Hashing, Search Applications

Acknowledgement: The course slides are adapted from the slides prepared by R. Sedgewick and K. Wayne of Princeton University.
ST implementations: summary

<table>
<thead>
<tr>
<th>implementation</th>
<th>worst-case cost (after N inserts)</th>
<th>average-case cost (after N random inserts)</th>
<th>ordered iteration?</th>
<th>key interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequential search</td>
<td>N</td>
<td>N/2</td>
<td>no</td>
<td>equals()</td>
</tr>
<tr>
<td>binary search</td>
<td>lg N</td>
<td>N/2</td>
<td>yes</td>
<td>compareTo()</td>
</tr>
<tr>
<td>BST</td>
<td>N</td>
<td>1.38 lg N</td>
<td>yes</td>
<td>compareTo()</td>
</tr>
<tr>
<td>red-black BST</td>
<td>2 lg N</td>
<td>1.00 lg N</td>
<td>yes</td>
<td>compareTo()</td>
</tr>
</tbody>
</table>

Q. Can we do better?
A. Yes, but with different access to the data (if we don’t need ordered ops).

ASCII Table to map Characters

- This example is easy as we have a table that maps each character to an index naturally.
- Can we extend this idea, as a general solution for Symbol Tables?
  
- First step:
  - Extend this idea to a subset of integers between 0 and M.
  - Simple, just create an array of size M

- Second step:
  - Can we generalise for integers between -Infinity and +Infinity
  - Not so feasible! Create an array of size Infinity.
  - Probably in the data we will only observe a small subset of the integers
  - So, the first problem with this approach is the Domain Size (number of valid inputs)

Motivation: Counting Characters

- Assume that you are coding a program to count the frequency of characters between a-z
- The algorithm is very easy as below
- Create an array for the frequencies, a character can be transformed to array index by: c - ‘a’.

```java
int size = 'z'-'a'+1;
int[] counts = new int[size];
String text = “Lorem ipsum…”;
for (int i = 0; i < text.length(); i++) {
    if(text.charAt(i)>='a' && text.charAt(i)<='z') {
        counts[text.charAt(i)-'a']++;
    }
}
for (int i = 0; i < counts.length; i++) {
    System.out.println((char)(i+'a') + " " + counts[i]);
}
```

Mapping a Larger Domain to Smaller

IDEA: Find a mapping from our real values to a smaller number of array indices.

Create a function h(x) that maps key values between 0 and 10^10 to values between 0 and 10 so that we can store them in an array of size 10

One example; h(x) = x % 10

If we are lucky the keys will map uniformly to these numbers, so if we have 10 numbers the array will store one value per each cell. But, if we have 11 keys, than a cell will certainly have multiple keys mapped to it. This is called as collision!
Hash Tables

- We will generalise our solution such that:
  - We can map any large domain to a feasible array
  - Able to map any object/data type to index, so a hash function working for any data type.
  - Accept the possibility of collisions and find a strategy to resolve them.

- There is a concept called as Minimal Perfect Hashing, which maps key values domain to array indices one-to-one. Our example for characters is a good example for this. But this will not be possible for most data types (we will have to settle for one-to-many mapping, i.e. collisions!).

### Hashing: basic plan

Save items in a key-indexed table (index is a function of the key).

**Hash function.** Method for computing array index from key.

<table>
<thead>
<tr>
<th>Index</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>“it”</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

**Issues.**
- Computing the hash function.
- Equality test: Method for checking whether two keys are equal.
- Collision resolution: Algorithm and data structure to handle two keys that hash to the same array index.

**Classic space-time tradeoff.**
- No space limitation: trivial hash function with key as index. Very large index table, few collisions
- No time limitation: trivial collision resolution with sequential search. Small table, lots of collisions, must search within the cell.
- Space and time limitations: hashing (the real world).

### Computing the hash function

**Idealistic goal.** Scramble the keys uniformly to produce a table index.
- Efficiently computable.
- Each table index equally likely for each key.

**Ex 1. Phone numbers.**
- Bad: first three digits.
- Better: last three digits.

**Ex 2. Social Security numbers.**
- Bad: first three digits.
- Better: last three digits.

**Practical challenge.** Need different approach for each key type.
Java’s hash code conventions

All Java classes inherit a method `hashCode()`, which returns a 32-bit `int`.

**Requirement.** If `x.equals(y)`, then `(x.hashCode() == y.hashCode())`.

**Highly desirable.** If `!x.equals(y)`, then `(x.hashCode() != y.hashCode())`.

**Default implementation.** Memory address of `x`.

**Legal (but poor) implementation.** Always return 17.

**Customized implementations.** `Integer`, `Double`, `String`, `File`, `URL`, `Date`, ...

**User-defined types.** Users are on their own.

Implementing hash code: integers, booleans, and doubles

**Java library implementations**

```
public final class Integer {
  private final int value;
  ...
  public int hashCode() {
    return value;
  }
}
```

```
public final class Double {
  private final double value;
  ...
  public int hashCode() {
    long bits = doubleToLongBits(value);
    return (int) (bits ^ (bits >>> 32));
  }
}
```

```
public final class Boolean {
  private final boolean value;
  ...
  public int hashCode() {
    if (value) return 1231;
    else       return 1237;
  }
}
```

**Implementing hash code: strings**

```
public final class String {
  private final char[] s;
  ...
  public int hashCode() {
    int hash = 0;
    for (int i = 0; i < length(); i++)
      hash = s[i] + (31 * hash);
    return hash;
  }
}
```

* Horner’s method to hash string of length `L`: `L` multiplies/adds.
* Equivalent to \( \tilde{h} = s[0] \cdot 31^{L-1} + \ldots + s[L-2] \cdot 31^1 + s[L-1] \cdot 31^0 \).

