Chapter 11
Hash Tables

• Introduction to Hashing
• Open Addressing
• Separate Chaining
• Hash Functions
• Hashing Efficiency
• Hashing and External Storage
• Summary

Overview

• A hash table:
  – a data structure that offers very fast insertion and searching.
• Hash tables sound almost too good to be true.
• No matter how many data items there are, insertion and searching (and sometimes deletion) can take close to constant time: \( O(1) \) in Big O notation.
  – In practice this is just a few machine instructions.
• For a human user of a hash table this is essentially instantaneous.
• It’s so fast that computer programs typically use hash tables when they need to look up tens of thousands of items in less than a second (as in spelling checkers).
• Significantly faster than trees, which operate in relatively fast \( O(\log N) \) time.
• Not only are they fast, hash tables are relatively easy to program.

Overview

• Hash tables do have several disadvantages.
  – They’re based on arrays, and arrays are difficult to expand once they’ve been created.
  – For some kinds of hash tables, performance may degrade catastrophically when the table becomes too full, so the programmer needs to have a fairly accurate idea of how many data items will need to be stored (or be prepared to periodically transfer data to a larger hash table, a time-consuming process).
  – Also, there’s no convenient way to visit the items in a hash table in any kind of order (such as from smallest to largest).
  – If you need this capability, you'll need to look elsewhere.
• However, if you don't need to visit items in order, and you can predict in advance the size of your database,
  – hash tables are unparalleled in speed and convenience.

Introduction to Hashing

• Hash tables and hashing.
• One important concept: How a range of key values is transformed into a range of array index values.
• In a hash table
  – this is accomplished with a hash function.
  – However, for certain kinds of keys, no hash function is necessary; the key values can be used directly as array indices.
  – We’ll look at this simpler situation first and then go on to show how hash functions can be used when keys aren’t distributed in such an orderly fashion.

Introduction to Hashing
Employee Numbers as Keys

• A program to access employee records for a small company with, say, 1,000 employees.
  – Each employee record requires 1,000 bytes of storage.
  – Thus you can store the entire database in only 1 megabyte, which will easily fit in your computer’s memory.
• The company’s personnel director has specified that she wants the fastest possible access to any individual record.
  – Thus, for the most recently hired worker.
  – In fact, access by other keys is deemed unnecessary.
• What sort of data structure should you use in this situation?

Introduction to Hashing
Employee Numbers as Keys

• A program to access employee records for a small company with, say, 1,000 employees.
  – Each employee record requires 1,000 bytes of storage.
  – Thus you can store the entire database in only 1 megabyte, which will easily fit in your computer’s memory.
• The company’s personnel director has specified that she wants the fastest possible access to any individual record.
  – Thus, for the most recently hired worker.
• Also, every employee has been given a number from 1 (for the founder) to 1,000 (for the most recently hired worker).
  – These employee numbers can be used as keys to access the records; in fact, access by other keys is deemed unnecessary.
• What sort of data structure should you use in this situation?

Keys Are Index Numbers

• One possibility is a simple array.
  – Each employee record occupies one cell of the array, and the index number of the cell is the employee number for that record.
  – This is shown in Figure 11.1.
• As you know, accessing a specified array element is very fast if you know its index number.
• The clerk looking up Herman Alcazar knows that he is employee number 72, so he enters that number, and the program goes instantly to index number 72 in the array.
• A single program statement is all that’s necessary:
  empRecord rec = databaseArray[72];
Introduction to Hashing

Employee Numbers as Keys

Keys Are Index Numbers

• It's also very quick to add a new item:
  – You insert it just past the last occupied element.
  – The next new record — for Jim Chan, the newly hired employee number 1,001 — would go in cell 1,001. Again, a single statement inserts the new record:
    `databaseArray[totalEmployees++] = newRecord;`
• Presumably the array is made somewhat larger than the current number of employees, to allow room for expansion; but not much expansion is anticipated.

Not Always So Orderly

• The speed and simplicity of data access using this array-based database make it very attractive.
  – However, it works in our example only because the keys are unusually well organized.
    • They run sequentially from 1 to a known maximum, and this maximum is a reasonable size for an array.
    • There are no deletions, so memory-wasting gaps don't develop in the sequence.
    • New items can be added sequentially at the end of the array, and the array doesn't need to be very much larger than the current number of items.

Converting Words to Numbers

• What we need is a system for turning a word into an appropriate index number. To begin, we know that computers use various schemes for representing individual characters as numbers. One such scheme is the ASCII code, in which a is 97, b is 98, and so on, up to 122 for z.
  • However, the ASCII code runs from 0 to 255, to accommodate capitals, punctuation, and so on. There are really only 26 letters in English words, so let's devise our own code—a simpler one that can potentially save memory space. Let's say a is 1, b is 2, c is 3, and so on up to 26 for z. We'll also say a blank is 0, so we have 27 characters. (Uppercase letters aren't used in this dictionary.)
  • How do we combine the digits from individual letters into a number that represents an entire word? There are all sorts of approaches. We'll look at two representative ones, and their advantages and disadvantages.

A Dictionary

• In many situations the keys are not so well behaved as in the employee database just described.
  • The classic example is a dictionary.
    – If you want to put every word of an English-language dictionary, from a to zyzyva, into your computer's memory, so they can be accessed quickly, a hash table is a good choice.
  • A similar widely used application for hash tables is in computer-language compilers, which maintain a symbol table in a hash table.
    – The symbol table holds all the variable and function names made up by the programmer, along with the addresses where they can be found in memory.
    – The program needs to access these names very quickly, so a hash table is the preferred data structure.
  • Let's say we want to store a 50,000-word English-language dictionary in main memory.
    – You would like every word to occupy its own cell in a 50,000-cell array, so you can access the word using an index number.
    – This will make access very fast. But what's the relationship of these index numbers to the words?
    – Given the word morphosis, for example, how do we find its index number?

Figure 11.1: Employee numbers as array indices

Introduction to Hashing

A Dictionary

Converting Words to Numbers

• What we need is a system for turning a word into an appropriate index number. To begin, we know that computers use various schemes for representing individual characters as numbers. One such scheme is the ASCII code, in which a is 97, b is 98, and so on, up to 122 for z.
  • However, the ASCII code runs from 0 to 255, to accommodate capitals, punctuation, and so on. There are really only 26 letters in English words, so let's devise our own code—a simpler one that can potentially save memory space. Let's say a is 1, b is 2, c is 3, and so on up to 26 for z.
  • However, the ASCII code runs from 0 to 255, to accommodate capitals, punctuation, and so on. There are really only 26 letters in English words, so let's devise our own code—a simpler one that can potentially save memory space. Let's say a is 1, b is 2, c is 3, and so on up to 26 for z. We'll also say a blank is 0, so we have 27 characters. (Uppercase letters aren't used in this dictionary.)
  • How do we combine the digits from individual letters into a number that represents an entire word? There are all sorts of approaches. We'll look at two representative ones, and their advantages and disadvantages.

