5.1 **STRING Sorts**

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
5.1 STRING SORTS

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
String processing

**String.** Sequence of characters.

**Important fundamental abstraction.**
- Genomic sequences.
- Information processing.
- Communication systems (e.g., email).
- Programming systems (e.g., Java programs).
- ...

“*The digital information that underlies biochemistry, cell biology, and development can be represented by a simple string of G's, A's, T's and C's. This string is the root data structure of an organism's biology.*” — M. V. Olson
The char data type

**C char data type.** Typically an 8-bit integer.
- Supports 7-bit ASCII.
- Can represent at most 256 characters.

<table>
<thead>
<tr>
<th>0 1 2 3 4 5 6 7 8 9 A B C D E F</th>
<th>0 1 2 3 4 5 6 7 8 9 A B C D E F</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUL SOH STX ETX EOT ENQ ACK BEL BS HT LF VT FF CR SO SI</td>
<td></td>
</tr>
<tr>
<td>DLE DC1 DC2 DC3 DC4 NAK SYN ETB CAN EM SUB ESC FS GS RS US</td>
<td></td>
</tr>
<tr>
<td>SP ! &quot; # $ % &amp; ' ( ) * + , - . /</td>
<td></td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7 8 9 : ; &lt; = &gt; ?</td>
<td></td>
</tr>
<tr>
<td>@ A B C D E F G H I J K L M N O</td>
<td></td>
</tr>
<tr>
<td>P Q R S T U V W X Y Z [ \ ] ^ _</td>
<td></td>
</tr>
<tr>
<td>` a b c d e f g h i j k l m n o</td>
<td></td>
</tr>
<tr>
<td>p q r s t u v w x y z {</td>
<td>} ~ DEL</td>
</tr>
</tbody>
</table>

Hexadecimal to ASCII conversion table

**Java char data type.** A 16-bit unsigned integer.
- Supports original 16-bit Unicode.
- Supports 21-bit Unicode 3.0 (awkwardly).
I ♥ Unicode
The String data type

String data type in Java. Immutable sequence of characters.

Length. Number of characters.
Indexing. Get the $i^{th}$ character.
Concatenation. Concatenate one string to the end of another.

```
String s = "ATTACKAKTDAWN"

s.length()  // 12
s.charAt(3)  // K
s.substring(7, 11)  // DAWN
```
The String data type: immutability

Q. Why immutable?

A. All the usual reasons.
   • Can use as keys in symbol table.
   • Don't need to defensively copy.
   • Ensures consistent state.
   • Supports concurrency.
   • Improves security.

public class FileInputStream
{
    private String filename;
    public FileInputStream(String filename)
    {
        if (!allowedToReadFile(filename))
            throw new SecurityException();
        this.filename = filename;
    }
    ...
}

attacker could bypass security if string type were mutable
The String data type: representation

Representation (Java 7). Immutable char[] array + cache of hash.

<table>
<thead>
<tr>
<th>operation</th>
<th>Java</th>
<th>running time</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>s.length()</td>
<td>1</td>
</tr>
<tr>
<td>indexing</td>
<td>s.charAt(i)</td>
<td>1</td>
</tr>
<tr>
<td>concatenation</td>
<td>s + t</td>
<td>$M + N$</td>
</tr>
</tbody>
</table>

...
String performance trap

Q. How to build a long string, one character at a time?

A. Use StringBuilder data type (mutable char[] array).
Comparing two strings

Q. How many character compares to compare two strings of length $W$?

![Example strings](https://via.placeholder.com/150)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>r</td>
<td>e</td>
<td>f</td>
<td>e</td>
<td>t</td>
<td>c</td>
<td>h</td>
</tr>
<tr>
<td>p</td>
<td>r</td>
<td>e</td>
<td>f</td>
<td>i</td>
<td>x</td>
<td>e</td>
<td>s</td>
</tr>
</tbody>
</table>

Running time. Proportional to length of longest common prefix.
- Proportional to $W$ in the worst case.
- But, often sublinear in $W$. 
### Alphabets

**Digital key.** Sequence of digits over fixed alphabet.

**Radix.** Number of digits $R$ in alphabet.

<table>
<thead>
<tr>
<th>name</th>
<th>$R()$</th>
<th>$\lg R()$</th>
<th>characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BINARY</td>
<td>2</td>
<td>1</td>
<td>01</td>
</tr>
<tr>
<td>OCTAL</td>
<td>8</td>
<td>3</td>
<td>01234567</td>
</tr>
<tr>
<td>DECIMAL</td>
<td>10</td>
<td>4</td>
<td>0123456789</td>
</tr>
<tr>
<td>HEXADECIMAL</td>
<td>16</td>
<td>4</td>
<td>0123456789ABCDEF</td>
</tr>
<tr>
<td>DNA</td>
<td>4</td>
<td>2</td>
<td>ACTG</td>
</tr>
<tr>
<td>LOWERCASE</td>
<td>26</td>
<td>5</td>
<td>abcdefghijklmnopqrstuvwxyz</td>
</tr>
<tr>
<td>UPPERCASE</td>
<td>26</td>
<td>5</td>
<td>ABCDEFGHIJKLMNOPQRSTUVWXYZ</td>
</tr>
<tr>
<td>PROTEIN</td>
<td>20</td>
<td>5</td>
<td>ACDEFGHIJKLMNOPQRSTUVWXYZWY</td>
</tr>
<tr>
<td>BASE64</td>
<td>64</td>
<td>6</td>
<td>ABCDEFGHIJKLMNOPQRSTUVWXYZ0123456789+/-</td>
</tr>
<tr>
<td>ASCII</td>
<td>128</td>
<td>7</td>
<td>ASCII characters</td>
</tr>
<tr>
<td>EXTENDED_ASCII</td>
<td>256</td>
<td>8</td>
<td>extended ASCII characters</td>
</tr>
<tr>
<td>UNICODE16</td>
<td>65536</td>
<td>16</td>
<td>Unicode characters</td>
</tr>
</tbody>
</table>
5.1 String Sorts