**Ex.** `String s = "call";`  
`int code = s.hashCode();`  
\[ 3045982 = 99 \cdot 31^3 + 97 \cdot 31^2 + 108 \cdot 31^1 + 108 \cdot 31^0 \]
\[ = 108 \cdot 31 \cdot (108 \cdot 31 \cdot (97 + 31) \cdot (99)) \]  
(Horner’s method)

**Performance optimization.**

* Cache the hash value in an instance variable.
* Return cached value.

```
public final class String {
  private int hash = 0;
  private final char[] s;
  ...
  public int hashCode() {
    int h = hash;
    if (h != 0) return h;
    for (int i = 0; i < length(); i++)
      h = s[i] + (31 * hash);
    hash = h;
    return h;
  }
}
```
Implementing hash code: user-defined types

```java
public final class Transaction implements Comparable<Transaction>
{
    private final String who;
    private final Date when;
    private final double amount;
    public Transaction(String who, Date when, double amount)
    { /* as before */ }
    ...
    public boolean equals(Object y)
    { /* as before */ }
    public int hashCode()
    {
        int hash = 17;
        hash = 31*hash + who.hashCode();
        hash = 31*hash + when.hashCode();
        hash = 31*hash + ((Double) amount).hashCode();
        return hash;
    }
}
```

Hash code design

“Standard” recipe for user-defined types.
- Combine each significant field using the $31 \times y$ rule.
- If field is a primitive type, use wrapper type `hashCode()`.
- If field is null, return 0.
- If field is a reference type, use `hashCode()`.
- If field is an array, apply to each entry. apply rule recursively or use `Arrays.deepHashCode()`

In practice. Recipe works reasonably well; used in Java libraries.
In theory. Keys are bitstring; "universal" hash functions exist.

Basic rule. Need to use the whole key to compute hash code;
consult an expert for state-of-the-art hash codes.

Modular hashing

Hash code. An `int` between $-2^{31}$ and $2^{31}-1$.
Hash function. An `int` between 0 and $M-1$ (for use as array index).

```java
    private int hash(Key key)
    { return key.hashCode() % M; }
```

- Bug

```java
    private int hash(Key key)
    { return Math.abs(key.hashCode()) % M; }
```

- 1-in-a-billion bug

```java
    private int hash(Key key)
    { return (key.hashCode() & 0x7fffffff) % M; }
```

- Correct

Uniform hashing assumption

Uniform hashing assumption. Each key is equally likely to hash to an integer between 0 and $M - 1$.

Bins and balls. Throw balls uniformly at random into $M$ bins.

Birthday problem. Expect two balls in the same bin after $\sim \pi M / 2$ tosses.

Coupon collector. Expect every bin has $\geq 1$ ball after $\sim M \ln M$ tosses.

Load balancing. After $M$ tosses, expect most loaded bin has $\Theta (\log M / \log \log M)$ balls.
**Uniform hashing assumption**

Uniform hashing assumption. Each key is equally likely to hash to an integer between 0 and \( M - 1 \).

**Bins and balls.** Throw balls uniformly at random into \( M \) bins.

Bins and balls graph.

**Collisions**

Collision. Two distinct keys hashing to same index.
- Birthday problem \( \Rightarrow \) can’t avoid collisions unless you have a ridiculous (quadratic) amount of memory.
- Coupon collector + load balancing \( \Rightarrow \) collisions will be evenly distributed.

Challenge. Deal with collisions efficiently.

**Separate chaining symbol table**

Use an array of \( M < N \) linked lists. [H. P. Luhn, IBM 1953]
- Hash: map key to integer \( i \) between 0 and \( M - 1 \).
- Insert: put at front of \( i \)th chain (if not already there).
- Search: need to search only \( i \)th chain.

Separate chaining graph.
public class SeparateChainingHashST<Key, Value> {
    private int M = 97;  // number of chains
    private Node[] st = new Node[M];  // array of chains

    private static class Node {
        private Object key;
        private Object val;
        private Node next;
    }

    private int hash(Key key) {
        return (key.hashCode() & 0x7fffffff) % M;
    }

    public Value get(Key key) {
        int i = hash(key);
        for (Node x = st[i]; x != null; x = x.next)
            if (key.equals(x.key)) return (Value) x.val;
        return null;
    }

    public void put(Key key, Value val) {
        int i = hash(key);
        for (Node x = st[i]; x != null; x = x.next)
            if (key.equals(x.key)) { x.val = val; return; }
        st[i] = new Node(key, val, st[i]);
    }
}

Separate chaining ST: Java implementation

Analysis of separate chaining

Proposition. Under uniform hashing assumption, probability that the number of keys in a list is within a constant factor of $N / M$ is extremely close to 1.