Introduction to Hashing

A Dictionary

Add the Digits

• A simple approach to converting a word to a number might be to simply add the code numbers for each character. Say we want to convert the word cats to a number. First we convert the characters to digits using our homemade code:
  `c = 3
  a = 1
  t = 20
  s = 19`
• Then we add them:
  `3 + 1 + 20 + 19 = 43`
• Thus in our dictionary the word cats would be stored in the array cell with index 43. All the other English words would likewise be assigned an array index calculated by this process.
• How well would this work?
Add the Digits

- How well would this work? For the sake of argument, let’s restrict ourselves to 10-letter words. Then (remembering that a blank is 0), the first word in the dictionary, a, would be coded by

\[ 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 1 = 1 \]

- The last potential word in the dictionary would be zzzzzzzzzz (ten Zs). Our code obtained by adding its letters would be

\[ 26 + 26 + 26 + 26 + 26 + 26 + 26 + 26 + 26 + 26 = 260 \]

- Thus the total range of word codes is from 1 to 260. Unfortunately, there are 50,000 words in the dictionary, so there aren’t enough index numbers to go around. Each array element will need to hold about 192 words (50,000 divided by 260).

Multiply by Powers

- Let’s try a different way to map words to numbers. If our array was too small before, let’s make sure it’s big enough. What would happen if we created an array in which every word, in fact every potential word, from a to zzzzzzzzzz, was guaranteed to occupy its own unique array element?

- To do this, we need to be sure that every character in a word contributes in a unique way to the final number.

Multiply by Powers

- In this system we break a number into its digits, multiply them by appropriate powers of 10 (because there are 10 possible digits), and add the products.

- In a similar way we can decompose a word into its letters, convert the letters to their numerical equivalents, multiply them by appropriate powers of 27 (because there are 27 possible characters, including the blank), and add the results. This gives a unique number for every word.

Multiply by Powers

- Say we want to convert the word cats to a number. We convert the digits to numbers as shown earlier. Then we multiply each number by the appropriate power of 27, and add the results:

\[ 3*19,683 + 1*729 + 20*27 + 19*1 \]

- Calculating the powers gives

\[ 59,049 + 729 + 540 + 19 \]

- and multiplying the letter codes times the powers yields

\[ 59,049 + 729 + 540 + 19 \]

which sums to 60,337.
Introduction to Hashing

Multiply by Powers

• Indeed generate a unique number for every potential word.
• We just calculated a four-letter word. What happens with larger words?
• Unfortunately the range of numbers becomes rather large.
• The largest 10-letter word, zzzzzzzzzz, translates into
  \[26 \times 27^9 + 26 \times 27^8 + 26 \times 27^7 + 26 \times 27^6 + 26 \times 27^5 + 26 \times 27^4 +
  26 \times 27^3 + 26 \times 27^2 + 26 \times 27^1 + 26 \times 27^0\]
• Just by itself, \(27^9\) is more than 7,000,000,000,000, so you can see that the sum will be huge.
• An array stored in memory can't possibly have this many elements.

Introduction to Hashing

Hashing

• Need a way to compress the huge range of numbers we obtain from the numbers-multiplied-by-powers system into a range that matches a reasonably sized array.
• How big an array for our English dictionary?
  – If we only have 50,000 words, you might assume our array should have approximately this many elements.
  – However, it turns out we're going to need an array with about twice this many cells. (It will become clear later why this is so.)
  – So we need an array with 100,000 elements.

Introduction to Hashing

Hashing

• A similar expression can be used to compress the really huge numbers that uniquely represent every English word into index numbers that fit in our dictionary array:
  \[
  \text{arrayIndex} = \text{hugeNumber} \mod \text{arraySize};
  \]
• This is an example of a hash function.
  – It hashes (converts) a number in a large range into a number in a smaller range.
  – This smaller range corresponds to the index numbers in an array.
• An array into which data is inserted using a hash function is called a hash table.
  (We'll talk more about the design of hash functions later in the chapter.)
Introduction to Hashing

Hashing

• In the huge range, each number represents a potential data item (an arrangement of letters), but few of these numbers represent actual data items (English words).
• A hash function transforms these large numbers into the index numbers of a much smaller array.
• In this array we expect that, on the average, there will be one word for every two cells.
  – Some cells will have no words, and some more than one.
• A practical implementation of this scheme runs into trouble because hugeNumber will probably overflow its variable size, even for type long.
• We'll see how to deal with this later.

Introduction to Hashing

Collisions

• We pay a price for squeezing a large range into a small one.
  – There’s no longer a guarantee that two words won’t hash to the same array index.
• Similar to what happened when we added the letter codes, but the situation is nowhere near as bad.
  – When we added the letters, there were only 260 possible results (for words up to 10 letters). Now we're spreading this out into 50,000 possible results.
  – Even so, it’s impossible to avoid hashing several different words into the same array location, at least occasionally.
  – We’d hoped that we could have one data item per index number, but this turns out not to be possible.
  – The best we can do is hope that not too many words will hash to the same index.

Introduction to Hashing

Collisions

• To insert the word *melioration* into the array.
  – You hash the word to obtain its index number, but find that the cell at that number is already occupied by the word *demystify*, which happens to hash to the exact same number (for a certain size array).
  – This situation, shown in Figure 11.4, is called a collision.
• It may appear that the possibility of collisions renders the hashing scheme impractical, but in fact we can work around the problem in a variety of ways.

Open Addressing

• When a data item can’t be placed at the index calculated by the hash function, another location in the array is sought.
  – We’ll explore three methods of open addressing, which vary in the method used to find the next vacant cell.
    • linear probing,
    • quadratic probing,
    • double hashing.

Open Addressing

Linear Probing

• In linear probing we search sequentially for vacant cells.
• If 5,421 is occupied when we try to insert cats there, we go to 5,422, then 5,423, and so on, incrementing the index until we find an empty cell.
• This is called linear probing because it steps sequentially along the line of cells.
Open Addressing
Linear Probing

The Hash Workshop Applet
• When you start this applet, you’ll see a screen similar to Figure 11.5.

Figure 11.5: The Hash Workshop applet

Open Addressing
Linear Probing

The Hash Workshop Applet
• In this applet the range of keys runs from 0 to 999.
• The initial size of the array is 60.
• The hash function has to squeeze the range of keys down to match the array size.
• It does this with the modulo (%) operator, as we’ve seen before:
  \[ \text{arrayIndex} = \text{key} \mod \text{arraySize} \]
• For the initial array size of 60, this is
  \[ \text{arrayIndex} = \text{key} \mod 60 \]
• This hash function is simple enough that you can solve it mentally.

Figure 11.5: The Hash Workshop applet

Open Addressing
Linear Probing

The Hash Workshop Applet
• Operations are carried out by repeatedly pressing the same button.
  – For example, to find a data item with a specified number,
    • click the Find button repeatedly.
  – Remember, finish a sequence with one button before using another button.
  – For example, don’t switch from clicking Fill to some other button until the Press any key message is displayed.
• All the operations require you to type a numerical value at the beginning of the sequence.
• The Find button requires you to type a key value, for example, while New requires the size of the new table.

Figure 11.5: The Hash Workshop applet

Open Addressing
Linear Probing

The Hash Workshop Applet
• You can create a new hash table of a size you specify by using the New button.
  – The maximum size is 60;
    • this limitation results from the number of cells Open Addressing that can be viewed in the applet window.
  – The initial size is also 60.
    • We use this number because it makes it easy to check if the hash values are correct.
    • But as we’ll see later, in a general-purpose hash table, the array size should be a prime number, so 59 would be a better choice.