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
Review: summary of the performance of sorting algorithms

Frequency of operations.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>guarantee</th>
<th>random</th>
<th>extra space</th>
<th>stable?</th>
<th>operations on keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>insertion sort</td>
<td>$\frac{1}{2} N^2$</td>
<td>$\frac{1}{4} N^2$</td>
<td>1</td>
<td>✓</td>
<td>compareTo()</td>
</tr>
<tr>
<td>mergesort</td>
<td>$N \lg N$</td>
<td>$N \lg N$</td>
<td>$N$</td>
<td>✓</td>
<td>compareTo()</td>
</tr>
<tr>
<td>quicksort</td>
<td>$1.39 N \lg N^*$</td>
<td>$1.39 N \lg N$</td>
<td>$c \lg N$</td>
<td></td>
<td>compareTo()</td>
</tr>
<tr>
<td>heapsort</td>
<td>$2 N \lg N$</td>
<td>$2 N \lg N$</td>
<td>1</td>
<td></td>
<td>compareTo()</td>
</tr>
</tbody>
</table>

* probabilistic

Lower bound. $\sim N \lg N$ compares required by any compare-based algorithm.

Q. Can we do better (despite the lower bound)?

A. Yes, if we don't depend on key compares.  

use array accesses to make R-way decisions (instead of binary decisions)
Key-indexed counting: assumptions about keys

Assumption. Keys are integers between 0 and \( R - 1 \).

Implication. Can use key as an array index.

Applications.
- Sort string by first letter.
- Sort class roster by section.
- Sort phone numbers by area code.
- Subroutine in a sorting algorithm. [stay tuned]

Remark. Keys may have associated data \( \Rightarrow \) can't just count up number of keys of each value.

<table>
<thead>
<tr>
<th>input name</th>
<th>section</th>
<th>sorted result (by section)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson</td>
<td>2</td>
<td>Harris 1</td>
</tr>
<tr>
<td>Brown</td>
<td>3</td>
<td>Martin 1</td>
</tr>
<tr>
<td>Davis</td>
<td>3</td>
<td>Moore 1</td>
</tr>
<tr>
<td>Garcia</td>
<td>4</td>
<td>Anderson 2</td>
</tr>
<tr>
<td>Harris</td>
<td>1</td>
<td>Martinez 2</td>
</tr>
<tr>
<td>Jackson</td>
<td>3</td>
<td>Miller 2</td>
</tr>
<tr>
<td>Johnson</td>
<td>4</td>
<td>Robinson 2</td>
</tr>
<tr>
<td>Jones</td>
<td>3</td>
<td>White 2</td>
</tr>
<tr>
<td>Martin</td>
<td>1</td>
<td>Brown 3</td>
</tr>
<tr>
<td>Martinez</td>
<td>2</td>
<td>Davis 3</td>
</tr>
<tr>
<td>Miller</td>
<td>2</td>
<td>Jackson 3</td>
</tr>
<tr>
<td>Moore</td>
<td>1</td>
<td>Jones 3</td>
</tr>
<tr>
<td>Robinson</td>
<td>2</td>
<td>Taylor 3</td>
</tr>
<tr>
<td>Smith</td>
<td>4</td>
<td>Williams 3</td>
</tr>
<tr>
<td>Taylor</td>
<td>3</td>
<td>Garcia 4</td>
</tr>
<tr>
<td>Thomas</td>
<td>4</td>
<td>Johnson 4</td>
</tr>
<tr>
<td>Thompson</td>
<td>4</td>
<td>Smith 4</td>
</tr>
<tr>
<td>White</td>
<td>2</td>
<td>Thomas 4</td>
</tr>
<tr>
<td>Williams</td>
<td>3</td>
<td>Thompson 4</td>
</tr>
<tr>
<td>Wilson</td>
<td>4</td>
<td>Wilson 4</td>
</tr>
</tbody>
</table>

Typical candidate for key-indexed counting:

- input
- sorted result (by section)
- keys are small integers
Key-indexed counting demo

Goal. Sort an array \( a[] \) of \( N \) integers between 0 and \( R - 1 \).
- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```java
int N = a.length;
int[] count = new int[R+1];

for (int i = 0; i < N; i++)
    count[a[i]+1]++;

for (int r = 0; r < R; r++)
    count[r+1] += count[r];

for (int i = 0; i < N; i++)
    aux[count[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
    a[i] = aux[i];
```

\[ R = 6 \]

<table>
<thead>
<tr>
<th>i</th>
<th>a[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>d</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
<tr>
<td>4</td>
<td>f</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>d</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
</tr>
<tr>
<td>8</td>
<td>f</td>
</tr>
<tr>
<td>9</td>
<td>b</td>
</tr>
<tr>
<td>10</td>
<td>e</td>
</tr>
<tr>
<td>11</td>
<td>a</td>
</tr>
</tbody>
</table>

use a for 0
b for 1
c for 2
d for 3
e for 4
f for 5
Key-indexed counting demo

**Goal.** Sort an array \( a[] \) of \( N \) integers between 0 and \( R - 1 \).

- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```java
int N = a.length;
int[] count = new int[R+1];

for (int i = 0; i < N; i++)
    count[a[i]+1]++;

for (int r = 0; r < R; r++)
    count[r+1] += count[r];

for (int i = 0; i < N; i++)
    aux[count[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
    a[i] = aux[i];
```