Pf sketch. Distribution of list size obeys a binomial distribution.

Consequence. Number of probes for search/insert is proportional to $N / M$.
- $M$ too large ⇒ too many empty chains.
- $M$ too small ⇒ chains too long.
- Typical choice: $M \sim N / 5$ ⇒ constant-time ops.

ST implementations: summary

<table>
<thead>
<tr>
<th>Implementation</th>
<th>Worst-case cost (after $N$ inserts)</th>
<th>Average case (after $N$ random inserts)</th>
<th>Ordered iteration?</th>
<th>Key interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequential search</td>
<td>N</td>
<td>N/2</td>
<td>N/2</td>
<td>no</td>
</tr>
<tr>
<td>binary search</td>
<td>lg N</td>
<td>lg N</td>
<td>lg N/2</td>
<td>yes</td>
</tr>
<tr>
<td>BST</td>
<td>N</td>
<td>1.38 lg N</td>
<td>1.38 lg N</td>
<td>yes</td>
</tr>
<tr>
<td>red-black tree</td>
<td>2 lg N</td>
<td>2 lg N</td>
<td>2 lg N</td>
<td>yes</td>
</tr>
<tr>
<td>separated chaining</td>
<td>lg N</td>
<td>lg N</td>
<td>lg N</td>
<td>yes</td>
</tr>
</tbody>
</table>

* under uniform hashing assumption
Hashing

- Hash functions
- Separate chaining
- Linear probing

Collision resolution: open addressing

Open addressing. [Amdahl-Boehme-Rocherster-Samuel, IBM 1953]
When a new key collides, find next empty slot, and put it there.

Linear probing hash table

Hash. Map key to integer \( i \) between 0 and \( M - 1 \).
Insert. Put at table index \( i \) if free; if not try \( i + 1 \), \( i + 2 \), etc.

Linear probing hash table

Hash. Map key to integer \( i \) between 0 and \( M - 1 \).
Insert. Put at table index \( i \) if free; if not try \( i + 1 \), \( i + 2 \), etc.

M = 16

Hash. Map key to integer \( i \) between 0 and \( M - 1 \).
Insert. Put at table index \( i \) if free; if not try \( i + 1 \), \( i + 2 \), etc.

M = 16
Linear probing hash table

**Hash.** Map key to integer $i$ between 0 and $M-1$.
**Insert.** Put at table index $i$ if free; if not try $i+1$, $i+2$, etc.

```plaintext
insert S
hash(S) = 6
```

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
M = 16
S
```

Linear probing hash table

**Hash.** Map key to integer $i$ between 0 and $M-1$.
**Insert.** Put at table index $i$ if free; if not try $i+1$, $i+2$, etc.

```plaintext
insert E
hash(E) = 10
```

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
M = 16
```

Linear probing hash table

**Hash.** Map key to integer $i$ between 0 and $M-1$.
**Insert.** Put at table index $i$ if free; if not try $i+1$, $i+2$, etc.

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
M = 16
```

Linear probing hash table

**Hash.** Map key to integer $i$ between 0 and $M-1$.
**Insert.** Put at table index $i$ if free; if not try $i+1$, $i+2$, etc.

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
M = 16
```
**Linear probing hash table**

*Hash.* Map key to integer $i$ between 0 and $M - 1$.

*Insert.* Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert E
hash(E) = 10
```

```
<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

$M = 16$

```
```

**Linear probing hash table**

*Hash.* Map key to integer $i$ between 0 and $M - 1$.

*Insert.* Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert E
hash(E) = 10
```

```
<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

$M = 16$

```
```

**Linear probing hash table**

*Hash.* Map key to integer $i$ between 0 and $M - 1$.

*Insert.* Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
linear probing hash table
```

```
<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

$M = 16$

```
```

**Linear probing hash table**

*Hash.* Map key to integer $i$ between 0 and $M - 1$.

*Insert.* Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert A
hash(A) = 4
```

```
<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

$M = 16$

```
```

```
```
Linear probing hash table

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1 \), \( i + 2 \), etc.

```
insert A
hash(A) = 4
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>st[]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( M = 16 \)

A

Linear probing hash table

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1 \), \( i + 2 \), etc.

```
insert R
hash(R) = 14
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>st[]</td>
<td></td>
<td></td>
<td>A</td>
<td>S</td>
<td>E</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>

\( M = 16 \)
**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>S</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$M = 16$

---

**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Insert R**

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$M = 16$

---

**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Insert C**

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$M = 16$
**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M − 1$.
**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert C
hash(C) = 5
```

```
M = 16

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>S</td>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M − 1$.
**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert H
hash(H) = 4
```

```
M = 16

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>C</td>
<td>S</td>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert H
hash(H) = 4
```

```
M = 16
H
```

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
A C S E R
```

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
A C S E R
```

```
M = 16
H
```

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
```

```
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
```

```
M = 16
H
```

```
M = 16
H
```
Linear probing hash table

Hash. Map key to integer $i$ between 0 and $M - 1$.
Insert. Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

$\text{hash}(H) = 4$

\begin{center}
\begin{tabular}{ccccccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
\hline
\end{tabular}
\end{center}