Figure 11.5: The Hash Workshop applet

Open Addressing
Linear Probing

The Hash Workshop Applet
• Initially the hash table contains 30 items, so it’s half full.
• However, you can also fill it with a specified number of data items using the Fill button.
• Keep clicking Fill, and when prompted, type the number of items to fill. Hash tables work best when they are not more than half or at most two-thirds full (40 items in a 60-cell table).
• We’ll see that the filled cells aren’t evenly distributed in the cells.
  – Sometimes there’s a sequence of several empty cells, and sometimes a sequence of filled cells.
• Let’s call a sequence of filled cells in a hash table a filled sequence.
• As you add more and more items, the filled sequences become longer.
  – This is called clustering, shown in Figure 11.6.
The Hash Workshop Applet

The Fill Button
• When you use the applet, note that
  – It may take a long time to fill a hash table if you try to fill it too full (for example, if you try to put 59 items in a 60-cell table).
  – You may think the program has stopped, but be patient. It’s extremely inefficient at filling an almost-full array.
• Also, note that if the hash table becomes completely full, the algorithms all stop working;
  – In this applet they assume that the table has at least one empty cell.

The Fill Routine in the Hash Applet

The Find Button
• This results in an array index. The cell at this index may be the key you’re looking for.
  – Why not?
  – Remember that during insertion the probe process steps along a series of cells, looking for a vacant cell.
  – If a cell is made empty in the middle of this sequence of full cells, the Find routine will give up when it sees the empty cell, even if the desired cell can eventually be reached.
  – For this reason a deleted item is replaced by an item with a special key value that identifies it as deleted.
• In this applet we assume all legitimate key values are positive, so the deleted value is chosen as –1.
• Deleted items are marked with the special key “Del”.
• The Insert button inserts a new item at the first available empty cell or in a “Del” item.
• The Find button will treat a “Del” item as an existing item for the purposes of searching for another item further along.
• If there are many deletions, the hash table fills up with these ersatz “Del” data items, which makes it less efficient.
  – For this reason many hash table implementations don’t allow deletion.
  – If it is implemented, it should be used sparingly.
• Following a collision, the Find algorithm simply stops along the array looking at each cell in sequence.
  – If it encounters an empty cell before finding the key it’s looking for, it knows the search has failed.
  – There’s no use looking further, because the insertion algorithm would have inserted the item at this cell (if not earlier).
• Figure 11.7 shows successful and unsuccessful linear probes.

The Ins Button
• The Ins button inserts a data item, with a key value that you type into the number box, into the hash table.
• It uses the same algorithm as the Find button to locate the appropriate cell.
• If the original cell is occupied, it will probe linearly for a vacant cell.
  – When it finds one, it inserts the item.
• Try inserting some new data items.
  – Type in a 3-digit number and watch what happens.
  – Most items will go into the first cell they try, but some will suffer collisions, and need to step along to find an empty cell.
  – The number of steps they take is the probe length. Most probe lengths are only a few cells long. Sometimes, however, you may see probe lengths of 4 or 5 cells, or even longer as the array becomes excessively full.
• Notice which keys hash to the same index.
  – If the array size is 60, the keys 7, 67, 127, 187, 247, and so on up to 967 all hash to index 7. Try inserting this sequence or a similar one. This will demonstrate the linear probe.

The Del Button
• Deleted items are marked with the special key “Del”.
• The Insert button will insert a new item at the first available empty cell or in a “Del” item.
• The Find button will treat a “Del” item as an existing item for the purposes of searching for another item further along.
• If there are many deletions, the hash table fills up with these ersatz “Del” data items, which makes it less efficient.
  – For this reason many hash table implementations don’t allow deletion.
  – If it is implemented, it should be used sparingly.

Duplicates Allowed?
• Can you allow data items with duplicate keys to be used in hash tables?
• The fill routine in the Hash applet doesn’t allow duplicates, but you can insert them with the Insert button if you like.
  – Then you’ll see that only the first one can be accessed.
  – The only way to access a second item with the same key is to delete the first one. This isn’t too convenient.
• You could rewrite the Find algorithm to look for all items with the same key instead of just the first one.
  – However, it would then need to search through all the cells of every linear sequence it encountered.
  – This wastes time for all table accesses, even when no duplicates are involved.
• In the majority of cases you probably want to forbid duplicates.
Open Addressing
Java Code for a Linear Probe Hash Table

The find() Method
• The find() method first calls hashFunc() to hash the search key to obtain the index number hashVal.
  – The hashFunc() method applies the % operator to the search key and the array size, as we’ve seen before.
• Next, in a while condition,
  – find() checks if the item at this index is empty (null).
  – If not, it checks if the item contains the search key.
  – If it does, find() increments hashVal and goes back to the top of the while loop to check if the next cell is occupied.

Open Addressing
Java Code for a Linear Probe Hash Table

The insert() Method
• The insert() method uses about the same algorithm as find() to locate where a data item should go.
• However, it’s looking for an empty cell or a deleted item (key –1), rather than a specific item.
• Once this empty cell has been located, insert() places the new item into it.

Open Addressing
Java Code for a Linear Probe Hash Table

The delete() Method
• The delete() method finds an existing item using code similar to find().
• Once the item is found, delete() writes over it with the special data item nonItem, which is predefined with a key of –1.

Open Addressing
Java Code for a Linear Probe Hash Table

The hash.java Program
• A DataItem object contains just one field, an integer that is its key.
  – These objects could contain more data, or a reference to an object of another class (such as employee or partNumber).
• The major field in class HashTable is an array called hashArray.
• Other fields are the size of the array and the special nonItem object used for deletions.

Open Addressing
Java Code for a Linear Probe Hash Table

The find() Method
• As hashVal steps through the array, it eventually reaches the end.
  – When this happens we want it to wrap around to the beginning.
  – We could check for this with an if statement, setting hashVal to 0 whenever it equaled the array size.
  – However, we can accomplish the same thing by applying the % operator to hashVal and the array size.
• Cautious programmers might not want to assume the table is not full, as is done here.
  – The table should not be allowed to become full, but if it did, this method would loop forever.
  – For simplicity we don’t check for this situation.
• A table with 10,000 cells and 6,667 items has a load factor of 2/3.
• It’s as if a quadratic probe became increasingly desperate as its search lengthened.
  – At first it calmly picks the adjacent cell.
  – If that’s occupied, it thinks it may be in a small cluster so it tries something 4 cells away.
  – If that’s occupied it becomes a little concerned, thinking it may be in a larger cluster, and tries 9 cells away.
  – If that’s occupied it feels the first tinges of panic and jumps 16 cells away.
  – Pretty soon it’s flying hysterically all over the place, as you can see if you try searching with the hashDouble Workshop applet when the table is almost full.
• Clusters can form even when the load factor isn’t high.
• When you’re asked to select double or quadratic probe, click the Quad button.
• You’ll need to go through the old array in sequence, inserting each item into the new array with the insert() method.
• Java offers a class Vector that is an array-like data structure that can be expanded.
• However, it’s not much help because of the need to rehash all data items when the table changes size.
• Expanding the array is only practical when there’s plenty of time available to carry it out.

The listing for hash.java

• To see what’s going on, it’s best to create tables with fewer than about 20 items, so all the items can be displayed on one line.