-offset by 1

<table>
<thead>
<tr>
<th>i</th>
<th>a[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>d</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
<tr>
<td>4</td>
<td>f</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>d</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
</tr>
<tr>
<td>8</td>
<td>f</td>
</tr>
<tr>
<td>9</td>
<td>b</td>
</tr>
<tr>
<td>10</td>
<td>e</td>
</tr>
<tr>
<td>11</td>
<td>a</td>
</tr>
</tbody>
</table>

- offset by 1
- stay tuned

<table>
<thead>
<tr>
<th>r</th>
<th>count[r]</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>3</td>
</tr>
<tr>
<td>d</td>
<td>1</td>
</tr>
<tr>
<td>e</td>
<td>2</td>
</tr>
<tr>
<td>f</td>
<td>1</td>
</tr>
<tr>
<td>_</td>
<td>3</td>
</tr>
</tbody>
</table>
Key-indexed counting demo

**Goal.** Sort an array `a[]` of `N` integers between 0 and `R - 1`.

- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```java
int N = a.length;
int[] count = new int[R+1];

for (int i = 0; i < N; i++)
    count[a[i]+1]++;

for (int r = 0; r < R; r++)
    count[r+1] += count[r];

for (int i = 0; i < N; i++)
    aux[count[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
    a[i] = aux[i];
```

<table>
<thead>
<tr>
<th>i</th>
<th>a[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>d</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
</tr>
<tr>
<td>4</td>
<td>f</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>d</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
</tr>
<tr>
<td>8</td>
<td>f</td>
</tr>
<tr>
<td>9</td>
<td>b</td>
</tr>
<tr>
<td>10</td>
<td>e</td>
</tr>
<tr>
<td>11</td>
<td>a</td>
</tr>
</tbody>
</table>

r | count[r] |
---|----------|
 a | 0        |
 b | 2        |
 c | 5        |
 d | 6        |
 e | 8        |
 f | 9        |
 g | 12       |

6 keys < d, 8 keys < e
so d's go in a[6] and a[7]
Key-indexed counting demo

**Goal.** Sort an array \( a[] \) of \( N \) integers between 0 and \( R - 1 \).
- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```
int N = a.length;
int[] count = new int[R+1];

for (int i = 0; i < N; i++)
    count[a[i]+1]++;

for (int r = 0; r < R; r++)
    count[r+1] += count[r];

for (int i = 0; i < N; i++)
    aux[count[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
    a[i] = aux[i];
```
Key-indexed counting demo

**Goal.** Sort an array \( a[] \) of \( N \) integers between 0 and \( R - 1 \).

- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```java
int N = a.length;
int[] count = new int[R+1];

for (int i = 0; i < N; i++)
    count[a[i]+1]++;

for (int r = 0; r < R; r++)
    count[r+1] += count[r];

for (int i = 0; i < N; i++)
    aux[count[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
    a[i] = aux[i];
```

<table>
<thead>
<tr>
<th>i</th>
<th>a[i]</th>
<th>i</th>
<th>aux[i]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a</td>
<td>0</td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>1</td>
<td>a</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>2</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>3</td>
<td>b</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>4</td>
<td>b</td>
</tr>
<tr>
<td>5</td>
<td>c</td>
<td>5</td>
<td>c</td>
</tr>
<tr>
<td>6</td>
<td>d</td>
<td>6</td>
<td>d</td>
</tr>
<tr>
<td>7</td>
<td>d</td>
<td>7</td>
<td>d</td>
</tr>
<tr>
<td>8</td>
<td>e</td>
<td>8</td>
<td>e</td>
</tr>
<tr>
<td>9</td>
<td>f</td>
<td>9</td>
<td>f</td>
</tr>
<tr>
<td>10</td>
<td>f</td>
<td>10</td>
<td>f</td>
</tr>
<tr>
<td>11</td>
<td>f</td>
<td>11</td>
<td>f</td>
</tr>
</tbody>
</table>
Key-indexed counting: analysis

**Proposition.** Key-indexed takes time proportional to $N + R$.

**Proposition.** Key-indexed counting uses extra space proportional to $N + R$.

**Stable? ✓**