$M = 16$

Linear probing hash table

Hash. Map key to integer $i$ between 0 and $M - 1$.
Insert. Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

$\text{hash}(X) = 15$

\begin{center}
\begin{tabular}{ccccccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
\hline
\end{tabular}
\end{center}

$M = 16$
**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert X
hash(X) = 15
```

```
<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

$M = 16$

**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

```
insert M
hash(M) = 1
```

```
<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

$M = 16$
**Linear probing hash table**

Hash. Map key to integer \( i \) between 0 and \( M - 1 \).

Insert. Put at table index \( i \) if free; if not try \( i + 1, i + 2, \) etc.

<table>
<thead>
<tr>
<th>( M )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{hash}(M) = 1 )</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M = 16 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>

**Linear probing hash table**

Hash. Map key to integer \( i \) between 0 and \( M - 1 \).

Insert. Put at table index \( i \) if free; if not try \( i + 1, i + 2, \) etc.

<table>
<thead>
<tr>
<th>( M )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{hash}(P) = 14 )</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M = 16 )</td>
<td></td>
<td>P</td>
<td></td>
<td></td>
<td></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>
**Linear probing hash table**

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1, i + 2, \) etc.

<table>
<thead>
<tr>
<th>Index</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( M = 16 \)

**Example:**

<table>
<thead>
<tr>
<th>Insert</th>
<th>Hash(P) = 14</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Linear probing hash table**

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1, i + 2, \) etc.

<table>
<thead>
<tr>
<th>Index</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( M = 16 \)

**Example:**

<table>
<thead>
<tr>
<th>Insert</th>
<th>Hash(L) = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Linear probing hash table**

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1, i + 2, \) etc.

<table>
<thead>
<tr>
<th>Index</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>E</td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( M = 16 \)
### Linear probing hash table

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1 \), \( i + 2 \), etc.

**Insert**

Insert \( L \)

**Hash**

Insert \( L \)

\[
\text{Hash}(L) = 6
\]

\[
\begin{array}{cccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
P & M & A & C & S & H & E & | & R & X \\
\end{array}
\]

\( M = 16 \)

\[
\begin{array}{cccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline
P & M & A & C & S & H & E & | & R & X \\
\end{array}
\]

\( M = 16 \)
**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Search.** Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Search.** Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

**Search hit** (return corresponding value)
**Linear probing hash table**

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Search.** Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

---

Linear probing hash table

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>L</td>
<td>E</td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$M = 16$

---

**Linear probing hash table**

Hash. Map key to integer $i$ between 0 and $M - 1$.

Insert. Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

Search. Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

---

Linear probing hash table

Hash. Map key to integer $i$ between 0 and $M - 1$.

Insert. Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

Search. Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

---

Linear probing hash table

Hash. Map key to integer $i$ between 0 and $M - 1$.

Insert. Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

Search. Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.
### Linear probing hash table

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Search.** Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

```plaintext
search L
hash(L) = 6
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w[]$</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>L</td>
<td>E</td>
<td></td>
<td></td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M = 16$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>

### Linear probing hash table

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Search.** Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

```plaintext
search K
hash(K) = 5
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w[]$</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>L</td>
<td>E</td>
<td></td>
<td></td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M = 16$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>

---

**Hash.** Map key to integer $i$ between 0 and $M - 1$.

**Insert.** Put at table index $i$ if free; if not try $i + 1$, $i + 2$, etc.

**Search.** Search table index $i$; if occupied but no match, try $i + 1$, $i + 2$, etc.

```plaintext
search K
hash(K) = 5
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w[]$</td>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>L</td>
<td>E</td>
<td></td>
<td></td>
<td>R</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M = 16$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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>
**Linear probing hash table**

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1, i + 2 \), etc.

**Search.** Search table index \( i \); if occupied but no match, try \( i + 1, i + 2 \), etc.

```
search K
hash(K) = 5
```

```
M = 16
```

```
\[ \begin{array}{c}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\
\hline
P & M & A & C & S & H & L & E & R & X \\
\end{array} \]
```

**Search miss (return null)**
**Linear probing - Summary**

**Hash.** Map key to integer \( i \) between 0 and \( M - 1 \).

**Insert.** Put at table index \( i \) if free; if not try \( i + 1, i + 2 \), etc.

**Search.** Search table index \( i \); if occupied but no match, try \( i + 1, i + 2 \), etc.

**Note.** Array size \( M \) must be greater than number of key-value pairs \( N \).

### M = 16

<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>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>M</td>
<td>A</td>
<td>C</td>
<td>S</td>
<td>H</td>
<td>L</td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Clustering**

**Cluster.** A contiguous block of items.

**Observation.** New keys likely to hash into middle of big clusters.

---

**Linear probing ST implementation**

```java
public class LinearProbingHashST<Key, Value>
{
    private int M = 30001;
    private Value[] vals = (Value[]) new Object[M];
    private Key[] keys = (Key[]) new Object[M];

    private int hash(Key key) {
        /* as before */
    }

    public void put(Key key, Value val)
    {
        int i;
        for (i = hash(key); keys[i] != null; i = (i+1) % M)
            if (keys[i].equals(key))
                break;
        keys[i] = key;
        vals[i] = val;
    }

    public Value get(Key key)
    {
        for (int i = hash(key); keys[i] != null; i = (i+1) % M)
            if (key.equals(keys[i]))
                return vals[i];
        return null;
    }
}
```

---

**Knuth’s parking problem**

**Model.** Cars arrive at one-way street with \( M \) parking spaces. Each desires a random space \( i \): if space \( i \) is taken, try \( i + 1, i + 2 \), etc.