Here’s some sample interaction with hash.java:

Enter size of hash table: 12
Enter first letter of show, insert, delete, or find: s
Enter initial number of items: 8
Enter size of hash table: 12
Figure 11.8: Quadratic probing

Open Addressing
Java Code for a Linear Probe Hash Table

Expanding the Array

• One option when a hash table becomes too full is to expand its array.
  – In Java, arrays have a fixed size and can’t be expanded.
  – Your program could create a new, larger array, and then rehash the contents of the old small array into the new large one.
  – However, this is a time-consuming process.
  – Remember that the hash function calculates the location of a given data item based on the array size, so the locations in the large array won’t be the same as those in a small array.
  – You can’t, therefore, simply copy the items from one array to the other.
  – You’ll need to go through the old array in sequence, inserting each item into the new array with the insert() method.
  – Java offers a class Vector that is an array-like data structure that can be expanded.
  – However, it’s not much help because of the need to rehash all data items when the table changes size.
  – Expanding the array is only practical when there’s plenty of time available to carry it out.

The Step Is the Square of the Step Number

• In a linear probe,
  – if the primary hash index is x,
    – subsequent probes go to x+1, x+2, x+3, and so on.
• In quadratic probing,
  – probes go to x+1, x+4, x+9, x+16, x+25, and so on.
  – The distance from the initial probe is the square of the step number: x+1², x+2², x+3², x+4², x+5², and so on.
• Figure 11.8 shows some quadratic probes.

The Step Is the Square of the Step Number

• It’s as if a quadratic probe became increasingly desperate as its search lengthened.
  – At first it calmly picks the adjacent cell.
  – If that’s occupied, it thinks it may be in a small cluster so it tries something 4 cells away.
  – If that’s occupied it becomes a little concerned, thinking it may be in a larger cluster, and tries 9 cells away.
  – If that’s occupied it feels the first tinges of panic and jumps 16 cells away.
  – Pretty soon it’s flying hysterically all over the place, as you can see if you try searching with the hashDouble Workshop applet when the table is almost full.
• Clusters can form even when the load factor isn’t high.
• Clusters can form even when the load factor isn’t high.
• Clusters can form even when the load factor isn’t high.

Expanding the Array

• One option when a hash table becomes too full is to expand its array.
  – In Java, arrays have a fixed size and can’t be expanded.
  – Your program could create a new, larger array, and then rehash the contents of the old small array into the new large one.
  – However, this is a time-consuming process.
  – Remember that the hash function calculates the location of a given data item based on the array size, so the locations in the large array won’t be the same as those in a small array.
  – You can’t, therefore, simply copy the items from one array to the other.
  – You’ll need to go through the old array in sequence, inserting each item into the new array with the insert() method.
  – Java offers a class Vector that is an array-like data structure that can be expanded.
  – However, it’s not much help because of the need to rehash all data items when the table changes size.
  – Expanding the array is only practical when there’s plenty of time available to carry it out.

The Step Is the Square of the Step Number

• In a linear probe,
  – if the primary hash index is x,
    – subsequent probes go to x+1, x+2, x+3, and so on.
• In quadratic probing,
  – probes go to x+1, x+4, x+9, x+16, x+25, and so on.
  – The distance from the initial probe is the square of the step number: x+1², x+2², x+3², x+4², x+5², and so on.
• Figure 11.8 shows some quadratic probes.

The Step Is the Square of the Step Number

• It’s as if a quadratic probe became increasingly desperate as its search lengthened.
  – At first it calmly picks the adjacent cell.
  – If that’s occupied, it thinks it may be in a small cluster so it tries something 4 cells away.
  – If that’s occupied it becomes a little concerned, thinking it may be in a larger cluster, and tries 9 cells away.
  – If that’s occupied it feels the first tinges of panic and jumps 16 cells away.
  – Pretty soon it’s flying hysterically all over the place, as you can see if you try searching with the hashDouble Workshop applet when the table is almost full.

Open Addressing
Java Code for a Linear Probe Hash Table

The Step Is the Square of the Step Number

• In a linear probe,
  – if the primary hash index is x,
    – subsequent probes go to x+1, x+2, x+3, and so on.
• In quadratic probing,
  – probes go to x+1, x+4, x+9, x+16, x+25, and so on.
  – The distance from the initial probe is the square of the step number: x+1², x+2², x+3², x+4², x+5², and so on.
• Figure 11.8 shows some quadratic probes.

The Step Is the Square of the Step Number

• It’s as if a quadratic probe became increasingly desperate as its search lengthened.
  – At first it calmly picks the adjacent cell.
  – If that’s occupied, it thinks it may be in a small cluster so it tries something 4 cells away.
  – If that’s occupied it becomes a little concerned, thinking it may be in a larger cluster, and tries 9 cells away.
  – If that’s occupied it feels the first tinges of panic and jumps 16 cells away.
  – Pretty soon it’s flying hysterically all over the place, as you can see if you try searching with the hashDouble Workshop applet when the table is almost full.
• Clusters can form even when the load factor isn’t high.
• Clusters can form even when the load factor isn’t high.
• Clusters can form even when the load factor isn’t high.
Open Addressing
Quadratic Probing

The HashDouble Applet with Quadratic Probes
• Incidentally, if you try to fill the hash table too full,
  you may see the message Can't complete fill. This occurs
  when the probe sequences get very long.
  Every additional step in the probe sequence makes a bigger step size.
  If the sequence is too long, the step size will eventually exceed the
  capacity of its integer variable, so the applet shuts down the fill
  process before this happens.
• Once the table is filled, select an existing key value and use the
  Find key to see if the algorithm can find it.
  Often it's located at the initial cell, or the one adjacent to it.
  If you're patient, however, you'll find a key that requires three or
  four steps, and you'll see the step size lengthen for each step.
  You can also use Find to search for a nonexistent key; this search
  continues until an empty cell is encountered.

Open Addressing
Quadratic Probing

The Problem with Quadratic Probes
• Quadratic probes eliminate the clustering problem we saw with the
  linear probe, which is called primary clustering.
  However, quadratic probes suffer from a different and more subtle
  clustering problem.
  This occurs because all the keys that hash to a particular cell follow
  the same sequence in trying to find a vacant space.
  Let's say 184, 302, 420, 544 all hash to 7 and are inserted in this
  order.
  Then 302 will require a one-step probe, 420 will require a 2-step probe,
  and 544 will require a 3-step probe.
  Each additional item with a key that hashes to 7 will require a longer
  probe.
  This phenomenon is called secondary clustering.
• Secondary clustering is not a serious problem, but quadratic
  probing is not often used because there's a slightly better solution.

Open Addressing
Double Hashing

• Experience has shown that this secondary hash
  function must have certain characteristics:
  It must not be the same as the primary hash function.
  It must never output a 0 (otherwise there would be no step;
  every probe would land on the same cell, and the algorithm
  would go into an endless loop).
• Experts have discovered that functions of the following
  form work well:
  \[ \text{stepSize} = \text{constant} - (\text{key} \% \text{constant}) \]
  where constant is prime and smaller than the array size.
  For example,
  \[ \text{stepSize} = 5 - (\text{key} \% 5) \]
  This is the secondary hash function used in the Workshop applet.

Open Addressing
Double Hashing

The HashDouble Applet with Quadratic Probes
• Important: Always make the array size a prime
  number.
  Use 59 instead of 60, for example.
  (Other primes less than 60 are 53, 47, 43, 41, 37,
  31, 29, 23, 19, 17, 13, 11, 7, 5, 3, and 2.)
  If the array size is not prime, an endless sequence
  of steps may occur during a probe.
  If this happens during a Fill operation, the applet will be
  paralyzed.