```plaintext
| a[0] | Anderson 2 | Harris 1 | aux[0] |
| a[12]| Robinson 2 | Taylor 3  | aux[12]|
| a[16]| Thompson 4 | Smith 4   | aux[16]|
| a[18]| Williams 3 | Thompson 4 | aux[18]|
```
5.1 String Sorts

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
Least-significant-digit-first string sort

LSD string (radix) sort.

- Consider characters from right to left.
- Stably sort using $d^{th}$ character as the key (using key-indexed counting).

<table>
<thead>
<tr>
<th>d=0</th>
<th>d=1</th>
<th>d=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
<tr>
<td>b</td>
<td>a</td>
<td>d</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>e</td>
</tr>
<tr>
<td>d</td>
<td>a</td>
<td>f</td>
</tr>
<tr>
<td>e</td>
<td>d</td>
<td>g</td>
</tr>
<tr>
<td>f</td>
<td>e</td>
<td>h</td>
</tr>
</tbody>
</table>

Sort is stable (arrows do not cross)
LSD string sort: correctness proof

**Proposition.** LSD sorts fixed-length strings in ascending order.

**Pf.** [by induction on $i$]

After pass $i$, strings are sorted by last $i$ characters.

- If two strings differ on sort key, key-indexed sort puts them in proper relative order.
- If two strings agree on sort key, stability keeps them in proper relative order.

**Proposition.** LSD sort is stable.

**Pf.** Key-indexed counting is stable.
public class LSD
{
    public static void sort(String[] a, int W)
    {
        int R = 256;
        int N = a.length;
        String[] aux = new String[N];

        for (int d = W-1; d >= 0; d--)
        {
            int[] count = new int[R+1];
            for (int i = 0; i < N; i++)
                count[a[i].charAt(d) + 1]++;
            for (int r = 0; r < R; r++)
                count[r+1] += count[r];
            for (int i = 0; i < N; i++)
                aux[count[a[i].charAt(d)]++] = a[i];
            for (int i = 0; i < N; i++)
                a[i] = aux[i];
        }
    }
}

fixed-length W strings
radix R
do key-indexed counting for each digit from right to left
key-indexed counting
# Summary of the performance of sorting algorithms

## Frequency of operations.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>guarantee</th>
<th>random</th>
<th>extra space</th>
<th>stable?</th>
<th>operations on keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>insertion sort</td>
<td>( \frac{1}{2} N^2 )</td>
<td>( \frac{1}{4} N^2 )</td>
<td>1</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>mergesort</td>
<td>( N \log N )</td>
<td>( N \log N )</td>
<td>( N )</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>quicksort</td>
<td>( 1.39 N \log N )</td>
<td>( 1.39 N \log N )</td>
<td>( c \log N )</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>heapsort</td>
<td>( 2 N \log N )</td>
<td>( 2 N \log N )</td>
<td>1</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>LSD sort †</td>
<td>( 2W (N + R) )</td>
<td>( 2W (N + R) )</td>
<td>( N + R )</td>
<td>✔</td>
<td>charAt()</td>
</tr>
</tbody>
</table>

* † probabilistic
  † fixed-length \( W \) keys

### Q. What if strings are not all of same length?
String sorting interview question

**Problem.** Sort one million 32-bit integers.

**Ex.** Google (or presidential) interview.

**Which sorting method to use?**

- Insertion sort.
- Mergesort.
- Quicksort.
- Heapsort.
- LSD string sort.
String sorting interview question

Google CEO Eric Schmidt interviews Barack Obama
How to take a census in 1900s?

1880 Census. Took 1500 people 7 years to manually process data.

Herman Hollerith. Developed counting and sorting machine to automate.

- Use punch cards to record data (e.g., gender, age).
- Machine sorts one column at a time (into one of 12 bins).
- Typical question: how many women of age 20 to 30?

Hollerith tabulating machine and sorter

punch card (12 holes per column)

1890 Census. Finished in 1 year (and under budget)!
How to get rich sorting in 1900s?

Punch cards. [1900s to 1950s]

- Also useful for accounting, inventory, and business processes.
- Primary medium for data entry, storage, and processing.

Hollerith's company later merged with 3 others to form Computing Tabulating Recording Corporation (CTRC); company renamed in 1924.
LSD string sort: a moment in history (1960s)

To sort a card deck
- start on right column
- put cards into hopper
- machine distributes into bins
- pick up cards (stable)
- move left one column
- continue until sorted

Lysergic Acid Diethylamide
(Lucy in the Sky with Diamonds)
5.1 String Sorts

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
Reverse LSD

- Consider characters from left to right.
- Stably sort using $d^{th}$ character as the key (using key-indexed counting).

### Table Examples

<table>
<thead>
<tr>
<th>d = 0</th>
<th>d = 1</th>
<th>d = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>2</td>
<td>c</td>
<td>a</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
<td>a</td>
</tr>
<tr>
<td>4</td>
<td>f</td>
<td>e</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
<td>d</td>
</tr>
<tr>
<td>6</td>
<td>d</td>
<td>a</td>
</tr>
<tr>
<td>7</td>
<td>b</td>
<td>e</td>
</tr>
<tr>
<td>8</td>
<td>f</td>
<td>d</td>
</tr>
<tr>
<td>9</td>
<td>b</td>
<td>e</td>
</tr>
<tr>
<td>10</td>
<td>e</td>
<td>b</td>
</tr>
<tr>
<td>11</td>
<td>a</td>
<td>c</td>
</tr>
</tbody>
</table>

Sort key (d = 0): 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Sort key (d = 1): 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Sort key (d = 2): 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

Not sorted!
Most-significant-digit-first string sort

MSD string (radix) sort.
- Partition array into $R$ pieces according to first character (use key-indexed counting).
- Recursively sort all strings that start with each character (key-indexed counts delineate subarrays to sort).

```
0  d  a  b
1  a  d  d
2  c  a  b
3  f  a  d
4  f  e  e
5  b  a  d
6  d  a  d
7  b  e  e
8  f  e  d
9  b  e  d
10 e  b  b
11 a  c  e
```

```
0  a  d  d
1  a  c  e
2  b  a  d
3  b  e  e
4  b  e  d
5  c  a  b
6  d  a  b
7  d  a  d
8  e  b  b
9  f  a  d
10 f  e  e
11 f  e  d
```