**Q.** What is mean displacement of a car?

**Half-full.** With \( M / 2 \) cars, mean displacement is \( \sim 3 / 2 \).

**Full.** With \( M \) cars, mean displacement is \( \sim \sqrt{\pi M / 8} \).
**Proposition.** Under uniform hashing assumption, the average number of probes in a linear probing hash table of size $M$ that contains $N = \alpha M$ keys is:

$$\frac{1}{2} \left( 1 + \frac{1}{1 - \alpha} \right) \approx \frac{1}{2} \left( 1 + \frac{1}{1 - \alpha^2} \right)$$

**Pf.**

Parameters.

- $M$ too large $\Rightarrow$ too many empty array entries.
- $M$ too small $\Rightarrow$ search time blows up.
- Typical choice: $\alpha = N / M \approx \frac{1}{2}$.

# probes for search hit is about $3/2$
# probes for search miss is about $5/2$

---

**War story: String hashing in Java**

- **hashCode()** in Java 1.1.
  - For long strings: only examine 8-9 evenly spaced characters.
  - Benefit: saves time in performing arithmetic.

  ```java
  public int hashCode()
  {
    int hash = 0;
    int skip = Math.max(1, length() / 8);
    for (int i = 0; i < length(); i += skip)
      hash = s[i] + (37 * hash);
    return hash;
  }
  ```

- Downside: great potential for bad collision patterns.

---

**War story: algorithmic complexity attacks**

Q. Is the uniform hashing assumption important in practice?

A. Obvious situations: aircraft control, nuclear reactor, pacemaker.

A. Surprising situations: denial-of-service attacks.

Real-world exploits. [Crosby-Wallach 2003]

- Bro server: send carefully chosen packets to DOS the server, using less bandwidth than a dial-up modem.
- Perl 5.8.0: insert carefully chosen strings into associative array.
- Linux 2.4.20 kernel: save files with carefully chosen names.
### Algorithmic complexity attack on Java

**Goal.** Find family of strings with the same hash code.

**Solution.** The base 31 hash code is part of Java's string API.

<table>
<thead>
<tr>
<th>key</th>
<th>hashCode()</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;aa&quot;</td>
<td>2112</td>
</tr>
<tr>
<td>&quot;bb&quot;</td>
<td>2112</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>key</th>
<th>hashCode()</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;AaAaAaAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaAaAaBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaAaBBAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaAaBBBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaBBAaAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaBBAaBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaBBBBAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;AaBBBBBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBAaAaAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBAaAaBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBAaBBAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBAaBBBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBBBAaAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBBBAaBB&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBBBBBAa&quot;</td>
<td>-540425984</td>
</tr>
<tr>
<td>&quot;BBBBBBBB&quot;</td>
<td>-540425984</td>
</tr>
</tbody>
</table>

$2^N$ strings of length $2N$ that hash to same value!

### Diversion: one-way hash functions

**One-way hash function.** "Hard" to find a key that will hash to a desired value (or two keys that hash to same value).

**Ex.** MD4, MD5, SHA-0, SHA-1, SHA-2, WHIRLPOOL, RIPEMD-160, ....

Known to be insecure

```java
String password = args[0];
MessageDigest sha1 = MessageDigest.getInstance("SHA1");
byte[] bytes = sha1.digest(password);
/* prints bytes as hex string */
```

**Applications.** Digital fingerprint, message digest, storing passwords.

**Caveat.** Too expensive for use in ST implementations.

### Separate chaining vs. linear probing

**Separate chaining.**
- Easier to implement delete.
- Performance degrades gracefully.
- Clustering less sensitive to poorly-designed hash function.

**Linear probing.**
- Less wasted space.
- Better cache performance.

**Q.** How to delete?
**Q.** How to resize?

### Hashing: variations on the theme

Many improved versions have been studied.

**Two-probe hashing.** (separate-chaining variant)
- Hash to two positions, insert key in shorter of the two chains.
- Reduces expected length of the longest chain to $\log \log N$.

**Double hashing.** (linear-probing variant)
- Use linear probing, but skip a variable amount, not just 1 each time.
- Effectively eliminates clustering.
- Can allow table to become nearly full.
- More difficult to implement delete.

**Cuckoo hashing.** (linear-probing variant)
- Hash key to two positions; insert key into either position; if occupied, reinsert displaced key into its alternative position (and recur).
- Constant worst case time for search.
Hash tables vs. balanced search trees

Hash tables.
- Simpler to code.
- No effective alternative for unordered keys.
- Faster for simple keys (a few arithmetic ops versus \(\log N\) compares).
- Better system support in Java for strings (e.g., cached hash code).

Balanced search trees.
- Stronger performance guarantee.
- Support for ordered ST operations.
- Easier to implement `compareTo()` correctly than `equals()` and `hashCode()`.