Open Addressing
Double Hashing

• To eliminate secondary clustering as well as primary clustering,
  another approach can be used: double hashing (sometimes called
  rehashing).
  Secondary clustering occurs because the algorithm that generates
  the sequence of steps in the quadratic probe always generates the
  same steps: 1, 4, 9, 16, and so on.
  What we need is a way to generate probe sequences that depend
  on the key instead of being the same for every key.
  Then numbers with different keys that hash to the same index will use
  different probe sequences.
  The solution is to hash the key a second time, using a different
  hash function, and use the result as the step size.
  For a given key the step size remains constant throughout a probe,
  but it's different for different keys.

Open Addressing
Double Hashing

• For any given key all the steps will
  be the same size, but different keys
  generate different step sizes.
  With this hash function the step
  sizes are all in the range 1 to 5.
  This is shown in Figure 11.9.
Open Addressing

Double Hashing

The HashDouble Applet with Double Hashing

• It starts up automatically in Double-hashing mode,
  – but if it's in Quadratic mode you can switch to Double by creating
    a new table with the New button and clicking the Double button
    when prompted.
• To best see probes at work you'll need to fill the table
  rather full;
  – say to about nine-tenths capacity or more.
• Even with such high load factors, most data items will be found
  in the cell found by the first hash function; only a few will require
  extended probe sequences.
• Try finding existing keys.
  – When one needs a probe sequence, you'll see how all the steps
    are the same size for a given key, but that the step size is
    different—between 1 and 5— for different keys.

Java Code for Double Hashing

• The listing for hashDouble.java,
  – Similar to the hash.java program, but uses two hash functions,
    • one for finding the initial index,
    • the second for generating the step size.
• As before, the user can
  • show the table contents,
  • insert an item,
  • delete an item,
  • find an item.
• Output and operation of this program are similar to those
  of hash.java.

Table Size a Prime Number

• Double hashing requires that the size of the hash table is
  a prime number.
• To see why, imagine a situation where the table size is
  not a prime number.
• For example, suppose the array size is 15 (indices from
  0 to 14), and that a particular key hashes to an initial
  index of 0 and a step size of 5.
  – The probe sequence will be 0, 5, 10, 0, 5, 10, and so on,
    repeating endlessly
  • Only these three cells are ever examined, so the algorithm will
    never find the empty cells that might be waiting at 1, 2, 3, and so on.
  • The algorithm will crash and burn.
• If the array size were 13, which is prime,
  – the probe sequence eventually visits every cell. It's 0, 5, 10, 0, 5, 12, 4, 9, 1, 6, 11, 3, and so on and on.
  – If there is even one empty cell, the probe will find it.
• Using a prime number as the array size makes it
  impossible for any number to divide it evenly, so the
  probe sequence will eventually check every cell.
• A similar effect occurs using the quadratic probe.
  – In that case, however, the step size gets larger with each step,
    and will eventually overflow the variable holding it, thus
    preventing an endless loop.
• In general, double hashing is the probe sequence of
  choice when open addressing is used.

Separate Chaining

• In open addressing, collisions are
  resolved by looking for an open cell in the
  hash table.
• A different approach is to install a linked
  list at each index in the hash table.
• A data item’s key is hashed to the index in
  the usual way, and the item is inserted
  into the linked list at that index.
• Other items that hash to the same index
  are simply added to the linked list; there’s
  no need to search for empty cells in the
  primary array.
• Figure 11.10 shows how separate
  chaining looks.
• Separate chaining is conceptually
  somewhat simpler than the various probe
  schemes used in open addressing.
  – However, the code is longer because it
    must include the mechanism for the linked
    lists, usually in the form of an additional
class.
Separate Chaining
The HashChain Workshop Applet

• To see how separate chaining works, start the HashChain Workshop applet.
• It displays an array of linked lists, as shown in Figure 11.11.

Separate Chaining
The HashChain Workshop Applet

• Each element of the array occupies one line of the display, and the linked lists extend from left to right.
• Initially there are 25 cells in the array (25 lists).
  – This is more than fits on the screen:
  – you can move the display up and down with the scrollbar to see the entire array.
  – The display shows up to six items per list.
  – You can create a hash table with up to 100 lists, and use load factors up to 2.0.
  Higher load factors may cause the linked lists to exceed six items and run off the right edge of the screen, making it impossible to see all the items.
  – (This may happen very occasionally even at the 2.0 load factor.)
• Experiment with the HashChain applet by inserting some new items with the Ins button.
  – You’ll see how the red arrow goes immediately to the correct list and inserts the item at the beginning of the list.
  – The lists in the HashChain applet are not sorted, so insertion does not require searching through the list.
  – (The example program will demonstrate sorted lists.)

Separate Chaining
The HashChain Workshop Applet

• Try to find specified items using the Find button.
• During a Find operation, if there are several items on the list, the red arrow must step through the items looking for the correct one.
• For a successful search, half the items in the list must be examined on the average, as we discussed in Chapter 5, “Linked Lists.”
• For an unsuccessful search all the items must be examined.

Separate Chaining
The HashChain Workshop Applet

Load Factors
• The load factor (the ratio of the number of items in a hash table to its size) is typically different in separate chaining than in open addressing.
• In separate chaining it’s normal to put N or more items into an N-cell array:
  – thus the load factor can be 1 or greater.
  – There’s no problem with this; some locations will simply contain two or more items in their lists.
  – Of course, if there are many items on the lists, access time is reduced because access to a specified item requires searching through an average of half the items on the list.
• Finding the initial cell takes fast $O(1)$ time, but searching through a list takes time proportional to the number of items on the list; $O(M)$ time.
Thus we don’t want the lists to become too full.

Separate Chaining
The HashChain Workshop Applet

Duplicates
• Duplicates are allowed and may be generated in the Fill process.
  – All items with the same key will be inserted in the same list, so if you need to discover all of them, you must search the entire list in both successful and unsuccessful searches.
  – This lowers performance.
• The Find operation in the applet only finds the first of several duplicates.

Separate Chaining
The HashChain Workshop Applet

Load Factors
• A load factor of 1, as shown in the Workshop applet, is common.
• With this load factor, roughly one third of the cells will be empty, one third will hold one item, and one third will hold two or more items.
• In open addressing, performance degrades badly as the load factor increases above one half or two thirds.
• In separate chaining the load factor can rise above 1 without hurting performance very much.
• This makes separate chaining a more robust mechanism, especially when it’s hard to predict in advance how much data will be placed in the hash table.
Separate Chaining
The HashChain Workshop Applet
Deletion
- In separate chaining, deletion poses no special problems as it does in open addressing.
- The algorithm hashes to the proper list and then deletes the item from the list.
- Because probes aren’t used, it doesn’t matter if the list at a particular cell becomes empty.
- We’ve included a Del button in the Workshop applet to show how deletion works.

Separate Chaining
The HashChain Workshop Applet
Table Size
- With separate chaining it’s not so important to make the table size a prime number, as it is with quadratic probes and double hashing.
- There are no probes in separate chaining, so there’s no need to worry that a probe will go into an endless sequence because the step size divides evenly into the array size.
- On the other hand, certain kinds of key distributions can cause data to cluster when the array size is not a prime number.
- We’ll have more to say about this when we discuss hash functions.

