix: Java implementation

Find longest key in symbol table that is a prefix of query string.
• Search for query string.
• Keep track of longest key encountered.

```java
public String longestPrefixOf(String query) {
    int length = search(root, query, 0, 0);
    return query.substring(0, length);
}

private int search(Node x, String query, int d, int length) {
    if (x == null) return length;
    if (x.val != null) length = d;
    if (d == query.length()) return length;
    char c = query.charAt(d);
    return search(x.next[c], query, d+1, length);
}
```
T9 texting

**Goal.** Type text messages on a phone keypad.

**Multi-tap input.** Enter a letter by repeatedly pressing a key until the desired letter appears.

**T9 text input.** ["A much faster and more fun way to enter text."]
- Find all words that correspond to given sequence of numbers.
- Press 0 to see all completion options.

**Ex.** hello
- **Multi-tap:** 4 4 3 3 5 5 5 5 5 6 6 6
- **T9:** 4 3 5 5 6

![T9 keypad diagram](www.t9.com)
To: info@t9support.com
Date: Tue, 25 Oct 2005 14:27:21 -0400 (EDT)

Dear T9 texting folks,

I enjoyed learning about the T9 text system from your webpage, and used it as an example in my data structures and algorithms class. However, one of my students noticed a bug in your phone keypad

http://www.t9.com/images/how.gif

Somehow, it is missing the letter s. (!)

Just wanted to bring this information to your attention and thank you for your website.

Regards,

Kevin
To: "'Kevin Wayne'" <wayne@CS.Princeton.EDU>
Date: Tue, 25 Oct 2005 12:44:42 -0700

Thank you Kevin.

I am glad that you find T9 o valuable for your cla. I had not noticed thi before. Thank for writing in and letting u know.

Take care,

Brooke nyder
OEM Dev upport
AOL/Tegic Communication
1000 Dexter Ave N. uite 300
eattle, WA 98109

ALL INFORMATION CONTAINED IN THIS EMAIL IS CONSIDERED CONFIDENTIAL AND PROPERTY OF AOL/TEGIC COMMUNICATIONS
Compressing a trie

Collapsing 1-way branches at bottom.
Internal node stores character; leaf node stores suffix (or full key).

Collapsing interior 1-way branches.
Node stores a sequence of characters.

Removing one-way branching in a trie

```c
put("shells", 1);
put("shellfish", 2);
```
A classic algorithm

**Patricia tries.** [Practical Algorithm to Retrieve Information Coded in Alphanumeric]

- Collapse one-way branches in binary trie.
- Thread trie to eliminate multiple node types.

**Applications.**
- Database search.
- P2P network search.
- IP routing tables: find longest prefix match.
- Compressed quad-tree for N-body simulation.
- Efficiently storing and querying XML documents.

**Implementation.** One step beyond this lecture.
Suffix tree

**Suffix tree.** Threaded trie with collapsed 1-way branching for string suffixes.

Applications.
- Linear-time longest repeated substring.
- Computational biology databases (BLAST, FASTA).

Implementation. One step beyond this lecture.
String symbol tables summary

A success story in algorithm design and analysis.

Red-black tree.
• Performance guarantee: \( \log N \) key compares.
• Supports ordered symbol table API.

Hash tables.
• Performance guarantee: constant number of probes.
• Requires good hash function for key type.

Tries. R-way, TST.
• Performance guarantee: \( \log N \) characters accessed.
• Supports extensions to API based on partial keys.

Bottom line. You can get at anything by examining 50-100 bits (!!!)