Java system includes both.
- Red-black BSTs: `java.util.TreeMap`, `java.util.TreeSet`.

---

**Search Applications**

- Sets
- Dictionary clients
- Indexing clients
- Sparse vectors

---

**Set API**

Mathematical set. A collection of distinct keys.

```java
public class SET<Key extends Comparable<Key>>

    SET(); // create an empty set
    void add(Key key); // add the key to the set
    boolean contains(Key key); // is the key in the set?
    void remove(Key key); // remove the key from the set
    int size(); // return the number of keys in the set
    Iterator<Key> iterator(); // iterator through keys in the set
```

**Q.** How to implement?
**A.** Remove “value” from any ST implementation
Exception filter

- Read in a list of words from one file.
- Print out all words from standard input that are { in, not in } the list.

```java
public class WhiteList {
    public static void main(String[] args) {
        Set<String> set = new HashSet<String>();
        In in = new In(args[0]);
        while (!in.isEmpty())
            set.add(in.readString());
        while (!StdIn.isEmpty()) {
            String word = StdIn.readString();
            if (set.contains(word))
                StdOut.println(word);
        }
    }
}
```

Exception filter applications

- Read in a list of words from one file.
- Print out all words from standard input that are { in, not in } the list.

<table>
<thead>
<tr>
<th>application</th>
<th>purpose</th>
<th>key</th>
<th>in-list</th>
</tr>
</thead>
<tbody>
<tr>
<td>spell checker</td>
<td>identify misspelled words</td>
<td>word</td>
<td>dictionary words</td>
</tr>
<tr>
<td>browser</td>
<td>mark visited pages</td>
<td>URL</td>
<td>visited pages</td>
</tr>
<tr>
<td>parental controls</td>
<td>block sites</td>
<td>URL</td>
<td>bad sites</td>
</tr>
<tr>
<td>chess</td>
<td>detect draw</td>
<td>board</td>
<td>positions</td>
</tr>
<tr>
<td>spam filter</td>
<td>eliminate spam</td>
<td>IP address</td>
<td>spam addresses</td>
</tr>
<tr>
<td>credit cards</td>
<td>check for stolen cards</td>
<td>number</td>
<td>stolen cards</td>
</tr>
</tbody>
</table>

Exception filter: Java implementation

- Read in a list of words from one file.
- Print out all words from standard input that are { in, not in } the list.

```java
public class BlackList {
    public static void main(String[] args) {
        Set<String> set = new HashSet<String>();
        In in = new In(args[0]);
        while (!in.isEmpty())
            set.add(in.readString());
        while (!StdIn.isEmpty()) {
            String word = StdIn.readString();
            if (!set.contains(word))
                StdOut.println(word);
        }
    }
}
```
Dictionary lookup

Command-line arguments.
- A comma-separated value (CSV) file.
- Key field.
- Value field.

Ex 1. DNS lookup.

```
% more ip.csv
www.princeton.edu,128.112.128.15
www.princeton.edu,128.112.128.16
www.math.princeton.edu,128.112.128.17
www.harvard.edu,149.206.50.127
www.harvard.edu,128.112.128.24
www.yale.edu,128.34.236.74
www.yale.edu,128.36.224.30
```

```
java LookupCSV ip.csv 0 1
adobe.com
999.999.999.99
```

```
java LookupCSV ip.csv 1 0
128.112.128.15
Not found
```

```
URL is key
IP is value
```

Ex 2. Amino acids.

```
% more amino.csv
TTT,Phe,F,Phenylalanine
TTC,Phe,F,Phenylalanine
TTA,Leu,L,Leucine
TTG,Leu,L,Leucine
TCT,Ser,S,Serine
TCC,Ser,S,Serine
TCA,Ser,S,Serine
TCG,Ser,S,Serine
TAT,Tyr,Y,Tyrosine
TAC,Tyr,Y,Tyrosine
TAA,Stop,Stop,Stop
TAG,Stop,Stop,Stop
TGT,Cys,C,Cysteine
TGC,Cys,C,Cysteine
TGA,Stop,Stop,Stop
TGG,Trp,W,Tryptophan
CTT,Leu,L,Leucine
CTC,Leu,L,Leucine
CTA,Leu,L,Leucine
CTG,Leu,L,Leucine
CCT,Pro,P,Proline
CCC,Pro,P,Proline
CCA,Pro,P,Proline
CCG,Pro,P,Proline
CAT,His,H,Histidine
CAC,His,H,Histidine
CAA,Gln,Q,Glutamine
CAG,Gln,Q,Glutamine
CGT,Arg,R,Arginine
CGC,Arg,R,Arginine
CGA,Arg,R,Arginine
```

```
% java LookupCSV amino.csv 0 3
ACT
Threonine
TAG
Stop
CAT
Histidine
```

```
login is key
first name is value
```