Separate Chaining
The HashChain Workshop Applet
Buckets
- Another approach similar to separate chaining is to use an array at each location in the hash table, instead of a linked list.
- Such arrays are called buckets.
- This approach is not as efficient as the linked list approach, however, because of the problem of choosing the size of the buckets.
- If they’re too small they may overflow, and if they’re too large they waste memory.
- Linked lists, which allocate memory dynamically, don’t have this problem.

Separate Chaining
Java Code for Separate Chaining
- The hashChain.java program includes a SortedList class and an associated Link class.
- Sorted lists don’t speed up a successful search, but they do cut the time of an unsuccessful search in half. (As soon as an item larger than the search key is reached, which on average is half the items in a list, the search is declared a failure.)
- Deletion times are also cut in half;
- However, insertion times are lengthened,
  - because the new item can’t just be inserted at the beginning of the list;
  - its proper place in the ordered list must be located before it’s inserted.
  - If the lists are short, the increase in insertion times may not be important.

Separate Chaining
Java Code for Separate Chaining
- In situations where many unsuccessful searches are anticipated, it may be worthwhile to use the slightly more complicated sorted list, rather than an unsorted list. However, an unsorted list is preferred if insertion speed is more important.
- The hashChain.java program, shown in Listing 11.1, begins by constructing a hash table with a table size and number of items entered by the user. The user can then insert, find, and delete items, and display the list. For the entire hash table to be viewed on the screen, the size of the table must be no greater than 16 or so.
- Listing 11.1 The hashChain.java Program

Separate Chaining
Java Code for Separate Chaining
The hashChain.java Program
Here’s the output when the user creates a table with 20 lists, inserts 20 items into it, and displays it with the s option.
Enter size of hash table: 20
Enter initial number of items: 20
Enter first letter of show, insert, delete, or find: s
0. List (first-->last): 240 1160
1. List (first-->last):
2. List (first-->last):
3. List (first-->last): 143
4. List (first-->last): 1004
5. List (first-->last): 1485 1585
6. List (first-->last):
7. List (first-->last): 87 1407
8. List (first-->last):
9. List (first-->last):
10. List (first-->last): 490
11. List (first-->last):
12. List (first-->last): 872
13. List (first-->last): 1073
14. List (first-->last): 594 954
15. List (first-->last): 335
16. List (first-->last): 1216
17. List (first-->last): 1057 1357
18. List (first-->last): 938 1818
19. List (first-->last):
- If you insert more items into this table, you’ll see the lists grow longer, but maintain their sorted order. You can delete items as well.
- We’ll return to the question of when to use separate chaining when we discuss hash table efficiency later in this chapter.
Hash Functions

• What makes a good hash function
• Improve the approach to hashing strings

Quick Computation

• A good hash function is simple, so it can be computed quickly.
• The major advantage of hash tables is their speed.
• If the hash function is slow, this speed will be degraded.
• A hash function with many multiplications and divisions is not a good idea.
  – (The bit-manipulation facilities of Java or C++, such as shifting bits right to divide a number by a multiple of 2, can sometimes be used to good advantage.)
• Purpose of a hash function: Take a range of key values and transform them into index values in such a way that the key values are distributed randomly across all the indices of the hash table.
  – Keys may be completely random or not so random.

Random Keys

• A so-called perfect hash function maps every key into a different table location.
  – This is only possible for keys
    • that are unusually well behaved,
    • and whose range is small enough to be used directly as array indices (as in the employee-number example at the beginning of this chapter).
  – In most cases neither of these situations exist,
    • and the hash function will need to compress a larger range of keys into a smaller range of index numbers.

• The distribution of key values in a particular database determines what the hash function needs to do.
  • In this chapter we've assumed that the data was randomly distributed over its entire range.
  – In this situation the hash function
    \[
    \text{index} = \text{key} \mod \text{arraySize};
    \]
    is satisfactory.
    – It involves only one mathematical operation, and if the keys are truly random the resulting indices will be random too, and therefore well distributed.

Non-Random Keys

• However, data is often distributed non-randomly.
  – Imagine a database that uses car-part numbers as keys. Perhaps these numbers are of the form
    \[03340000394050535\]
    – This is interpreted as follows:
      Supplier 033 Supplier number (1 to 9, currently up to 70)
      Supplier 033 Category code (100, 150, 200, 250, up to 950)
      Supplier 033 Date of introduction (1 to 12)
      Supplier 033 Year of introduction (00 to 99)
      Supplier 033 Month of introduction (1 to 12)
      Supplier 033 Percentage of parts in each group (100)
    – The key used for the part number shown would be 0,334,000,394,050,535.
    – However, such keys are not randomly distributed.
    • The majority of numbers from 0 to 9,999,999,999,999 can't actually occur.
      • (For example, supplier numbers above 70, category codes from that aren't multiples of 50, and months from 13 to 98. )
    • Also, the checksum is not independent of the other numbers.
    – Some work should be done to these part numbers to help ensure that they form a range of more truly random numbers.

Don't Use Non-Data

• The key fields should be squeezed down until every bit counts.
  – For example, the category codes should be changed to run from 0 to 15.
  – Also, the checksum should be removed because it doesn't add any additional information; it's deliberately redundant.
• Various bit-twiddling techniques are appropriate for compressing the various fields in the key.
Hash Functions

Non-Random Keys

Use All the Data

• Every part of the key (except non-data, as described above) should contribute to the hash function.
• Don’t just use the first 4 digits or some such expurgation.
• The more data that contributes to the key, the more likely it is that the keys will hash evenly into the entire range of indices.
• Sometimes the range of keys is so large it overflows type int or type long variables.
  – We’ll see how to handle overflow when we talk about hashing strings.

To summarize: The trick is to find a hash function that’s simple and fast, yet excludes the non-data parts of the key and uses all the data.

Hash Functions

Non-Random Keys

Use a Prime Number for the Modulo Base

• Often the hash function involves using the modulo operator (%) with the table size.
  – We’ve already seen that it’s important for the table size to be prime number when using a quadratic probe or double hashing.
• However, if the keys themselves may not be randomly distributed, it’s important for the table size to be a prime number no matter what hashing system is used.
  – This is because, if many keys share a divisor with the array size, they may tend to hash to the same location, causing clustering. Using a prime table size eliminates this possibility.
  – For example, if the table size is a multiple of 50 in our car part example, the category codes will all hash to index numbers that are multiples of 50.
  – However, with a prime number such as 53, you are guaranteed that no keys will divide into the table size.
• The moral is to examine your keys carefully, and tailor your hash algorithm to remove any irregularity in the distribution of the keys.