```
0  a  d  d
1  a  c  e
```

```
2  b  a  d
3  b  e  e
4  b  e  d
```

```
5  c  a  b
6  d  a  b
```

```
7  d  a  d
8  e  b  b
9  f  a  d
10 f  e  e
11 f  e  d
```

```
0  a  d  d
1  a  c  e
```

```
2  b  a  d
3  b  e  e
4  b  e  d
```

```
5  c  a  b
6  d  a  b
```

```
7  d  a  d
8  e  b  b
9  f  a  d
10 f  e  e
11 f  e  d
```

```
0  a  d  d
1  a  c  e
```

```
2  b  a  d
3  b  e  e
4  b  e  d
```

```
5  c  a  b
6  d  a  b
```

```
7  d  a  d
8  e  b  b
9  f  a  d
10 f  e  e
11 f  e  d
```
### MSD string sort: example

**input**

<table>
<thead>
<tr>
<th>she</th>
<th>sells</th>
<th>seashells</th>
<th>by</th>
<th>the</th>
<th>sea</th>
<th>shore</th>
<th>the</th>
<th>shells</th>
<th>she</th>
<th>sells</th>
<th>surely</th>
<th>the</th>
<th>seashells</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
</tr>
</thead>
<tbody>
<tr>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
</tr>
<tr>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
</tr>
<tr>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
</tr>
<tr>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
</tr>
<tr>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
</tr>
<tr>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
</tr>
<tr>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
</tr>
<tr>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
</tr>
<tr>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
</tr>
<tr>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
</tr>
</tbody>
</table>

**output**

<table>
<thead>
<tr>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
<th>are</th>
</tr>
</thead>
<tbody>
<tr>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
<td>by</td>
</tr>
<tr>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
<td>sea</td>
</tr>
<tr>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
<td>seashells</td>
</tr>
<tr>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
</tr>
<tr>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
<td>sells</td>
</tr>
<tr>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
</tr>
<tr>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
<td>shore</td>
</tr>
<tr>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
<td>shells</td>
</tr>
<tr>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
<td>she</td>
</tr>
<tr>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
<td>surely</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
<td>the</td>
</tr>
</tbody>
</table>

**Trace of recursive calls for MSD string sort (no cutoff for small subarrays, subarrays of size 0 and 1 omitted)**

---

**Note:**

- The MSD string sort algorithm is used to sort strings lexicographically.
- The trace shows the recursive calls made during the sort.
- The algorithm uses a divide-and-conquer strategy, where the string is divided into substrings based on the first character and sorted recursively.
- The output is the sorted list of strings.

---

**End of string goes before any char value**
**Variable-length strings**

Treat strings as if they had an extra char at end (smaller than any char).

<table>
<thead>
<tr>
<th></th>
<th>sea</th>
<th>a</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>sea</td>
<td>s</td>
<td>he l l s</td>
</tr>
<tr>
<td>2</td>
<td>se l l s</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>she</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>she</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>she l l s</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>sh o r e</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>s u r e l y</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

```
private static int charAt(String s, int d) {
    if (d < s.length()) return s.charAt(d);
    else return -1;
}
```

**C strings.** Have extra char '\0' at end \[no extra work needed.\]
public static void sort(String[] a) {
    aux = new String[a.length];
    sort(a, aux, 0, a.length - 1, 0);
}

private static void sort(String[] a, String[] aux, int lo, int hi, int d) {
    if (hi <= lo) return;
    int[] count = new int[R+2];
    for (int i = lo; i <= hi; i++)
        count[charAt(a[i], d) + 2]++;
    for (int r = 0; r < R+1; r++)
        count[r+1] += count[r];
    for (int i = lo; i <= hi; i++)
        aux[count[charAt(a[i], d) + 1]] = a[i];
    for (int i = lo; i <= hi; i++)
        a[i] = aux[i - lo];

    for (int r = 0; r < R; r++)
        sort(a, aux, lo + count[r], lo + count[r+1] - 1, d+1);
}
MSD string sort: potential for disastrous performance

**Observation 1.** Much too slow for small subarrays.
- Each function call needs its own `count[]` array.
- ASCII (256 counts): 100x slower than copy pass for $N = 2$.
- Unicode (65,536 counts): 32,000x slower for $N = 2$.

**Observation 2.** Huge number of small subarrays because of recursion.
Cutoff to insertion sort

**Solution.** Cutoff to insertion sort for small subarrays.

- Insertion sort, but start at $d^{th}$ character.

```java
private static void sort(String[] a, int lo, int hi, int d)
{
    for (int i = lo; i <= hi; i++)
        for (int j = i; j > lo && less(a[j], a[j-1], d); j--)
            exch(a, j, j-1);
}
```

- Implement `less()` so that it compares starting at $d^{th}$ character.