Ex 3. Class list.

```
% more classlist.csv
13,Berl,Ethan Michael,P01,eberl
11,Bourque,Alexander Joseph,P01,abourque
11,Chen,
Christel,P01,cchenb
10,Douglas,Malia Morioka,P01,malia
12,Goodrick,Sara Lynn,P01,sgoodrick
11,Hyde,Gregory Thomas,P01,ghythe
13,Hyde,Gregory Thomas,P01,ghythe
13,Kim,Hyunmoon,P01,hkim
11,Kleinfeld,Ivan Maximillian,P01,ikleinfe
12,Korac,Damjan,P01,dkorac
11,MacDonald,Graham David,P01,gmacdona
10,Michel,Brian Thomas,P01,bmichel
13,Partridge,Brenton Alan,P01,bpartridge
13,Rilee,Alexander,P01,arilee
13,Roopakalu,Ajay,P01,aroopaka
11,Sheng,Ben C,P01,bsheng
12,Webb,Natalie Sue,P01,nwebb
```

```
% java LookupCSV classlist.csv 4 1
eberl
Ethan
nwebb
Natalie
```

```
% java LookupCSV classlist.csv 4 3
dpan
P01
```

```
login is key
first name is value
precept is value
```
public class LookupCSV
{
    public static void main(String[] args)
    {
        In in = new In(args[0]);
        int keyField = Integer.parseInt(args[1]);
        int valField = Integer.parseInt(args[2]);

        ST<String, String> st = new ST<String, String>();
        while (!in.isEmpty())
        {
            String line = in.readLine();
            String[] tokens = database[i].split(",");
            String key = tokens[keyField];
            String val = tokens[valField];
            st.put(key, val);
        }

        while (!StdIn.isEmpty())
        {
            String s = StdIn.readString();
            if (!st.contains(s)) StdOut.println("Not found");
            else StdOut.println(st.get(s));
        }
    }
}
File indexing

**Goal.** Given a list of files specified, create an index so that you can efficiently find all files containing a given query string.

```bash
% ls *.txt
assop.txt magna.txt moby.txt sawyer.txt tale.txt
% ls *.java
% java FileIndex *.java
BlackList.java Concordance.java DeDup.java FileIndex.java ST.java SET.java WhiteList.java
import FileIndex.java SET.java ST.java
Comparator null
```

**Solution.** Key = query string; value = set of files containing that string.

```java
class FileIndex
{
    public static void main(String[] args)
    {
        ST<String, SET<File>> st = new ST<String, SET<File>>();
        for (String filename : args)
        {
            File file = new File(filename);
            In in = new In(file);
            while (!(in.isEmpty()))
            {
                String word = in.readString();
                if (!st.contains(word))
                    st.put(word, new SET<File>());
                SET<File> set = st.get(word);
                set.add(file);
            }
        }
        while (!StdIn.isEmpty())
        {
            String query = StdIn.readString();
            StdOut.println(st.get(query));
        }
    }
}
```

---

Book index

**Goal.** Index for an e-book.

```java
class Concordance
{
    public static void main(String[] args)
    {
        for (String filename : args)
        {
            File file = new File(filename);
            In in = new In(file);
            if (!in.isEmpty())
            {
                String word = in.readString();
                if (!st.contains(word))
                    st.put(word, new SET<File>());
                SET<File> set = st.get(word);
                set.add(file);
            }
        }
        while (!StdIn.isEmpty())
        {
            String query = StdIn.readString();
            StdOut.println(st.get(query));
        }
    }
}
```

---

Concordance

**Goal.** Preprocess a text corpus to support concordance queries: given a word, find all occurrences with their immediate contexts.
public class Concordance {
    public static void main(String[] args) {
        In in = new In(args[0]);
        String[] words = StdIn.readAll().split(“\s+”);
        ST<String, SET<Integer>> st = new ST<String, SET<Integer>>();
        for (int i = 0; i < words.length; i++)
            if (!st.contains(words[i]))
                st.put(words[i], new SET<Integer>());
                SET<Integer> pages = st.get(words[i]);
                pages.add(i);
        while (!StdIn.isEmpty())
            String query = StdIn.readString();
            SET<Integer> set = st.get(query);
            for (int k : set)
                // print words[k-5] to words[k+5]
    }
}

Search Applications

- Sets
- Dictionary clients
- Indexing clients
- Sparse clients

Vectors and matrices

**Vector.** Ordered sequence of N real numbers.

**Matrix.** N-by-N table of real numbers.

**vector operations**

\[
\begin{align*}
a &= [0 \ 3 \ 15] \quad b = [-1 \ 2 \ 2] \\
a + b &= [-1 \ 5 \ 17] \\
a \cdot b &= (0 \cdot -1) + (3 \cdot 2) + (15 \cdot 2) = 36 \\
\|a\| &= \sqrt{a \cdot a} = \sqrt{0^2 + 3^2 + 15^2} = 3\sqrt{26}
\end{align*}
\]

**matrix-vector multiplication**

\[
\begin{bmatrix}
0 & 1 & 1 \\
2 & 4 & -2 \\
0 & 3 & 15
\end{bmatrix}
\begin{bmatrix}
-1 \\
2 \\
2
\end{bmatrix}
= \begin{bmatrix}
4 \\
2 \\
36
\end{bmatrix}
\]