Hash Functions

Hashing Strings

• We saw at the beginning of this chapter
  – how to convert short strings to key numbers by multiplying digit codes by powers of a constant.
  – In particular, we saw that the four-letter word cats could turn into a number by calculating
    
    \[ \text{key} = 3 \times 27^3 + 1 \times 27^2 + 20 \times 27^1 + 19 \times 27^0 \]

  • This approach has the desirable attribute of involving all the characters in the input string.
  – The calculated key value can then be hashed into an array index in the usual way:
    
    \[ \text{index} = (\text{key}) \mod \text{arraySize}; \]

Here’s a Java method that finds the key value of a word:

```java
public static int hashFunc1(String key)
{
    int hashVal = 0;
    int pow27 = 1; // 1, 27, 27*27, etc
    for(int j=key.length()-1; j>=0; j--) // right to left
    {
        int letter = key.charAt(j) - 96; // get char code
        hashVal += pow27 * letter; // times power of 27
        pow27 *= 27; // next power of 27
    }
    return hashVal % arraySize; // mod
}
```

• The loop starts at the rightmost letter in the word.
  – If there are N letters, this is N–1.
  – The numerical equivalent of the letter, according to the code we devised at the beginning of this chapter (a=1 and so on), is placed in letter.
  – This is then multiplied by a power of 27, which is 1 for the letter at N–1, 27 for the letter at N–2, and so on.
• The `hashFunc1()` method is not as efficient as it might be.
  – Aside from the character conversion, there are two multiplications and an addition inside the loop.
  – We can eliminate a multiplication by taking advantage of a mathematical identity called Horner’s method.
  – Homer was an English mathematician, 1773–1827.
  • This states that an expression like
    \[ \text{var4} \times n^4 + \text{var3} \times n^3 + \text{var2} \times n^2 + \text{var1} \times n + \text{var0} \]
  can be written as
    \[ (((\text{var4} \times n + \text{var3}) \times n + \text{var2}) \times n + \text{var1}) \times n + \text{var0} \]
  • To evaluate this, we can start inside the innermost parentheses and work outward.

Here’s a Java method that finds the key value of a word:

```java
public static int hashFunc2(String key)
{
    int hashVal = 0;
    for(int j=0; j<key.length(); j++) // left to right
    {
        int letter = key.charAt(j) - 96; // get char code
        hashVal = hashVal * 27 + letter; // multiply and add
    }
    return hashVal % arraySize; // mod
}
```

• If we translate this to a Java method we have the following code:

```java
public static int hashFunc2(String key)
{
    int hashVal = 0;
    for(int j=0; j<key.length(); j++) // left to right
    {
        int letter = key.charAt(j) - 96; // get char code
        hashVal = hashVal + letter; // multiply and add
    }
    return hashVal % arraySize; // mod
}
```
• Here we start with the leftmost letter of the word (which is somewhat more natural than staring on the right), and we have only one multiplication and one addition each time through the loop (aside from extracting the character from the string).
  – The `hashFunc2()` method unfortunately can’t handle strings longer than about 7 letters.
Hash Functions

Hashing Strings

• Longer strings cause the value of hashVal to exceed the size of type int.
  – (If we used type long, the same problem would still arise for somewhat longer strings.)
• Can we modify this basic approach so we don’t overflow any variables?
  – Notice that the key we eventually end up with is always less than the array size, because we apply the modulo operator
  – It’s not the final index that’s too big; it’s the intermediate key values.
• It turns out that with Horner’s formulation we can apply the modulo (%) operator at each step in the calculation.
  – This gives the same result as applying the modulo operator once at the end, but avoids overflow. (It does add an operation inside the loop.)

Hash Functions

Hashing Strings

• The hashFunc3() method shows how this looks:
  public static int hashFunc3(String key)
  {
    int hashVal = 0;
    for(int j=0; j<key.length(); j++) // left to right
      { // get char code
        hashVal = (hashVal * 27 + key.charAt(j) - 96) % arraySize; // mod
      }
    return hashVal; // no mod
  } // end hashFunc3()

• This approach or something like it is normally taken to hash a string.
• Various bit-manipulation tricks can be played as well, such as using a base of 32 (or a larger power of 2) instead of 27, so that multiplication can be effected using the shift (>>) operator, which is faster than the modulo (%) operator.
• We can use an approach similar to this to convert any kind of string to a number suitable for hashing.
  – The strings can be words, names, or any other concatenation of characters.

Hashing Efficiency

• We’ve noted that insertion and searching in hash tables can approach O(1) time.
  – If no collision occurs, only a call to the hash function and a single array reference are necessary to insert a new item or find an existing item.
  – This is the minimum access time.
  – If collisions occur, access times become dependent on the resulting probe lengths.
  – Each cell accessed during a probe adds another time increment to the search for a vacant cell (for insertion) or for an existing cell.
  – During an access, a cell must be checked to see if it’s empty, and—in the case of searching or deletion—if it contains the desired item.
• Thus an individual search or insertion time is proportional to the length of the probe.
  – This is in addition to a constant time for the hash function.
  – The average probe length (and therefore the average access time) is dependent on the load factor (the ratio of items in the table to the size of the table).
  – As the load factor increases, probe lengths grow longer.
  – We’ll look at the relationship between probe lengths and load factors for the various kinds of hash tables we’ve studied.

Hashing Efficiency

Open Addressing

• The loss of efficiency with high load factors is more serious for the various open addressing schemes than for separate chaining.
• In open addressing, unsuccessful searches generally take longer than successful searches.
• During a probe sequence, the algorithm can stop as soon as it finds the desired item, which is, on the average halfway through the probe sequence.
• On the other hand, it must go all the way to the end of the sequence before it’s sure it can’t find an item.

Hashing Efficiency

Open Addressing

Linear Probing

• The following equations show the relationship between probe length (P) and load factor (L) for linear probing.
  – For a successful search it’s
    \[ P = \frac{1 + L}{2(1 - L)} \]
  – and for an unsuccessful search it’s
    \[ P = \frac{1 + 1}{2(1 - L)} \]
• These formulas are from Knuth (see Appendix B, “Further Reading”), and their derivation is quite complicated.
• Figure 11.12 shows the result of graphing these equations.

Hashing Efficiency

Open Addressing

Linear Probing

• At a load factor of 1/2, a successful search takes 1.5 comparisons and an unsuccessful search takes 2.5
• At a load factor of 2/3, the numbers are 2.0 and 5.0.
• At higher load factors the numbers become very large.
  – The moral: The load factor must be kept under 2/3 and preferably under 1/2.
  – On the other hand, the lower the load factor, the more memory is needed for a given amount of data.
  – The optimum load factor in a particular situation depends on the tradeoff between memory efficiency, which decreases with lower load factors, and speed, which increases.
On the average, half the items must be examined before the correct one is located.

When in doubt, use separate chaining.

Increasing the load factor causes major performance penalties in open addressing, but performance degrades only linearly in separate chaining.

When in doubt, use separate chaining.

In a successful search, the algorithm hashes to the appropriate list and then searches along the list for the item.

Insertion

If the lists are ordered, insertion is always immediate, in the sense that no comparisons are necessary.

The hash function must still be computed, so let's call the insertion time 1.

If the lists are ordered, then, as with an unsuccessful search, an average of half the items in each list must be examined, so the insertion time is $1 + \frac{loadFactor}{2}$.

For open addressing, only half the items must be examined in an unsuccessful search, so the time is the same as for a successful search.

Separate Chaining

Separate chaining is preferable to open addressing.

If open addressing is to be used, double hashing seems to be the preferred system by a small margin over quadratic probing.

The exception is the situation where plenty of memory is available and the data won't expand after the table is created.

If the number of items that will be inserted in a hash table isn't known when the table is created, separate chaining is preferable to open addressing.

Increasing the load factor causes major performance penalties in open addressing, but performance degrades only linearly in separate chaining.

When in doubt, use separate chaining.

In separate chaining, it's typical to use a load factor of about 1.0 (the number of data items equals the array size).

Smaller load factors don't improve performance significantly, but the time for all operations increases linearly with load factor, so going beyond 2 or so is generally a bad idea.

Separate Chaining

Separate chaining is preferable to open addressing.

If the number of items that will be inserted in a hash table isn't known when the table is created, separate chaining is preferable to open addressing.