```java
private static boolean less(String v, String w, int d)
{
    for (int i = d; i < Math.min(v.length(), w.length()); i++)
    {
        if (v.charAt(i) < w.charAt(i)) return true;
        if (v.charAt(i) > w.charAt(i)) return false;
    }
    return v.length() < w.length();
}
```
MSD string sort: performance

Number of characters examined.

- MSD examines just enough characters to sort the keys.
- Number of characters examined depends on keys.
- Can be sublinear in input size!

Characters examined by MSD string sort

<table>
<thead>
<tr>
<th>Random (sublinear)</th>
<th>Non-random with duplicates (nearly linear)</th>
<th>Worst case (linear)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1EI0402</td>
<td>are</td>
<td>1DNB377</td>
</tr>
<tr>
<td>1HYL490</td>
<td>by</td>
<td>1DNB377</td>
</tr>
<tr>
<td>1R0Z572</td>
<td>sea</td>
<td>1DNB377</td>
</tr>
<tr>
<td>2HXE734</td>
<td>seashells</td>
<td>1DNB377</td>
</tr>
<tr>
<td>2IYE230</td>
<td>seashells</td>
<td>1DNB377</td>
</tr>
<tr>
<td>2X0R846</td>
<td>sells</td>
<td>1DNB377</td>
</tr>
<tr>
<td>3CDB573</td>
<td>sells</td>
<td>1DNB377</td>
</tr>
<tr>
<td>3CVP720</td>
<td>she</td>
<td>1DNB377</td>
</tr>
<tr>
<td>3IGJ319</td>
<td>she</td>
<td>1DNB377</td>
</tr>
<tr>
<td>3KNA382</td>
<td>shells</td>
<td>1DNB377</td>
</tr>
<tr>
<td>3TAV879</td>
<td>shore</td>
<td>1DNB377</td>
</tr>
<tr>
<td>4CQP781</td>
<td>surely</td>
<td>1DNB377</td>
</tr>
<tr>
<td>4QGI284</td>
<td>the</td>
<td>1DNB377</td>
</tr>
<tr>
<td>4YHV229</td>
<td>the</td>
<td>1DNB377</td>
</tr>
</tbody>
</table>
### Summary of the performance of sorting algorithms

#### Frequency of operations.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>guarantee</th>
<th>random</th>
<th>extra space</th>
<th>stable?</th>
<th>operations on keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>insertion sort</td>
<td>( \frac{1}{2} N^2 )</td>
<td>( \frac{1}{4} N^2 )</td>
<td>1</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>mergesort</td>
<td>( N \lg N )</td>
<td>( N \lg N )</td>
<td>( N )</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>quicksort</td>
<td>1.39 ( N \lg N )</td>
<td>1.39 ( N \lg N )</td>
<td>( c \lg N )</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>heapsort</td>
<td>2 ( N \lg N )</td>
<td>2 ( N \lg N )</td>
<td>1</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>LSD sort ( ^\dagger )</td>
<td>2 ( W(N + R) )</td>
<td>2 ( W(N + R) )</td>
<td>( N + R )</td>
<td>✔</td>
<td>charAt()</td>
</tr>
<tr>
<td>MSD sort ( ^\ddagger )</td>
<td>2 ( W(N + R) )</td>
<td>( N \log_R N )</td>
<td>( N + DR )</td>
<td>✔</td>
<td>charAt()</td>
</tr>
</tbody>
</table>

D = function-call stack depth (length of longest prefix match)

* probabilistic
\( ^\dagger \) fixed-length W keys
\( ^\ddagger \) average-length W keys
MSD string sort vs. quicksort for strings

Disadvantages of MSD string sort.
• Extra space for aux[].
• Extra space for count[].
• Inner loop has a lot of instructions.
• Accesses memory "randomly" (cache inefficient).

Disadvantage of quicksort.
• Linearithmic number of string compares (not linear).
• Has to rescan many characters in keys with long prefix matches.

Goal. Combine advantages of MSD and quicksort.
Optimization 0. Cutoff to insertion sort.

Optimization 1. Replace recursion with explicit stack.
- Push subarrays to be sorted onto stack.
- Now, one count[] array suffices.

Optimization 2. Do $R$-way partitioning in place.
- Eliminates aux[] array.
- Sacrifices stability.

---

American national flag problem

Dutch national flag problem

---

Engineering a radix sort (American flag sort)

ABSTRACT: Radix sorting methods have excellent asymptotic performance on string data, for which comparison is not a unit-time operation. Attractive for use in large byte-addressable memories, these methods have nevertheless long been eclipsed by more easily programmed algorithms. Three ways to sort strings by bytes left to right—a stable list sort, a stable two-array sort, and an in-place “American flag” sort—are illustrated with practical C programs. For heavy-duty sorting, all three perform comparably, usually running at least twice as fast as a good quicksort. We recommend American flag sort for general use.
5.1 STRING SORTS

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
3-way string quicksort (Bentley and Sedgewick, 1997)

Overview. Do 3-way partitioning on the $d^{th}$ character.
- Less overhead than $R$-way partitioning in MSD string sort.
- Does not re-examine characters equal to the partitioning char.
  (but does re-examine characters not equal to the partitioning char)
3-way string quicksort: trace of recursive calls

Trace of first few recursive calls for 3-way string quicksort (subarrays of size 1 not shown)
3-way string quicksort: Java implementation

```java
private static void sort(String[] a) {
    sort(a, 0, a.length - 1, 0);
}