Sparse vectors and matrices

**Sparse vector.** An N-dimensional vector is sparse if it contains \(O(1)\) nonzeros.

**Sparse matrix.** An N-by-N matrix is sparse if it contains \(O(N)\) nonzeros.

**Property.** Large matrices that arise in practice are sparse.

\[
\begin{bmatrix}
0 & 0 & .36 & .36 & .18 \\
0 & 0 & .36 & .36 & .18 \\
0 & 0 & .90 & 0 & 0 \\
.90 & 0 & 0 & 0 & 0 \\
.47 & .47 & 0 & 0 & 0
\end{bmatrix}
\]
Matrix-vector multiplication (standard implementation)

```java
// double[,] a = new double[N][N];
// double[] x = new double[N];
// double[] b = new double[N];

// initialize a[,] and x[]

for (int i = 0; i < N; i++)
{
    sum = 0.0;
    for (int j = 0; j < N; j++)
        sum += a[i][j] * x[j];
    b[i] = sum;
}
```

Sparse matrix-vector multiplication

**Problem.** Sparse matrix-vector multiplication.

**Assumptions.** Matrix dimension is 10,000; average nonzeros per row ~ 10.

Sparse vector data type

```java
public class SparseVector
{
    private HashST<Integer, Double> v;
    public SparseVector()
    {  v = new HashST<Integer, Double>();  }
    public void put(int i, double x)
    {  v.put(i, x);  }
    public double get(int i)
    {
        if (!v.contains(i)) return 0.0;
        else return v.get(i);
    }
    public Iterable<Integer> indices()
    {  return v.keys();  }
    public double dot(double[] that)
    {
        double sum = 0.0;
        for (int i : indices())
            sum += that[i] * this.get(i);
        return sum;
    }
}
```

Vector representations

**ID array (standard) representation.**
- Constant time access to elements.
- Space proportional to N.

```
  0  1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19
  0  36  0  0  34  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

**Symbol table representation.**
- Key = index, value = entry.
- Efficient iterator.
- Space proportional to number of nonzeros.
Matrix representations

2D array (standard) matrix representation: Each row of matrix is an array.
- Constant time access to elements.
- Space proportional to $N^2$.

Sparse matrix representation: Each row of matrix is a sparse vector.
- Efficient access to elements.
- Space proportional to number of nonzeros (plus $N$).

Sample searching challenge

**Problem.** Rank pages on the web.

**Assumptions.**
- Matrix-vector multiply
- 10 billion+ rows
- Sparse

Which “searching” method to use to access array values?
1. Standard 2D array representation
2. Symbol table
3. Doesn’t matter much.

Sparse matrix-vector multiplication

---

```
SparseVector[] a = new SparseVector[N];
double[] x = new double[N];
double[] b = new double[N];
...
// Initialize a[] and x[]
...
for (int i = 0; i < N; i++)
  b[i] = a[i].dot(x);
```
### Sparse vector data type

```java
public class SparseVector {
    private int N; // length
    private ST<Integer, Double> st; // the elements

    public SparseVector(int N) {
        this.N = N;
        this.st = new ST<Integer, Double>();
    }

    public void put(int i, double value) {
        if (value == 0.0) st.remove(i);
        else st.put(i, value);
    }

    public double get(int i) {
        if (st.contains(i)) return st.get(i);
        else return 0.0;
    }

    public double dot(SparseVector that) {
        double sum = 0.0;
        for (int i : this.st)
            if (that.st.contains(i))
                sum += this.get(i) * that.get(i);
        return sum;
    }

    public double norm() {
        return Math.sqrt(this.dot(this));
    }

    public SparseVector plus(SparseVector that) {
        SparseVector c = new SparseVector(N);
        for (int i : this.st)
            c.put(i, this.get(i));
        for (int i : that.st)
            c.put(i, that.get(i) + c.get(i));
        return c;
    }
}
```

### Sparse matrix data type

```java
public class SparseMatrix {
    private final int N; // length
    private SparseVector[] rows; // the elements

    public SparseMatrix(int N) {
        this.N = N;
        this.rows = new SparseVector[N];
        for (int i = 0; i < N; i++)
            this.rows[i] = new SparseVector(N);
    }

    public void put(int i, int j, double value) {
        rows[i].put(j, value);
    }

    public double get(int i, int j) {
        return rows[i].get(j);
    }

    public SparseVector times(SparseVector x) {
        SparseVector b = new SparseVector(N);
        for (int i = 0; i < N; i++)
            b.put(i, rows[i].dot(x));
        return b;
    }
}
```

### Compressed row storage (CRS)

**Compressed row storage.**

- Store nonzeros in a 1D array `val[]`.
- Store column index of each nonzero in parallel 1D array `col[]`.
- Store first index of each row in array `row[]`.

```
A = [
    11 0 0 41
    0 22 0 0
    0 33 41
    14 0 34 44
    0 25 0 0
    16 26 36 46
]
```
Compressed row storage (CRS)

**Benefits.**
- Cache-friendly.
- Space proportional to number of nonzeros.
- Very efficient matrix-vector multiply.

```java
double[] y = new double[N];
for (int i = 0; i < n; i++)
    for (int j = row[i]; j < row[i+1]; j++)
        y[i] += val[j] * x[col[j]];
```

**Downside.** No easy way to add/remove nonzeros.

**Applications.** Sparse Matlab.