Increasing the load factor causes major performance penalties in open addressing, but performance degrades only linearly in separate chaining.

When in doubt, use separate chaining.

In a successful search, the algorithm hashes to the appropriate list and then searches along the list for the item.

Insertion

If the lists are ordered, insertion is always immediate, in the sense that no comparisons are necessary.

The hash function must still be computed, so let's call the insertion time 1.

If the lists are ordered, then, as with an unsuccessful search, an average of half the items in each list must be examined, so the insertion time is $1 + \frac{loadFactor}{2}$.

For an unsuccessful search, an average of half the items must be examined in an unsuccessful search, so the time is the same as for a successful search.

Separate Chaining

Separate chaining is preferable to open addressing.

If the number of items that will be inserted in a hash table isn't known when the table is created, separate chaining is preferable to open addressing.

Increasing the load factor causes major performance penalties in open addressing, but performance degrades only linearly in separate chaining.

When in doubt, use separate chaining.

In a successful search, the algorithm hashes to the appropriate list and then searches along the list for the item.

Insertion

If the lists are ordered, insertion is always immediate, in the sense that no comparisons are necessary.

The hash function must still be computed, so let's call the insertion time 1.

If the lists are ordered, then, as with an unsuccessful search, an average of half the items in each list must be examined, so the insertion time is $1 + \frac{loadFactor}{2}$.

For an unsuccessful search, an average of half the items must be examined in an unsuccessful search, so the time is the same as for a successful search.

Separate Chaining

Separate chaining is preferable to open addressing.

If the number of items that will be inserted in a hash table isn't known when the table is created, separate chaining is preferable to open addressing.

Increasing the load factor causes major performance penalties in open addressing, but performance degrades only linearly in separate chaining.

When in doubt, use separate chaining.

In a successful search, the algorithm hashes to the appropriate list and then searches along the list for the item.

Insertion

If the lists are ordered, insertion is always immediate, in the sense that no comparisons are necessary.

The hash function must still be computed, so let's call the insertion time 1.

If the lists are ordered, then, as with an unsuccessful search, an average of half the items in each list must be examined, so the insertion time is $1 + \frac{loadFactor}{2}$.

For an unsuccessful search, an average of half the items must be examined in an unsuccessful search, so the time is the same as for a successful search.

Separate Chaining

Separate chaining is preferable to open addressing.

If the number of items that will be inserted in a hash table isn't known when the table is created, separate chaining is preferable to open addressing.

Increasing the load factor causes major performance penalties in open addressing, but performance degrades only linearly in separate chaining.

When in doubt, use separate chaining.

In a successful search, the algorithm hashes to the appropriate list and then searches along the list for the item.

Insertion

If the lists are ordered, insertion is always immediate, in the sense that no comparisons are necessary.

The hash function must still be computed, so let's call the insertion time 1.

If the lists are ordered, then, as with an unsuccessful search, an average of half the items in each list must be examined, so the insertion time is $1 + \frac{loadFactor}{2}$.

For an unsuccessful search, an average of half the items must be examined in an unsuccessful search, so the time is the same as for a successful search.
Hashing and External Storage

Table of File Pointers

- The central feature in external hashing is a hash table containing block numbers, which refer to blocks in external storage.
- The hash table is sometimes called an index (in the sense of a book's index).
  - It can be stored in main memory, or, if it is too large, stored externally on disk, with only part of it being read into main memory at a time.
  - Even if it fits entirely in main memory, a copy will probably be maintained on the disk, and read into memory when the file is opened.

Hashing and External Storage

Non-Full Blocks

- Let's reuse the example from the last chapter in which the block size is 8,192 bytes, and a record is 512 bytes.
  - Thus a block can hold 16 records.
  - Every entry in the hash table points to one of these blocks. Let's say there are 100 blocks in a particular file.
- In external hashing it's important that blocks don't become full.
  - Thus we might store an average of 8 records per block.
  - Some blocks would have more records, and some fewer.
  - There would be about 800 records in the file.
  - This arrangement is shown in Figure 11.15.

Full Blocks

- Even with a good hash function, a block will occasionally become full.
- This can be handled using variations of the collision-resolution schemes discussed for internal hash tables: open addressing and separate chaining.
  - In open addressing, if during insertion one block is found to be full, the algorithm inserts the new record in a neighboring block.
    - In linear probing this is the next block, but it could also be selected using a quadratic probe or double hashing.
  - In separate chaining, special overflow blocks are made available; when a primary block is found to be full, the new record is inserted in the overflow block.
  - Full blocks are undesirable because an additional disk access is necessary for the second block; this doubles the access time.
    - However, this is acceptable if it happens rarely.
- We've discussed only the simplest hash table implementation for external storage.
  - There are many more complex approaches that are beyond the scope of this book.

Summary (I)

- A hash table is based on an array.
- The range of key values is usually greater than the size of the array.
- A key value is hashed to an array index by a hash function.
- An English-language dictionary is a typical example of a database that can be efficiently handled with a hash table.
- The hashing of a key to an already filled array cell is called a collision.
- Collisions can be handled in two major ways: open addressing and separate chaining.
- In open addressing, data items that hash to a full array cell are placed in another cell in the array.

Summary (II)

- In separate chaining, each array element consists of a linked list. All data items hashing to a given array index are inserted in that list.
- We discussed three kinds of open addressing: linear probing, quadratic probing, and double hashing.
- In linear probing the step size is always 1, so if $x$ is the array index calculated by the hash function, the probe goes to $x, x+1, x+2, x+3,$ and so on.
- The number of such steps required to find a specified item is called the probe length.
Summary (III)

• In linear probing, contiguous sequences of filled cells appear. These are called primary clusters, and they reduce performance.
• In quadratic probing the offset from x is the square of the step number, so the probe goes to x, x+1, x+4, x+9, x+16, and so on.
• Quadratic probing eliminates primary clustering, but suffers from the less severe secondary clustering.
• Secondary clustering occurs because all the keys that hash to the same value follow the same sequence of steps during a probe.
• All keys that hash to the same value follow the same probe sequence because the step size does not depend on the key, but only on the hash value.
• In double hashing the step size depends on the key, and is obtained from a secondary hash function.

Summary (IV)

• If the secondary hash function returns a value s in double hashing, the probe goes to x, x+s, x+2s, x+3s, x+4s, and so on, where s depends on the key, but remains constant during the probe.
• The load factor is the ratio of data items in a hash table to the array size.
• The maximum load factor in open addressing should be around 0.5. For double hashing at this load factor, searches will have an average probe length of 2.
• Search times go to infinity as load factors approach 1.0 in open addressing.
• It’s crucial that an open-addressing hash table does not become too full.
• A load factor of 1.0 is appropriate for separate chaining.
• At this load factor a successful search has an average probe length of 1.5, and an unsuccessful search, 2.0.

Summary (V)

• Probe lengths in separate chaining increase linearly with load factor.
• A string can be hashed by multiplying each character by a different power of a constant, adding the products, and using the modulo (%) operator to reduce the result to the size of the hash table.
• To avoid overflow, the modulo operator can be applied at each step in the process, if the polynomial is expressed using Horner’s method.
• Hash table sizes should generally be prime numbers. This is especially important in quadratic probing and separate chaining.
• Hash tables can be used for external storage. One way to do this is to have the elements in the hash table contain disk-file block numbers.