private static void sort(String[] a, int lo, int hi, int d) {
    if (hi <= lo) return;
    int lt = lo, gt = hi;
    int v = charAt(a[lo], d);
    int i = lo + 1;
    while (i <= gt) {
        int t = charAt(a[i], d);
        if (t < v) exch(a, lt++, i);
        else if (t > v) exch(a, i, gt--);
        else i++;
    }
    sort(a, lo, lt-1, d);
    if (v >= 0) sort(a, lt, gt, d+1);
    sort(a, gt+1, hi, d);
}
```
3-way string quicksort vs. standard quicksort

**Standard quicksort.**
- Uses $\sim 2N \ln N$ string compares on average.
- Costly for keys with long common prefixes (and this is a common case!)

**3-way string (radix) quicksort.**
- Uses $\sim 2N \ln N$ character compares on average for random strings.
- Avoids re-comparing long common prefixes.

---

Fast Algorithms for Sorting and Searching Strings

Jon L. Bentley*  Robert Sedgewick#

**Abstract**
We present theoretical algorithms for sorting and searching multikey data, and derive from them practical C implementations for applications in which keys are character strings. The sorting algorithm blends Quicksort and radix sort, it is competitive with the best known C sort codes. The searching algorithm blends tries and binary that is competitive with the most efficient string sorting programs known. The second program is a symbol table implementation that is faster than hashing, which is commonly regarded as the fastest symbol table implementation. The symbol table implementation is much more space-efficient than multiway trees, and supports more advanced searches.
3-way string quicksort vs. MSD string sort

MSD string sort.

- Is cache-inefficient.
- Too much memory storing count[].
- Too much overhead reinitializing count[] and aux[].

3-way string quicksort.

- Is cache-friendly.
- Is in-place.
- Has a short inner loop.

Bottom line. 3-way string quicksort is method of choice for sorting strings.
Summary of the performance of sorting algorithms

Frequency of operations.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>guarantee</th>
<th>random</th>
<th>extra space</th>
<th>stable?</th>
<th>operations on keys</th>
</tr>
</thead>
<tbody>
<tr>
<td>insertion sort</td>
<td>$\frac{1}{2} N^2$</td>
<td>$\frac{1}{4} N^2$</td>
<td>1</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>mergesort</td>
<td>$N \lg N$</td>
<td>$N \lg N$</td>
<td>$N$</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>quicksort</td>
<td>$1.39 N \lg N$*</td>
<td>$1.39 N \lg N$</td>
<td>$c \lg N$</td>
<td>✔</td>
<td>compareTo()</td>
</tr>
<tr>
<td>heapsort</td>
<td>$2 N \lg N$</td>
<td>$2 N \lg N$</td>
<td>1</td>
<td></td>
<td>compareTo()</td>
</tr>
<tr>
<td>LSD sort †</td>
<td>$2W(N + R)$</td>
<td>$2W(N + R)$</td>
<td>$N + R$</td>
<td>✔</td>
<td>charAt()</td>
</tr>
<tr>
<td>MSD sort ‡</td>
<td>$2W(N + R)$</td>
<td>$N \log R N$</td>
<td>$N + DR$</td>
<td>✔</td>
<td>charAt()</td>
</tr>
<tr>
<td>3-way string</td>
<td>$1.39 WN \lg R$*</td>
<td>$1.39 WN \lg N$</td>
<td>$\log N + W$</td>
<td></td>
<td>charAt()</td>
</tr>
<tr>
<td>3-way string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* probabilistic
† fixed-length W keys
‡ average-length W keys
5.1 String Sorts

- strings in Java
- key-indexed counting
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays
Keyword-in-context search

Given a text of $N$ characters, preprocess it to enable fast substring search (find all occurrences of query string context).

```
% more tale.txt
it was the best of times
it was the worst of times
it was the age of wisdom
it was the age of foolishness
it was the epoch of belief
it was the epoch of incredulity
it was the season of light
it was the season of darkness
it was the spring of hope
it was the winter of despair
```

Applications. Linguistics, databases, web search, word processing, ....
Keyword-in-context search

Given a text of \( N \) characters, preprocess it to enable fast substring search (find all occurrences of query string context).

Applications. Linguistics, databases, web search, word processing, ....
Suffix sort

**input string**

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
it was best it was w
```

**form suffixes**

```
0 it was best it was w
1 twas best it was w
2 was best it was w
3 as best it was w
4 sbest it was w
5 best it was w
6 est it was w
7 st it was w
8 tit was w
9 it was w
10 twas w
11 was w
12 as w
13 s w
14 w
```

**sort suffixes to bring query strings together**

```
0 it was best it was w
9 it was w
4 sbest it was w
7 st it was w
8 tit was w
9 it was w
12 as w
10 twas w
11 was w
```

**array of suffix indices in sorted order**
Keyword-in-context search: suffix-sorting solution

- Preprocess: **suffix sort** the text.
- Query: **binary search** for query; scan until mismatch.

KWIC search for "search" in Tale of Two Cities

<table>
<thead>
<tr>
<th>Position</th>
<th>Search Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>632698</td>
<td>sealed my letter and...</td>
</tr>
<tr>
<td>713727</td>
<td>seamstress is lifted...</td>
</tr>
<tr>
<td>660598</td>
<td>seamstress of twenty...</td>
</tr>
<tr>
<td>67610</td>
<td>seamstress who was with...</td>
</tr>
<tr>
<td>4430</td>
<td>search for contraband...</td>
</tr>
<tr>
<td>42705</td>
<td>search for your father...</td>
</tr>
<tr>
<td>499797</td>
<td>search of her husband...</td>
</tr>
<tr>
<td>182045</td>
<td>search of impoverish...</td>
</tr>
<tr>
<td>143399</td>
<td>search of other carry...</td>
</tr>
<tr>
<td>411801</td>
<td>search the straw hold...</td>
</tr>
<tr>
<td>158410</td>
<td>seared marking about...</td>
</tr>
<tr>
<td>691536</td>
<td>seas and madame defar...</td>
</tr>
<tr>
<td>536569</td>
<td>seas a terrible pass...</td>
</tr>
<tr>
<td>484763</td>
<td>seas that had brought...</td>
</tr>
</tbody>
</table>
Q. How to efficiently form (and sort) suffixes?

```java
String[] suffixes = new String[N];
for (int i = 0; i < N; i++)
    suffixes[i] = s.substring(i, N);
Arrays.sort(suffixes);
```
public final class String implements Comparable<String> {
    private char[] value;  // characters
    private int offset;    // index of first char in array
    private int length;    // length of string
    private int hash;      // cache of hashCode()
    ...
}

String s = "Hello, World"

String t = s.substring(7, 12);
The String data type: Java 7u6 implementation

```java
public final class String implements Comparable<String>
{
    private char[] value;  // characters
    private int hash;      // cache of hashCode()
    ...
}
```

String `s = "Hello, World"

```
  value[]
  H   E   L   L   L   O   ,   W   O   R   L   D
    0   1   2   3   4   5   6   7   8   9   10  11
```

String `t = s.substring(7, 12);

```
  value[]
  W   O   R   L   D
    0   1   2   3   4
```
The String data type: performance

String data type (in Java). Sequence of characters (immutable).
Java 7u5. Immutable char[] array, offset, length, hash cache.
Java 7u6. Immutable char[] array, hash cache.

<table>
<thead>
<tr>
<th>operation</th>
<th>Java 7u5</th>
<th>Java 7u6</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>indexing</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>substring extraction</td>
<td>1</td>
<td>$N$</td>
</tr>
<tr>
<td>concatenation</td>
<td>$M + N$</td>
<td>$M + N$</td>
</tr>
<tr>
<td>immutable?</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>memory</td>
<td>$64 + 2N$</td>
<td>$56 + 2N$</td>
</tr>
</tbody>
</table>
I'm the author of the substring() change. As has been suggested in the analysis here there were two motivations for the change:

- Reduce the size of String instances. Strings are typically 20-40% of common apps footprint.
- Avoid memory leakage caused by retained substrings holding the entire character array.

Changing this function, in a bugfix release no less, was totally irresponsible. It broke backwards compatibility for numerous applications with errors that didn't even produce a message, just freezing and timeouts... All pain, no gain. Your work was not just vain, it was thoroughly destructive, even beyond its immediate effect.

http://www.reddit.com/r/programming/comments/1qw73v/til_oracle_changed_the_internal_string
Suffix sort

Q. How to efficiently form (and sort) suffixes in Java 7u6?
A. Define Suffix class ala Java 7u5 String.

```java
public class Suffix implements Comparable<Suffix>
{
    private final String text;
    private final int offset;
    public Suffix(String s, int offset)
    {
        this.text = text;
        this.offset = offset;
    }
    public int length() { return text.length() - offset; }
    public char charAt(int i) { return text.charAt(offset + i); }
    public int compareTo(Suffix that) { /* see textbook */ }
}
```

text[]

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>E</th>
<th>L</th>
<th>L</th>
<th>O</th>
<th>,</th>
<th>W</th>
<th>O</th>
<th>R</th>
<th>L</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

offset
Q. How to efficiently form (and sort) suffixes in Java 7u6?
A. Define Suffix class ala Java 7u5 String.

```java
String[] suffixes = new String[N];
for (int i = 0; i < N; i++)
    suffixes[i] = new Suffix(s, i);
Arrays.sort(suffixes);
```
Lessons learned

Lesson 1. Put performance guarantees in API.
Lesson 2. If API has no performance guarantees, don't rely upon any!

Corollary. May want to avoid String data type for huge strings.
  • Are you sure charAt() and length() take constant time?
  • If lots of calls to charAt(), overhead for function calls is large.
  • If lots of small strings, memory overhead of String is large.

Ex. Our optimized algorithm for suffix arrays is 5x faster and uses 32x less memory than our original solution in Java 7u5!
**Suffix Arrays: theory**

Q. What is worst-case running time of our suffix arrays algorithm?
- Quadratic.
- Linearithmic.
- Linear.
- None of the above.  \[ N^2 \log N \]
Suffix Arrays: theory

Q. What is complexity of suffix arrays?
   • Quadratic.
   • Linearithmic.  Manber-Myers algorithm (see video)
   ✔  Linear.  suffix trees (beyond our scope)
   • Nobody knows.

Suffix arrays:
A new method for on-line string searches

Udi Manber¹
Gene Myers²
Department of Computer Science
University of Arizona
Tucson, AZ 85721

May 1989
Revised August 1991

Abstract

A new and conceptually simple data structure, called a suffix array, for on-line string searches is introduced in this paper. Constructing and querying suffix arrays is reduced to a sort and search paradigm that employs novel algorithms. The main advantage of suffix arrays over suffix trees is that, in practice, they use three to five times less space. From a complexity standpoint, suffix arrays permit on-line string searches of the type, “Is W a substring of A?” to be answered in time $O(P + \log N)$, where $P$ is the length of $W$ and $N$ is the length of $A$, which is competitive with (and in some cases slightly better than) suffix trees. The only drawback is that in those instances where the underlying alphabet is finite and small, suffix trees can be constructed in $O(N)$ time in the worst case, versus $O(N \log N)$ time for suffix arrays. However, we give an augmented algorithm that, regardless of the alphabet size, constructs suffix arrays in $O(N)$ expected time, albeit with lesser space efficiency. We believe that suffix arrays will prove to be better in practice than suffix trees for many applications.
Suffix Arrays: practice

Applications. Bioinformatics, information retrieval, data compression, ...

Many ingenious algorithms.

- Memory footprint very important.
- State-of-the-art still changing.

<table>
<thead>
<tr>
<th>year</th>
<th>algorithm</th>
<th>worst case</th>
<th>memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>Manber–Myers</td>
<td>$N \log N$</td>
<td>$8 N$</td>
</tr>
<tr>
<td>1999</td>
<td>Larsson–Sadakane</td>
<td>$N \log N$</td>
<td>$8 N$</td>
</tr>
<tr>
<td>2003</td>
<td>Kärkkäinen–Sanders</td>
<td>$N$</td>
<td>$13 N$</td>
</tr>
<tr>
<td>2003</td>
<td>Ko–Aluru</td>
<td>$N$</td>
<td>$10 N$</td>
</tr>
<tr>
<td>2008</td>
<td>divsufsort2</td>
<td>$N \log N$</td>
<td>$5 N$</td>
</tr>
<tr>
<td>2010</td>
<td>sais</td>
<td>$N$</td>
<td>$6 N$</td>
</tr>
</tbody>
</table>
String sorting summary

We can develop linear-time sorts.
- Key compares not necessary for string keys.
- Use characters as index in an array.

We can develop sublinear-time sorts.
- Input size is amount of data in keys (not number of keys).
- Not all of the data has to be examined.

3-way string quicksort is asymptotically optimal.
- $1.39 \, N \, \lg \, N$ chars for random data.

Long strings are rarely random in practice.
- Goal is often to learn the structure!
- May need specialized algorithms.